Investigating the Role of Linguistic Knowledge in Vision and Language Tasks

Arushi Goel

School of Informatics

University of Edinburgh

This dissertation is submitted for the degree of

Doctor of Philosophy

November 2023

To my loving grandparents . . .

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Arushi Goel
November 2023

Acknowledgements

With immense gratitude and a heart brimming with appreciation, I extend my sincerest acknowledgments to all those who have played an instrumental role in the completion of this journey. This endeavor has been a testament to the power of collaboration, love, and support.

First and foremost, I owe an immeasurable debt of gratitude to my supervisor, Hakan Bilen for his unwavering guidance, insightful feedback, and unrelenting encouragement that have been the pillars of strength throughout this journey. Special thanks to my co-supervisors Frank Keller and Basura Fernando for being a part of this journey with me. Your dedication to fostering intellectual growth and scholarly excellence has left an indelible mark on my academic and personal development. I would also like to thank Kartic Subr who assisted each year in the annual review process and motivated me to think outside of the box. Many thanks to my examiners Marcus Rohrbach and Changjian Li for agreeing to examine my thesis. Their constructive feedback during the viva session has greatly helped in improving this work.

To all the collaborators during the course of my PhD internships. Special thanks to the amazing crew at Amazon Alexa, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, Serhi Havrylov and other incredible interns Giorgio Giannone, Hanchen Wang, and Fangyu Liu. Even during the hard remote working days, you all made it a memorable experience that I will cherish forever. To all those who shaped my experience at Google Research and in Zurich. My heartiest thanks to Jasper Uijilings, Thomas Mensink, Lluis Castrejon, Andre Araujo, and Vittorio Ferrari for the amazing collaborative experience and for providing the environment that molded me into a better researcher. All the other folks at the Zurich office and the interns, Christoph, Denys, Emanuelle, and Teddy you made it an experience of a lifetime.

To the Women in Computer Vision, who have shown me the power of a team, love, and unconditional support. Thank you for letting me be a part of this wonderful community and I will be indebted forever for the lessons I have learned, the fun we had while organizing workshops, and the mentorship sessions.

My parents, grandparents, brother and sister-in-law, and all of my family, my rock and foundation, I am eternally grateful for your boundless love, understanding, and unwavering support in all the good times and the bad. The sacrifices you have made for me and the constant encouragement have been the driving force behind every accomplishment, and I carry your values with me in every endeavor I undertake.

To all those who started as colleagues and office-mates and turned out to be my best friends, thank you for creating an intellectually stimulating environment and instilling love and friendship. Panagiotis and Kiyoon, thank you for all the beers and curries. Kasia, thank you for reminding me that I am never alone and how grateful I am. Thank you for all the laughter and love. Nikita and Adarsh, thanks for reminding me how important cakes and cannolis are for surviving the thesis. Octave, thanks for listening to my complaints on all the long lunches, coffee breaks, baking, and wine. Raman and Ruchika, thanks for bringing your youthful energy and enthusiasm into my life. I would also like to thank Wei-Hong, Bo, Lucas, Konda, Xialei, Davide, Anil, Rohit, Parag, Sal, Thomas, Adrian, Antreas, and Titas for the exchange of ideas and collaborative efforts that have enriched my understanding and broadened my horizons. Edinburgh has blessed me with a lot of love from people outside of work, a special thanks to Tom, Sakshi, Robin, Swapnil, Anjali, Pratyush, Nidhi, Prakrati, Ronish, Naomi, Peter, Sam, Babita, Isha for putting up with me during the hard times and supporting me always.

I am also very grateful to my friends that I made more than a decade ago at home, in university, and also during my amazing time in Singapore. Thank you to all of you, Hirina, Palash, Vanshika, Aashini, Aliza, Sumaiyya, Agrima, Shirin, Raavi, Karnika, Yash, Manish, Bharat, Aditya, Rishabh, Vartul, Yashaswa, Malay, Palak, Vinay, Apoorva, Rakesh, Rahul, Dipesh, Krishna, Charu and Dhiraj. All of you have supported me unconditionally through the different stages of my life and that has made me into who I am today. Thank you and I owe a lot to you all.

To all those whose names are written here and those whose support may not be explicitly mentioned, your impact on my journey is immeasurable, and for that, I am deeply grateful. As I turn the page to the next chapter, I carry forward the lessons learned, the bonds forged, and the memories created during this incredible experience. The completion of this thesis marks not an end, but a new beginning, and I am excited to see where this journey takes me.

Abstract

Artificial Intelligence (AI) has transformed the way we interact with technology *e.g.,* chatbots, voice-based assistants, smart devices, and so on. One particular area that has gained tremendous attention and importance is learning through multimodal data sources within AI systems. By incorporating multimodal learning into AI systems, we can bridge the gap between human and machine communication, enabling more intuitive and natural interactions. Multimodal learning is the integration of multiple sensory modalities, such as text, images, speech, and gestures, to enable machines to understand and interpret humans and the world around us more comprehensively. In this thesis we develop strategies to exploit multimodal data (specifically text and images) along with linguistic knowledge, making multimodal systems more reliable and accurate for various vision and language tasks.

In the first part of the thesis, we focus on developing AI systems that can understand the visual world around us and respond in a more natural and human-like manner. This task is popularly known as *image captioning*. Despite the significant progress in this task, the image captions generated by the models are extremely *generic* and *template-like* for visually similar images. We address this limitation and generate detailed and image-specific captions by exploiting prior and implicit linguistic knowledge, without the need for more labeled data or computational overhead. Unlike previous work, our proposed method generates captions that reflect the image in detail.

To further allow AI models to better understand and interpret context, in the second part of the thesis we leverage information from multiple modalities to gather a more

comprehensive understanding of the visual data by generating *scene graphs*. Unlike image captioning that provides a high-level interpretation of the scene, in this setting a key question is – how do different objects/entities in the scene interact with each other? Collecting large amounts of labeled data that can capture every possible interaction is very expensive and infeasible. Hence, we propose an efficient training strategy that generates complete and informative scene graphs from incomplete and missing labels using the knowledge of label informativeness from linguistics.

In the third part of the thesis, we study the narrative descriptions of images generated from human speech *i.e.,* natural language, to enable natural interaction between humans and machines. One fundamental and challenging problem when dealing with natural language is the task of *coreference resolution*. For example, in the sentence "John saw a dog. He petted it," coreference resolution determines that "he" refers to "John" and "it" refers to the "dog." While coreference resolution may seem straightforward to humans, it poses several significant challenges for AI systems. Without proper coreference resolution, models will struggle to derive the correct meaning and produce coherent outputs. To address this important and complex problem, we propose a novel benchmark dataset for *multimodal coreference resolution* to evaluate *coreference resolution* in text and *narrative grounding* in images. We also propose a weakly supervised method with rule-based linguistic knowledge to address *multimodal coreference resolution* without a large supervised training dataset.

Finally, we address the limitations of the weakly supervised learning setup in *multimodal coreference resolution* by proposing a semi-supervised learning strategy. By using a small labeled and a large unlabeled dataset with robust self-supervised and pseudo-labeled loss functions, we achieve strong performance gains for coreference resolution and narrative grounding in a data-efficient way.

Our work addresses important aspects in vision and language and paves the way for interesting future avenues. In the last part of the thesis, we discuss in more detail directions for the future that are important for advancing the field and unlocking its full potential. Hence, continued research is needed to push the boundaries of multimodal learning.

List of Abbreviations

ADAM	Adaptive moment estimation
AI	Artificial intelligence
ALBEF	Align before fuse
AP	Average Precision
BBR	Bounding box regression
BCE	Binary corss entropy
BERT	Bidirectional encoder representations from transformers
BLANC	BiLateral assessment of noun-phrase coreference
BLEU	Bilingual evaluation understudy
BLIP	Bootstrapping language image pre-training
CE	Cross entropy
CEAF	Constrained entity aligned F-Measure
CIDER	Consensus-based image description evaluation
CIN	Coreferenced image narratives
CLIP	Contrastive language image pre-training

CLTA	Conditional latent topic attention
CNN	Convolutional neural networks
CR	Coreference resolution
EBM	Energy based modeling
FC	Fully connected
FFN	Feed forward network
FPN	Feature pyramid networks
FREQ	Frequency
GAN	Generative adversarial networks
GCN	Graph convolutional networks
GD	Groudning
GLIP	Grounded language image pre-training
GPT	Generative pretrained transformer
GPU	Graphical processing unit
GRU	Gated recurrent units
GT	Ground truth
HTML	Hyper text markup language
IC	Image captioning
IMP	Iterative message passing
ITC	Image text contrastive
KD	Knowledge distillation

KERN	Knowledge embedded routing network
KL	Kullback leibler divergence
LDA	Latent dirichlet allocation
LLM	Large language models
LSTM	Long short term memory units
MAF	Multimodal alignment framework
MCR	Multimodal coreference resolution
MD	Movie descriptions
MLM	Masked language modeling
MLP	Multi layer perceptron
MSCOCO	Microsoft common objects in context
MSE	Mean squared error
MT	Mouse traces
NARE	Not all relations are equal
NLL	Negative log likelihood
NLP	Natural language processing
NLVR	Natural language for visual reasoning
OSCAR	Object semantics aligned pre-training
PCR	Pseudo coreference resolution loss
PGD	Pseudo grounding loss
POS	Positional

RCNN	Region-based convolutional neural network
RGBD	Red green blue depth
RNN	Recurrent neural networks
ROUGE	Recall-oriented understudy for gisting evaluation
SAE	Sentence auto-encoder
SGD	Stochastic gradient descent
SGG	Scene graph generation
SPICE	Semantic Propositional Image Caption Evaluation
TDE	Total direct effect
TV	Total variation
UNITER	Universal image-text representation
VALSE	Vision and language structured evaluation
VG	Visual genome
VL	Vision language
VLM	Vision language models
VQA	Visual question answering
WS	Weakly supervised

Contents

Chapter 1

Introduction and Background

Multimodal learning, which combines vision (*i.e.,* images, videos) and language (*i.e.,* text), plays a crucial role in advancing artificial intelligence (AI) systems (Sanderson, 2023). Multimodal learning can be divided into discriminative and generative tasks. For the discriminative tasks, the main goal is to either find matching between the two modalities of image and text (Jia et al., 2021; Lee et al., 2018b; Li et al., 2019b; Radford et al., 2021; Wang et al., 2018) or develop understanding of one modality based on the other (Chen et al., 2019a, 2020; Li et al., 2021; Singh et al., 2022; Wang et al., 2015) *e.g.,* image-to-text retrieval, image search based on text. In generative tasks, the aim is to generate one modality conditioned on the other, for instance, generate text based on the image (Cho et al., 2021; Pan et al., 2004; You et al., 2016) or generate image based on text (Nichol et al., 2021; Qiao et al., 2019a,b; Xu et al., 2018). By leveraging both visual and textual information, multimodal learning enables machines to perceive the world more comprehensively and understand it in a way that is closer to human cognition (Wagman, 1991). Below are some key reasons why multimodal learning is important in AI:

- **Enhanced perception:** Vision and language are two primary modalities through which humans perceive and understand the world. By integrating visual and textual inputs, AI models can gain a more holistic perception of their environment, allowing them to better recognize objects, scenes, and people. This improves the accuracy and robustness of various AI applications such as object detection (Girshick, 2015; Lin et al., 2014; Ren et al., 2015), image classification (Deng et al., 2009; Simonyan and Zisserman, 2014), and scene understanding (Kazemzadeh et al., 2014; Krishna et al., 2017; Plummer et al., 2015).

- **Contextual understanding:** Combining vision and language enables AI systems to understand the context of the visual content. Language provides valuable semantic information that complements visual cues, helping machines grasp the relationships, attributes, and characteristics of objects or scenes (Krishna et al., 2017; Yang et al., 2018; Zhang et al., 2017c, 2019). For instance, when a model learns to associate the word "riding" with visual representations of "person riding bicycle", it develops a deeper contextual understanding of the concept.

- **Richer representations:** Multimodal learning enables the creation of richer and more expressive representations of data. By fusing visual and textual features, AI models can capture intricate details and nuanced relationships that may not be fully captured by a single modality alone. This can lead to more powerful and meaningful representations that can be leveraged in a wide range of downstream tasks such as image captioning (Anderson et al., 2018b; Cho et al., 2021; Huang et al., 2019b; Pan et al., 2004; You et al., 2016), visual question answering (Alayrac et al., 2022; Anderson et al., 2018b; Antol et al., 2015; Chen et al., 2020; Lu et al., 2016b; Shih et al., 2016; Singh et al., 2022), and visual storytelling (Ferraro et al., 2016; Hsu et al., 2020; Huang et al., 2016).

- **Improved communication and interaction:** The ability to comprehend both vision and language facilitates more effective communication between AI systems and humans. For example, multimodal models can interpret and respond to natural language queries about visual content, enabling intuitive and interactive interfaces. This opens up possibilities for applications like virtual assistants (White, 2018), chatbots (Yang et al., 2022b), and human-robot interaction systems (Kosuge and Hirata, 2004; Sheridan, 2016) that can understand and generate multimodal inputs.

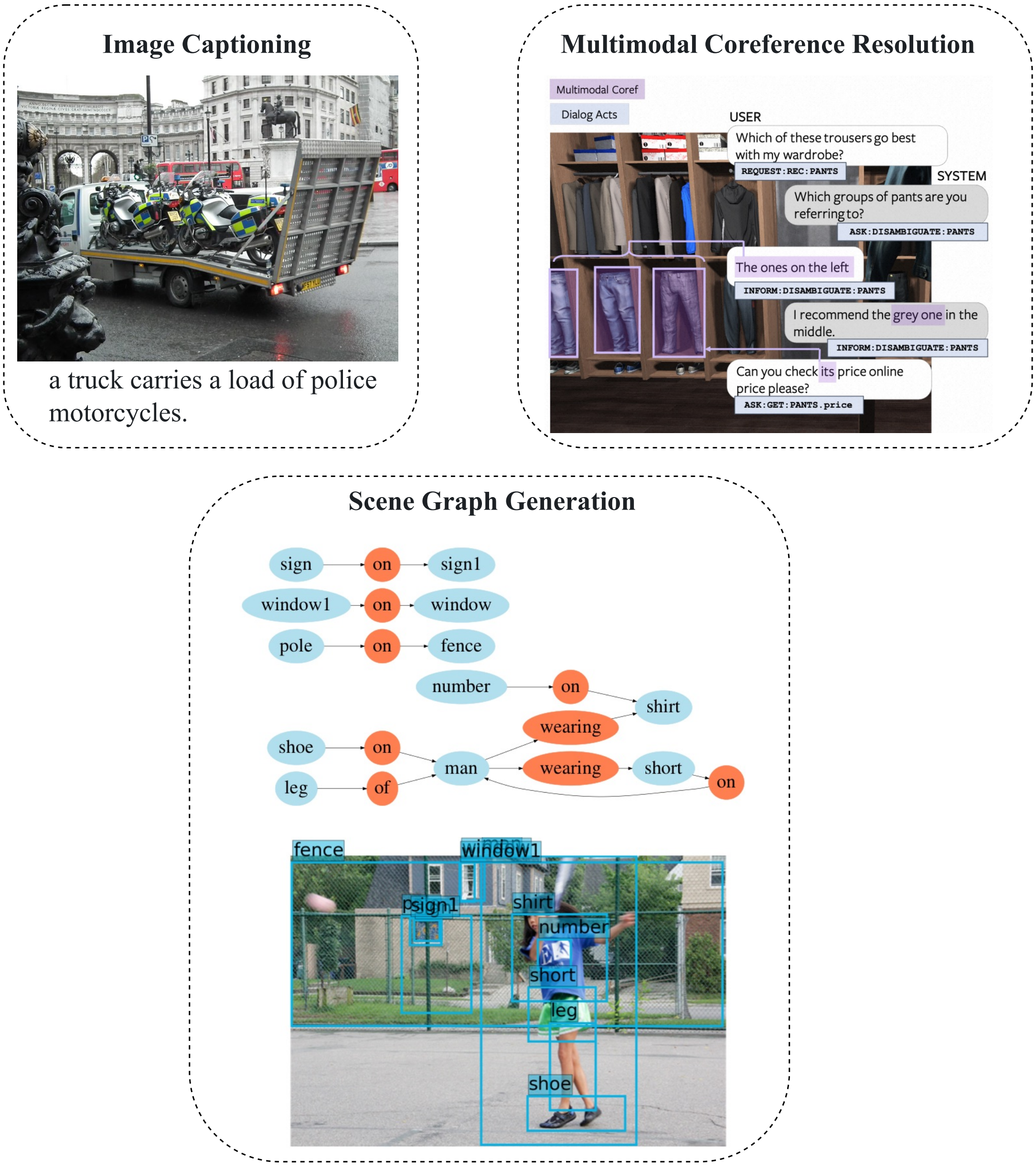
- **Bridging the semantic gap:** The semantic gap refers to the mismatch between low-level sensory data (*e.g.,* pixels) and high-level semantic concepts (*e.g.,* objects, actions, relationships). Multimodal learning helps bridge this gap by learning meaningful representations that connect visual and textual information (Chen et al., 2019a, 2020; Li et al., 2021; Singh et al., 2022; Wang et al., 2015). This enables AI models to bridge the semantic divide and generate more accurate and semantically coherent outputs.

Hence, multimodal learning, which combines vision and language, is vital in AI as it enhances perception, fosters contextual understanding, enables richer representations, improves communication and interaction, and helps bridge the semantic gap. By leveraging the synergy between vision and language, AI systems can achieve a more comprehensive understanding of the world, resulting in more sophisticated and capable AI applications. Next, we present a preliminary background on the main goals of this thesis which is centered around developing robust vision and language models.

1.1 Background

1.1.1 Vision and language tasks

Our goal in this thesis is to develop multi-modal learning systems *i.e.,* combine visual and textual modalities for enhancing the capabilities of AI models. Real-world data available on the web is highly multi-modal, for instance, images associated with text (Changpinyo et al., 2021b; Krishna et al., 2017; Lin et al., 2014; Plummer et al., 2015). The progress in vision and language tasks holds great promise for applications in areas like robotics and autonomous systems (Anderson et al., 2018c, 2021; Kosuge and Hirata, 2004; Lynch et al., 2023; Sheridan, 2016; Tellex et al., 2020), content understanding (Cho et al., 2021; Nichol et al., 2021; Pan et al., 2004; Qiao et al., 2019a,b; Xu et al., 2018; You et al., 2016), assistive technologies (Huang et al., 2022; Lancioni and Singh, 2014), and human-computer interaction (Gupta et al., 2014; Karray et al., 2008; Pustejovsky and Krishnaswamy, 2021). Hence, it is crucial to develop AI systems that can learn from multi-modal information sources. By enabling machines to understand and generate language in the context of visual information, these tasks contribute to building more intelligent and intuitive systems that can effectively interact with and interpret the world around us.



a truck carries a load of police motorcycles.

Which of these trousers go best with my wardrobe?
REQUEST:REC:PANTS

Which groups of pants are you referring to?
ASK:DISAMBIGUATE:PANTS

The ones on the left
INFORM:DISAMBIGUATE:PANTS

I recommend the grey one in the middle.
INFORM:DISAMBIGUATE:PANTS

Can you check its price online price please?
ASK:GET:PANTS.price

Figure 1.1 Overview of the vision-language tasks of (a) image captioning (Lin et al., 2014), (b) multimodal coreference resolution (Guo et al., 2022) and, (c) scene graph generation (Krishna et al., 2017).

In this thesis, we are interested in developing accurate and reliable models for several vision and language tasks. Several works have introduced the tasks that focus on using both the vision and text modalities (Fig. 1.1 presents a visualization of a few such tasks) such as visual question answering (Alayrac et al., 2022; Anderson et al., 2018b; Antol et al., 2015; Chen et al., 2020; Lu et al., 2016b; Shih et al., 2016; Singh et al., 2022), image captioning (Anderson et al., 2018b; Cho et al., 2021; Huang et al., 2019b; Pan et al., 2004; You et al., 2016), referring expression comprehension and grounding (Kazemzadeh et al., 2014; Plummer et al., 2015; Yu et al., 2016, 2018b), visual commonsense reasoning (Zellers et al., 2019), image-text retrieval (Chen et al., 2019a, 2020; Li et al., 2021; Singh et al., 2022; Wang et al., 2015), visual dialog (Das et al., 2017), scene graph generation and visual relationship understanding (Krishna et al., 2017; Yang et al., 2018; Zhang et al., 2017c, 2019), multimodal coreference resolution (Kong et al., 2014; Ramanathan et al., 2014; Rohrbach et al., 2017), visual generation from text (Nichol et al., 2021; Qiao et al., 2019a,b; Xu et al., 2018) and visual storytelling (Ferraro et al., 2016; Hsu et al., 2020; Huang et al., 2016). These tasks involve extracting information from visual data and integrating it with textual information to enable machines to comprehend and generate language that corresponds to visual content.

Initial attempts to solve vision and language tasks involved processing the image with gist and sift based features (Chu and Zhao, 2014; Oliva and Torralba, 2006; Xie et al., 2018) or by extracting high-level concepts from images such as objects, actions, and attributes (Krishna et al., 2017). The textual data was processed either using dependency parsing (Nivre, 2005) or markovian-based language models (Bengio et al., 1999). With the advent of deep learning, the above methods were replaced by modern methods such as processing images with a Convolutional Neural Network (CNN) (Simonyan and Zisserman, 2014) or a Vision Transformer (ViT) (Dosovitskiy et al., 2020) and processing text using sequential models such as recurrent neural networks (RNNs) (Sutskever et al., 2011), long short term memory networks (LSTMs) (Graves and Graves, 2012) and gated recurrent units (GRUs) (Chung et al., 2014). These methods are really powerful and generalized for learning robust representations.

Subsequently, the vision and text features are fused using either early fusion or late fusion techniques (Moens et al., 2018) depending on the downstream task. However, these sequence-based models *e.g.,* RNNs struggle to encode long sentences and retain context from language effectively due to vanishing/exploding gradients. This led to the surge in efficient transformer based models such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and GPT (Generative Pre-Trained Transformer) (Radford et al., 2019) which do not rely on sequential processing due to the self-attention mechanism (Vaswani et al., 2017) and this laid the foundation for many vision and language models (Chen et al., 2020; Li et al., 2021; Singh et al., 2022; Su et al., 2019).

1.1.2 Role of linguistic knowledge

Several tasks in vision and language (VL) (Anderson et al., 2018b; Antol et al., 2015; Lin et al., 2014) require a source of knowledge to learn and extract concepts (Liu and Singh, 2004), entities (Sun et al., 2018), facts (Wang et al., 2017) and commonsense knowledge (Ilievski et al., 2021). For instance, imagine an image of a person drowning and screaming in a swimming pool. Given a question associated with this image: “Why is the person screaming?”, requires developing models that can rely on linguistic knowledge (specifically, commonsense knowledge and facts in this example) along with visual and text feature encodings to correctly answer the question. As humans we are good at reasoning about this and can tell that the person is afraid of drowning and does not know how to swim, that’s why he is screaming. In this thesis, we wish to incorporate varied sources of knowledge during learning for improving generalization and contextual understanding in a variety of VL tasks. Next, we discuss briefly all the sources of linguistic knowledge and their importance for VL tasks.

External knowledge is the encyclopedic knowledge necessary for addressing vision and language tasks that require factual knowledge and cannot be solely addressed using signals from vision and text data. For example, answering the given question – *Where and when was Barack Obama born?* requires extracting the relevant knowledge source from Wikipedia. Sources of text-based external knowledge include Wikipedia

(Voss, 2005), Wikidata (Vrandečić and Krötzsch, 2014), DBPedia (Auer et al., 2007) and ConceptNet (Liu and Singh, 2004). All of these knowledge sources contain a vast amount of information that is useful for tasks requiring fact-based knowledge such as knowledge-based visual question answering (Marino et al., 2019; Shah et al., 2019). For instance, in VQA, given an image of a tomato and avocado salad and the question "What is the number of calories in this salad", the aim is to answer the question by processing the image and text and if needed extracting knowledge from an external database.

Commonsense knowledge is the type of knowledge that humans have perceived about the world. For example, *dogs bark* and *lemons are sour*. Commonsense knowledge is not only limited to physical commonsense but also includes visual, social, temporal, and behavioral commonsense (Ilievski et al., 2021). One of the largest sources of commonsense knowledge sources is ConceptNet (Liu and Singh, 2004) which contains millions of commonsense facts about a variety of entities. This led to the surge in the development of linguistic tasks (Bao et al., 2014; Hermjakob et al., 2000) that rely on ConceptNet for reasoning and understanding. More recently, there has been an emergence in vision and language tasks that require high-level cognition and commonsense reasoning about the world, which is quite an easy task for humans but not for machines (as illustrated in the example of a person drowning in water above). A few such tasks include visual commonsense reasoning (Zellers et al., 2019), Visual COMET (Park et al., 2020), and Perceptual IQ (An et al., 2022).

Internal or implicit knowledge does not rely on external knowledge sources but rather obtains extra knowledge from the existing data (Davison et al., 2019) which can be knowledge of the language, world knowledge, physical knowledge, or commonsense knowledge. This can be achieved by pre-training models on large-scale data without any human supervision and then using them as unstructured knowledge bases. There are different family of models and architectures such as autoencoders (Hinton and Salakhutdinov, 2006; Vincent et al., 2008) or transformers (Devlin et al., 2018; Vaswani et al., 2017) that can be trained with a self-supervised objective function (Radford et al., 2019; Vaswani et al., 2017) on a unlabeled dataset, and subsequently the knowledge stored in the weights of the network can be used as sources of internal knowledge. Recently, large language models (LLMs) (Sanderson, 2023) are a strong example of this type of internal knowledge, which if queried or prompted effectively, leads to strong gains in performance for knowledge-intensive tasks (Bao et al., 2014; Hermjakob et al., 2000). Apart from the importance of implicit knowledge in knowledge-intensive tasks, it is very effective for learning useful priors for transfer learning and generalization (Raina et al., 2006).

Rule-based or lexical knowledge refers to linguistic information and rules that govern the structure, meaning, and usage of words and language. This form of knowledge helps in generalizing the knowledge contained in millions and billions of examples by defining preset rules in language. It encompasses the knowledge of vocabulary, grammar, syntax, semantics, and linguistic rules that guide the composition and interpretation of language (Fiez and Raichle, 1997; Pustejovsky and Boguraev, 1993). It encompasses aspects such as word definitions, synonyms, antonyms, word senses, part-of-speech (POS) tags, morphological properties, and lexical semantic knowledge (Voutilainen, 2003). Lexical knowledge allows models to understand and reason about individual words and their meanings.

It also includes knowledge about word order, sentence structures, phrase structures, syntactic categories, and syntactic dependencies (Nivre, 2005). Rule-based linguistic knowledge enables models to parse and understand the grammatical structure of sentences, facilitating syntactic analysis and generation of grammatically correct language (Ding et al., 2017). Rule-based linguistic knowledge includes various linguistic rules and constraints that govern the formation and interpretation of language. These rules encompass aspects such as agreement rules, tense and aspect rules, negation rules, subject-verb agreement, and other syntactic, morphological, and phonological rules (Anderson, 1982). Knowledge of these rules enables models to generate and interpret language according to the specific rules and constraints of the language. In natural language processing, these resources serve as valuable references and tools for building models that can process, understand, and generate human language accurately.

Discussion. In particular, to make AI systems that mimic human learning, relying on raw inputs from vision and text modalities alone is not sufficient. Aside from sensory modalities, humans also acquire knowledge from external sources, from experience in the form of commonsense, and from learning and memorizing. Hence, research efforts have been made to incorporate these sources of knowledge (Section 1.1.2) into learning of vision and language models (Gui et al., 2021; Marino et al., 2021; Yang et al., 2022c). To test these knowledge-enhanced VL models, several datasets and tasks have been proposed that particularly focus on evaluating the importance of knowledge in learning. Some of these tasks include knowledge-based visual question answering (Marino et al., 2019; Schwenk et al., 2022; Shah et al., 2019), visual commonsense reasoning with commonsense facts (Zellers et al., 2019), knowledge in visual reasoning (Zhang et al., 2021c) and compositional reasoning (Hudson and Manning, 2019; Johnson et al., 2017) and knowledge in visual dialog (Lu et al., 2017; Zhang et al., 2022). It is important to note that all the above mentioned tasks have been designed to evaluate the use of external and commonsense knowledge sources. However, in this thesis, we do not focus on specific reasoning and querying over these knowledge sources for VL tasks but instead focus on how to use both internal and rule-based lexical knowledge for improving the learning of VL models.

1.2 Motivation

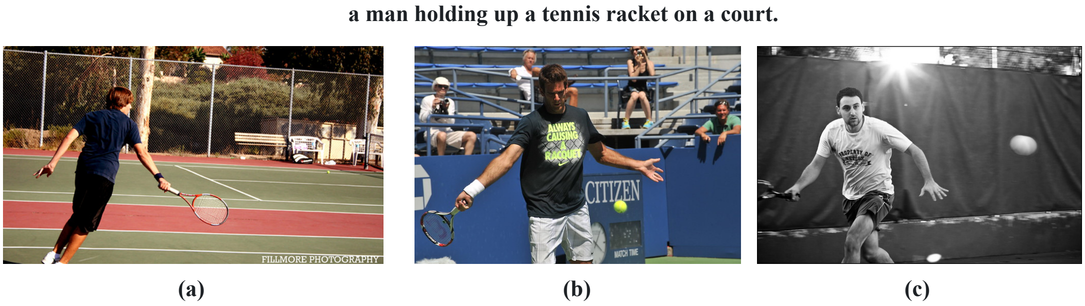
The success of existing vision and language models hinges predominantly on large labeled datasets with bounding box annotations for grounding text in images (Krishna et al., 2017; Lin et al., 2014; Plummer et al., 2015) or images paired with high-quality textual descriptions (Changpinyo et al., 2021b; Lin et al., 2014; Schuhmann et al., 2022). However, creating these datasets can be prohibitively expensive and does not scale well. Human labelers are either needed to write detailed captions and descriptions for tens or hundreds of thousands of images or annotate images exhaustively with bounding boxes corresponding to words in the text, which is costly and time-consuming. This process also inevitably introduces subjective human biases into the training data. For example, labelers may describe the same image differently based on their own backgrounds and perspectives. The expense and effort required to produce these large labeled datasets pose a major bottleneck to developing powerful multimodal AI systems. Motivated by these limitations, we propose to exploit internal and lexical knowledge for vision and language (VL) tasks, especially with limited labeled data. In Section 1.3, we discuss in detail the challenges and limitations that motivate us to develop knowledge-enhanced vision and language models.

Specifically, we investigate these challenges for the following vision and language tasks: image captioning (Lin et al., 2014), scene graph generation (Krishna et al., 2017), and multimodal coreference resolution (Guo et al., 2022; Rohrbach et al., 2017) (see Fig. 1.1 for visualization of these tasks). We first briefly describe these three tasks: 1) *image captioning* involves generating a natural language description describing the content of the image, 2) *scene graph generation* requires understanding the image in finer detail by detecting objects, relationships between objects and the attributes of these objects and then generating a well-connected graph of the image and, 3) *multimodal coreference resolution* involves comprehending the contextual information in the language *i.e.,* links between nouns and pronouns and then establishing connections with specific regions in an image *i.e.,* visual grounding.

These three tasks require the VL model to understand the visual content in the image in detail (*scene graphs*), generate semantically and contextually appropriate captions for the visual content (*image captioning*), and link expressions or references across modalities, such as connecting pronouns in text with corresponding visual entities (*multimodal coreference resolution*). By bridging the gap between visual and textual modalities effectively we develop more reliable and robust models that can be used for real-world applications such as human-robot interaction and assistive agents.

1.3 Challenges

Variability in visual content. Visual content varies from one object to another (*e.g.,* different types of dog) and from one scene to another (*e.g.,* different types of beach). To capture this visual variability, we need to collect precise and fine-grained annotations, describing the content of the image in detail. Figure 1.2 shows examples of a set of images with different visual attributes, such as varied actions and backgrounds but all are labeled with the same caption “A man holding a tennis racket on a court”. In Chapter 2 of this thesis, we identify this limitation and how difficult it is to capture the variability in visual content in natural language hence making it challenging to generate a single detailed, precise, and correct caption. Hence, we propose to generate captions for an image with varied linguistic expressions or levels of detail, helping models to handle visual variability and not suffer from collapsing to producing a single type of caption for every image, also known as “mode collapse”.

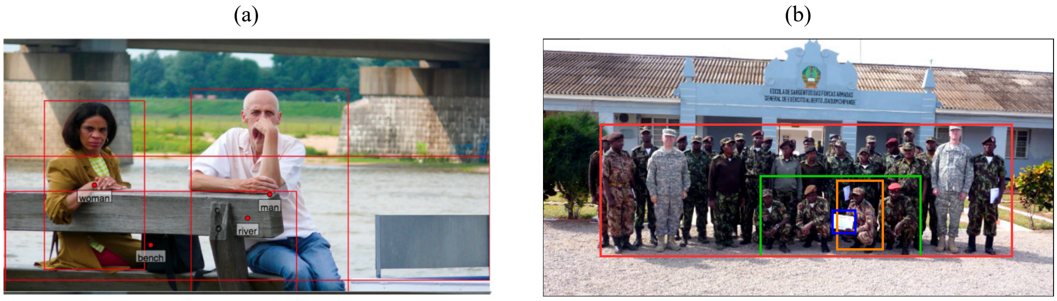


(a) (b) (c)

Figure 1.2 Example of images with visual variability (attributes, actions, foreground, and background) labeled with the common caption “a man holding up a tennis racket on a court” (Lin et al., 2014).

Variability in language: Natural language is inherently ambiguous and, different interpretations or descriptions are possible for the same visual input. This ambiguity makes it challenging for learning a comprehensive and holistic vision and language system, that can generalize to different annotator styles and perspectives. In Chapter 3 to Chapter 5, we address this problem for the tasks of scene graph generation and multimodal coreference resolution. Language variability affects the representation of relationships and structures in scene graphs. Different annotators may describe object relationships using different linguistic patterns (for instance, as shown in Fig. 1.3(a), one annotator could say “man sits on bench” and the other one could say “man on bench”), resulting in variations in the structure and organization of the scene graph.

Moreover, language variability introduces ambiguity and contextual dependencies in coreference resolution as shown in Fig. 1.3(b). The interpretation of references depends on the surrounding linguistic and visual context, for instance, as shown in Fig. 1.3(b), **this man** with a white object in his hands is the man squatting down and not the one standing on the extreme right. Variations in language usage and expressions make it challenging to determine the correct referents, resulting in potential errors in multimodal coreference resolution.



man **sits on** bench
man **on** bench

Few people are standing and **these people** are sitting like squat position and **this man** holding **an object** in **his** hands **which** is white in color and we can see plants. In the background we can see building, grass and sky.

Figure 1.3 (a) Example of language ambiguity on scene graph generation (Krishna et al., 2017) and (b) Example of narrative language for multimodal coreference resolution (Pont-Tuset et al., 2020; Rohrbach et al., 2017).

Missing annotations: Missing annotations are prevalent in several vision tasks such as image classification where an image is only labeled with one positive label but has multiple positive labels (Cole et al., 2021). This incompleteness hampers the ability of the models to fully understand and reason about the image, potentially leading to inaccurate or incomplete predictions. We particularly study the problem of missing annotations in scene graph generation in Chapter 3 of this thesis as shown in Figure 1.3(a). The dataset is either labeled with the relation "on" or "sits on". Missing annotations may introduce biases in the scene graph representation. If certain types of relationships are missing from the annotations, the resulting scene graphs may not adequately represent the diversity and complexity of real-world scenes. Biases in annotations can lead to biased models and biased downstream applications, impacting fairness and generalization. Hence, in Chapter 3 of this thesis, we particularly address this limitation where certain types of relationships are overrepresented or underrepresented in the dataset due to annotation limitations.

Difficulty in getting large-scale annotated data: Obtaining large-scale annotated data for detailed and fine-grained annotations is very time-consuming and resource intensive. For instance, take the example of scene graph generation (Krishna et al., 2017), the annotators need to identify objects, their attributes, and relationships between objects and draw bounding boxes for each object, making the annotation process very costly. In Chapter 4 and Chapter 5 of this thesis, we propose data efficient approaches for the challenging task of multimodal coreference resolution, where it is very costly to get fully annotated data. For tasks such as this, annotators need to possess a solid understanding of both visual and linguistic concepts and requires a combination of domain knowledge, linguistic skills, and familiarity with annotation guidelines. Hence, hiring annotators in bulk is not the most effective solution. Moroever, the cost associated with collecting and annotating such datasets is very high, especially when considering the expertise required and the need for iterative annotation processes. To address these challenges, we propose methods that do not rely on large scale annotated training data whilst still achieving impressive performances for multimodal coreference resolution.

1.4 Contributions

The main contribution of this thesis is to utilize different forms of prior knowledge to improve vision and language tasks. Specifically, we incorporate implicit knowledge and rule-based lexical knowledge to enhance contextual understanding.

- In Chapter 2, we focus on addressing bias in generating natural language descriptions from images, known as image captioning. We leverage ***implicit knowledge*** about language structure and semantics to address mode collapse and improve generalization from limited data. Our method relies on latent variables *i.e.,* topics to capture contextual relationships between image regions and caption words.

Moroever, linguistic knowledge captured via autoencoding captions provides regularization on the syntactic and semantic structure. Together, these allow for generating more accurate and coherent captions.

- In Chapter 3, we target to address subjective human biases in training data for generating scene graphs from images. We employ ***rule-based lexical knowledge*** to learn informative object relations without full supervision. Our method exploits lexical properties of relation labels, obviating the need for complete relation annotations. This facilitates learning complex inter-object relationships from biased data.

- In Chapter 4, we propose an evaluation dataset for multimodal coreference resolution (MCR). Resolving the task of MCR requires comprehending the contextual dependencies between linguistic references and visual context. This contextual understanding is a fundamental aspect of general AI, as it allows systems to capture the nuanced and interdependent nature of real-world scenes and language.

- Training the model for MCR in a fully-supervised setting is extremely expensive due to the nature of annotations required to train models for MCR. In Chapter 4, we propose a novel weakly supervised approach by exploiting the ***rule-based lexical knowledge*** for coreferences by relying on only paired image-caption data, without any bounding-box labels and coreference chain annotations. The lexical rules act as a regularizer to align textual entities with visual context.

- Finally, in Chapter 5, we extend the capabilities of the model to learn coreferences and the integration of the two modalities. Specifically, we propose a semi-supervised vision and language learning framework that relies on a small set of labeled data and on a large set of unlabeled data. The model builds on the capabilities of transferring prior ***implicit knowledge*** from the labeled set to the unlabeled set for inferring referential relationships, understanding contextual information, and making logical associations between linguistic expressions and visual entities.

1.5 Thesis outline

This thesis consists of the following chapters including this introduction – Chapter 2, which proposes the use of prior linguistic knowledge for improving image captioning and generating visually detailed captions. Chapter 3 focuses on scene graph generation and presents a new method that efficiently leverages the lexical knowledge to learn detailed scene graphs from partially annotated data. Chapter 4 focuses on multimodal coreference resolution. In this chapter, we present a new benchmark dataset for evaluating multimodal coreference resolution and a data efficient approach – a weakly supervised method for resolving coreferences by leveraging lexical knowledge. Chapter 5 evaluates the dataset proposed in Chapter 4 by addressing the limitations of the weakly supervised method and proposing a semi-supervised method that relies on a small labeled set with annotations for coreference resolution in text and narrative grounding in images. Chapter 6 summarizes our contributions and introduces ideas for future research.

Chapter 2 to Chapter 5 each contain a paper that has been peer-reviewed and accepted for publication in a conference. The publications included in this thesis are:

- Chapter 2 is based on “Injecting prior knowledge into image caption generation.” **Arushi Goel**, Basura Fernando, Thanh-Son Nguyen, and Hakan Bilen. In Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pp. 369-385. Springer International Publishing, 2020 (*Oral Presentation*) (Goel et al., 2020).

- Chapter 3 is based on “Not all relations are equal: Mining informative labels for scene graph generation.” **Arushi Goel**, Basura Fernando, Frank Keller, and Hakan Bilen. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15596-15606. 2022 (Goel et al., 2022a).

- Chapter 4 is based on “Who are you referring to? Coreference resolution in image narrations.” **Arushi Goel**, Basura Fernando, Frank Keller, and Hakan Bilen. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15247-15258. 2023 (Goel et al., 2023b).

- Chapter 5 is based on “Semi-supervised multimodal coreference resolution in image narrations.” **Arushi Goel**, Basura Fernando, Frank Keller, and Hakan Bilen. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2023 (Goel et al., 2023a).
- Chapter 6 summarizes our contributions and proposes directions for future work.

Chapter 2

Injecting prior knowledge into image caption generation

Automatically generating natural language descriptions from an image is a challenging problem in artificial intelligence that requires a good understanding of the visual and textual signals and the correlations between them. The state-of-the-art methods in image captioning struggle to approach human-level performance, especially when data is limited. In this chapter, we propose to improve the performance of the state-of-the-art image captioning models by incorporating two sources of prior knowledge: (i) conditional latent topic attention, that uses a set of latent variables (topics) as an anchor to generate highly probable words and, (ii) a regularization technique that exploits the inductive biases in the syntactic and semantic structure of captions and improves the generalization of image captioning models. Our experiments validate that our method produces more human interpretable captions and also leads to significant improvements on the MSCOCO dataset in both the full and low data regimes.

This chapter begins with an introduction of the background of image captioning models, limitations of these methods and the key idea of the proposed method for incorporating knowledge in captioning models in Section 2.1, reviewing related work in Section 2.2, elaborating the proposed method in Section 2.3, evaluating and analyzing the proposed method on multiple image captioning backbones in Section 2.5 and conclusion and discussion in Section 2.6.

2.1 Introduction

In recent years there has been a growing interest to develop end-to-end learning algorithms in computer vision tasks. Despite the success in many problems such as image classification (He et al., 2016) and person recognition (Joon Oh et al., 2015), the state-of-the-art methods struggle to reach human-level performance in solving more challenging tasks such as image captioning within limited time and data which involves understanding the visual scenes and describing them in a natural language. This is in contrast to humans who are effortlessly successful in understanding the scenes which they have never seen before and communicating them in a language. It is likely that this efficiency is due to the strong prior knowledge of structure in the visual world and language (Chomsky, 2014).

Motivated by this observation, in this chapter we ask "How can such prior knowledge be represented and utilized to learn better image captioning models with deep neural networks?". To this end, we look at the state-of-the-art encoder-decoder image captioning methods (Anderson et al., 2018a; Vinyals et al., 2015; Xu et al., 2015) where a Convolutional Neural Network (CNN) encoder extracts an embedding from the image, and a Recurrent Neural Network (RNN) decoder generates the text based on the embedding. This framework typically contains two *dynamic* mechanisms to model the sequential output: i) an attention module (Bahdanau et al., 2014; Xu et al., 2015) that identifies the relevant parts of the image embedding based on the previous word and visual features and ii) the RNN decoder that predicts the next words based on its previous state and attended visual features. While these two components are very powerful to model complex relations between the visual and language cues, we hypothesize that they are also capable of and at the same time prone to overfitting to wrong correlations, thus leading to poor generalization performance when the data is limited. Hence, we propose to regulate these modules with two sources of prior knowledge.

First, we propose an attention mechanism that accurately attends to relevant image regions and better cope with complex associations between words and image regions. For instance, in the example of a "man playing tennis", the input visual attention

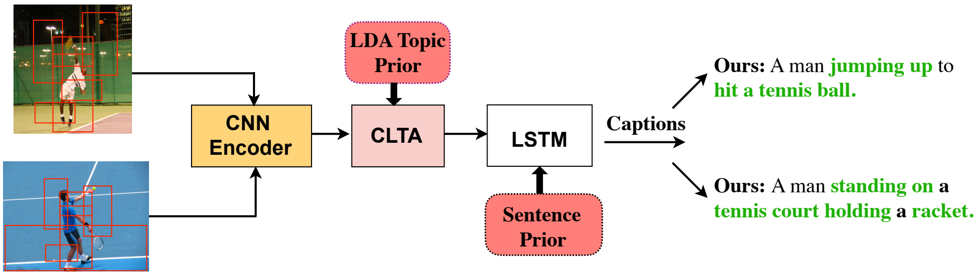

LDA Topic Prior
CNN Encoder
CLTA
LSTM
Sentence Prior
Captions
Ours: A man jumping up to hit a tennis ball.
Ours: A man standing on a tennis court holding a racket.

Figure 2.1 Our final model with Conditional Latent Topic Attention (CLTA) injected with Latent Dirichlet Allocation (LDA) Topic Prior and Sentence Prior (Sentence Auto-Encoder (SAE) regularizer) both rely on prior knowledge to find relevant words and generate non-template like and generalized captions compared to the same baseline caption for both images - *A man hitting a tennis ball with a racket.*

encoder might only look at the local features (*tennis ball*) leaving out the global visual information (*tennis court*). Hence, it generates a trivial caption as "A man is hitting a tennis ball", which is not the full description of the image in context (as shown in Fig. 2.1).

We solve this ambiguity by incorporating prior knowledge of context via latent topic models (Blei et al., 2003), which are known to identify semantically meaningful topics (Chang et al., 2009), into our attention module. In particular, we introduce a Conditional Latent Topic Attention (CLTA) module that models relationship between a word and image regions through a latent shared space *i.e.,* latent topics to find salient regions in an image. Some image regions and word representations may project to the same set of latent topics more than others and therefore more likely to co-occur. *Tennis ball* steers the model to associate this word with the latent topic, "tennis", which further is responsible for localizing *tennis court* in the image. If a region-word pair has a higher probability with respect to a latent topic and if the same topic has a higher probability with respect to some other regions, then it is also a sa. ient region and will be highly weighted. Therefore, we compute two sets of probabilities conditioned on the current word of the captioning model. We use conditional-marginalized probability where marginalization is done over latent topics to find salient image regions to generate the next word. Our CLTA is modeled as a neural network where marginalized probability is used to weight the image region features to obtain a context vector that is passed to an image captioning decoder to generate the next word.

Second, the complexity in the structure of natural language makes it harder to generate fluent sentences while preserving a higher amount of encoded information (high Bleu-4 scores). Although current image captioning models are able to model this linguistic structure, the generated captions follow a more template-like form, for instance, “A man hitting a tennis ball with a racket.” As shown in Fig. 2.1, visually similar images have template-like captions from the baseline model. This limitation might be due to the challenge of learning an accurate mapping from a high-dimensional input (millions of pixels) to an exponentially large output space (all possible word combinations) with limited data. As the sentences have certain structures, it would be easier to learn the mapping to a lower dimensional output space. Inspired by sequence-to-sequence (seq2seq) machine translation (Gehring et al., 2017; Luong et al., 2015a; Sutskever et al., 2014; Wiseman and Rush, 2016), we introduce a new regularization technique for captioning models coined SAE Regularizer. In particular, we design and train an additional seq2seq sentence auto-encoder model (“SAE”) that first reads in a whole sentence as input, generates a lower fixed dimensional vector and, then the vector is further used to reconstruct the input sentence. Our SAE is trained to learn the structure of the input (sentence) space in an offline manner by exploiting the regularity of the sentence space.

Specifically, we use SAE-Dec as an auxiliary decoder branch (see Fig. 2.3). Adding this regularizer forces the representation from the image encoder and language decoder to be more representative of the visual content and less likely to overfit. SAE-Dec is employed along with the original image captioning decoder (“IC-Dec”) to output the target sentence during training, however, we do not use SAE regularizer at test time reducing additional computations.

Both of the proposed improvements also help to overcome the problem of training on large image-caption paired data (Lin et al., 2014; Liu and Singh, 2004) by incorporating prior knowledge which is learned from unstructured data in the form of latent topics and SAE. These priors – also known as “inductive biases” – help the models make inferences that go beyond the observed training data. Through an extensive set of experiments, we demonstrate that our proposed CLTA module and SAE-Dec regularizer improves the image captioning performance both in the limited data and full data training regimes on the MSCOCO dataset (Lin et al., 2014).

2.2 Related work

Here, we first discuss related attention mechanisms and then the use of knowledge transfer in image captioning models.

Attention mechanisms in image captioning. The pioneering work in neural machine translation (Bahdanau et al., 2014; Cho et al., 2014a; Luong et al., 2015b) has shown that attention in encoder-decoder architectures can significantly boost the performance in sequential generation tasks. Visual attention is one of the biggest contributors in image captioning (Anderson et al., 2018a; Fang et al., 2015; Huang et al., 2019a; Xu et al., 2015). Soft attention and hard attention variants for image captioning were introduced in (Xu et al., 2015) based on whether the attention has access to the entire image or only a patch. In soft-attention, the alignment weights are learned and placed “softly" over all patches in the query image. Whereas, in hard attention, the models selects only one patch of image to attend at a time. Anderson et al. (2018a) propose a variant of the self-attention mechanism (Vaswani et al., 2017) that has shown impressive performance for sequence understanding tasks. The bottom-up and top-down self-attention proposed by Anderson et al. (2018a) looks at a set of salient image regions using a bottom-up attention. The top-down mechanism uses task-specific context to predict an attention distribution over the image regions. This combined mechanism enables attention to be calculated at the level of objects and other salient image regions. While this attention mechanism is highly useful, for certain words the attended region may however be irrelevant. To address this, Huang et al. (2019a), propose attention on attention (AoA). Interestingly, they use attention at both the encoder and the decoder step of the captioning process. In the encoder, AoA helps to better model relationships among different objects in the image and in the decoder, AoA filters out irrelative attention results and keeps only the useful ones.

Our proposed attention significantly differs in comparison to these attention mechanisms. First, the traditional attention methods, soft-attention (Bahdanau et al., 2014) and scaled dot product attention (Vaswani et al., 2017) aim to find features or regions in an image that highly correlates with a word representation (Anderson et al., 2018a; Bahdanau et al., 2014; Sharma et al., 2018). In contrast, our *conditional-latent topic attention* uses latent variables *i.e.,* topics as anchors to find the relationship between word representations and image regions (features). Some image regions and word representations may project to the same set of latent topics more than others and therefore more likely to co-occur. Our method learns to model these relationships between word-representations and image region features using our latent space. We allow competition among regions and latent topics to compute two sets of probabilities to find salient regions. This competing strategy and our latent topics guided by pre-trained LDA topics (Blei et al., 2003) allow us to better model relationships between visual features and word representations. Hence, the neural structure and our attention mechanism are quite different from all prior work (Anderson et al., 2018a; Bahdanau et al., 2014; Huang et al., 2019a; Xu et al., 2015).

Knowledge transfer in image captioning. It is well known that language consists of semantic and syntactic biases (Bao et al., 2019; Marcheggiani et al., 2018). We exploit these biases by first training a recurrent caption auto-encoder to capture this useful information using (Sutskever et al., 2014). Our captioning auto-encoder is trained to reconstruct the input sentence and hence, this decoder encapsulates the structural, syntactic, and semantic information of input captions. During captioning process, we regularize the captioning RNN with this pre-trained caption-decoder to exploit biases in the language domain and transfer them to the visual-language domain. To the best of our knowledge, no prior work has attempted such knowledge transfer in image captioning. (Zhou et al., 2019b) encodes external knowledge in the form of knowledge graphs using Concept-Net (Liu and Singh, 2004) to improve image captioning. The closest to ours is the work of (Yang et al., 2019) where they propose to generate scene graphs from both sentences and images and then encode the scene graphs to a common dictionary before decoding them back to sentences. However, the generation of scene graphs from images itself is an extremely challenging task. Besides obtaining an encoding for a graph is more challenging than obtaining a representation for sentences. Finally, we propose to transfer syntactic and semantic information as a regularization technique during the image captioning process as an auxiliary loss. Our experiments suggest that this leads to considerable improvements, especially in more structured measures such as CIDEr (Vedantam et al., 2015).

2.3 Method

We first review image captioning with attention in Section 2.3.1, introduce our CLTA mechanism in Section 2.3.2, and then our sentence auto-encoder (SAE) regularizer in Section 2.3.3.

2.3.1 Image captioning with attention

Image captioning models are based on encoder-decoder architecture (Xu et al., 2015) that use a CNN as the image encoder and a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) as the decoder – see Fig.2.1.

The encoder takes an image as input and extracts a feature set $v = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_R\}$ corresponding to R regions of the image, where $\boldsymbol{v}_i \in \mathbb{R}^D$ is the D-dimensional feature vector for the i^{th} region. The decoder outputs a caption y by generating one word at each time step. At time step t, the feature set v is combined into a single vector $\boldsymbol{v}_a^t$ by taking weighted sum as follows:

$$\boldsymbol{v}_a^t = \sum_{i=1}^{R} \alpha_i^t \boldsymbol{v}_i \tag{2.1}$$

where α_i^t is the CLTA weight for region i at time t, that is explained in the next section. The decoder LSTM ϕ then takes a concatenated vector $[\boldsymbol{v}_a^t | \boldsymbol{y}_{t-1}]$ and the previous hidden state $\mathbf{h_{t-1}}$ as input and generates the next hidden state $\mathbf{h_t}$:

$$\mathbf{h_t} = \phi([\boldsymbol{v}_a^t|E\boldsymbol{y}_{t-1}], \mathbf{h_{t-1}}, \Theta_\phi) \tag{2.2}$$

where, $|$ denotes concatenation, $\boldsymbol{y}_{t-1} \in \mathbb{R}^K$ is the one-hot vector of the word generated at time $t-1$, K is the vocabulary size, $\boldsymbol{h}^t \in \mathbb{R}^n$ is the hidden state of the LSTM at time t, n is the LSTM dimensionality, and Θ_ϕ are trainable parameters of the LSTM. The word embedding matrix $E \in \mathbb{R}^{m \times K}$ is trained to translate one-hot vectors to word embeddings as in (Xu et al., 2015), where m is the word embedding dimension. Finally, the decoder predicts the output word by applying a linear mapping ψ on the hidden state and $\boldsymbol{v}_a^t$ as follows:

$$\boldsymbol{y}_t = \psi([\mathbf{h_t}|\boldsymbol{v}_a^t], \Theta_\psi) \tag{2.3}$$

where Θ_ψ are trainable parameters. Our LSTM implementation closely follows the formulation in (Zaremba et al., 2014). In the next section, we describe our proposed CLTA mechanism.

2.3.2 CLTA: Conditional latent topic attention

At time step t, our CLTA module takes the previous LSTM hidden state ($\boldsymbol{h}^{t-1}$) and image features to output the attention weights α^t. Specifically, we use a set of latent topics to model the associations between textual ($\boldsymbol{h}^{t-1}$) and visual features ($\boldsymbol{v}$) to compute the attention weights. The attention weight for region i is obtained by taking the conditional-marginalization over the latent topic l as follows:

$$\alpha_i^t = P(\text{region} = i|h^{t-1}, \boldsymbol{v}) = \sum_{l=1}^{C} P(\text{region} = i|h^{t-1}, \boldsymbol{v}, l) P(l|h^{t-1}, \boldsymbol{v}_i) \tag{2.4}$$

where l is a topic variable in the C-dimensional latent space. To compute $P(l|h^{t-1}, \boldsymbol{v}_i)$, we first project both textual and visual features to a common C-dimensional shared latent space, and obtain the associations by summing the projected features as follows:

$$\boldsymbol{q}_i^t = W_{sc}\boldsymbol{v}_i + W_{hc}\boldsymbol{h}^{t-1} \tag{2.5}$$

where $W_{sc} \in \mathbb{R}^{C\times D}$ and $W_{hc} \in \mathbb{R}^{C\times n}$ are the trainable projection matrices for visual and textual features, respectively. Then the latent topic probability is given by:

$$P_L = P(l|\boldsymbol{h}^{t-1}, \boldsymbol{v}_i) = \frac{\exp(\boldsymbol{q}_{il}^t)}{\sum_{k=1}^{C}\exp(\boldsymbol{q}_{ik}^t)} \tag{2.6}$$

Afterwards, we compute the probability of a region given the textual, vision features, and latent topic variable as follows:

$$\boldsymbol{r}_i^t = W_{sr}\boldsymbol{v}_i + W_{hr}\boldsymbol{h}^{t-1} \tag{2.7}$$

$$P(\text{region} = i|\boldsymbol{h}^{t-1}, v, l) = \frac{\exp(\boldsymbol{r}_{il}^t)}{\sum_{k=1}^{R}\exp(\boldsymbol{r}_{kl}^t)} \tag{2.8}$$

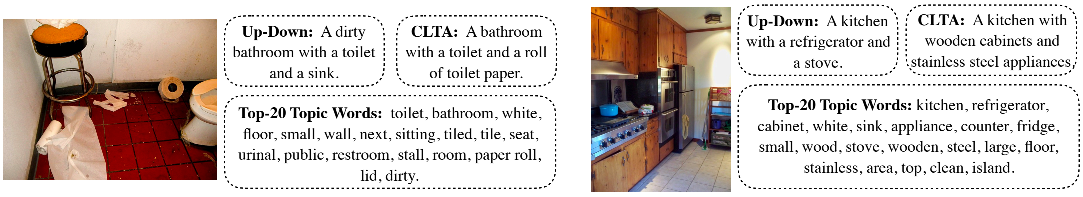
where $W_{sr} \in \mathbb{R}^{C\times D}$ and $W_{hr} \in \mathbb{R}^{C\times n}$ are the trainable projection matrices for visual and textual features, respectively.

The latent topic posterior in Eq. (2.6) is pushed to the pre-trained LDA topic prior by adding a KL-divergence term to the image captioning objective. We apply Latent Dirichlet Allocation (LDA) (Blei et al., 2003) on the caption data. Then, each caption has an inferred topic distribution Q_T from the LDA model which acts as a prior on the latent topic distribution, P_L. For doing this, we take the average of the C-dimensional latent topics at all time steps from $0,\ldots,t-1$ as:

$$P_{L_{avg}} = \frac{1}{t}\sum_{k=0}^{t-1} P(l|\boldsymbol{h}^k, \boldsymbol{v}_i) \tag{2.9}$$

Hence, the KL-divergence objective is defined as:

$$D_{KL}(P_{L_{avg}}||Q_T) = \sum_{c\in C} P_{L_{avg}}(c) \times log(\frac{P_{L_{avg}}(c)}{Q_T(c)}) \tag{2.10}$$



Up-Down: A dirty bathroom with a toilet and a sink.

CLTA: A bathroom with a toilet and a roll of toilet paper.

Top-20 Topic Words: toilet, bathroom, white, floor, small, wall, next, sitting, tiled, tile, seat, urinal, public, restroom, stall, room, paper roll, lid, dirty.

Up-Down: A kitchen with a refrigerator and a stove.

CLTA: A kitchen with wooden cabinets and stainless steel appliances.

Top-20 Topic Words: kitchen, refrigerator, cabinet, white, sink, appliance, counter, fridge, small, wood, stove, wooden, steel, large, floor, stainless, area, top, clean, island.

Figure 2.2 Image-Caption pairs generated from our CLTA module with 128 dimensions and visualization of Top-20 words from the latent topics.

This learned latent topic distribution captures the semantic relations between the visual and textual features in the form of visual topics, and therefore we also use this latent posterior, P_L as a source of meaningful information during the generation of the next hidden state. The modified hidden state $\mathbf{h_t}$ in Eq. (2.2) is now given by:

$$\mathbf{h_t} = \phi([\boldsymbol{v}_a^t | E\boldsymbol{y}_{t-1} | P_L], \mathbf{h_{t-1}}, \Theta_\phi) \tag{2.11}$$

We visualize the distribution of latent topics in Figure 2.2. While traditional “soft-max" attention exploits simple correlations among textual and visual information, we make use of latent topics to model associations between them.

2.3.3 SAE regularizer

Encoder-decoder methods are widely used for translating one language to another (Bahdanau et al., 2014; Cho et al., 2014b; Sutskever et al., 2014). When the input and target sentences are the same, these models function as auto-encoders by first encoding an entire sentence into a fixed-(low) dimensional vector in a latent space, and then reconstructing it. Autoencoders are commonly employed for unsupervised training in text classification (Dai and Le, 2015) and machine translation (Luong et al., 2015a).

In this chapter, our SAE regularizer has two advantages: i) acts as a soft constraint on the image captioning model to regularize the syntactic and semantic space of the captions for better generalization and, ii) encourages the image captioning model to extract more context information for better modeling long-term memory. These two properties of the SAE regularizer generate semantically meaningful captions for

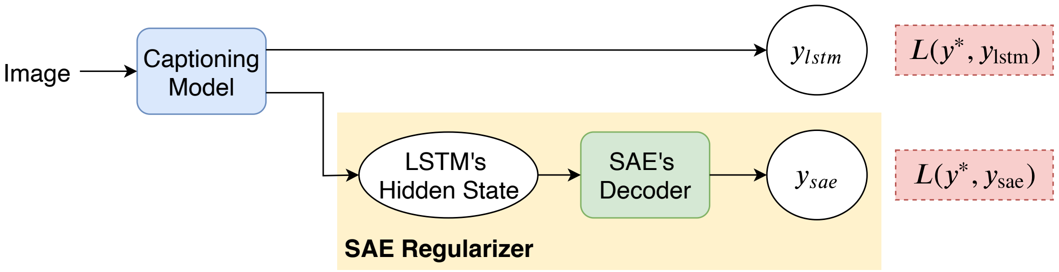


Figure 2.3 Illustration of our proposed Sentence Auto-Encoder (SAE) regularizer with the image captioning decoder. The captioning model is trained by adding the SAE decoder as an auxiliary branch and thus acting as a regularizer.

an image with syntactic generalizations and prevents the generation of naive and template-like captions.

Our SAE model uses network architecture of (Sutskever et al., 2014) with Gated Recurrent Units (GRU) (Chung et al., 2014). Let us denote the parameter of the decoder GRU by Θ_{D}. A stochastic variation of the vanilla sentence auto-encoders is de-noising auto-encoders (Vincent et al., 2008) which are trained to "de-noise" corrupted versions of their inputs. To inject such input noise, we drop each word in the input sentence with a probability of 50% to reduce the contribution of a single word to the semantics of a sentence. We train the SAE model in an offline stage on the training set of the captioning dataset. After the SAE model is trained, we discard its encoder and integrate only its decoder to regularize the captioning model.

As depicted in Figure 2.3, the pre-trained SAE decoder takes the last hidden state vector of captioning LSTM $\boldsymbol{h}$ as input and generates an extra caption (denoted as y_{sae}) in addition to the output of the captioning model (denoted as y_{lstm}). We use the output of the SAE decoder only in train time to regulate the captioning model ϕ by implicitly transferring the previously learned latent structure with SAE decoder.

Our integrated model is optimized to generate two accurate captions (*i.e.*, y_{sae} and y_{lstm}) by minimizing a weighted average of the two loss values:

$$\arg\min_{\Omega} \quad \lambda \mathscr{L}(y^*, y_{\mathrm{lstm}}) + (1-\lambda)\mathscr{L}(y^*, y_{\mathrm{sae}}) \tag{2.12}$$

where $\mathscr{L}$ is the cross-entropy loss computed for each caption, word by word against the ground truth caption y^*, λ is the trade-off parameter, and Ω are the parameters of our model. We consider two scenarios that we use during our experimentation.

- First, we set the parameters of the SAE decoder Θ_D to be the weights of the pre-trained SAE decoder and freeze them while optimizing Equation (2.12) in terms of $\Omega = \{\Theta_\phi, \Theta_\psi, E\}$.
- Second, we initialize Θ_D with the weights of the pre-trained SAE decoder and fine-tune them along with the LSTM parameters, *i.e.,* $\Omega = \{\Theta_\phi, \Theta_\psi, E, \Theta_\mathrm{D}\}$.

As discussed in Section 2.3.2, we also minimize the KL divergence in Eq. (2.10) along with the final regularized objective in Eq. (2.12) as:

$$\arg\min_{\Omega} \quad \lambda \mathscr{L}(y^*, y_{\mathrm{lstm}}) + (1-\lambda)\mathscr{L}(y^*, y_{\mathrm{sae}}) + \gamma D_{KL}(P_{L_{avg}} || Q_T) \tag{2.13}$$

where γ is the weight for the KL divergence loss.

Discussion. An alternative way of exploiting the information from the pre-trained SAE model is to bring the representations from the captioning decoder closer to the encodings of the SAE encoder by minimizing the Euclidean distance between the hidden state from the SAE encoder and the hidden state from the captioning decoder at each time-step. However, we found this setting is too restrictive on the learned hidden state of the LSTM and resulting in poor captioning performance.

2.4 Experiments

Dataset. Our models are evaluated on the standard MSCOCO 2014 image captioning dataset (Lin et al., 2014). For fair comparisons, we use the same data splits for training, validation and testing as in (Karpathy and Fei-Fei, 2015) which have been used extensively in prior works. This split has 113,287 images for training, 5k images for validation and testing respectively with 5 captions for each image. We perform evaluation on all relevant metrics for generated sentence evaluation - CIDEr (Vedantam et al., 2015), Bleu (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE-L (Lin and Och, 2004) and, SPICE (Anderson et al., 2016).

Implementation details. For training our image captioning model, we compute the image features based on the Bottom-Up architecture proposed by (Anderson et al., 2018a), where the model is trained using a Faster-RCNN model (Ren et al., 2015) on the Visual-Genome Dataset (Krishna et al., 2017) with object and attribute information. These features are extracted from R regions and each region feature has D dimensions, where R and D are 36 and 2048 respectively as proposed in (Anderson et al., 2018a). We use these 36×2048 image features in all our experiments.

2.4.1 Experimental setup

LDA topic models. The LDA (Blei et al., 2003) model is learned in an offline manner to generate a C dimensional topic distribution for each caption. Briefly, the LDA model treats the captions as word documents and group these words to form C topics (cluster of words), learns the word distribution for each topic $(C \times V)$ where V is the vocabulary size and also generates a topic distribution for each input caption, Q_T where each C^{th} dimension denotes the probability for that topic.

Sentence auto-encoder. The Sentence Auto-encoder is trained offline on the MSCOCO 2014 captioning dataset (Lin et al., 2014) with the same splits as discussed above. For the architecture, we have a single layer GRU for both the encoder and the decoder. The word embeddings are learned with the network using an embedding layer and the dimension of both the hidden state and the word embeddings is 1024. During training, the decoder is trained with teacher-forcing (Bengio et al., 2015) with a probability of 0.5. For inference, the decoder decodes till it reaches the end of the caption token. The learning rate for this network is 2e-3 and it is trained using the ADAM (Kingma and Ba, 2014) optimizer.

Image captioning decoder with SAE regularizer. The architecture of our image captioning decoder is the same as the Up-Down model (Anderson et al., 2018a) with their "soft-attention" replaced by our CLTA module and trained with the SAE regularizer. We also retrain the AoANet model proposed by Huang et al. (2019a) by incorporating our CLTA module and the SAE regularizer. In the results section, we show improvements over the Up-Down and AoANet models using our proposed approaches. Note, the parameters for training Up-Down and AoANet baselines are same as the original setting. While training the captioning models together with the SAE-decoder, we jointly learn an affine embedding layer (dimension 1024) by combining the embeddings from the image captioning decoder and the SAE-decoder. During inference, we use beam search to generate captions from the captioning decoder using a beam size of 5 for Up-Down and a beam-size of 2 for AoANet. For training the overall objective function as given in Equation 2.13, the value of λ is initialized by 0.7 and increased by a rate of 1.1 every 5 epochs until it reaches a value of 0.9 and γ is fixed to 0.1. We use the ADAM optimizer with a learning rate of 2e-4. Our code is implemented using PyTorch (pyt) and is available at https://github.com/goelarushi/WS-IC.

2.5 Results

First, we study the caption reconstruction performance of vanilla and denoising SAE, then report our model's image captioning performance on MS-COCO dataset with full and limited data, investigate multiple design decisions, and analyze our results qualitatively.

2.5.1 Sentence auto-encoder results

An ideal SAE must learn mapping its input to a fixed low-dimensional space such that a whole sentence can be summarized and reconstructed accurately. To this end, we experiment with two SAEs, Vanilla-SAE and Denoising-SAE, and report their reconstruction performances in terms of Bleu4 and cross-entropy (CE) loss in fig.2.4.

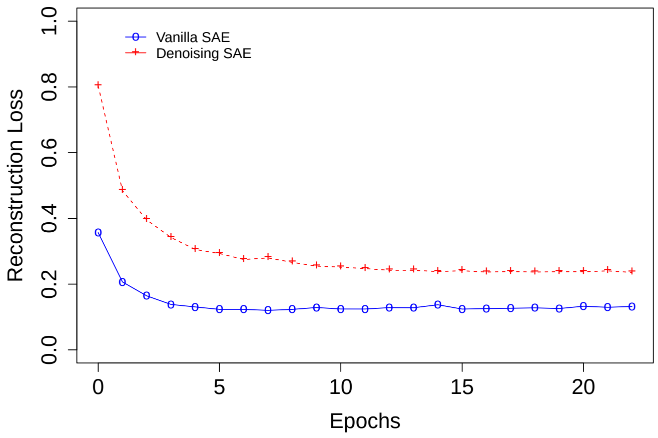


Figure 2.4 Error Curve for the Sentence Auto-Encoder on the Karpathy test split. The error starts increasing approximately after 20 epochs.

Models	Bleu-4 ↑	CE-Loss ↓
Vanilla SAE	**96.33**	**0.12**
Denoising SAE	89.79	0.23

Table 2.1 Bleu-4 Evaluation and Reconstruction Cross-Entropy Loss for the Sentence Auto-Encoder on the Karpathy test split of MSCOCO 2014 caption dataset (Lin et al., 2014).

The vanilla model, when the input words are not corrupted, outperforms the denoising one in both metrics. This is expected as the denoising model is only trained with corrupted input sequences. The loss for both the Vanilla and Denoising SAE start from a relatively high value of approximately 0.8 and 0.4 respectively, and converge to a significantly low error of 0.1 and 0.2. For a better analysis, we also compute the Bleu-4 metrics on our decoded caption against the 5 ground-truth captions. As reported in Table 2.1, both models obtain significantly high Bleu-4 scores. This indicates that an entire caption can be compressed in a low-dimensional vector (1024) and can be successfully reconstructed.

2.5.2 Image captioning results

Here we incorporate the proposed CLTA and SAE regularizer to recent image-captioning models including Up-Down (Anderson et al., 2018a) and AoANet (Huang et al., 2019a) and report their performance on MS-COCO dataset in multiple metrics (see Table 2.2).

Models	cross-entropy loss						cider optimization					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
LSTM-A (Yao et al., 2017)	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
RFNet (Jiang et al., 2018)	76.4	35.8	27.4	56.8	112.5	20.5	79.1	36.5	27.7	57.3	121.9	21.2
Up-Down (Anderson et al., 2018a)	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM (Yao et al., 2018)	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	22.0
AoANet (Huang et al., 2019a)	77.4	37.2	28.4	57.5	119.8	21.3	80.2	38.9	29.2	58.8	129.8	22.4
Up-Down†	75.9	36.0	27.3	56.1	113.3	20.1	79.2	36.3	27.7	57.3	120.8	21.2
Up-Down† + CLTA + SAE-Reg	**76.7**	**37.1**	**28.1**	**57.1**	**116.2**	**21.0**	**80.2**	**37.4**	**28.4**	**58.1**	**127.4**	**22.0**
Relative Improvement	+0.8	+1.1	+0.8	+1.0	+2.9	+0.9	+1.0	+1.1	+0.7	+0.8	+6.6	+0.8
AoANet*	77.3	36.9	**28.5**	57.3	118.4	21.6	80.5	39.1	29.0	58.9	128.9	22.7
AoANet† + CLTA + SAE-Reg	**78.1**	**37.9**	28.4	**57.5**	**119.9**	**21.7**	**80.8**	**39.3**	**29.1**	**59.1**	**130.1**	**22.9**
Relative Improvement	+0.8	+1.0	-0.1	+0.2	+1.5	+0.1	+0.3	+0.2	+0.1	+0.2	+1.2	+0.2

Table 2.2 Image captioning performance on the “Karpathy” test split of the MSCOCO 2014 caption dataset (Lin et al., 2014) from other state-of-the-art methods and our models. Our Conditional Latent Topic Attention with the SAE regularizer significantly improves across all the metrics using both *cross-entropy loss* and *cider optimization*. † denotes our trained models and * indicates the results obtained from the publicly available pre-trained model.

The tables report the original results of these methods from their publications in the top block and the rows in cyan show relative improvement of our models when compared to the baselines.

The baseline models are trained for two settings - 1) Up-Down†, is the model re-trained on the architecture of Anderson et al. (2018a) and, 2) AoANet†, is the Attention-on-Attention model re-trained as in Huang *et al.*Huang et al. (2019a). Note that for both Up-Down and AoANet, we use the original source code to train them in our own hardware. We replace the “soft-attention" module in our Up-Down baseline by CLTA directly. The AoANet model is based on the powerful Transformer (Vaswani et al., 2017) architecture with the multi-head dot attention in both the encoder and decoder. For AoANet, we replace the dot attention in the decoder of AoANet at each head by the CLTA which results in multi-head CLTA. The SAE-decoder is added as a regularizer on top of these models as also discussed in Section 2.4.1. As discussed later in Section 2.5.5, we train all our models with 128 dimensions for the CLTA and with the Denoising SAE decoder (initialized with $\boldsymbol{h}^{last}$).

We evaluate our models with the cross-entropy loss training and also by using the CIDEr score optimization (Rennie et al., 2017) after the cross-entropy pre-training stage (Table 2.2). For the cross-entropy one, our combined approach consistently

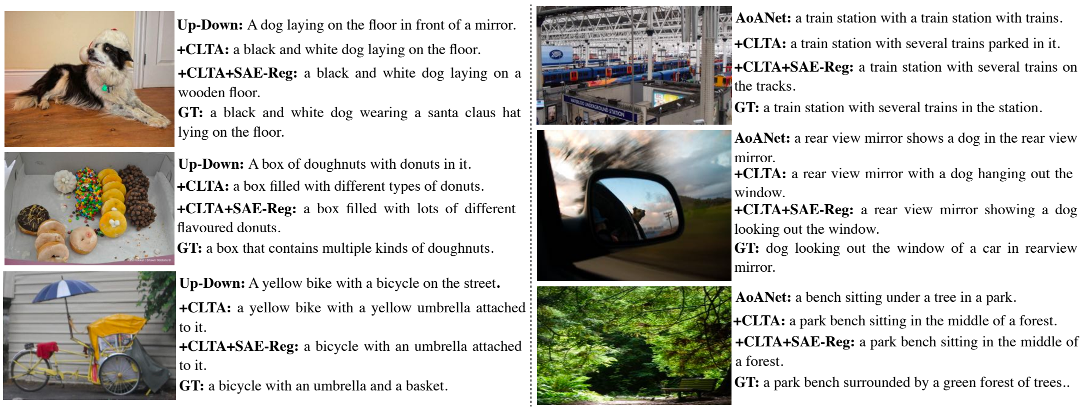
Models	50% data		75% data		100% data	
	Bleu-4	CIDEr	Bleu-4	CIDEr	Bleu-4	CIDEr
Up-Down	35.4	112.0	35.8	112.7	36.0	113.3
Up-Down+CLTA	36.3	113.7	36.3	114.5	36.5	115.0
Up-Down+CLTA+SAE-Reg	**36.6**	**114.8**	**36.8**	**115.6**	**37.1**	**116.2**
AoANet	36.6	116.1	36.8	118.1	36.9	118.4
AoANet+CLTA	36.9	116.7	37.1	118.4	37.4	119.1
AoANet+CLTA+SAE-Reg	**37.2**	**117.5**	**37.6**	**118.9**	**37.9**	**119.9**

Table 2.3 Evaluation of our CLTA and SAE-Regularizer methods by training on a subset of the MSCOCO “Karpathy” Training split.

improves over the baseline performances across all metrics. It is clear from the results that improvements in CIDEr and Bleu-4 are quite significant which shows that our approach generates more human-like and accurate sentences. It is interesting to note that AoANet with CLTA and SAE-regularizer also gives consistent improvements despite having a strong transformer language model. We show in Section 2.5.4 the differences between our captions and the captions generated from Up-Down and AoANet. Our method is modular and improves on state-of-the-art models despite the architectural differences. Moreover, the SAE decoder is discarded after training and hence it brings no additional computational load during test time but with a significant performance boost. For CIDEr optimization, our models based on Up-Down and AoANet also show significant improvements in all metrics for our proposed approach.

2.5.3 Learning to caption with less data

Table 2.3 evaluates the performance of our proposed models for a subset of the training data, where *x*% is the percentage of the total data that is used for training. All these subsets of the training samples are chosen randomly. Our CLTA module is trained with 128 dimensions for the latent topics along with the Denoising SAE Regularizer initialized with the last hidden state of the LSTM (Up-Down+CLTA+SAE-Reg). Despite the number of training samples, our average improvement with CLTA and SAE-Regularizer is around 1% in Bleu-4 and 2.9% in CIDEr for the Up-Down



Up-Down: A dog laying on the floor in front of a mirror.
+CLTA: a black and white dog laying on the floor.
+CLTA+SAE-Reg: a black and white dog laying on a wooden floor.
GT: a black and white dog wearing a santa claus hat lying on the floor.

Up-Down: A box of doughnuts with donuts in it.
+CLTA: a box filled with different types of donuts.
+CLTA+SAE-Reg: a box filled with lots of different flavoured donuts.
GT: a box that contains multiple kinds of doughnuts.

Up-Down: A yellow bike with a bicycle on the street.
+CLTA: a yellow bike with a yellow umbrella attached to it.
+CLTA+SAE-Reg: a bicycle with an umbrella attached to it.
GT: a bicycle with an umbrella and a basket.

AoANet: a train station with a train station with trains.
+CLTA: a train station with several trains parked in it.
+CLTA+SAE-Reg: a train station with several trains on the tracks.
GT: a train station with several trains in the station.

AoANet: a rear view mirror shows a dog in the rear view mirror.
+CLTA: a rear view mirror with a dog hanging out the window.
+CLTA+SAE-Reg: a rear view mirror showing a dog looking out the window.
GT: dog looking out the window of a car in rearview mirror.

AoANet: a bench sitting under a tree in a park.
+CLTA: a park bench sitting in the middle of a forest.
+CLTA+SAE-Reg: a park bench sitting in the middle of a forest.
GT: a park bench surrounded by a green forest of trees..

Figure 2.5 Example of generated captions from the baseline Up-Down, AoANet, our proposed CLTA and, our final models with both CLTA and SAE Regularizer.

model and 0.8% in Bleu-4 and 1.2% in CIDEr for the AoANet model. The significant improvements in Bleu-4 and CIDEr scores with only 50% and 75% of the data compared to the baseline validates our proposed methods as a form of rich prior.

2.5.4 Qualitative results

In Fig. 2.5, we show examples of images and captions generated by the baselines Up-Down and AoANet along with our proposed methods, CLTA and SAE-Regularizer. The baseline models have repetitive words and errors while generating captions (*in front of a mirror*, *a dog in the rear view mirror*). Our models correct these mistakes by finding relevant words according to the context and putting them together in a human-like caption format (*a rear view mirror shows a dog* has the same meaning as *a rear view mirror shows a dog in the rear view mirror* which is efficiently corrected by our models by bringing in the correct meaning). From all the examples shown, we can see that our model overcomes the limitation of overfitting in current methods by completing a caption with more semantic and syntactic generalization (*e.g., different flavoured donuts* and *several trains on the tracks*).

2.5.5 Ablation study

Conditional latent topic attention (CLTA). Table 2.4a depicts the results for the CLTA module that is described in Section 2.3.2. Soft-attention is used as a baseline

Models	Baseline	CLTA		
	Soft-Attention	128	256	512
Bleu-4	36.0	36.5	36.6	**36.7**
CIDEr	113.3	115.0	115.2	**115.3**

(a) Evaluation scores for the Up-Down model with soft-attention and ablations of our CLTA module.

Models	SAE-Decoder	$\boldsymbol{h}$	Bleu-4	CIDEr
Baseline	No	-	36.0	113.3
CLTA-128	Vanilla	First	36.9	115.8
		Last	36.8	115.3
	Denoising	First	36.8	116.1
		Last	37.1	**116.2**
CLTA-512	Denoising	Last	**37.2**	115.9

(b) Additional quantitative evaluation results from different settings of the SAE decoder when trained with image captioning decoder. $\boldsymbol{h}$ denotes the hidden state.

Table 2.4 Ablative Analysis for different settings on our (a) CLTA module and, (b) SAE regularizer training.

and corresponds to the attention mechanism in (Xu et al., 2015) which is the main attention module in Up-Down image captioning model by Anderson et al. (2018a). We replace this attention with the CLTA and evaluate its performance for different numbers of latent dimensions, *i.e.,* topics (C). The models trained with latent topic dimensions of 128, 256, and 512 all outperform the baseline significantly. The higher CIDEr and Bleu-4 scores for these latent topics show the model's capability to generate more descriptive and accurate human-like sentences. As we increase the dimensions of latent topics from 128 to 512, we predict more relevant keywords as new topics learned by the CLTA module with 512 dimensions are useful in encoding more information and hence generating meaningful captions.

Image captioning decoder with SAE regularizer. Table 2.4b reports ablations for our full image captioning model (Up-Down with CLTA) and the SAE regularizer. As discussed in Section 2.3.3, SAE decoder (parameters defined by Θ_D) is initialized with the hidden state of the image captioning decoder. During training, we test different settings of how the SAE decoder is trained with the image captioning decoder: (1) Vanilla vs Denoising SAE and, (2) $\boldsymbol{h}^{\text{first}}$ vs $\boldsymbol{h}^{\text{last}}$, whether the SAE decoder is initialized with the first or last hidden state of the LSTM decoder. For all the settings, we fine-tune the parameters of GRU_D (Θ_D) when trained with the image captioning model (the parameters are initialized with the weights of the pre-trained Vanilla or Denoising SAE decoder).

The results in Table 2.4b are reported on different combinations from the settings described above, with the CLTA having 128 and 512 dimensions in the image captioning decoder. Adding the auxiliary branch of SAE decoder significantly improves over the baseline model with CLTA and in the best setting, Denoising SAE with $\boldsymbol{h}^{\text{last}}$ improves the CIDEr and Bleu-4 scores by 1.2 and 0.6 respectively. As the SAE decoder is trained for the task of reconstruction, fine-tuning it to the task of captioning improves the image captioning decoder.

Initializing the Vanilla SAE decoder with $\boldsymbol{h}^{\text{last}}$ does not provide enough gradient during training and quickly converges to a lower error, hence this brings lower generalization capacity to the image captioning decoder. As $\boldsymbol{h}^{\text{first}}$ is less representative of an entire caption compared to $\boldsymbol{h}^{\text{last}}$, vanilla SAE with $\boldsymbol{h}^{\text{first}}$ is more helpful to improve the captioning decoder training. On the other hand, the Denoising SAE being robust to noisy summary vectors provide enough training signal to improve the image captioning decoder when initialized with either $\boldsymbol{h}^{\text{first}}$ or $\boldsymbol{h}^{\text{last}}$ but slightly better performance with $\boldsymbol{h}^{\text{last}}$ for Bleu-4 and CIDEr as it forces $\boldsymbol{h}^{\text{last}}$ to have an accurate lower-dim representation for the SAE and hence better generalization. It is clear from the results in Table 2.4b, that Denoising SAE with $\boldsymbol{h}^{\text{last}}$ helps to generate accurate and generalizable captions. From our experiments, we found that CLTA with 128 topics and Denoising SAE (with $\boldsymbol{h}^{\text{last}}$) has better performance than even its counterpart with 512 topics. Hence, for all our experiments in Section 2.5.2 and Section 2.5.3 our topic dimension is 128 with Denoising SAE initialized with $\boldsymbol{h}^{\text{last}}$.

2.6 Conclusion and discussion

In this chapter, we have introduced two novel methods for image captioning that exploit prior knowledge and hence help to improve state-of-the-art models even when the data is limited. The first method exploits the association between visual and textual features by learning latent topics via an LDA topic prior and obtains robust attention weights for each image region. The second one is an SAE regularizer that is pre-trained in an autoencoder framework to learn the structure of the captions and is plugged into the image captioning model to regulate its training. Using these modules, we obtain consistent improvements on two investigate models, bottom-up top-down and the AoANet image captioning model, indicating the usefulness of our two modules as a strong prior.

Although the proposed method addresses major shortcomings in image captioning, there are certain limitations that can be important considerations for future work. First, as the SAE regularizer is trained with a self-supervised objective *i.e.,* reconstruction of the caption, it can be pre-trained on a much larger dataset such as Conceptual Captions (Changpinyo et al., 2021b) or Visual Genome (Krishna et al., 2017) in addition to the MSCOCO dataset (Lin et al., 2014). Pre-training on a large dataset would be useful when the image captioning model is trained on limited data, especially on domains where paired image-caption data is expensive to obtain such as fashion (Yang et al., 2020), news articles (Ramisa et al., 2017), etc. Second, in our proposed method the model uses the topics learned from the set of captions as a prior. While this acts as a very good regularizer to find important image regions, there is no explicit information injected from the topics during inference time. We hypothesize that training a standalone image-to-topic model and then using the learned topics as inputs to the image captioning model can help generate highly specific image captions.

Ever since this work, the models used for image captioning have seen a paradigm shift due to two reasons – 1) release of large image-caption datasets such as Visual Genome (Krishna et al., 2017), Conceptual Captions (Changpinyo et al., 2021b) and LAION- 5B (Schuhmann et al., 2022) and, 2) the replacement of sequential models such as recurrent neural networks (Zaremba et al., 2014) by transformer-based generative methods (Vaswani et al., 2017). This led to the development of powerful vision and language models such as VisualBERT (Su et al., 2019), UNITER (Chen et al., 2020), OSCAR (Li et al., 2020) (to name a few) where the visual transformer encoder and the text transformer encoder interact with each other using a cross-attention mechanism (Devlin et al., 2018). These models are then trained on large-scale image caption paired datasets with self-supervised objectives showing good generalization capabilities for tasks such as image captioning. This form of large-scale pretraining (Chen et al., 2020; Mokady et al., 2021) followed by fine-tuning on the target task such as image captioning can be seen as a way of using internal or implicit knowledge as discussed in this chapter with the SAE regularizer but with more model parameters and trained on much larger data (Changpinyo et al., 2021b).

More recently, with the growth in large language models such as GPT2 / GPT3 (Sanderson, 2023) and large vision encoders such as Vision Transformers (ViTs) (Dosovitskiy et al., 2020), there has been increasing amounts of work in developing methods that rely on frozen backbones of these large vision and language models and learn a transformation from the image space to the text space (Chen et al., 2020; Li et al., 2022a, 2023a). Such family of methods leverage all forms of knowledge stored in the parameters of these large models and show great improvements in the task of image captioning. This aligns with the takeaways of this chapter indicating the usefulness of using knowledge for developing robust models for image captioning.

Chapter 3

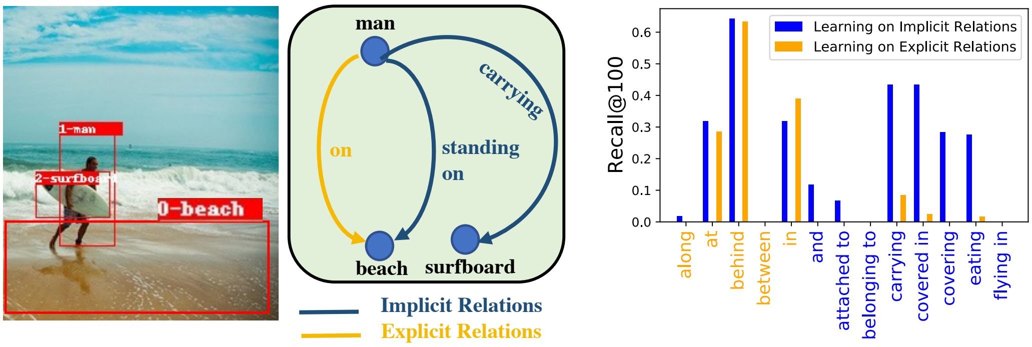
Not all relations are equal: Mining informative labels for scene graph generation

In the previous chapter, we study the task of image captioning for providing fine-grained details of an image – an ideal task for encouraging interaction between visual and language domains. However, describing an image in language goes beyond providing a high-level understanding of the scene. It also requires spatially localizing the objects present in the scene, defining the attributes (color, shape, etc.) of these objects, and what relationships they have with each other. For instance, imagine a human giving a command to the robot, (*Place the "knife"* ***on*** *the "table"*). For the robot to act on this command, it has to develop an understanding of the objects (*knife* and *table*), where they are located in the scene, and then understand the relationship between them (*on*). In this chapter, we move from captions to graphs, an intermediate task between low-level object recognition and high-level scene understanding (Sadeghi and Farhadi, 2011) also known as *scene graph generation*. The task of scene graph generation is to predict triplets of the form *(subject, predicate, object)* where subject and object are the *objects* in an image and predicate is the *relation* between them. Hence, predicting multiple triplets from an image *i.e.,* a scene graph would not only help in a detailed understanding of a scene but also aid in generating image captions that capture visual variability of images (Nguyen et al., 2021).

Specifically, in this chapter, we study scene graph generation (SGG) that aims to capture a wide variety of interactions between pairs of objects, which is essential for full scene understanding. We investigate the use of lexical knowledge for learning fine-grained visual details from images from incomplete and biased labeled data.

Existing SGG methods trained on the entire set of relations fail to acquire complex reasoning about visual and textual correlations due to various biases in training data. Learning on trivial relations that indicate generic spatial configurations like ‘on’ instead of informative relations such as ‘parked on’ does not enforce this complex reasoning, harming generalization. To address this problem, we propose a novel framework for SGG training that exploits relation labels based on their informativeness. Our model-agnostic training procedure imputes missing informative relations for less informative samples in the training data and trains an SGG model on the imputed labels along with existing annotations. We show that this approach can successfully be used in conjunction with state-of-the-art SGG methods and improves their performance significantly in multiple metrics on the standard Visual Genome benchmark. Furthermore, we obtain considerable improvements for unseen triplets in a more challenging zero-shot setting.

This chapter begins with an introduction of the background of scene graph generation methods, limitations, and the key ideas of our proposed method for addressing bias in scene graph in Section 3.1, related work in Section 3.2, elaborating the way to acquire lexical knowledge and incorporate that in our proposed method in Section 3.3, discussing the experimental details in Section 3.4, evaluating and analyzing the proposed method on multiple SGG backbones in Section 3.5 and conclusion and discussion in Section 3.6.



(a)

(b)

Figure 3.1 (a) An example of a scene graph with *implicit* and *explicit* relations. (b) Per-class Recall@100 for an SGG model trained either on only explicit (orange) or implicit (blue) relations.

3.1 Introduction

In this chapter, we look at a structured vision-language problem, scene graph generation (Krishna et al., 2017; Xu et al., 2017), which aims to capture a wide variety of interactions between pairs of objects in images. SGG can be seen as a step towards comprehensive scene understanding and benefits several high-level visual tasks such as object detection/segmentation (Girshick, 2015; He et al., 2017; Ren et al., 2015), image captioning (Anderson et al., 2018b; Huang et al., 2019b; Lin et al., 2014), image/video retrieval (Faghri et al., 2017), and visual question answering (Antol et al., 2015; Lu et al., 2016b). In the literature, SGG is typically formulated as predicting a triplet of a localized subject-object pair connected by a relation (*e.g., person* ***wear*** *shirt*). Broadly, recent advances in SGG have been obtained by extracting local and global visual features in convolutional neural networks (Hung et al., 2019; Simonyan and Zisserman, 2014; Zhang et al., 2019) or graph neural networks (Lin et al., 2020; Xu et al., 2017; Yang et al., 2018) combined with language embeddings (Mikolov et al., 2013; Pennington et al., 2014) or statistical priors (Zellers et al., 2018) for predicting relations between objects.

Despite the remarkable progress in this task, various factors (long-tail data distribution, language or reporting bias (Misra et al., 2016)) in the established SGG benchmarks (*e.g.,* Visual Genome (Krishna et al., 2017)) have been shown to drive existing methods towards biased and inaccurate relation predictions (Tang et al., 2019, 2020). One major cause is that each subject-object pair is annotated with *only one positive* relation, which typically depends on the annotator's preference and is *subjective*, while other plausible relations are treated as negative.[1] For instance, a subject-object pair *man–beach* in Fig. 3.1a is only annotated with one relation as *man* ***on*** *beach*, even though other plausible relations are available such as *standing on*. Hence, the models (Hung et al., 2019; Li et al., 2017b; Tang et al., 2019; Zellers et al., 2018) trained on this data become biased towards more frequently occurring labels, as reported in Tang et al. (2020). To alleviate the biased training, Tang et al. (2020) employs counterfactual causality to force the SGG model to base its predictions only on visual evidence rather than the data bias. Wang et al. (2020b) propose a semi-supervised technique that jointly imputes the missing labels of subject–object pairs with no annotated relations to obtain more balanced triplet distributions. Suhail et al. (2021) propose an energy-based method that can learn to model the joint probability of triplets from a few samples and thereby avoids generating biased graphs with inconsistent structure.

While the recent methods (Chiou et al., 2021; Suhail et al., 2021; Tang et al., 2020; Wang et al., 2020b) successfully tackle the bias towards more frequently occurring labels, this chapter studies another type of bias related to *label informativeness*. It also manifests itself in missing annotations and has not been addressed in SGG before. In particular, we hypothesize that certain relation labels (implicit labels) are more informative than others (explicit labels) and training on implicit labels improves a model's ability to reason over complex correlations in visual and textual data.

Our key intuition comes from prior computational and cognitive models (Logan and Sadler, 1996) and recent work (Collell et al., 2018) that categorize relations into *explicit (or spatial)* and *implicit*, based on whether the relation defines the relative spatial configuration between the two objects implicitly or explicitly (*e.g., man* ***standing on*** *beach vs. man* ***on*** *beach* in Fig. 3.1a). Explicit relations are often *easy to learn*, *e.g.,* from the spatial coordinates of subject–object pairs, thanks to their highly deterministic spatial arrangements, while implicit ones are often challenging due to

[1] Note that both training and test sets of the standard benchmarks are subject to the similar biases.

the relative spatial variation and require deliberate reasoning. To test our hypothesis, we conduct experiments where we train an SGG model either only on explicit, or only implicit relations and evaluate them on a test set including both types (*i.e.,* zero-shot implicit or explicit relation classification), see Fig. 3.1b. Surprisingly, training only on implicit relations obtains good performance not only over implicit ones but also unseen explicit ones (only 2% lower in average training on explicit relations and 4% lower when trained on all labels), while training only on explicit relations performs poorly on implicit relations (where the performance drops to 0.1% from 24.3%).[2] In other words, training on implicit labels enables the model to better generalize to unseen explicit labels. However, due to partially annotated training data, many subject-object pairs are only labeled by explicit relations, and their implicit relations are missing and obscured by explicit ones.

Motivated by our analysis, we design a novel model-agnostic training procedure for SGG that jointly extracts more information from partially labeled data by mining the missing implicit labels, trains an SGG model on them, and boosts its performance. In particular, our method involves a two-stage training pipeline. The first stage trains an SGG model on a subset of training data including only annotated implicit relations, which allows the model to learn rich correlations in the data and encourages it to predict more informative implicit labels in the next stage. The second stage includes an alternating procedure that imputes missing implicit labels on the subset of samples annotated with explicit relations, followed by training on both the annotated and imputed labels, called *label refinement*. In this stage, a model is prone to confirm to its own (wrong) predictions to achieve a lower loss as observed in semi-supervised learning (Arazo et al., 2020; Tarvainen and Valpola, 2017). To prevent such overfitting, we regularize the model with a latent space augmentation strategy. We demonstrate that our method yields significant performance gains in the SGG task for the standard and zero-shot settings on the Visual Genome (Krishna et al., 2017) when applied to several existing scene graph generation models.

[2] We provide the full analysis in Section 3.5 and Chapter B.

In short, our contributions are as follows. We identify a previously unexplored issue, *missing informative labels* in the standard SGG benchmark, and address this through a model agnostic training procedure based on alternating label imputation and model training with effective regularization strategies. This method can be incorporated into state-of-the-art SGG models and boosts their performance by a significant margin.

3.2 Related work

Scene graph generation. SGG has been extensively studied in the past few years with the goal of better understanding the object relations in an image by either focusing on architecture designs (Chen et al., 2019b; Li et al., 2017b; Tang et al., 2019; Xu et al., 2017; Yang et al., 2018; Zellers et al., 2018) or feature fusion methods (Dornadula et al., 2019; Gu et al., 2019; Hung et al., 2019; Li et al., 2018; Yan et al., 2020; Zareian et al., 2020a,b; Zhang et al., 2019). Lu et al. (2016a) introduced the VRD task and dataset and proposed a combination of visual features with language priors to learn a relationship triplet. Bin et al. (2019) and Zhan et al. (2019) exploit mutual relations and undetermined relationships respectively for visual relationship detection. Liang et al. (2018) design a ranking-based objective function to give high relevance to the annotated relationships. A limitation of their approach is that the model will fail to produce multiple plausible predicates for some pairs and overfit to the training data.

Kolesnikov et al. (2019) augment the object detection pipeline with a novel box attention mechanism to model pairwise interactions between objects. Li et al. (2017a) formulate the visual relationship triplets as a visual phrase instead of treating each entity separately. Some related works leverage graph-based formulations for message passing between objects and triplets (Li et al., 2017b, 2018; Mi and Chen, 2020; Xu et al., 2017; Yin et al., 2018; Zellers et al., 2018). Zhuang et al. (2017) propose a context-aware interaction recognition framework that leads to zero-shot generalizations.

To further improve relationship detection Plummer et al. (2016) and Yu et al. (2017) use multiple-cues and linguistic priors from external Wikipedia data respectively. The fusion of semantic, spatial and visual features also play an important role. VTransE

(Zhang et al., 2017a) and its extension UVTransE (Hung et al., 2020) project different transformations of appearance features fused with language and spatial features to a predicate space. Inayoshi et al. (2020) combine image feature maps with semantic and spatial features before flattening them thus keeping the spatial information in the region proposals. Zhang et al. (2019) design graphical contrastive losses to address incorrectly paired objects in images and shows strong performance on frequently occurring relation triplets. Zhou et al. (2019a) propose RLM-Net a two-stage pipeline where first they suppress object pairs by learning relevance probabilities for pairs and then learn a graph-based predicate recognition model with multi-modal features. They cluster predicates with similar spatial positions and then smooth out the classification scores using a graph neural network.

Recently, Tang et al. (2019, 2020) reported that the performance gains from these methods largely come from improved performance only on the head classes (frequently occurring relations) while the performance on most other relations is poor. They propose replacing the biased evaluation metric Recall@k with mean-Recall@k to assign equal importance to all labels. The same authors Tang et al. (2020) report that the bias in the data often drives SGG models to predict frequent labels and propose to use counterfactual causality *i.e.,* measure the difference in predictions between the original scene and a counterfactual one to remove the effect of context bias and focus on the main visual effects of the relation. Chiou et al. (2021) addresses the bias in scene graph generation by learning from positive and unlabeled object pairs. Suhail et al. (2021) propose an energy based loss that learns the joint likelihood of objects and relations instead of learning them individually. This helps to incorporate commonsense structure (*man* ***riding*** *horse* and *man* ***feeding*** *horse* occurring together are highly improbable) and in better context aggregation. Unlike (Suhail et al., 2021; Tang et al., 2020), our focus is on extracting more information from the training data through mining informative labels. In fact, we show that our model is orthogonal to theirs and boosts performance when incorporated into theirs. The most similar to ours, Wang et al. (2020b) proposes a semi-supervised method that employs two deep networks, where the auxiliary one imputes missing labels of unlabeled pairs and self-trains on them and transfers its knowledge to the main network. Unlike Wang et al. (2020b), who treat all the labels equally, our method only imputes informative implicit labels. This is crucial, as shown in Table 3.3, because, without such consideration, imputing labels cannot extract any substantial information from unlabeled samples leading to only minor gains. In addition, our framework is more efficient as it involves only a single network that jointly infers labels and trains on them, outperforming Wang et al. (2020b) significantly.

Label completion. There is a rich body of work in the literature that focuses on learning from partial/missing labels in a multi-label learning setting where each image is labelled for multiple categories with some missing labels (Bucak et al., 2011; Cabral et al., 2011; Cole et al., 2021; Durand et al., 2019; Huang et al., 2021; Mahajan et al., 2018). Common strategies to address this can be divided into two categories: 1) graph-based methods (Huynh and Elhamifar, 2020; Wu et al., 2018) that exploit the similarity between samples to predict missing labels, and 2) low-rank matrix completion which extracts label correlations (Cabral et al., 2011; Durand et al., 2019; Xu et al., 2014; Yang et al., 2016) to complete missing labels. There is another setting in which some instances miss all the labels, also called semi-supervised learning in multi-label classification (Tan et al., 2017; Zhao and Guo, 2015). In this setting, the classifier is trained for unseen data. While related to our setting, we look at the classification of relations conditioned on subject–object pairs (rather than on the image level), with each pair already labeled with one relation (rather than unlabeled images). Finally, we group the label set in two groups and treat them asymmetrically in our training.

Semi-supervised learning. Semi-supervised learning methods exploit unlabeled data via either pseudo-labeling or imputation with small amounts of labeled data (Rizve et al., 2021; Shi et al., 2018) or by enforcing consistency regularization on the unlabeled data to produce consistent predictions over various perturbations of the same input (Tarvainen and Valpola, 2017; Verma et al., 2021) by applying several augmentation strategies such as Mixup (Zhang et al., 2017b), RandAugment (Cubuk et al., 2020), AutoAugment (Cubuk et al., 2018) or combine both pseudo-labeling and consistency regularization (Berthelot et al., 2019; Sohn et al., 2020). Inspired by pseudo-labeling in semi-supervised learning, the main motivation of our method is to impute informative missing labels to improve generalization and learn complex features by relying on partially labeled data and still predict more accurate labels on the biased test set.

3.3 Method

3.3.1 Revisiting SGG pipeline

In SGG, we seek to localize and classify the objects in an image followed by labeling the visual relations between each pair of objects (or subject and object). Concretely, let C and P denote the object and relation classes respectively. Each subject or object $e = (e^b, e^c) \in \mathscr{E}$ consists of a bounding box $e^b \in \mathbb{R}^4$ and a class label $e^c \in C$. A relation tuple is a triplet of the form $r = (s, p, o)$ where the subject s and the object o $(s, o \in \mathscr{E})$ are joined by the relation $p \in P$, *e.g.*, *man* ***wearing*** *shirt*. Given an image I, we can then use a set of objects $E = \{e_i\}_{i=1}^m$ and a set of relations $R = \{r_j\}_{j=1}^n$, where m and n are the number of subject/objects and relation triplets in an image respectively, to define a scene graph $S = (E, R)$. A scene graph can also be written as a combination of a set of bounding boxes $B = \{e_i^b\}_{i=1}^m$, a set of class labels $Y = \{e_i^c\}_{i=1}^m$ and a set of relations R.

The SGG models can be decomposed as:

$$P(S|I) = P(B|I)\ P(Y|B, I)\ P(R|B, Y, I) \tag{3.1}$$

where $P(B|I)$ is the object detector or bounding box prediction model, $P(Y|B, I)$ is an object class model and $P(R|B, Y, I)$ is a relation prediction model.

Existing methods (Hung et al., 2020; Li et al., 2017a; Lu et al., 2016a; Tang et al., 2019; Xu et al., 2017; Zellers et al., 2018; Zhang et al., 2017a; Zhuang et al., 2017) often employ a two-step process for the scene graph generation task. First, bounding-box proposals $(P(B|I))$ with class predictions and confidence scores $(P(Y|B, I))$ are extracted using off-the-shelf object detectors (Girshick, 2015; Ren et al., 2015). Then, a multimodal feature fusion model combines visual, language, and spatial features to predict the relation for a given subject-object pair ($P(R|B,Y,I)$). Several methods adopt BiLSTMs (Zellers et al., 2018), Bi-TreeLSTMs (Tang et al., 2019) or fully connected layers (Hung et al., 2020; Zhang et al., 2017a) to encode the co-occurrence between object pairs for relation prediction.

3.3.2 Missing relation labels

Many visual relations are hypernyms, hyponyms, or synonyms (Ramanathan et al., 2015; Yang et al., 2021) and hence are non-mutually exclusive. The standard SGG datasets (*e.g.,* Visual Genome (Krishna et al., 2017)) ignore this fact and only assume *one* annotated label per subject–object pair. Which one is assumed strongly depends on the annotator (manifesting as labeling or language/reporting bias (Misra et al., 2016)).

One way to circumvent this problem is to collect multiple labels for each triplet, which is however expensive and time consuming. Another potential solution is to use linguistic sources such as WordNet (Miller, 1995) or VerbNet (Schuler, 2005) to automatically obtain the missing labels by exploiting the linguistic dependencies between relations. However, this is not trivial, as some of the relation and spatial vocabulary in the SGG datasets are not included in WordNet. Moreover, the context of relations in these language resources does not always allow the right inferences. For instance, in WordNet, *person **riding** horse* does not imply *person **on** horse* (no hyponymy relation), but this is the visual implication in the SGG datasets.

While one can use existing methods (Chiou et al., 2021; Wang et al., 2020b) to infer the missing labels, the estimated labels can be noisy and uninformative such that re-training a model on them may not improve the generalization performance. Here, inspired by previous work (Collell et al., 2018), we propose to group visual relations into two sets: explicit and implicit.[3] Explicit relations encode spatial information between two objects such as ‘on’, ‘in front of’, or ‘under’ and are typically easy to learn, even only based on subject–object locations (Collell et al., 2018). The implicit ones are normally verbs such as ‘riding’, ‘walking’ and for learning them the model

[3] Further details about the explicit and implicit relations are in Section 3.4 and Chapter B.

has to find complex correlations in visual and textual data. In existing SGG datasets, some object pairs are annotated with implicit labels, while other pairs are labeled only with explicit ones and their implicit labels are missing. We propose to divide the set of predicate labels P into two sets, *explicit* and *implicit* and denote them by $\mathbb{E}$ and $\mathbb{I}$ respectively.

3.3.3 Proposed method

In this section, we explain our proposed method for training the relation classifier f_θ to implement $P(R|B,Y,I)$. For each image I, we assume that an object detector provides a set of candidate subject–object pairs $\{(s-o)\}$ and each pair is represented by a d-dimensional joint embedding $\boldsymbol{x} \in \mathbb{R}^d$ including its visual, semantic and spatial features. Note that we apply our method to various existing SGG models which use different object detector and joint embedding functions, and we provide these details in Section 3.4. In particular, f_θ is instantiated as a deep neural network parameterized by θ, takes in a joint feature embedding $\boldsymbol{x}$ for a subject–object pair $s-o$ and outputs a softmax probability over $|P|$ relations, *i.e.,* $f_\theta(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^{|P|}$. Our goal is to learn a relation classifier f_θ that can correctly estimate the relation label of a subject-object pair in an unseen image.

Given a training set $\mathscr{D}$ with $|\mathscr{D}|$ samples, each including tuples of subject-object pairs $s-o$ along with the relation label p and the joint feature embedding $\boldsymbol{x}$, which we denote with $X = \{(s,p,o,\boldsymbol{x})\}_{i=1}^{|X|}$ with $|X|$ tuples. We formulate the learning problem as a minimization of two loss terms:

$$\min_\theta \frac{1}{|\mathscr{D}|} \sum_{i=1}^{|\mathscr{D}|} \left(\mathscr{L}^{\mathbb{I}}(X_i;\theta) + \mathscr{L}^{\mathbb{E}}(X_i;\theta)\right) \tag{3.2}$$

where $\mathscr{L}^{\mathbb{I}}(X_i;\theta)$ and $\mathscr{L}^{\mathbb{E}}(X_i;\theta)$ are the loss terms defined over implicit and explicit relations respectively.

For a given image I and its tuple X, we pick the tuples whose relation is annotated only with implicit relation label (*i.e.,* $X^{\mathbb{I}} = \{(s,p,o,\boldsymbol{x}) \mid p \in \mathbb{I}\}$) and define the implicit

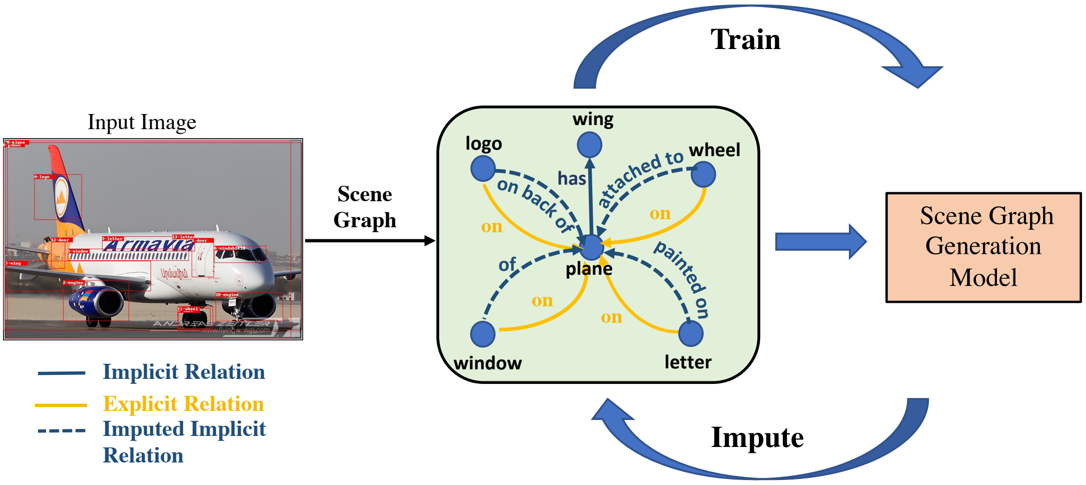


Figure 3.2 Pipeline of our proposed framework for mining informative labels for scene graph generation. An input image can be represented as a scene graph (in green). The yellow and blue solid arrows in the scene graph denote ground-truth explicit and implicit relations, respectively. The scene graph generation model is first trained on a subset of the implicit relations. In an iterative fashion, we then impute implicit relations (dashed blue arrows) for the triplets annotated with explicit relations and train the model with all relations (implicit, explicit, and imputed).

loss term as:

$$\mathscr{L}^{\mathbb{I}}(X^{\mathbb{I}};\theta) = \frac{1}{|X^{\mathbb{I}}|} \sum_{(s,p,o,\boldsymbol{x})\in X^{\mathbb{I}}} \mathscr{L}_{CE}(f_\theta(\boldsymbol{x}),p) \tag{3.3}$$

where $\mathscr{L}_{CE}$ is the cross-entropy loss. In other words, for the implicit relations, we follow the standard practice and compute its loss by using its ground-truth implicit relation label p which is a one-hot vector, as each subject–object pair is annotated with only one label.

Similarly, we formulate the explicit term $\mathscr{L}^{\mathbb{E}}(X^{\mathbb{E}};\theta)$ for the tuples with explicit labels:

$$\mathscr{L}^{\mathbb{E}}(X^{\mathbb{E}};\theta) = \frac{1}{|X^{\mathbb{E}}|} \sum_{(s,p,o,\boldsymbol{x})\in X^{\mathbb{E}}} \mathscr{L}_{KL}(f_\theta(\boldsymbol{x}),\hat{p}) \tag{3.4}$$

where $\mathscr{L}_{KL}$ is the Kullback-Leibler divergence, $X^{\mathbb{E}} = \{(s,p,o,\boldsymbol{x}) \mid p \in \mathbb{E}\}$ and $\hat{p}$ is the *imputed* relation label for the subject-object pair, which is a vector with soft probabilities. Next, we discuss our method to obtain $\hat{p}$.

Label imputation. The subject–object pairs that are annotated with explicit relations only can often be labeled with more informative implicit relation labels. For instance, the ground truth label may be *person* ***beside*** *table*, but *person* ***eating at*** *table* might also be correct in this case, and more informative. To impute the missing implicit relation for a subject–object pair, which is originally annotated with an explicit relation label, we follow a two-step procedure.

First, we take each subject–object pair annotated with an explicit label from $X^{\mathbb{E}}$ along with its joint embedding and impute its implicit label through the relation classifier f_θ as:

$$\bar{p} = \underset{i \in \mathbb{I}}{\arg\max} \left[\frac{\exp(f_\theta^i(\boldsymbol{x}))}{\sum_{j \in \mathbb{I}} \exp(f_\theta^j(\boldsymbol{x}))} \right] \tag{3.5}$$

where f^i denotes its logit for the i-th relation class. In words, we compute softmax probabilities only over implicit relation labels and pick the highest scoring implicit label to obtain a one-hot vector $\bar{p}$. Note that one can also obtain a soft probability vector over all implicit label classes, however, we empirically show that the former works better.

Second, as the subject–object pair is originally labelled with an explicit label p, we also use this information and average the imputed label $\bar{p}$ with its original label p as:

$$\hat{p} = (p + \bar{p})/2. \tag{3.6}$$

We call this step as ***Label Refinement***. As $\hat{p}$ includes equal probabilities (*i.e.,* 0.5) for each of the explicit and implicit labels, it is not a one-hot vector. Thus, we use KL divergence Eq. (3.4) to encourage the model to predict both labels. Compared to standard cross-entropy loss, the KL divergence loss increases the model entropy by reducing overconfidence, resulting in smoother predictions. Most of the traditional methods for SGG trained with cross-entropy may get confused by inconsistent annotations, where the same relation is labeled with less informative spatial relation in some images while more informative labels are used in some other images. Our loss function formulation and multi-label nature of the targets addresses this inconsistency, unlike previous work (Suhail et al., 2021; Tang et al., 2019).

Latent space augmentation. Our training pipeline is illustrated in Fig. 3.2 which follows an alternating optimization and consists of two alternating steps where we employ the relation classifier f_θ to impute the implicit labels and simultaneously optimize the classifier parameters θ. The main challenge here is that the model parameters can overfit to its own imputed labels quickly, resulting in a local optimum solution. This problem is notoriously known as confirmation bias that also occurs in many semi-supervised problems (Arazo et al., 2020; Sohn et al., 2020). To prevent overfitting to the wrong imputed labels, existing solutions include applying various kinds of data augmentation including standard geometric and color transformations (DeVries and Taylor, 2017), their combinations (Cubuk et al., 2018, 2020; Yun et al., 2019) and also generating augmented version of samples (Verma et al., 2019; Zhang et al., 2017b).

As many SGG methods build on the feature space of an object detector, many augmentation strategies that are applied to raw pixels are not applicable to our case. Hence, we use Manifold Mixup (Verma et al., 2019) that generates augmented embeddings (in the manifold space rather than in the pixel space) by taking a convex combination of different pairs of embeddings ($\boldsymbol{x}$ and $\boldsymbol{x}'$) and also their labels (p and p'):

$$\begin{aligned}\tilde{\boldsymbol{x}} &= \lambda.\boldsymbol{x} + (1-\lambda).\boldsymbol{x}' \\ \tilde{p} &= \lambda.p + (1-\lambda).p'\end{aligned} \tag{3.7}$$

where λ is sampled from a beta distribution, *i.e.,* $\lambda \sim \text{Beta}(\alpha, \alpha)$ with α as a hyper-parameter. Note that we apply this augmentation to the whole training set and allow mixing embeddings across samples from both the implicit and explicit set of relations. This augmentation acts as a regularizer and accounts for overfitting to the incorrect imputed labels while training.

3.3.4 Algorithm

In Algorithm 1, we detail our training pipeline. To obtain the initial parameters θ_0, we first train our model on the tuples with only implicit labels as following (Line 2):

$$\theta_0 = \arg\min_{\theta} \frac{1}{|\mathscr{D}|} \sum_{i=1}^{|\mathscr{D}|} \mathscr{L}^{\mathbb{I}}(X_i; \theta). \tag{3.8}$$

The key intuition behind this is that model learned on only the implicit relations are more likely to produce confident predictions over implicit labels and hence not 'distracted' by explicit relation labels.

After the training on implicit relations, we iteratively impute implicit relation labels for the subject-object pairs annotated with the explicit relations (Line 5) and update the model parameters using Eq. (3.2) (Line 7). The model parameters optimized in Eq. (3.2) take as input the augmented versions of the sample and label pair (as in Eq. (3.7)) for both the implicit and explicit set of relation labels (Line 6).

Algorithm 1 Our proposed optimization of the SGG model

```
1: Input: Training set $\mathscr{D}$ with $|\mathscr{D}|$ samples, each including a set of tuples $X$ with
   $(s, p, o, \boldsymbol{x})$ with subject, object and relation label, joint embedding resp., Explicit
   and implicit relation sets $\mathbb{E}$ and $\mathbb{I}$ resp., relation classifier $f_\theta$, $T$ number of iterations,
   $\eta$ learning rate.
2: Initialize $\theta$ as in Eq. (3.8)
3: for $t = 0, \ldots, T$ do
4:     Sample a minibatch $B = (X_1, \ldots, X_{|B|}) \sim \mathscr{D}$,
5:     Impute $\hat{p}$: Impute implicit labels $\hat{p}_t$ for $B^{\mathbb{E}}$ by using Eq. (3.5) and Eq. (3.6),
6:     Augment $B$ by applying manifold mixup in Eq. (3.7),
7:     Update $\theta$: $\theta_{t+1} \leftarrow \theta_t + \eta\Delta_\theta$ where $\Delta_\theta$ is the update for $\theta$ obtained from
   Eq. (3.2),
8: end for
9: return $\theta$
```

3.4 Experiments

Dataset and evaluation settings. We evaluate our proposed method for scene graph generation on the Visual Genome (VG) (Krishna et al., 2017) dataset. We use the pre-processed version of the VG dataset as proposed in Xu et al. (2017). The dataset consists of 108k images with 150 object categories and 50 relation categories. The training, test, and validation split used in the experiments also follow previous work (Suhail et al., 2021; Tang et al., 2020; Xu et al., 2017).

For evaluation on the **Visual Genome dataset**, we follow Xu et al. (2017) and report performance on three settings: **(1) Predicate Classification (PredCls).** This task measures the accuracy of relation (also termed as predicate in literature) prediction when the ground truth object classes and boxes are given. It is not affected by the object detector accuracy. **(2) Scene Graph Classification (SGCls).** In this setting, we know the ground truth boxes and we have to predict the object classes and the relations between them. **(3) Scene Graph Detection (SGDet).** This is the most challenging setting and the models are used to predict object bounding boxes, object classes and the relations between them. We measure Mean Recall@K (mR@K) (Tang et al., 2019, 2020) to evaluate scene graph generation models. More recent work has preferred mR@K over regular Recall@K (Xu et al., 2017) due to data imbalance (Tang et al., 2020). Mean Recall@K treats each relation separately and then averages Recall@K over all relations. We also measure the **zero-Shot Recall**, zsR@K, for three settings of PredCls, SGCls, and SGDet, which helps to evaluate the generalization ability of the model in predicting subject–relation–object triplets not seen during training.

Model generalization. Our proposed framework has the flexibility to be trained with any scene graph generation model. Hence, we train with different model architectures to demonstrate the generalizability of our approach: Iterative Message Passing (IMP) (Xu et al., 2017), Neural-Motifs (Zellers et al., 2018) and VCTree (Tang et al., 2019). We also train with two other debiasing methods that build upon these models, Energy-based Modeling (EBM) (Suhail et al., 2021), where the authors propose to train with an additional energy-based loss and Total Direct Effect (TDE). where counterfactual reasoning is used during inference.

Explicit and Implicit relation labels. For the Visual Genome dataset (Krishna et al., 2017), Xu et al. (2017) released a version of the dataset with 50 relations which are: *above, across, against, along, and, at, attached to, behind, belonging to, between,* *carrying, covered in, covering, eating, flying in, for, from, growing on, hanging from, has, holding, in, in front of, laying on, looking at, lying on, made of, mounted on, near, of, on, on back of, over, painted on, parked on, part of, playing, riding, says, sitting on, standing on, to, under, using, walking in, walking on, watching, wearing, wears, with.*

Inspired by (Collell et al., 2018), we divide the relation label space into a set of explicit and implicit relations. We define explicit relations when the spatial arrangement of objects is implied by the label itself *e.g.,* “on", “below", “next to” and so on. For implicit relations, the spatial arrangement of objects is only indirectly implied, “riding”, “walking”, “holding" etc. More specifically, the explicit relations are *above, across, against, along, at, behind, between, in, in front of, near, on, over, under*, and the rest are treated as implicit relations.

Implementation details. Following previous work (Tang et al., 2019, 2020), we train the scene graph generation models on top of the pre-trained Faster R-CNN object detector with ResNetXt-101-FPN backbone (Ren et al., 2015). The weights of the SGG models' object detector are frozen during training in all three settings – PredCls, SGCls, and SGDet. The mAP of the object detector on the Visual Genome dataset is 28% using 0.5 IoU. For training each scene graph generation model using our proposed method, we use the default training settings as in (Suhail et al., 2021; Tang et al., 2020) for fair comparisons. The models are trained with the SGD optimizer with a batch size of 12, an initial learning rate of 10^{-2} and 0.9 momentum.

The models are trained for the first 30,000 batch iterations on the implicit label subset with the standard cross-entropy loss. After label imputation, the model is trained on the rest of the imputed data and the implicit subset for another 20,000 batch iterations. The value of α is set to 4 from which λ is sampled for the mixing function in the latter half of training. Our code is available here[4].

3.5 Results

[4] https://groups.inf.ed.ac.uk/vico/research/NARE

		Predicate Classification			Scene Graph Classification			Scene Graph Detection		
Models	Method	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
IMP (Chen et al., 2019b; Xu et al., 2017)	-	-	9.8	10.5	-	5.8	6.0	-	3.8	4.8
FREQ (Tang et al., 2019; Zellers et al., 2018)	-	8.3	13.0	16.0	5.1	7.2	8.5	4.5	6.1	7.1
Motifs (Zellers et al., 2018)	-	10.8	14.0	15.3	6.3	7.7	8.2	4.2	5.7	6.6
KERN (Chen et al., 2019b)	-	-	17.7	19.2	-	9.4	10.0	-	6.4	7.3
VCTree (Tang et al., 2019)	-	14.0	17.9	19.4	8.2	10.1	10.8	5.2	6.9	8.0
VCTree-L2+cKD (Wang et al., 2020b)	-	14.4	18.4	20.0	9.7	12.1	13.1	5.7	7.7	9.0
IMP (Xu et al., 2017)	Baseline	8.9	11.0	11.8	5.4	6.4	6.7	2.2	3.3	4.1
	Ours	**12.3**	**14.6**	**15.3**	**7.1**	**8.0**	**8.3**	**6.9**	**7.8**	**8.1**
Motif-TDE-Sum (Tang et al., 2020; Zellers et al., 2018)	Baseline	17.9	24.8	28.7	9.8	13.2	15.1	6.6	8.9	11.0
	Ours	**21.3**	**27.1**	**29.7**	**11.3**	**14.3**	**15.7**	**8.4**	**10.4**	**11.8**
VCTree (Tang et al., 2019)	Baseline	13.1	16.5	17.8	8.5	10.5	11.2	5.3	7.2	8.4
	Ours	**18.0**	**21.7**	**23.1**	**11.9**	**14.1**	**15.2**	**7.1**	**8.2**	**8.7**
VCTree-EBM (Suhail et al., 2021)	Baseline	14.2	18.2	19.7	10.4	12.5	13.4	5.7	7.7	9.1
	Ours	**21.0**	**24.9**	**26.5**	**14.0**	**16.2**	**17.1**	**7.8**	**10.1**	**11.8**
VCTree-TDE (Tang et al., 2020)	Baseline	16.3	22.9	26.3	11.9	15.8	18.0	6.6	9.0	10.8
	Ours	**22.2**	**28.1**	**30.6**	**17.8**	**22.0**	**23.6**	**8.4**	**10.3**	**11.5**

Table 3.1 Scene Graph Generation performance comparison on mean Recall@K (Tang et al., 2020) under all three experimental settings. We compare the results of our proposed framework (Ours) with the original model (Baseline) using different SGG architectures.

		PredCls	SGCls	SGDet
Models	Method	zsR@20/50	zsR@20/50	zsR@20/50
IMP (Xu et al., 2017)	Baseline	**12.17/17.66**	**2.09/3.3**	0.14/0.39
	Ours	7.12/10.50	1.57/2.32	**1.52/2.48**
Motif-TDE-Sum (Tang et al., 2020)	Baseline	8.28/14.31	**1.91**/2.95	1.54/2.33
	Ours	**9.33/14.43**	1.87/**2.99**	**2.06/3.05**
VCTree (Tang et al., 2019)	Baseline	1.43/**4.0**	**0.39/1.2**	0.19/0.46
	Ours	**1.51**/3.7	0.36/1.0	**0.43/0.95**
VCTree-TDE (Tang et al., 2020)	Baseline	8.98/**14.52**	3.16/4.97	1.47/2.3
	Ours	**9.11**/13.52	**4.26/6.20**	**2.24/3.25**

Table 3.2 Zero shot recall performance for our proposed method compared with the original model (baseline).

3.5.1 Quantitative results

Table 3.1 compares the performance of the state-of-the-art methods when trained with our proposed training framework on the Visual Genome dataset (Krishna et al., 2017). Our method consistently improves performance on all three evaluation settings (PredCls, SGCls, SGDet) when trained with existing methods. With IMP (Xu et al., 2017) and Motif-TDE-Sum (Tang et al., 2020; Zellers et al., 2018), we obtain an absolute improvement of 3.4% and 1.7% in PredCls and SGCls, respectively. For the most challenging setting of SGDet, there is an improvement of 4.7% with the IMP model, showing the generalization ability of our approach on a visual-only model (no language/textual features). When debiasing approaches such as TDE (Tang et al., 2020) and EBM (Suhail et al., 2021) are incorporated, we obtain consistent improvements

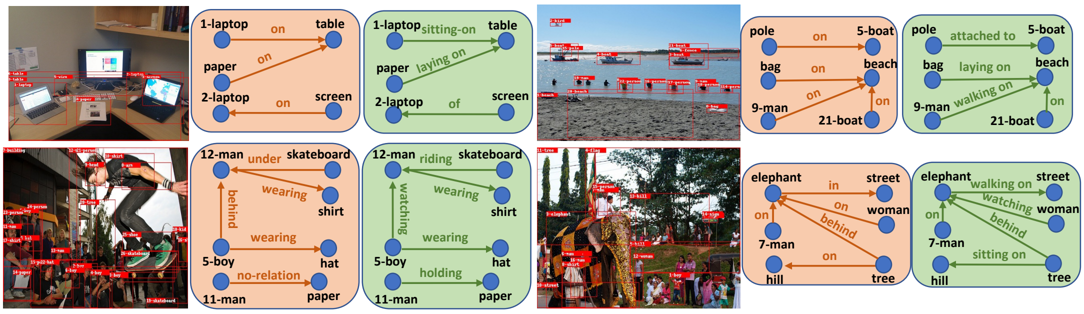


Figure 3.3 Visualization of scene graphs generated by the VCTree-EBM (Suhail et al., 2021) based learning framework (in orange) and our proposed method using VCTree-EBM as the backbone (in green).

for VCTree. In both cases, we gain significantly in all the settings and achieve a new state-of-the-art performance in scene graph generation by combining our proposed method with VCTree-TDE.

In Table 3.2, we report the zero-shot recall performance and compare it with baselines. Our proposed method achieves improvements in most of the settings with different SGG backbones, except for IMP in PredCls and SGCls. IMP being a visual-only model fails to learn correlations via textual features for different relation classes and hence performs poorly in zero-shot, due to low recall on explicit relations. However, the multi-modal nature of Motif and VCTree brings out the strength of our method in generalizing to unseen triplets during test time.

	Predicate Classification			
Method	Train Label	Imputed with	Imputed on	mR@20/50
Baseline	All	-	-	17.85/24.75
	Explicit	-	-	14.06/20.34
	Random	-	-	16.99/23.33
	Implicit	-	-	18.24/24.93
Ours	Random	Top1	Random	17.11/23.56
	Explicit	Top1-Explicit	Implicit	14.23/20.51
	Implicit	Top1-Implicit	Explicit	**21.26/27.14**

Table 3.3 Experimental results on the Predicate Classification setting with different ways of label imputation.

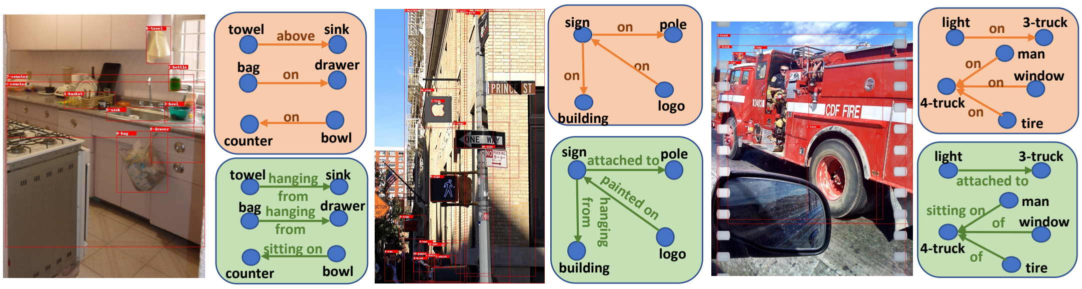


Figure 3.4 Scene Graph Visualization results for the missing *implicit* relations imputed for the triplets annotated with *explicit* relations. The scene graphs in orange are the ground-truth triplets with their corresponding imputed *implicit* relations in green.

3.5.2 Ablative analysis

In this section, we study different components of our method separately to validate their effectiveness. Table 3.3 evaluates the contribution of our proposed label imputation strategy. All the models are trained using the Motif-Sum (Zellers et al., 2018) backbone with TDE (Tang et al., 2020) at inference time, Motif-TDE-Sum. *Random* implies that we randomly divide the relation labels into two sets without the knowledge of *explicit* and *implicit* relations. The results in the rows for the method *Baseline* show the effect of training with *implicit* relations on scene graph generation performance. If the model is trained with the *explicit* relations only, mean recall drops. There is also a marginal drop in performance compared to training with all relations when trained on a random subset of the relations. Even training on a subset of the data with only implicit relations, we can achieve at-par performance compared to training with all relations.

We also compare the results under settings when the model is trained on a particular subset of *Train Label* with different imputation strategies (results in rows with the method *Ours*). The relation labels are imputed on the hold-out relation set with the top-1 (best) in the set of labels that it was trained on. Training and imputing with the random or explicit set of labels decreases performance compared to the baseline, with a significant drop when only explicit relations are considered. This shows the importance of learning with and imputing implicit relations: they provide useful information about interactions between object pairs not captured by explicit relations.

	PredCls			
Method	Mixup	Soft/Hard Labels	Refinement	mR@20/50
Baseline	✗	✗	✗	17.85/24.75
	✓	✗	✗	17.43/24.20
Ours	✗	Hard	✗	18.26/24.23
	✓	✗	✓	18.90/25.32
	✗	Hard	✓	20.81/26.78
	✓	Hard	✗	19.90/26.35
	✓	Soft	✓	20.76/27.10
	✓	Hard	✓	**21.26/27.14**

Table 3.4 Ablation study on our proposed training framework with Motif-TDE-Sum (Tang et al., 2020; Zellers et al., 2018) as the SGG backbone.

In Table 3.4, we study whether our method improves over the baselines mostly due to training with *Latent Space Augmentation* (Manifold Mixup) or combination of multiple components. Applying manifold mixup on either the baseline (row 2) or with only explicit ground-truth labels without imputation (row 4) does not provide significant gains. This indicates that mixup is effective only when applied to the imputed implicit labels. This hints that mixup can reduce overfitting to noisy imputations and allow for imputing more informative labels. We also show that the performance of soft imputed labels (row 7) is very similar to our proposed method (with hard labels, row 8). Using hard labels reduces noise in the predictions and encourages the model to learn from more high-confident predictions.

In Table 3.5, we show the results for training on relation subsets and our final method on the explicit and implicit set of relations separately. As discussed in Section 3.1, when the model is only trained on explicit relations, it fails to generalize to implicit relations. This is in contrast to training only on the implicit set of relations. This indicates that implicit relations are rich in information and perhaps learn complex and generalizable features. Our final method outperforms training on all relations (original model) and the subset of relations by a significant margin, showing the strength of mining informative labels for less informative samples.

Method	PredCls - mR@50/100		
	Train Relations ↓/Test Relations →	Explicit	Implicit
MOTIF-TDE-Sum	All Relations	24.47/28.79	24.96/28.74
	Explicit only	22.89/25.34	0.08/0.09
	Implicit only	20.10/22.89	24.34/26.03
	Ours (final)	**24.83/27.80**	**27.99/30.38**

Table 3.5 Performance Comparison on the Explicit and Implicit set of relations separately with different subsets of training labels.

3.5.3 Qualitative results

Figure 3.3 visualizes the scene graphs predicted from the baseline VCTree-EBM (Suhail et al., 2021) model (in orange) and compares it to the scene graphs obtained via our proposed training framework (in green). Our method consistently predicts more informative relations such as ***laying on, walking on, holding*** instead of simple prepositional relations such as ***on, in***. Moreover, our method also effectively identifies triplets with relations that were missed in the baseline. For instance, in the bottom-left image, our method localizes *man* ***holding*** *paper* correctly. Our method also corrects relations that are incorrectly predicted in the baseline, *woman* ***watching*** *elephant* as opposed to *woman* ***on*** *elephant* in the bottom-right image.

In Figure 3.4, we visualize the imputed *implicit* relations for the triplets annotated with explicit relations. In orange, we show the ground-truth scene graphs and the corresponding imputed scene graphs in green. It can be clearly observed from the ground-truth scene graphs that there is annotator bias towards spatial relations. Our label imputation strategy is able to find alternate and missing implicit relations for these triplets and exploit them during training. For instance, our method imputes important relations such as ***attached to, hanging from, sitting on*** which are more descriptive than their explicit counterpart ***on***. This shows the importance of label imputation and generating descriptive scene graphs for comprehensive scene understanding.

3.6 Conclusion and discussion

In this chapter, we proposed a novel model-agnostic training framework for scene graph generation. We introduced the concept of label informativeness, which had not been explored in SGG before. A model trained on informative relations is able to model the visual and textual context better compared to training on simple spatial relations. We showed how to impute informative relations from the partially labeled data and jointly train with imputed and ground truth relations. We improved the performance across models and tasks, including in a zero-shot setting.

There are a few limitations that should be investigated in future work. One limitation of our approach is its limited ability in predicting implicit relations with very few samples. More specifically, the distribution of labels for the implicit subset of relations is long-tailed, with some predicates (*e.g.,* wearing) having more samples than others (*e.g.,* growing on). Hence, when the model is trained on the implicit relation set, it might not always assign the low-frequency relations as pseudo-labels for the explicit set. This can lead to bias towards the frequently occurring implicit relations, leading to low generalizability for the entire set of informative predicates. Secondly, the division of the label space into implicit and explicit relations is based on heuristic rules and is specific to a particular dataset, in our case the Visual Genome dataset. For training and inference on a different dataset with a unique set of relations, the implicit and explicit relations will need to be re-defined. In the future, learning this split of relations would be a more scalable and effective solution.

Addressing some of the limitations mentioned above, our work has steered and impacted the progress in scene graph generation by acting as strong baselines on a lot of follow-up work (Adaimi et al., 2023; Li et al., 2023b; Yu et al., 2023). Recent works have proposed an orthogonal approach to relying on lexical rules for label space learning by generating more informative triplets using Large Language Models (LLMs) (Sanderson, 2023) and subsequently train a scene graph generation method on these newly generated informative labels (Yu et al., 2023).

Chapter 4

Who are you referring to? Coreference resolution in image narrations

In Chapter 2 and Chapter 3, we develop models for the tasks of image captioning and scene graph generation. While both of these tasks are undoubtedly crucial for bridging the gap between the modalities of vision and language, they are limited in their capabilities to establish a connection between these two modalities, which is indispensable for building AI systems with advanced capabilities. Take the conversation below between a human and an agent:

Human: Do you see a **refrigerator**?
Agent: Yes, I do.
Human: In **it** there is a **can of soda**, can you bring me **that**?

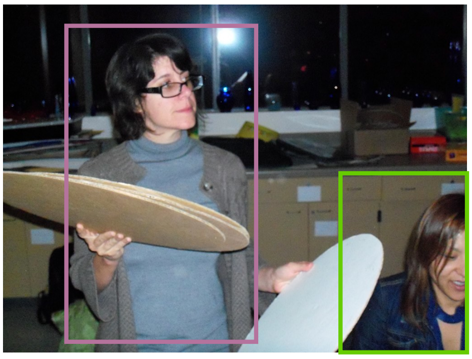
To understand the context in the above example, we need to resolve the context in language *i.e., it* refers to the *refrigerator* and *that* refers to the *can of soda.* Not only this but once we have established the context in language, we also need to link the phrases in the language (*e.g.,* can of soda) to the visual regions for the agent to act on it. Hence, to further enhance the capabilities of vision and language models, in this chapter, we move from graphs to narrations by learning correlations in the language (coreference resolution) and the association between language and images *i.e.,* grounding, also known as multimodal coreference resolution.

Coreference resolution aims to identify words and phrases which refer to the same entity in a text, a core task in natural language processing. In this chapter, we extend this task to resolving coreferences in long-form narrations of visual scenes. First, we introduce a new dataset with annotated coreference chains and their bounding boxes, as most existing image-text datasets only contain short sentences without coreferring expressions or labeled chains. We propose a new technique that learns to identify coreference chains using weak supervision, only from image-text pairs and a regularization using prior linguistic knowledge. Our model yields large performance gains over several strong baselines in resolving coreferences. We also show that coreference resolution helps improve grounding narratives in images.

This chapter begins with an introduction of multimodal coreference resolution and the key ideas of our proposed method in Section 4.1, related work in Section 4.2, our proposed Coreferenced Image Narratives (CIN) dataset in Section 4.3, details of our proposed weakly supervised method with lexical knowledge in Section 5.3, experimental details in Section 4.5, evaluating and analyzing the proposed method on the CIN dataset in Section 4.6 and conclusion and discussion in Section 4.7.

4.1 Introduction

Consider the image paired with the long-form description in Figure 4.1, an example from the Localized Narratives (Pont-Tuset et al., 2020). Can you tell whether *the woman* who is wearing spectacles refers to ***a person*** or ***another woman*** in the text? We are remarkably good at identifying referring expressions (or mentions) and determining which of them corefer to the same entity, a task that we regularly perform when we read text or engage in conversation. The text-only version of this problem is known as coreference resolution (CR) (Lee et al., 2011, 2017; Sukthanker et al., 2020), a core task in natural language processing (NLP) with a large literature. While solving text-only CR requires a very good understanding of the syntactic and semantic properties of the language, the visual version of CR shown in the example also demands an understanding of the visual scene. In our example, a language model has to figure



In the image we can see there is a person who is standing and holding cardboard sheets in her hand and she is wearing ash colour jacket and there is another woman sitting and at the back on the table there are wine bottles and cardboard boxes and books and the woman is wearing spectacles.

Figure 4.1 Coreference resolution from an image and narration pair. Each highlighted block of text is referred to as a *mention*. The mentions in the same color corefer to the same entity, and belong to the same coreference chain.

out that *a person* can be a woman, has hands, and correctly match it with *her [hand]* and *the woman*, but not with *another woman*. However, a language model alone cannot answer whether *the woman* refers to *a person* or *the woman*. This can only be disambiguated after visually inspecting which of the two *is wearing spectacles*.

Text-only CR has been a crucial component for a range of NLP applications including question answering (Das et al., 2017; Kwiatkowski et al., 2019), sentiment analysis (Cambria et al., 2017; Medhat et al., 2014), summarization (Gupta and Lehal, 2010; Shi et al., 2021) and machine translation (Bahdanau et al., 2014; Lopez, 2008; Wu et al., 2016). Most text-only CR methods are either rule-based (Lee et al., 2011; Raghunathan et al., 2010) using heuristics such as pronoun match or exact match based on part of speech tagging or are learned on large annotated text datasets from domains such as news text or Wikipedia articles (Bengtson and Roth, 2008; Joshi et al., 2020; Lee et al., 2017, 2018a). State-of-the-art methods (Joshi et al., 2020; Lee et al., 2017) fail to resolve coreferences correctly in image narrations for a few reasons. First, CR in image narrations often requires image understanding (see Fig. 4.1). Neural networks trained on text datasets (Chen et al., 2018; Pradhan et al., 2012) suffer from poor transferability and a significant performance drop when applied to image narrations because of domain shift. Image narrations are unstructured and can be noisy, unlike the well-edited text used during training (such as news or Wikipedia). Moreover, standard image-text datasets (Changpinyo et al., 2021b; Krishna et al., 2017; Lin et al., 2014;

Plummer et al., 2015) only contain short descriptions with very few or no cases of coreference, thus, are not suitable for training text-only CR models.

Some prior work have looked at visual CR for specific tasks. Ramanathan et al. (2014) and Rohrbach et al. (2017) link character mentions in TV shows or movie descriptions to character occurrences in videos. More recently, the Who's Waldo dataset (Cui et al., 2021) links person names in the caption to their occurrence in the image. However, these methods rely on a limited set of object categories and referring expression types (see Table 4.3 discussed below), require annotated training data, and therefore cannot be applied to long-form unconstrained image narrations that include open-world object categories and multiple types of referring expressions such as pronouns (*she*), common nouns (*another woman*), or proper nouns (*Peter*).

In this chapter, we look at *the problem of CR in image narrations*, *i.e.,* resolving the coreference of mentions in narrative text that is paired with an image. As the prior benchmarks in this domain are limited to either a small vocabulary of objects or specific referring expression types, we introduce a new dataset, Coreferenced Image Narratives , *CIN* which augments the rich long-form narrations in the existing Localized Narratives dataset (Pont-Tuset et al., 2020). We add coreference chain annotations and ground each chain by linking it to a bounding box in the corresponding image.

Manually annotating the whole dataset (Pont-Tuset et al., 2020) is expensive, hence these annotations are provided only for evaluation and are not available for training. To cope with this setting, we propose a weakly supervised CR method that learns to predict coreference chains from only paired image-text data. Our key idea is to learn the linking of the mentions to image regions in a joint multi-modal embedding space and use the links to form coreference chains during training. To this end, we propose a multimodal pipeline that represents each modality (image regions, text mentions and also mouse traces, additionally provided by Pont-Tuset et al. (2020)) with a modality-specific encoder and then exploit the cross-modal correlations between them to resolve coreference. Finally, inspired by the rule-based CR (Lee et al., 2011), we incorporate linguistic rules into our learning formulation in a principled way. We report extensive experiments on CIN and demonstrate that our method brings significant improvements in CR and in weakly supervised narrative grounding, a form of disambiguation that has been underexplored in visual grounding[1].

To summarize our contributions, we introduce (1) the new task of resolving coreferences in multimodal long-form textual descriptions (narrations), (2) a new dataset, CIN , that enables the evaluation of coreference chains in text and the localization of bounding boxes in images, which is provided with multiple baselines and detailed analysis for future work, (3) a new method that learns to resolve coreferences while jointly grounding them from weak supervision and exploiting linguistic knowledge, (4) a rigorous experimental evaluation showing significant improvement over the prior work not only in CR but also in weakly supervised grounding of complex phrases in narrative text.

4.2 Related work

Text-only CR in NLP has a long history of rule-based and machine learning-based approaches. Early methods (Hobbs, 1978; Raghunathan et al., 2010) used hand-engineered rules to parse dependency trees, which outperformed all learning-based methods at the time. Recently, neural network methods (Clark and Manning, 2016; Joshi et al., 2020; Lee et al., 2017; Wiseman et al., 2016, 2015) have achieved significant performance gains. The key idea is to identify all mentions in a document using a parser and then learn a distribution over all the possible antecedents for each mention. SpanBERT (Joshi et al., 2020) uses a span-based masked prediction objective for pre-training and shows improvements on the downstream task of CR. Stolfo et al. (2022), on the other hand, transfer the pre-trained representations using rules for CR. It is worth noting that all these learning-based approaches either require large pretraining data or training data annotated with gold standard (ground-truth) coreference chains, such as OntoNotes (Pradhan et al., 2012) or PreCo (Chen et al., 2018).

Visual CR includes learning to associate people or characters mentioned in the text with images or videos (Cui et al., 2021; Ramanathan et al., 2014; Rohrbach et al.,

[1] Our code and dataset is available at https://github.com/VICO-UoE/CIN.

2017). Kong et al. (2014) exploit CR to relate texts to 3D scenes. Another direction is to resolve coreferences in visual dialog (Kottur et al., 2018) for developing better question-answering systems. Unlike these works, we focus on learning coreferences from long unconstrained image narrations using weak supervision. A related group of work (Deng et al., 2021; Li and Sigal, 2021; Yu et al., 2018b,c) aims to ground phrases in image parts. In visual phrase grounding (Chen et al., 2017; Deng et al., 2021; Kamath et al., 2021; Li et al., 2022b; Liu et al., 2019a; Yu et al., 2016, 2018b), the main objective is to localize a single image region given a textual query. These models are trained on visual grounding datasets such as ReferItGame (Kazemzadeh et al., 2014), Flickr30K Entities (Plummer et al., 2015), or RefCOCO (Yu et al., 2016). However, due to short captions, the grounding of text boils down to mostly salient objects in images. In contrast, grounding narrations which aim at capturing all image regions is significantly more challenging and cannot be effectively solved with those prior methods.

Weakly supervised grounding, learning to ground from image-text pairs only, has recently been used in (Liu et al., 2019b,c, 2022, 2021; Wang et al., 2021) for referring expression grounding. These methods use phrase reconstruction from visual region features as a training signal. Other methods (Datta et al., 2019; Gupta et al., 2020; Wang et al., 2020a) use contrastive learning by creating many negative queries (based on word replacement) or by mining negative image regions for a given query. Wang et al. (2020a) is a strong method in this domain, hence we establish it as a baseline in our experiments. Liu et al. (2021) parses sentences to scene graphs for capturing visual relation between mentions to improve phrase grounding. However, this cannot be directly applied to our task, as parsing scene graphs from narrations are typically very noisy and incomplete. Wang et al. (2021) aims to learn/predict object class labels from the object detector during training and inference respectively. Due to the open-vocabulary setting in our dataset, we directly rely on predictions from the detector and use them as features to avoid the complexity of open-vocabulary object detection. Furthermore, as we show in the experiments that grounding is useful to anchor mentions but it is not sufficient to resolve coreferences without prior linguistic knowledge. Thus, our method also employs contrastive learning but for learning CR from weak supervision.

4.3 Coreferenced Image Narratives

Our CIN dataset contains 1880 images from the Localized Narratives dataset (Pont-Tuset et al., 2020) that come with long-form text descriptions (narrations) and mouse traces. These images are originally a subset of the test and validation set of the Flickr30k dataset (Plummer et al., 2015). We annotated this subset with coreference chains and bounding boxes in the image that are linked with the textual coreference chains, and use them only for validation and testing. Note that we also include singletons (*i.e.,* coreference chains of length one). Fig. 4.1 shows an example image from CIN .

Annotation procedure. The annotation involved three steps: (1) marking the mentions (sequences of words) that refer to a localized region in the image, (2) identifying coreference chains for the marked mentions, including (a) pronominal words such as *him* or *who* that are used to refer to other mentions, (b) mentions that refer to the same entity (*e.g., a lady* and *that person*), and (c) mentions that do not have any links (*e.g., another woman*), (3) drawing bounding boxes in the image for the coreference chains/mentions identified in steps (1) and (2). We created an annotation interface based on LabelStudio (lab), an HTML-based tool that allows us to combine text, image, and bounding box annotation. More details are provided in Chapter C.

Dataset	#noun phrases	#pronouns	#coreference chains	#bounding boxes
Flickr30k Entities (Plummer et al., 2015)	15,252	✗	✗	17,234
RefCOCO (Yu et al., 2016)	10,668	✗	✗	10,668
CIN (Ours)	19,587	1,659	3,310	21,246

Table 4.1 Statistics of relevant noun phrases, pronouns, coreference chains and bounding boxes on Flickr30k Entities (Plummer et al., 2015), RefCOCO Yu et al. (2016) and CIN .

Dataset statistics. We split the 1880 images in the dataset into a test and validation set using the pre-defined split of Plummer et al. (2015). More specifically, we have 1000

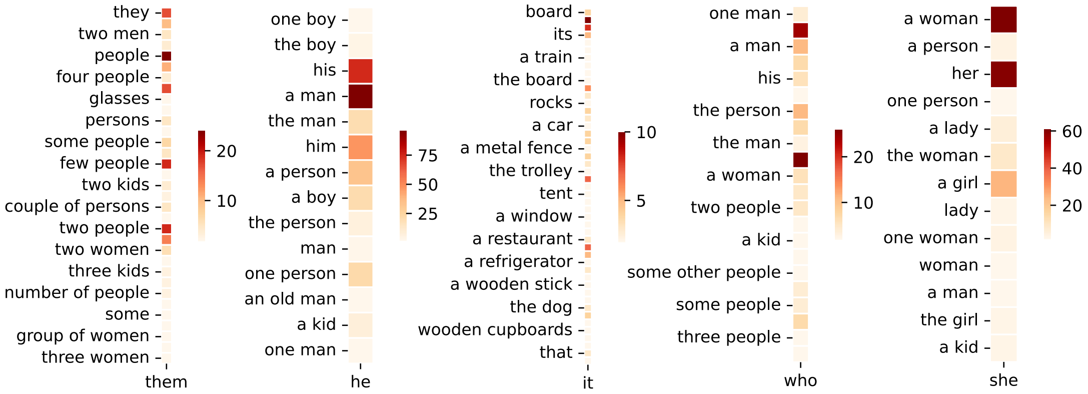


Figure 4.2 Numbers of mentions as part of the coreference chain for pronouns *them, he, it, who, she* in CIN .

images in the test set and 880 images in the validation set. It is important to note that the narrations have a lot of first person pronouns such as *I can see* . . . . We specifically instruct the annotators to exclude such mentions that are not a part of any coreference chain and at the same time cannot be grounded on the image. We elaborate more on the filtering process for these mentions in Chapter C.

Overall, the dataset has 19,587 noun phrase mentions, 1,659 pronouns, 3,310 coreference chains, and 21,246 bounding boxes. In Table 4.1, we compare the statistics of CIN with the test/val splits of other related datasets. In Fig. 4.2, we show the distribution over the frequency and types of mention such as *a metal fence* or *few people* that are referred to using a particular pronoun (*them, he, it, who* and *she*). There is a huge diversity in (1) the categories of the mentions and (2) how many times they form a part of the coreference chain.

In Table 4.2, we also report the frequency of the top-6 mentions for each pronoun category to check the diversity of pronoun words in the dataset.

he	she	who	them	it
a man (0.28)	a woman (0.33)	a person (0.18)	people (0.11)	a board (0.06)
his (0.21)	her (0.32)	one person (0.16)	few people (0.08)	a dog (0.05)
him (0.13)	a girl (0.13)	a man (0.07)	two people (0.08)	a building (0.04)
a person (0.09)	the woman (0.04)	the person (0.07)	they (0.08)	which (0.04)
one person (0.06)	a lady (0.03)	him (0.04)	a few people (0.08)	a wall (0.04)
a boy (0.06)	a person (0.02)	her (0.04)	their (0.06)	a vehicle (0.03)

Table 4.2 Frequency of top-N mentions for each pronoun category.

Comparison to existing CR datasets. In Table 4.3, we compare our proposed CIN dataset to other CR datasets. This comparison shows that most of the other datasets are either from a restricted domain (*shopping, indoor scenes, etc.*), have limited mention types referring to either only *people* or *limited object categories*, or do not cover all possible referring expression types such as common nouns (*a person*) and pronouns (*he*).

Dataset	Modality	Domain	Object categories	Referring expression types
NYU-RGBD v2 (Kong et al., 2014)	Images	Indoor home scenes	Household objects	Common nouns
SIMMC 2.0 (Guo et al., 2022)	Images	Shopping	Clothing	Common nouns
MPII-MD (Rohrbach et al., 2017)	Videos	Movies	People	Proper names, Pronouns
Who's Waldo (Cui et al., 2021)	Images	WikiMedia	People	Proper names
CIN (Ours)	Images	Open-world	General objects	Common nouns and Pronouns

Table 4.3 Comparison to existing datasets.

4.4 Method

4.4.1 Text-only CR

Given a sentence containing a set of mentions (*i.e.,* referential words or phrases), the task of CR is to identify which mentions refer to the same entity. This is fundamentally a clustering problem (Sukthanker et al., 2020). In this work, we use an off-the-shelf NLP parser (spa) to obtain the mentions. Formally, let $S = \{m_1, m_2, \dots, m_{|S|}\}$ denote a sentence with $|S|$ mentions, where each mention m contains a sequence of words, $\{w_1, w_2, \dots, w_{|m|}\}$. We assign a label y_{ij} to each mention pair (m_i, m_j), which is set to 1 when the pair refers to the same entity, and to -1 otherwise. We wish to learn a compatibility function, a deep network f that scores high if a pair refers to the same entity, and low otherwise.

Given a training set $\mathscr{D}$ that contains $|\mathscr{D}|$ sentences with their corresponding labels, one can learn f by optimizing a binary cross-entropy loss:

$$\min_f \sum_{S\in\mathscr{D}} \sum_{i=0}^{|S|-1} \sum_{j=i+1}^{|S|} \log(y_{ij}(\sigma(f(m_i, m_j))) - \frac{1}{2}) + \frac{1}{2}) \tag{4.1}$$

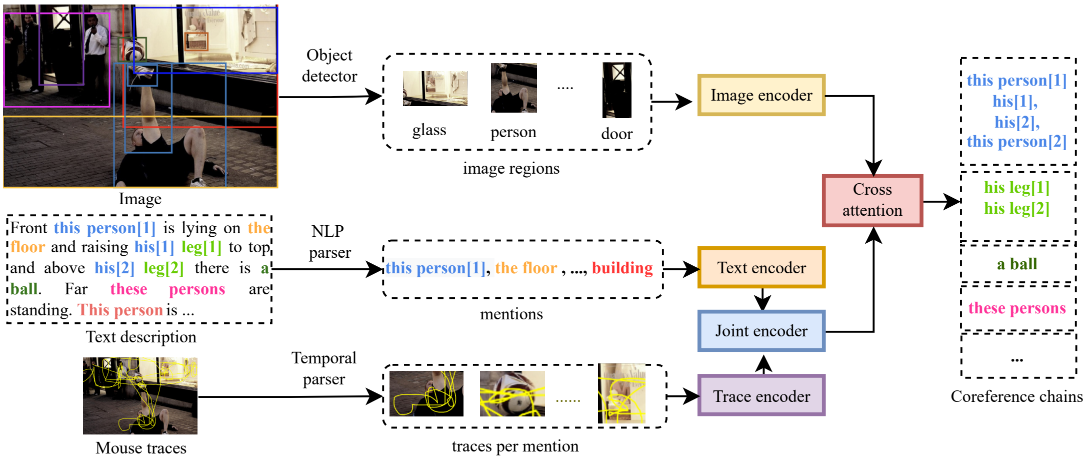


Figure 4.3 Overview of our pipeline. Our model encodes the image regions obtained from an object detector using the image encoder. We parse text mentions and mouse traces from the sentence description, which are then encoded using a text and trace encoder respectively. Finally, a joint text-trace encoder learns a joint embedding for text and traces. A cross-attention module attends to the words given an image region and then we compute the joint probability of the paired mentions, thus forming coreference chains.

where σ is the sigmoid function. Note that prior methods (Joshi et al., 2020; Lee et al., 2011, 2017) require large labeled datasets for training and are limited to only a single modality, text. These methods typically also combine the learning with fixed rules based on recency and grammatical principles (Lee et al., 2011).

4.4.2 CR in image narrations

Problem definition. Next, we extend the text-only CR to image-text data in the absence of coreference labels. Let (I,S) denote an image-text pair where S describes an image I as illustrated in Figure 4.1, and assume that coreference labels for mention pairs are not present. As in Section 4.4.1, our goal is to identify the mentions that refer to the same entity in an image-text pair. Each image is defined by $|I|$ regions $I = \{r_1, r_2, \ldots, r_{|I|}\}$ which are obtained by running the pre-trained object detector (trained on the COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017) dataset) in (Ren et al., 2015) on the image. Each region r is described by its bounding box coordinates $\boldsymbol{b}$, the text embedding for the detected object category $\boldsymbol{o}$, and the visual features $\boldsymbol{v}$. More details are provided in Section 4.4.3.

Weak supervision. We use 'weak supervision' to refer to a setting where no coreference label for mention pairs and no grounding of mentions (*i.e.,* bounding boxes are not linked to phrases in the text) are available. Moreover, in contrast to the output space of the object detector (a restricted set of object categories), the sentences describing our images come from the unconstrained vocabulary. Hence, an object instance in a sentence can be referred to with a synonym or may not even be present in the object detector vocabulary (Kuznetsova et al., 2020; Lin et al., 2014). Finally, the object detector can only output *category-level* labels and hence cannot localize object instances based on the more specific *instance-level* descriptions provided by the sentences. For instance in Figure 4.1, *a person* and *the woman* both are labeled as *person* by the object detector.

In addition to image and text, we explore the use of an auxiliary modality, mouse trace segments provided in (Pont-Tuset et al., 2020). Each mouse trace includes a sequence of 2D points over time that relate to a region in the image when describing the scene. As the text in Localized Narratives is transcription of the speech of the annotators, the mouse traces are synced with spoken words, which we denote as $T = \{t_1, t_2, \dots, t_{|T|}\}$ where $|T| = |S|$. These features are stacked with textual features (see Section 4.4.3).

In the weakly supervised setting, the key challenge is to replace the coreference label supervision with an alternative one. We hypothesize that each mention in a coreferring pair corresponds to (approximately) the same image region, and it is possible to learn a joint image-text space that is sufficiently rich to capture such correlations. Concretely, let $g(m,r)$ denote an auxiliary function that is instantiated as a deep network and outputs a score for the mention m being located at region r in image I. This grounding score for each mention can be converted into probability values by normalizing them over all regions in the image:

$$\bar{g}(m,r) = \frac{\exp\left(g(m,r)\right)}{\sum_{r' \in I} \exp\left(g(m,r')\right)}. \tag{4.2}$$

The compatibility function f can be defined as a sum product of a pair's grounding probabilities over all regions:

$$f(m,m') = \sum_{r \in I} \bar{g}(m,r)\bar{g}(m',r). \tag{4.3}$$

In words, mention pairs with similar region correlations yield bigger compatibility scores and are hence more likely to corefer to each other. The key idea is that we employ the grounding for mentions as anchors to relate coreferring mentions (*e.g., a person* and *the woman*). At test time, we compute $f(m,m')$ for all the pairs and threshold them to predict their pairwise coreference labels.

As no ground-truth bounding box for each mention is available for learning the grounding g, we pose grounding as a weakly supervised localization task as in (Gupta et al., 2020; Wang et al., 2020a). To this end, we impute the missing bounding boxes by taking the highest scoring region for a given mention m at each training iteration:

$$r_m = \arg\max_{r \in I} g(m,r). \tag{4.4}$$

Then we use r_m as the pseudo-truth to learn g as following:

$$\min_g \sum_{(I,S) \in \mathscr{D}} \sum_{m \in S} -\log\left(\frac{\exp(g(m,r_m))}{\sum_{I' \in \mathscr{D} \setminus I} \exp(g(m,r'_m))}\right) \tag{4.5}$$

where $r'_m = \arg\max_{r \in I'} g(m,r)$ is the highest scoring region in image I' for mention m. For each mention, we treat the highest scoring region in the original image as positive and other highest scoring regions across different images as negatives and optimize g for discriminating between the two. However, as the denominator requires computing g over all training samples at each iteration, which is not computationally feasible, we instead sample the negatives only from the randomly sampled minibatch.

Linguistic constraints. Learning the associations between textual and visual features helps with disambiguating coreferring mentions, especially when mentions contain visually salient and discriminative features. However, resolving coreferences when it comes to pronouns (*e.g., her, their*) or ambiguous phrases (*e.g., one man* or *another man*) remains challenging. To address such cases, we propose to incorporate a regularizer into the compatibility function $f(m,m')$ based on various linguistic priors. Concretely, we construct a look-up table for each mention pair $q(m,m')$ based on the following set of rules (Lee et al., 2011):

(a) Exact string match. Two mentions corefer if they exactly match and are noun phrases (not pronouns).

(b) Pronoun resolution. Based on the part-of-speech tags for the mentions, we set $q(m,m')$ to 1 if m is a pronoun and m' is the antecedent noun that occurs before the pronoun.

(c) Distance between mentions. Smaller distance is more likely to indicate coreference since mentions occur close together if they refer to the same entity.

(d) Last word match. In certain cases, the entire phrases might not match but only the last word of the phrases.

(e) Overlap between mentions. If two mentions have one or more overlapping words, then they are likely to corefer.

Finally, we include $q(m,m')$ as a regularizer in Eq. (4.5):

$$\begin{aligned}\min_{g} \sum_{(I,S)\in\mathscr{D}} \sum_{m\in S} \Big(-\log\big(\frac{\exp(g(m,r_m))}{\sum_{I'\in\mathscr{D}\setminus I}\exp(g(m,r'_m))}\big) \\ +\lambda \sum_{m'\in S} \|f(m,m') - q(m,m')\|_F^2 \Big)\end{aligned} \tag{4.6}$$

where λ is a scalar weight for the Frobenius norm term. Note that f is a function of g (see Eq. (4.3)). We show in Section 4.6 that incorporating this term results in steady and significant improvements in CR performance.

4.4.3 Network modules

Our model (illustrated in Figure 4.3) consists of an image encoder e_i and text encoder e_t to extract visual and linguistic information respectively and a cross-attention module a for their fusion.

Image encoder e_i takes in a d_r-dimensional vector for each region r that consists of a vector consisting of bounding box coordinates $\boldsymbol{b} \in R^4$, text embedding for the detected object category $\boldsymbol{o} \in R^{d_o}$ and visual features $\boldsymbol{v} \in R^{d_v}$. The regions are extracted from a pre-trained object detector (Ren et al., 2015) for the given image I. The image encoder applies a nonlinear transformation to this vector to obtain a d-dimensional embedding for each region r.

Text encoder e_t takes in the multiple mentions from a parsed multi-sentence image description S produced by an NLP parser (spa) and outputs a d-dimensional embedding for each word in the parsed mentions. Note that the parser does not only extract nouns but also pronouns as mentions.

Mouse trace encoder e_m takes in the mouse traces for each mention parsed above after it is preprocessed into a 5D vector of coordinates and area, $(x_{\text{min}}, x_{\text{max}}, y_{\text{min}}, y_{\text{max}}, \text{area})$ (Meng et al., 2021) and outputs a d_m-dimensional embedding. In Changpinyo et al. (2021a); Pont-Tuset et al. (2020), mouse trace embeddings have been exploited for image retrieval, however, we use them to resolve coreferences. We concatenate each mention embedding extracted from e_t with the mouse trace encoding e_m, denoted as e_{tm} and apply additional nonlinear transformations (Joint encoder in Fig. 4.3) before feeding into the cross-attention module.

Cross-attention module a takes in the joint text-trace embeddings for all the words in each mention and returns a d-dimensional vector for each m by taking a weighted average of them based on their correlations with the image regions. Concretely, in this module, we first compute the correlation between each word w (or joint word-mouse trace) and all regions, take the highest correlation over the regions through an auxiliary function $\bar{a}$:

$$\bar{a}(w) = \max_{r \in I} \Big(\frac{\exp(e_{tm}(w) \cdot e_i(r))}{\sum_{r' \in I} \exp(e_{tm}(w) \cdot e_i(r'))} \Big) \tag{4.7}$$

where $\cdot$ is the dot product. The transformation can be interpreted as the probability of word w being present in image I. Then we compute a weighted average of the word embeddings for each mention m:

$$a(m) = \sum_{w \in m} \bar{a}(w) e_{tm}(w). \tag{4.8}$$

Similarly, $a(m)$ can be seen as the probability of mention m being present in image I.

Scoring function $g(m,r)$ can be written as a dot product between the output of the attention module and region embeddings:

$$g(m,r) = a(m) \cdot e_i(r). \tag{4.9}$$

While taking a dot product between the two embeddings seemingly ignores the correlation between text and image data, the region embedding $e_i(r)$ encodes the semantic information about the detected object category in addition to other visual features and hence results in a high score only when the mention and region are semantically close. Further implementation details about the modules can be found in Section 4.5 and Chapter C.

4.5 Experiments

We train our models on the Flickr30k subset of the Localized Narratives (Pont-Tuset et al., 2020) which consists of 30k image-narration pairs, and evaluate on the proposed **CIN** dataset, which contains 1000 and 800 pairs for test and validation respectively.

Evaluation metrics. To evaluate the CR performance, we use the standard link-based metrics MUC (Vilain et al., 1995) and BLANC (Recasens and Hovy, 2011):[2]

[2] Refer to (Lee et al., 2017; Luo and Pradhan, 2016) for a more detailed discussion of CR metrics.

(a) MUC F-measure counts the coreference links (pairs of mentions) common to the predicted chain R and the ground-truth chain K by computing MUC-R (recall) and MUC-P (precision).
(b) BLANC measures the precision (BLANC-P) and recall (BLANC-R) between the ground-truth and predicted coreference links and also between non-coreferent links.
(c) Narrative grounding. For evaluating narrative grounding in images, we consider a prediction to be correct if the IoU (Intersection over Union) between the predicted bounding box and the ground truth box is larger than 0.5 (Gupta et al., 2020; Wang et al., 2020a). We report percentage accuracy for evaluating narrative grounding for both noun phrases and pronouns. Further details about the metrics are in Chapter C.

Inputs and modules. For the image modeling, we extract bounding box regions, visual features, and object class labels using the Faster-RCNN object detector (Ren et al., 2015). For the text modeling, we use Glove embeddings (Pennington et al., 2014) to encode the object class labels and the mentions from the textual branch. For the mouse traces, we follow (Pont-Tuset et al., 2020) and extract the trace for each word in the sentence and then convert it into bounding box coordinates for the initial representation. The model discussed in Section 4.4 referred to as 'Ours' in Section 4.6 uses the transformer backbone for the image, text, and trace encoders (more details in Chapter C).

Baselines. We consider the following baselines to fairly compare and evaluate our proposed method:
(a) Text-only CR: For all these methods, we directly evaluate the coreference chains using the narration only without the image. (1) *Rule-based* (Lee et al., 2011): In this method, a multi-sieve rule-based system is used to find mentions in the sentence and the coreference chains, (2) *Neural-Coref* (Lee et al., 2017): Instead of rules, this method is trained end-to-end using a neural network on a large corpus of Wikipedia data to detect mentions and coreferences, and (3) *Similarity-based:* We compute the cosine similarity between mentions using Glove word features and threshold them to get coreference chains.

Method	Text	Image	MT	MUC-R	MUC-P	MUC-F1	BLANC-R	BLANC-P	BLANC-F1
Rule-Based (Lee et al., 2011)	✓	✗	✗	5.6	10.13	6.4	3.3	4.1	4.9
Neural-Coref (Lee et al., 2017)	✓	✗	✗	0.11	0.17	0.13	1.59	36.99	3.23
Similarity-based	✓	✗	✗	7.07	14.43	9.06	37.48	65.17	45.98
GLIP (Li et al., 2022b)	✓	✓	✗	0.13	0.12	0.12	21.71	61.40	31.66
VisualBERT (Su et al., 2019)	✓	✓	✗	18.17	6.08	8.06	23.90	44.07	22.38
UNITER (Chen et al., 2020)	✓	✓	✗	16.92	7.15	8.83	27.64	55.30	29.20
VinVL (Zhang et al., 2021b)	✓	✓	✗	16.76	8.60	9.75	34.56	58.07	36.56
MAF† (Wang et al., 2020a)	✓	✓	✗	**25.86**	10.18	13.21	37.68	61.14	38.17
MAF++	✓	✓	✗	19.07	15.62	15.65	41.25	65.04	47.21
Ours	✓	✓	✗	22.07	17.10	17.58	42.72	65.92	48.29
	✓	✓	✓	24.87	**18.34**	**19.19**	**43.81**	**66.35**	**48.53**

Table 4.4 CR performance on CIN dataset. *MT denotes mouse trace and † denotes our trained model.*

(b) Visual-text models: The baselines discussed below are not trained for CR and hence we post-process their output in order to evaluate for CR. (1) VisualBert (Su et al., 2019), UNITER (Chen et al., 2020) and VinVL (Zhang et al., 2021b) are vision-language models trained on image-caption data and fine-tuned on downstream tasks such as VQA, NLVR. To test it for CR, we compute the cosine similarity for the multi-modal mention embeddings in a zero-shot way. (2) *GLIP* (Li et al., 2022b): GLIP is trained on large-scale image-text paired data with bounding box annotations and shows improvement in object detection and visual phrase grounding. To evaluate it for CR, we predict bounding boxes for the mentions in the narrations from GLIP. If the IoU overlap between the mentions is greater than 0.7, then we consider them to form a coreference chain, (3) *MAF*† (Wang et al., 2020a): MAF is a weakly supervised phrase grounding method, originally trained on the Flickr30k-Entities (Plummer et al., 2015). We train this model on narrations data and evaluate CR by computing Eq. (4.3). (4) *MAF++:* We retrain the MAF† model on the narrations with our regularization term. Architecturally our method differs from the MAF† in two aspects: i) we employ a transformer to encode visual and text features unlike the MLP in theirs and ii) we attend to the mouse traces when present (not present in MAF) and word features jointly whereas they directly compute the similarity function.

4.6 Results

Coreference resolution. In Table 4.4, we report CR performance of the baselines and our method. Our method significantly outperforms all the text-only and visual-text baselines on all the metrics. The text-only CR baselines in the first three rows fail to effectively resolve conferences from narrations. It is important to note that a relatively high number in BLANC scores (compared to MUC) occurs because this measure also counts non-coreferent links (*i.e.,* mentions that are not paired with anything), whereas MUC only measures pairs that are resolved.

The rule-based method (Lee et al., 2011) uses exact match noun phrases, pronoun-noun matches, and the distance between mentions as hard constraints. It achieves low scores on all metrics and especially on BLANC. The reason for this is the limitation of the rule-based heuristics: For instance, in long narrations, if a pronoun such as *she* occurs farther to its referent (*e.g., the woman*) than the predefined distance, it will not form a coreference chain. In contrast, as we apply rules as a soft constraint, we are able to make more flexible decisions in our method. Neural-Coref (Lee et al., 2017), a deep network on a pre-trained large-corpus of labeled CR data, obtains low scores on CIN for both MUC and BLANC. This is due to the large domain gap between the source and target data as well as the ambiguity in resolving the mentions without visual cues. Similar observations are made when pre-trained CR methods are applied to other domains such as biomedical text (Lu and Poesio, 2021) or social media (Aktaş et al., 2020). Lastly, the similarity-based baseline performs poorly, as the utilized off-the-shelf word vectors are not trained to cluster corefering mentions. The relatively high scores on BLANC are due to the frequent non-coreferents in our narratives. This kind of approach clusters words with similar meaning together *e.g., woman* and *another woman* (both representing female entities) or *he* and *she* (both pronouns).

Next, we compare our method to the visual grounding baselines that use both image and text input. Our method also outperforms these baselines: Though GLIP is pre-trained on large-scale data with ground-truth boxes for each object in captions, these captions are usually short and do not contain multiple mentions of entities, unlike in our data. Hence GLIP acts more like an object detector, fails to link coreferring pairs (low MUC scores), and merely identifies singletons (higher BLANC scores). While it is nontrivial to finetune GLIP on our data without groundtruth boxes, we finetuned MAF on our data, as its training does not require groundtruth boxes; we denote this as MAF†. This is the strongest baseline on our task, as training it on narrations including the pronouns reduces the domain gap and enables it to resolve coreferences well. However, this method obtains low precision by incorrectly linking visually similar mentions (that do not belong together) such as *trees*, *plant*, *flowers*. When the training is regularized with the linguistic priors from our method, denoted as MAF++, its performance significantly improves on both MUC and BLANC. The constraint helps to push away the negative mentions (*trees*, *plant*, etc.) and encourages the model to learn unique embeddings for them. Due to the self-attention in the transformer architectures, Ours without mouse traces (MT) achieves better performance than MAF++, a simple MLP baseline. The performance difference between our method without using mouse-traces and MAF++ can be explained by the better architecture described previously. Finally, our method achieves the best performance gains in CR thanks to the mouse traces and improved architecture over MAF.

Ablation on mouse traces In Table 4.4, we also analyze the contribution of modeling mouse traces (second last row). Adding the mouse traces improves performance on CR across all metrics. We hypothesize that the mouse traces provide a strong discriminative location prior to the textual mentions, which helps the model to learn a better compatibility score. To visualize qualitatively, consider the example in Figure 4.3, the same mention *this person* points to two different visual regions – one with the person holding the ball and the other person standing next to the door. In such cases, mouse traces provide a strong signal for disambiguation. But in many cases, mouse traces are noisy and can link mentions that are very close to each other in the image, referring to two different regions. In the above example, mouse traces for *these persons* and *this person* have a significant overlap and hence act as a noisy prior. Therefore, without the visual/image region features, it is very challenging to address the problem with mouse traces alone.

Method	Reg	Noun Phrases	Pronouns	Overall
MAF† (Wang et al., 2020a)	✗	21.60	18.31	20.91
MAF++	✓	25.58	22.36	24.91
Ours	✗	27.62	23.46	26.75
	✓	**30.27**	**25.96**	**29.36**

Table 4.5 Grounding accuracy (%) for noun phrases and pronouns and the overall accuracy on the CIN dataset.

Narrative grounding Not only does our method show performance gains on CR but also outperforms the baselines on another challenging task of narrative grounding. Table 4.5 compares results from our methods and baselines. We directly compare with the weakly supervised method for a fair comparison. MAF† (Wang et al., 2020a) is originally evaluated on the Flickr30k-Entities (Plummer et al., 2015) dataset where the textual descriptions are significantly shorter (*i.e.,* single sentence) than the image narrations in our dataset. The performance of MAF on our dataset is significantly lower (21% vs 61% on Flickr30k-Entities), which indicates that narrative grounding is a challenge in itself and cannot be addressed off-the-shelf by phrase grounding methods. When trained with the regularizer, the localization performance improves for both nouns and pronouns with our method and MAF++. With the help of regularization, the model learns to attend to different regions of the image for semantically similar mentions as they might be two separate entities (*e.g., five people* and *the people* in Fig. 4.4).

		CR						Grounding
MT	Regularization	MUC-R	MUC-P	MUC-F1	BLANC-R	BLANC-P	BLANC-F1	Acc (%)
✗	✗	21.84	12.29	14.09	40.15	62.82	43.69	25.97
✓	✗	20.19	15.79	16.26	41.91	65.42	47.82	26.75
✓	L1	20.76	15.47	16.05	41.73	64.94	47.09	27.65
✓	MSE	21.58	16.40	17.00	42.19	65.37	47.60	28.50
✗	Frobenius Norm	22.07	17.10	17.58	42.72	65.92	48.29	28.31
✓		**24.87**	**18.34**	**19.19**	**43.81**	**66.35**	**48.53**	29.36

Table 4.6 Ablation study with different regularizer types and with/without mouse traces.

Further ablations. In Table 4.6, we start by exploring the impact of training using our proposed architecture without incorporating mouse traces and the regularizer. This

	CR		Grounding
Attention Type	MUC-F1	BLANC-F1	Acc(%)
Average	17.02	48.26	28.83
Cross attention	**19.19**	**48.53**	**29.36**

Table 4.7 Our method with/without cross attention.

leads to a decrease in both coreference (CR) and grounding performance. Although the model manages to capture certain coreference links, it also generates a noticeable number of incorrect associations (resulting in lower precision scores, row 1) when compared to the model trained with mouse traces (row 2).

In subsequent rows, we investigate the effects of using different regularizers during training. Notably, using the Frobenius norm as a constraint brings improvements in performance, in contrast to using L1 and mean squared error (MSE) regularizers. This improvement can be attributed to the Frobenius norm's ability to impose a more solid constraint on the learned coreference matrix. It's important to note that the final row in the table corresponds to our proposed model, which combines mouse traces and the Frobenius norm regularizer (MT+Frobenius norm).

In Table 4.7 we compare the performance of our final method under two settings: (1) directly averaging the word features or (2) attending over the words by using the image as the query as discussed in Section 4.4. The MUC-F1 and narrative grounding scores are 17.02 and 28.83 respectively for the (1) setting and 19.19 and 29.36 respectively for the (2) setting. Both the narrative grounding accuracy and the coreference evaluation get a boost in performance for visually aware word features. More often than not, the word phrases are relatively short (*e.g., the machine*) and hence the model does not always learn to disambiguate better with attention to the grounding. On the other hand, this technique is especially useful for CR because the flow of visual information to the word features acts as a prior to cluster mentions that refer to the same region but are referred to with different mentions/entities in the text (*e.g., the machine* and *an equipment*).

Qualitative results Figure 4.4 qualitatively analyzes CR and narrative grounding. We visualize the narrative grounding results from our proposed method on the images.

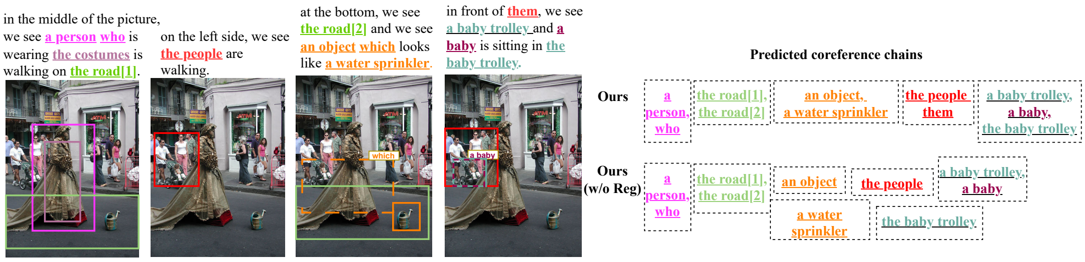

in the middle of the picture, we see a person who is wearing the costumes is walking on the road[1].
on the left side, we see the people are walking.
at the bottom, we see the road[2] and we see an object which looks like a water sprinkler.
in front of them, we see a baby trolley and a baby is sitting in the baby trolley.
which
a baby
Predicted coreference chains
Ours
a person, who
the road[1], the road[2]
an object, a water sprinkler
the people them
a baby trolley, a baby, the baby trolley
Ours (w/o Reg)
a person, who
the road[1], the road[2]
an object
the people
a baby trolley, a baby
a water sprinkler
the baby trolley

Figure 4.4 Qualitative results of predictions on the CIN dataset. The colored mentions in the text indicate the ground truth coreference chains. The solid and dotted bounding boxes on the image denote the correct and incorrect grounding respectively for our proposed method. We also show the predicted coreference chains for our final method with and without regularizer.

The model correctly resolves and localizes phrases such as *a person, who*, *the people, them*, and *a baby trolley, the baby trolley*. Whereas, the model fails to ground and chain the instance *a baby*. It is interesting to note that our model pairs *an object* and *water sprinkler*, thereby resolving ambiguity in what *the object* might refer to. But it fails to add *which* to this coreference chain. Moreover, without the language regularizer, our method fails to link *them* to *the people*. It is very hard to learn coreferences for these pronouns as they come with a weak language prior and hence are difficult for the model to disambiguate. Our model (without regularization) misses the referring expression of *the baby trolley* to refer to the instance of the trolley before. With the help of rules (*e.g.,* last token match), we can resolve these pairs more often than not. Hence, we clearly show the challenging problem of coreferences we are dealing with and indicate the great potential for developing models with strong contextual reasoning.

4.7 Conclusion and discussion

In this chapter, we introduced a novel task of resolving coreferences in image narrations, clustering mention pairs referring to the same entity. For benchmarking and enabling the progress, we introduce a dataset – CIN – that contains images with narrations annotated with coreference chains and their grounding in the images. We formulate the problem of learning CR by using weak supervision from image-text pairs to disambiguate coreference chains and linguistic priors to avoid learning grammatically wrong chains. We demonstrate strong experimental results in multiple settings on our proposed dataset for coreference resolution and narrative grounding in images.

While our method shows considerable improvements when compared to strong baselines in multimodal coreference resolution, there are certain limitations. Firstly, our model relies on input from mouse traces during both training and inference which are challenging to obtain and are not universally available for standard datasets. While mouse traces provide an important link between the semantic knowledge from sentences and the spatial information in images, developing a model without mouse traces will be a more flexible approach. Secondly, the language rules for coreferences are either too strict such as exact match of phrases, or too generic such as pronoun match. Relying on such pre-defined rules is harmful for generalization. Hence, directly learning coreferences from data could be highly effective in addressing the noise induced by the language rules during learning.

In the next chapter, we show (1) build a standalone model without the need for mouse traces during training or inference and (2) how to directly learn coreferences from image-narration data without any pre-defined linguistic rules.

Chapter 5

Semi-supervised multimodal coreference resolution in image narrations

In the previous chapter, we introduced a new benchmark dataset, Coreferenced Image Narratives (CIN), to benchmark multimodal coreference resolution. We also developed a weakly supervised method to predict coreference chains using rule-based/lexical language priors and simultaneously perform narrative grounding in images. In this chapter, we seek to further explore the modeling of multimodal coreference resolution *i.e.,* where a series of sentences *i.e., narration* is paired with an image. This poses significant challenges due to fine-grained image-text alignment, the inherent ambiguity present in narrative language, and the unavailability of large annotated training data. Contrary to Chapter 4, we present a data-efficient semi-supervised approach to tackle these challenges, that utilizes image-narration pairs to resolve coreferences and narrative grounding in a multimodal context. Our approach incorporates losses for both labeled and unlabeled data within a cross-modal framework. Through rigorous evaluation, we demonstrate that our proposed approach outperforms strong baselines both quantitatively and qualitatively for the tasks of coreference resolution and narrative grounding.

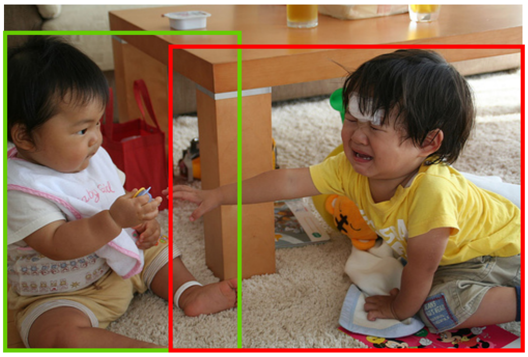
This chapter begins with revisiting the problem of multimodal coreference resolution as discussed in Chapter 4, limitations of the weakly supervised method proposed in Chapter 4 and the key ideas of our proposed method in Section 5.1, related work in

Section 5.2, elaborating our proposed method in Section 5.3, discussing the experimental details in Section 5.4, evaluating and analyzing the proposed method on the CIN dataset in Section 5.5 and conclusion and discussion in Section 5.6.

5.1 Introduction

In linguistic processing, coreference resolution is a standard task that aims to identify referring expressions such as noun-phrases and pronouns, which refer to the same entity. It is fundamental to many standard problems including question answering (Das et al., 2017; Kwiatkowski et al., 2019), sentiment analysis (Cambria et al., 2017; Medhat et al., 2014), summarization (Gupta and Lehal, 2010; Shi et al., 2021) and machine translation (Bahdanau et al., 2014; Lopez, 2008; Wu et al., 2016). In this chapter (similar to Chapter 4), we focus on a multimodal coreference resolution (MCR) scenario where the coreferences occur in a narration paired with an image and also link to an image part as shown in Fig. 5.1. Here resolving coreferences is challenging, as mentions referring to different entities can be very similar when encoded by a language model, *e.g.*, *one boy*, *the other boy*, *the boy*. Hence it demands a fine-grained understanding of each modality and as well as across them. In particular, it requires simultaneously grounding instances by identifying fine-grained visual details (*e.g.*, disambiguating them by recognizing the action ‘crying’, spotting ‘white color t-shirt and cream color short’ or ‘a white color sticker on the head’), and capturing long-range dependency across sentences (*e.g.*, ***two small boys*** and ***their***).

MCR has recently gained increasing attention, with several notable studies (Cui et al., 2021; Das et al., 2017; Guo et al., 2022; Hong et al., 2023; Huang et al., 2018; Parcalabescu et al., 2021; Ramanathan et al., 2014). However, many of them focus on images with simple short sentences, such as ‘A woman is driving a motorcycle. Is she wearing a helmet?’ (Das et al., 2017; Parcalabescu et al., 2021), or are limited to identifying movie characters or people (Cui et al., 2021; Ramanathan et al., 2014). In Chapter 4, we introduced a challenging and unconstrained MCR problem (see Fig. 5.1) including a dataset, Coreferenced Image Narratives (CIN), with both people



we can see **two small boys** are sitting on a white color mat and in that **one boy** is crying and **he** is wearing a yellow color t-shirt and grey color short. **The other boy** is wearing white color t-shirt and cream color short and **he** is also holding some object in **his** hand. On the head of **the boy** we can see a white color sticker and on **their** t-shirts we can see some text and designs also.

Figure 5.1 Example image-narration pair from the Coreferenced Image Narratives dataset. Phrases marked in the same color corefer to the same entity which is also grounded in the image. We do not show singletons for brevity.

and objects as referents with long sentence narrations. As manually annotating a large dataset with coreferencing and grounding labels is expensive, the authors provide annotations only for evaluation purposes. We also proposed a weakly supervised method that learns to jointly ground mentions in images and use them as anchors along with prior linguistic rules (Lee et al., 2011) to group coreferring mentions from only image and narration pairs without the annotations. Nevertheless, this method has multiple shortcomings: 1) weakly supervised grounding fails to disambiguate multiple instances of the same object class, boy (*one boy*, *the other boy*), 2) language rules such as *exact match of phrases* are either too strict or too generic *e.g., pronoun match*, linking pronouns to one antecedent (*one boy*, *he*, *he*, *his*) and, (3) they require an additional modality, mouse traces to learn coreferences which can be expensive to obtain.

Motivated by these limitations, we argue that it is not guaranteed to successfully resolve coreferences from only image-narration pairs in cases where multiple instances from the same object category are present, which is more often than not referred to in the narration. Since full manual annotations of coreference and bounding boxes is expensive, we propose to resolve coreferences and ground mentions in a semi-supervised setting where only a few data samples are labeled. Our approach involves a customized multi-modal fusion model that combines image region features and mention features from narrations through cross-attention (Li et al., 2021; Vaswani et al., 2017). We investigate different task-specific losses for training on labeled and unlabeled data and show that naively combining the training on the labeled and pseudo-labeled data suffers from severe overfitting (Arazo et al., 2020). Hence, we propose a robust loss function and thresholding-based training scheme to effectively learn from the unlabeled set. This unique approach results in consistent performance improvements with the inclusion of unlabeled data during training.

Our main contributions are 1) a vision-language framework for MCR trained on a small labeled and an unlabeled dataset, 2) novel task-specific losses (on both labeled and pseudo-labeled data) for learning joint multi-modal embeddings for coreference resolution while simultaneously improving narrative grounding, 3) extensive evaluation of our proposed method on the CIN dataset and ablation studies to validate our design choices, showing consistent performance gains compared to baselines on coreference resolution and narrative grounding.

5.2 Related work

Multimodal coreference resolution. MCR involves comprehending the contextual information in the language and establishing connections with specific regions in an image. Recently, considerable efforts have been dedicated to developing datasets that can effectively address this intricate task. Parcalabescu et al. (2021) introduced the VALSE dataset, which encompasses various coreference scenarios. However, this dataset focuses on the downstream task of visual question answering without evaluating coreference resolution or grounding. Hence, we evaluate our method in CIN dataset proposed in Chapter 4 that contains coreference chain and grounding annotations. Another approach to MCR datasets involves linking people's names mentioned in the text to corresponding images and resolving pronouns that connect to those specific names (Cui et al., 2021; Hong et al., 2023; Ramanathan et al., 2014). However, our main focus is to resolve coreferences in a generic scenario (with visual complexity) unlike the others that are either limited to only people names/characters

(Cui et al., 2021; Hong et al., 2023; Ramanathan et al., 2014) or have simple sentences (Das et al., 2017; Parcalabescu et al., 2021).

Vision-Language learning. Existing work on vision and language understanding employ either pre-trained object detector features (He et al., 2017; Ren et al., 2015) as an image encoder, ViT (Dosovitskiy et al., 2020) or a CNN (Simonyan and Zisserman, 2014) combined with a transformer-based text encoder (Devlin et al., 2018). To model cross-modal interaction between the image and text encoders, UNITER (Chen et al., 2020), ALBEF (Li et al., 2021) and VinVL (Zhang et al., 2021b) employ a multimodal encoder. They are pre-trained on large-scale image-caption pairs such as COCO (Lin et al., 2014), Conceptual captions (Changpinyo et al., 2021b; Sharma et al., 2018), Visual Genome (Krishna et al., 2017). The pre-training objectives are implemented with image-text contrastive loss, masked language modeling, and image-text matching loss. Our method is inspired by the same family of architectures and is trained using a set of self-supervised and task-based objectives in a semi-supervised learning fashion.

Semi-Supervised learning. There is a large body of work in semi-supervised learning (Ouali et al., 2020; Van Engelen and Hoos, 2020; Zhai et al., 2019). These methods typically exploit unlabeled data via either pseudo-labeling with small amounts of labeled data (Arazo et al., 2020; Lee et al., 2013; Rizve et al., 2021; Sohn et al., 2020; Zhang et al., 2021a) or by enforcing consistency regularization (Abuduweili et al., 2021; Berthelot et al., 2019) on the unlabeled data to produce consistent predictions over various perturbations of the same input by applying several augmentation strategies (Cubuk et al., 2018, 2020; Zhang et al., 2017b). Our method draws inspiration from pseudo-labeling literature and uses a robust loss function and thresholding to counter overfitting to pseudo-labels.

5.3 Method

5.3.1 Task overview

Our goal is a) to group mentions (*i.e.,* referential words or phrases) in the narration that corefer to the same entity and, b) ground each mention to a region in an image.

Formally, let $N = \{m_1, m_2, \ldots, m_{|N|}\}$ denote a narration with $|N|$ mentions for an image I with $|I|$ regions where $I = \{r_1, r_2, \ldots, r_{|I|}\}$. We wish to learn an embedding function f that takes in an image I and its narration N, parsed to contain a set of mentions, and outputs a score for a mention pair (m, m'):

$$\frac{f(m) \cdot f(m')}{|f(m)||f(m')|} \tag{5.1}$$

The mention pair m and m' corefer if the score in Eq. (5.1) is high, otherwise they do not.

For grounding of the mention m on the image region r, we also learn another function g that outputs a score for the mention m being located at region r in image I. Next, we describe in detail our methodology to learn the two functions f and g.

5.3.2 Model architecture

In Fig. 5.2, we illustrate our model architecture. Each image is parsed into a set of regions through a pre-trained object detector (Ren et al., 2015), where each region r is represented by a d-dimensional joint embedding $\boldsymbol{v}_r \in \mathbb{R}^d$ including its visual, semantic and spatial features. In particular, the visual encoder f_v is instantiated as a transformer block that takes in a joint feature embedding $\boldsymbol{v}_r$ for the object region r and outputs a D dimensional embedding, *i.e.,* $f_v(\boldsymbol{v}_r) : \mathbb{R}^d \to \mathbb{R}^D$.

Furthermore, we encode the words in each narration N using a tokenizer (Devlin et al., 2018) to get a set of tokens for the words $w \in \mathbb{R}^V$ where V is the vocabulary size. The text encoder f_t which is also a transformer block that takes in the word token w and outputs a D dimensional embedding, *i.e.,* $f_t(w) : \mathbb{R}^V \to \mathbb{R}^D$. The mention embeddings are computed by averaging its corresponding word representations as: $f_t(m) = \frac{1}{|m|} \sum_{w \in m} f_t(w)$ where, $|m|$ indicates the mention length in words, and the embeddings $f_t(m)$ have the same dimensionality as the visual features.

Next, the multi-modal encoder f fuses the visual features from the visual encoder $f_v(\boldsymbol{v}_r)$ with the mention features from the text encoder $f_t(m)$. Similar to the cross-modal architectures (Li et al., 2021; Zhang et al., 2021b), the embeddings from the

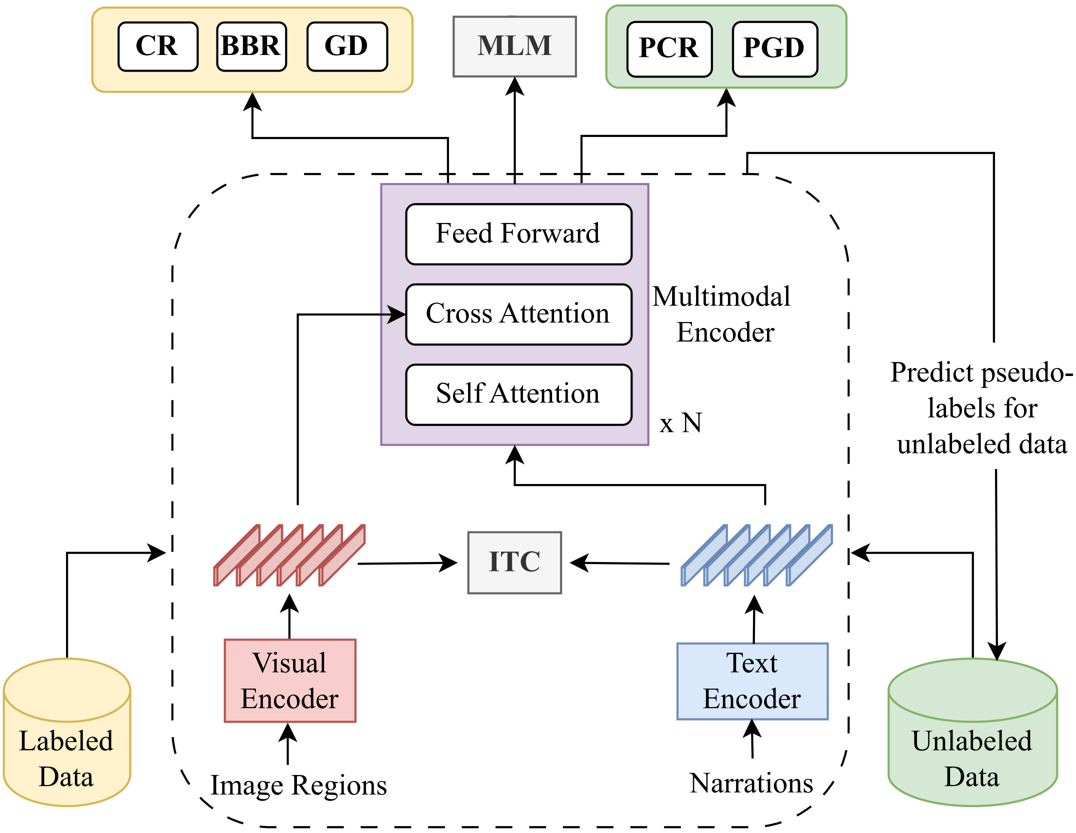


Figure 5.2 Illustration of our model architecture and training methodology. The pre-extracted image regions are fed into the visual encoder, the narrations are fed into the text encoder, and both modalities are fused using a multimodal encoder. The model is optimized using self-supervised objectives (in grey) and specialized task-based losses on both the labeled data (in yellow boxes) and the pseudo-labeled data (in green boxes).

text encoder are first encoded using self-attention layers (Vaswani et al., 2017). Then, a multi-head cross-attention module integrates the textual and visual features. In the cross-attention module, the self-attended mention embeddings $f_t(m)$ are treated as the query, while the image representations $f_v(\boldsymbol{v}_r)$ are treated as keys and values. The attention weights between the mention m and the region r are given as:

$$g(m,r) = \frac{\exp(\frac{f_t(m)^T.f_v(\boldsymbol{v}_r)}{\sqrt{d}})}{\sum_{r' \in I} \exp(\frac{f_t(m)^T.f_v(\boldsymbol{v}_{r'})}{\sqrt{d}})} \tag{5.2}$$

where, softmax is computed over the image regions for each mention. This attention matrix (or the grounding function) g from the multi-head cross attention learns fine-grained mention to region alignment scores. Finally, the vision-aware mention embedding is represented as:

$$f(m) = g(m,r).f_v(\boldsymbol{v}_r) \tag{5.3}$$

where, $f(m) \in \mathbb{R}^D$. This weighted embedding is then passed to a feed-forward module (Li et al., 2021) with an MLP and layer normalization. All the transformer encoders/blocks are based on the architecture proposed by Li et al. (2021). It is important to note that compared to the previous chapter, in this chapter the proposed model fuses vision and text features with a multimodal encoder, unlike theirs.

5.3.3 Semi-Supervised learning

Concretely, we aim to learn the parameters of the modules f_v, f_t, and f given a training dataset $\mathscr{D}$ with $|\mathscr{D}|$ samples of image-narration pairs. Specifically, we use a small labeled set $\mathscr{D}_s = \{x_i, y_i\}_{i=1}^{|\mathscr{D}_s|}$ where $x_i = \{I, N\}$ is the image-narration input pair and $y_i = \forall_{m \in N}\{P(m), A(m), b_m\}$ is the label for the input pair. In particular, the label for each mention m in the narration is given as: $P(m)$ and $A(m)$, the set of positive and negative mentions respectively for the mention m and b_m, the bounding-box coordinates of the region corresponding to the mention m.

Due to unavailability of a large labeled training set, we leverage the unlabeled data $\mathscr{D}_u$ = $\mathscr{D} \setminus \mathscr{D}_s$ where $\mathscr{D}_u = \{x_i\}_{i=1}^{|\mathscr{D}_u|}$ with only image-narration pairs as inputs. Our overall training objective is the joint loss function as follows:

$$\sum_{(x,y) \in \mathscr{D}_s} \frac{1}{|\mathscr{D}_s|} \mathscr{L}_s(x,y) + \sum_{x \in \mathscr{D}_u} \frac{1}{|\mathscr{D}_u|} \mathscr{L}_u(x) \tag{5.4}$$

where, $\mathscr{L}_s$ is the supervised loss and $\mathscr{L}_u$ is the unsupervised loss. First, we discuss how to formulate task-based supervised losses on the dataset $\mathscr{D}_s$.

(S1) Coreference loss (CR) Specifically, we propose to learn the similarity between the mention embeddings using a supervised contrastive loss (Khosla et al., 2020) which is defined as:

$$\mathcal{L}_{cr} = \sum_{m \in N} \frac{-1}{|P(m)|} \sum_{p \in P(m)} \\ \log \frac{exp(f(m).f(p)/\tau)}{\sum_{a \in A(m)} exp(f(m).f(a)/\tau)} \tag{5.5}$$

where τ is the temperature. This loss helps to cluster embeddings for coreferring mentions together and push the embeddings of non-referrants away from each other.

(S2) Grounding loss (`GD`) To align the mention m and region r, we use the grounding function g defined in Eq. (5.2). In particular, we first define the ground-truth binary alignment on the labeled training set $\mathcal{D}_s$. For the ground-truth bounding box b_m for a mention m we compute the intersection over union (IoU) between this bounding-box and the R pre-extracted image regions. This is crucial because we don't have the exact region-mention match for the detections from the object detector. Following this, we get the binary alignment function, $h(m,r)$ which is 1 for the mention m and the detected image region r if the region r has the maximum IoU overlap with the ground-truth bounding box b_m otherwise 0. Once, we have the ground-truth alignment $h(m,r)$, we compute the cross entropy loss as:

$$\mathcal{L}_{gd} = -\sum_{m \in N} \sum_{r \in I} h(m,r) \log(g(m,r)) \tag{5.6}$$

(S3) Bounding box regression loss (`BBR`) We further propose to add additional supervision to refine the object proposals from the detector for a mention. For each mention m, the ground-truth bounding box localization is represented as $b_m = (x,y,w,h)$. To learn refinements, we predict the bounding-deltas from the model as $\delta_m = (\delta_x, \delta_y, \delta_w, \delta_h)$ for each mention m. We then take the highest scoring region for a given mention m as:

$$r_m = \arg\max_{r \in I} g(m,r). \tag{5.7}$$

Our goal is to learn a transformation that maps a proposed box r_m to a ground-truth box b_m. We then apply the smooth-L1 loss following Ren et al. (2015) denoted as $\mathcal{L}_{bbr}$. Further details about this loss are given in Chapter D.

Next, we discuss how to train on the unlabeled subset of the dataset by generating pseudo-labels for the coreference and grounding tasks.

(U1) Pseudo coreference loss (PCR) Given the unlabeled dataset $\mathscr{D}_u$, we compute the pseudo coreferring pairs for the mentions in N. More specifically, we compute pseudo-positives $\hat{P}(m)$ and pseudo-negatives $\hat{A}(m)$ for a mention m by computing the cosine similarity between the embeddings as in Eq. (5.1). For each mention m, if the similarity with another mention m' is greater than a threshold then we label it as a positive otherwise a negative. Finally, we compute the triplet loss as:

$$\begin{aligned}\mathscr{L}_{pcr} = \sum_{m \in N} \max(||f(m) - \frac{1}{|\hat{P}(m)|} \sum_{p \in \hat{P}(m)} f(p)||^2 \\ -||f(m) - \frac{1}{|\hat{A}(m)|} \sum_{a \in \hat{A}(m)} f(a)||^2 + \alpha, 0)\end{aligned} \tag{5.8}$$

where α is the margin, $f(m)$ is the embeddings for the query mention m, $\frac{1}{|\hat{P}(m)|} \sum_{p \in \hat{P}(m)} f(p)$ is the mean of embeddings of the pseudo-positive labels $\hat{P}(m)$ and $\frac{1}{|\hat{A}(m)|} \sum_{a \in \hat{A}(m)} f(a)$ is the mean of embeddings of the pseudo-negative labels $\hat{A}(m)$.

The key intuition behind using the mean in a triplet loss formulation is to reduce overfitting to the noise in the pseudo labels. This works better in practice compared to the contrastive loss formulation in Eq. (5.5) or mining a random positive/negative label for the standard triplet loss, especially when dealing with pseudo labels.

(U2) Pseudo grounding loss (PGD) Furthermore, we compute the pseudo grounding loss on the unlabeled training dataset. Specifically, we impute the pseudo-labels from the grounding function, $g(m,r)$. We only consider the samples when the grounding score is greater than a certain confidence threshold t which is set to 0.9 in our experiments. The high threshold value ensures to consider only confident samples in the unlabeled set and eliminates learning from noisy samples. We denote this label after binary thresholding as $\hat{h}(m,r)$. The pseudo grounding alignment loss is denoted as:

$$\mathscr{L}_{pgd} = \sum_{m \in N} \sum_{r \in I} -\hat{h}(m,r) \log(g(m,r)) \tag{5.9}$$

Apart from the above mentioned task-based losses, we combine the standard image-text pretraining losses (Li et al., 2021; Vaswani et al., 2017). These losses help to learn better unimodal representations before fusion.

(U3) Image-Text contrastive loss (`ITC`) Following the training paradigm in Chapter 4, we incorporate the contrastive loss to align the image and narration pairs to learn better representations before fusion. This loss is defined as:

$$\mathscr{L}_{itc} = \sum_{m \in N} -\log\left(\frac{\exp(f_v(\boldsymbol{v}_r)f_t(m))}{\sum_{r' \in I}\exp(f_v(\boldsymbol{v}_{r'})f_t(m)))}\right) \tag{5.10}$$

where $f_v(\boldsymbol{v}_r)f_t(m)$ is the mention-region matching score from the visual and text representations before fusing in the multi-modal encoder and $\boldsymbol{v}_r$ are the raw features for the highest scoring region for a mention m.

(U4) Masked language modeling loss (`MLM`) To fine-tune the pre-trained BERT model (Devlin et al., 2018) on the image-narration data, we also use the pre-trained task of masked language modeling as proposed in Vaswani et al. (2017). In particular, the input word tokens are randomly masked and are replaced by a special masking token. The model needs to predict the mask token based on the unmasked words. This task is trained with a cross-entropy loss, $\mathscr{L}_{mlm}$.

Hence, our overall training objective in Eq. (5.4) is a combination of specialized task losses on the labeled training set $\mathscr{D}_s$ ($\mathscr{L}_{cr}$, $\mathscr{L}_{gd}$ and $\mathscr{L}_{bbr}$) and the unlabeled training set $\mathscr{D}_u$ ($\mathscr{L}_{pcr}$ and $\mathscr{L}_{pgd}$) and global pre-training objectives on the entire training dataset $\mathscr{D}$ ($\mathscr{L}_{itc}$ and $\mathscr{L}_{mlm}$).

5.3.4 Inference

To obtain the coreference scores, we form chains by measuring the cosine similarity between the mentions as described in Eq. (5.1), consider the pairs with similarity higher than a predefined threshold as positive. When evaluating narrative grounding, we extract the cross-attention scores from the last layer of the multimodal encoder. For each mention, we identify the region with the highest softmax score as the positively referred region.

5.4 Experiments

Method	Modality		MUC			B^3			CEAF$_{\phi4}$			CoNLL
	Text	Image	R	P	F1	R	P	F1	R	P	F1	F1
Neural Coref (Lee et al., 2017)	✓	✗	0.11	0.17	0.13	-	-	-	-	-	-	-
longdoc (Toshniwal et al., 2021)	✓	✗	7.79	8.43	7.24	62.27	76.10	67.69	48.77	84.95	61.02	45.31
VisualBERT (Su et al., 2019)	✓	✓	18.17	6.08	8.06	69.01	36.08	41.03	21.25	57.10	28.67	25.92
UNITER (Chen et al., 2020)	✓	✓	16.92	7.15	8.83	68.34	44.29	50.22	28.12	72.78	38.91	32.65
VinVL (Zhang et al., 2021b)	✓	✓	16.76	8.60	9.75	68.49	62.32	61.30	42.88	80.81	53.69	41.58
MAF (Wang et al., 2020a)	✓	✓	19.07	15.62	15.65	-	-	-	-	-	-	-
WS-MCR (Goel et al., 2023b)	✓	✓	24.87	18.34	19.19	-	-	-	-	-	-	-
Ours	✓	✗	13.30	14.12	12.55	67.91	79.48	72.41	56.05	86.20	67.05	50.67
	✓	✓	**31.11**	**35.25**	**31.86**	**70.63**	**87.85**	**78.06**	**63.99**	**93.44**	**75.47**	**61.79**

Table 5.1 Coreference resolution results on the CIN dataset from our proposed method and other state-of-the-art unimodal and multi-modal baselines.

Datasets. We evaluate our proposed method on the CIN dataset proposed in Chapter 4 that consists of 1000 test and 880 validation image-narration pairs from the Flickr30k split of the Localized Narratives dataset (Pont-Tuset et al., 2020) annotated with coreference chains and bounding boxes. We use the test split of the CIN dataset to report the performance on CR and narrative grounding. The annotations from the validation split are used as the small labeled set during training. The unlabeled dataset is the Flickr30k training subset of the Localized Narratives dataset (Pont-Tuset et al., 2020). This training split consists of 50k image-narration pairs but is not annotated with bounding boxes or coreference chains.

Implementation details. For the image regions, we extract bounding box regions, visual features, and object class labels using the Faster-RCNN object detector (Ren et al., 2015) as discussed in Chapter 4. We use a 4-layer transformer architecture for the text encoder and the multi-modal encoder similar to the ALBEF (Li et al., 2021) framework. The weights of the transformer encoders are initialized with BERT (Devlin et al., 2018). The visual encoder is a stack of two transformer encoder layers. Each transformer encoder layer includes a multi-head self-attention layer and an FFN. There are two heads in the multi-head attention layer, and two FC layers followed by ReLU activation layers in the FFN. Training details are in Chapter D.

Evaluation. We report results for coreference resolution and narrative grounding. For the former, we use the standard CoNLL F1 score which is the average of three coreference based metrics: MUC, B^3, and $CEAF_{\phi 4}$. For the latter, we follow the same procedure as discussed in Chapter 4 and report the grounding accuracy for both noun phrases and pronouns. More precisely, if the overlap between the ground-truth box and the predicted box is greater than 0.5, then it is considered to be a correct prediction.

5.5 Results

5.5.1 Coreference resolution

Table 5.1 reports the coreference resolution performance on the CIN dataset for our method and the baselines. Further details about the baselines are given in Chapter D. The text-based baselines Neural Coref (Lee et al., 2017) and longdoc (Toshniwal et al., 2021) are evaluated in a zero-shot way on the task. Their low CoNLL F1 scores indicate the incapability of the model to generalize to new domains which is in line with what has been evaluated extensively in the coreference literature (Porada et al., 2023; Toshniwal et al., 2021; Yang et al., 2022b). Moreover, it validates the need for multi-modal models for the task of multimodal coreference resolution.

We further compare to strong multi-modal baselines by directly evaluating the VLMs in a zero-shot way on the CIN dataset. Interestingly, all three methods, VisualBERT (Su et al., 2019), UNITER (Chen et al., 2020) and VinVL (Zhang et al., 2021b) perform better in MUC and B^3 compared to the text-based baseline, longdoc (Toshniwal et al., 2021) but drop in performance on the average CoNLL F1 scores. These results show the inability of these models to effectively find singletons, hence leading to poor performance in the precision scores. Moreover, we can conclude that the vision and language pre-trained models fail to generalize for MCR.

We also compare two weakly supervised methods that are trained on the CIN dataset, MAF (Wang et al., 2020a) and the weakly supervised method proposed in Chapter 4, WS-MCR. MAF is a weakly supervised grounding method trained with `ITC` that is evaluated for CR and WS-MCR learns weakly-supervised grounding and

CR combining the ITC loss and prior linguistic rules. Both of these methods improve significantly in MUC scores compared to other zero-shot unimodal and multi-modal baselines. We directly report results for these baselines from the previous chapter.

Finally, we compare with the text-only variant (without any image) of our method. This method improves over the baselines on the CoNLL F1 scores. The significant gains in performance of our final method, with both text and image, combined with label supervision shows the importance of carefully tuning the model with a small amount of labeled data and large amounts of pseudo-labeled data.

Method	Noun Phrases	Pronouns	Overall
MAF (Wang et al., 2020a)	21.60	18.31	20.91
WS-MCR (Goel et al., 2022b)	30.27	25.96	29.36
Ours (ITC + MLM)	27.44	22.77	26.45
Ours (Full)	**32.58**	**28.45**	**31.71**

Table 5.2 Comparison of narrative grounding performance on the CIN dataset.

5.5.2 Narrative grounding

In Table 5.2, we present a comprehensive comparison between the baselines and our proposed approach to the task of narrative grounding. This task is both challenging and crucial, as it evaluates the precise alignment between image regions and phrases in textual data. Notably, our proposed method goes beyond the traditional alignment of noun phrases and also addresses the critical aspect of grounding pronouns, which is vital for multimodal coreference resolution. We measure noun phrase grounding, pronoun grounding, and overall accuracy to measure performance, similar to Section 4.5.

Remarkably, our proposed method exhibits superior performance compared to weakly supervised baselines, showcasing a remarkable margin of approximately 2% and 2.5% in noun phrase and pronoun grounding accuracy, respectively. Furthermore, when compared to our unsupervised baseline, namely “Ours (ITC + MLM)", the inclusion of labeled and pseudo-labeled data yields a significant performance boost of approximately 6%. These results clearly demonstrate the significance of training with both grounding alignment and coreference resolution loss, highlighting the mutual benefits derived from this approach.

5.5.3 Ablation study

Data type	% Samples	CoNLL F1
Labeled	20%	60.04
	50%	61.24
Unlabeled	20%	56.82
	50%	59.11

Table 5.3 CR performance by changing the amount of labeled and unlabeled data during training.

Varying labeled and unlabeled data. We study the impact of labeled data on the learning process, allowing us to showcase the strengths of our approach. In Table 5.3, we measure the model's performance on CoNLL F1 scores at different proportions of labeled data (20% and 50%). Remarkably, despite the limited amount of labeled data samples, the model demonstrates consistently high performance without any significant drop. This highlights the exceptional ability of our model to effectively learn from a small labeled set, without relying on a large number of annotated training samples.

Furthermore, to validate the efficacy of our proposed method, we also investigate the influence of unlabeled data samples during training. Following the same data split as in the supervised experiments, we observe the changes in performance indicated by row 2 in Table 5.3. As the quantity of unlabeled samples increases, the model exhibits enhanced coreference resolution performance. This result reinforces the ability of our proposed method to leverage and effectively learn from pseudo-labeled data. Detailed results are in Chapter D.

Impact of different loss functions. In Table 5.4, we assess the performance of coreference resolution by incorporating various losses proposed in Section 5.3. Throughout the training process, the model consistently integrates the self-supervised objectives of `ITC` and `MLM`, results in first row of Table 5.4.

Losses							MUC			B^3			$CEAF_{\phi 4}$			CoNLL
ITC	MLM	CR	BBR	GD	PCR	PGD	R	P	F1	R	P	F1	R	P	F1	F1
✓	✓	✗	✗	✗	✗	✗	23.81	25.83	23.12	69.32	85.87	76.41	61.00	89.69	72.05	57.19
✓	✓	✓	✗	✗	✗	✗	22.70	21.40	20.23	69.05	80.03	73.66	55.52	87.09	67.20	53.70
		✓	✓	✗			23.86	24.52	22.31	69.31	84.15	75.67	59.50	89.15	70.80	56.26
		✓	✓	✓			27.68	29.04	26.66	69.93	85.43	76.61	60.92	90.61	72.26	58.51
✓	✓	✓	✓	✓	✓	✗	30.66	32.82	30.31	70.70	86.09	77.33	62.64	92.92	74.27	60.64
					✓	✓	**31.11**	**35.25**	**31.86**	**70.63**	**87.85**	**78.06**	**63.99**	**93.44**	**75.47**	**61.79**

Table 5.4 Ablation study on our proposed method with the combination of the proposed losses.

Integrating the supervised contrastive coreference resolution loss, CR, in addition to ITC and MLM, results in a significant performance drop. Due to the limited availability of labeled data, the model struggles to effectively generalize for coreference resolution, leading to overfitting and consequently lower F1 scores. However, by progressively incorporating the bounding box regression loss, BBR, and the grounding alignment loss GD, we get a much stronger training signal even with a small labeled set. This multi-task training objective contributes to an impressive improvement of approximately 1.5% in the CoNLL F1 score.

Subsequently, we investigate the impact of incorporating loss on pseudo-labeled data. By introducing the pseudo coreference loss, denoted as PCR, we observe a remarkable improvement of approximately 2% in the CoNLL F1 scores. This result highlights the significance of leveraging pseudo clusters and underscores the effectiveness of our proposed robust triplet loss, which computes the triplet loss using the mean of positive and negative embeddings. Notably, this approach successfully incorporates pseudo-labeled data without leading to overfitting while achieving substantial performance gains. Consequently, our final proposed method, which integrates the pseudo grounding loss, PGD, exhibits the most superior overall performance, validating the potency of pseudo-labels for both coreference resolution and grounding.

Choice of coreference resolution loss. In Table 5.5, we examine the impact of different types of coreference resolution losses. We present a comparison of the following loss combinations: 1) Binary cross-entropy loss (BCE) applied to both

CR Loss		MUC			B^3			CEAF$_{\phi 4}$			CoNLL
on $\mathscr{D}_s$	on $\mathscr{D}_u$	R	P	F1	R	P	F1	R	P	F1	F1
BCE	BCE	23.63	23.55	21.57	69.50	81.94	74.47	57.68	88.01	68.95	55.00
CR	CR	28.20	22.47	23.08	70.10	76.29	72.40	52.32	87.22	64.71	53.40
CR	RTC	29.24	32.41	29.46	70.37	86.71	77.45	63.14	92.59	74.55	60.49
CR	PCR	**31.11**	**35.25**	**31.86**	**70.63**	**87.85**	**78.06**	**63.99**	**93.44**	**75.47**	**61.79**

Table 5.5 Performance comparison with the choice of coreference resolution loss on the labeled dataset $\mathscr{D}_s$ and the unlabeled dataset $\mathscr{D}_u$.

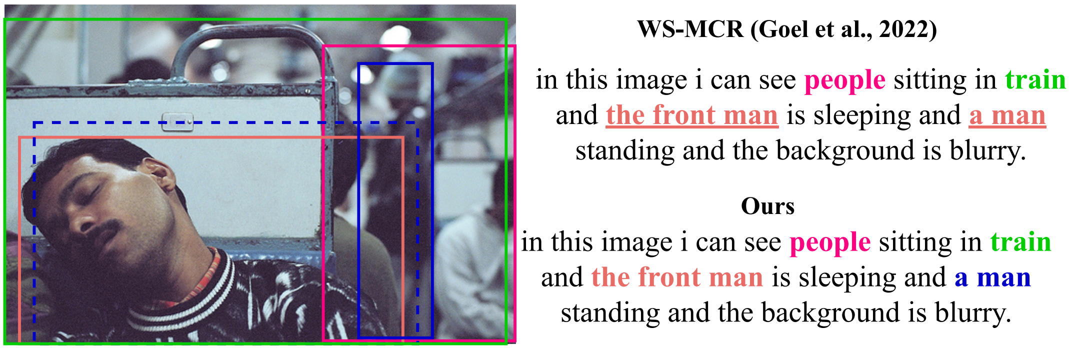
$\mathscr{D}_s$ and $\mathscr{D}_u$, 2) Supervised contrastive loss (CR) applied to both $\mathscr{D}_s$ and $\mathscr{D}_u$, and 3) Supervised contrastive loss (CR) on $\mathscr{D}_s$ and random triplet mining loss (RTC) on $\mathscr{D}_u$.

We observed a significant performance drop when training with the BCE loss, compared to utilizing the supervised contrastive loss. The supervised contrastive loss effectively promotes the learning of more discriminative embeddings for the clusters, unlike the binary cross-entropy loss. Consequently, the embeddings become more robust for CR, contributing to improved performance.

Interestingly, when applying the supervised contrastive loss to $\mathscr{D}_u$ (row 2), we observed a drop in performance. Our hypothesis is that the contrastive loss tends to overfit in the presence of noisy pseudo labels, leading to a degradation in performance. In contrast, our pseudo triplet loss formulation PCR is softer in penalizing noisy pseudo labels. This allows the model to gradually adapt and become more resilient to such noise, resulting in a more efficient clustering of mentions. We also compare to another ablation where instead of taking the mean of the embeddings for pseudo-positive labels and pseudo-negative labels, we sample a random positive and negative label (results in row 3) abbreviated as RTC. Randomly sampling the labels generalizes better than the other ablations but the mean cluster embeddings outperform than randomly selecting samples.

5.5.4 Qualitative results

In Fig. 5.3, we qualitatively visualize the performance of our method and compare it with the weakly supervised baseline from Chapter 4. Our model correctly separates



WS-MCR (Goel et al., 2022)

in this image i can see **people** sitting in **train** and **the front man** is sleeping and **a man** standing and the background is blurry.

Ours

in this image i can see **people** sitting in **train** and **the front man** is sleeping and **a man** standing and the background is blurry.

Figure 5.3 Visualization for grounding and coreference resolution. The colored boxes in the image correspond to the mentions with the same color in the sentence.

the mentions *the front man* and the *the man* both during CR and grounding, whereas the WS-MCR method as discussed in Chapter 4 incorrectly assigns the mention *the man* to the *the front man* and grounds it incorrectly too (denoted by the blue dotted line). Hence, our semi-supervised method can effectively learn to disambiguate the instances based on the visual details which is also helpful for coreference resolution.

5.6 Conclusion and discussion

In conclusion, this chapter addresses the fundamental and challenging task of multimodal coreference resolution where a narration is accompanied by an image. We propose a data efficient semi-supervised approach that incorporates task-based losses for both labeled and unlabeled data, operating within a cross-modal framework. Our method achieves remarkable results for CR and narrative grounding tasks on the CIN dataset, showcasing its effectiveness in handling the complexities of MCR.

We also outline some limitations of this work that are important considerations for future work. First, the current model's performance in coreference resolution and grounding is limited by the use of a pre-trained object detector. The detectors pretrained for object detection task have a limited object category vocabulary and lack in fine-grained properties including adjectives, human actions, and the open vocabulary found in narrations. This forces the model to rely on a predetermined set of regions and object classes, preventing it from directly learning region coordinates for a mention on an image. To improve performance, we envision the development of an end-to-end approach that eliminates this reliance on pre-defined regions.

Second, our model currently depends on ground-truth mentions to resolve coreferences and ground them. In the future, one promising direction would be to detect mentions simultaneously with coreference resolution and grounding. This would significantly improve the applicability of our proposed method and reduce dependence on off-the-shelf mention detectors or ground-truth annotations.

Third, we evaluate our proposed method without performing pre-training using large-scale image-narration datasets. Future work could consider how multimodal CR improves when incorporating specialized fine-grained pre-training objectives. As opposed to relying on image–text matching and masked language modeling, alternative methods such as fine-grained grounding alignment, context modeling or sentence completion could also be used as better pre-training strategies. Then, either the methods could be evaluated in a zero-shot way as shown in our experiments with VLMs, or fine-tuned using our proposed semi-supervised framework.

Chapter 6

Discussion and Future work

In this chapter, we summarize our contributions and propose directions for future research.

6.1 Summary of contributions

In Section 1.3, we highlighted the core challenges while training vision and language models for a variety of important tasks, especially image captioning, scene graph generation, and multimodal coreference resolution. Here, we review the contributions in the light of the challenges discussed in Section 1.3:

- **variability in visual content:** in Chapter 2, we introduce a model by incorporating two sources of implicit knowledge – 1) using latent language topics as anchors to find more salient image regions and, 2) transferring knowledge from a pre-trained sentence auto-encoder to the image captioning model. This enables us to learn the visual content in detail and generate semantically diverse image descriptions capturing the variations in images. Our experiments show that training with implicit knowledge helps to improve performance even when the data is limited.

- **variability in language:** in Chapter 3, we propose a model that captures the subjectivity in annotator interpretations for scene graph generation by using lexical knowledge. We propose a strategy to divide the label space according to linguistic rules and then use them to learn diverse and informative relations between objects, an important component in a scene graph. In Chapter 4, we investigated the question of variability in language for resolving coreferences. The complexity in narrative language and its relations to visual content was resolved by a joint vision-language model and in particular thanks to lexical knowledge rules that provided a strong language prior to resolve referrant expressions in language.

- **missing annotations:** an important contribution in Chapter 3 is to address missing annotations that arise due to variability in language and annotator subjectivity. Our proposed method takes care of the incomplete annotations with label relations learned from lexical knowledge. Incorporating this prior during learning helps to predict informative and plausible relations for scene graphs instead of overfitting to a single and uninformative relation.

- **difficulty in getting large-scale annotated training data:** we make two contributions to address this challenge. First, in Chapter 4 by relying on weak supervision from image-narration pairs and prior lexical knowledge to learn multimodal coreference resolution *i.e.,* clustering co-referrant mentions in text and linking these mentions to image regions. Second, in Chapter 5, we learn coreferences using a semi-supervised approach where we use a small labeled training set with coreference and bounding box labels along with a large unlabeled set. Hence, we benefit from transferring knowledge from a small subset of labels to the unlabeled data for learning complex associations in language and its association to images.

6.2 Future work

In this section, we propose possible directions for future work. First, we present some ideas to improve fine-grained modeling and reasoning in vision and language, and then we outline the importance of learning robust image and text correlations as discussed in this thesis for different tasks and problems.

6.2.1 Improving reasoning in vision and language

Generating narrations for images. Recent advances in multimodal foundational models (Li et al., 2022a, 2023a; Singh et al., 2022), have shown remarkable progress in generating concise captions for images which is also the main focus in Chapter 2. However, creating detailed narratives, such as extensive textual descriptions for an image, is equally vital. This task requires a deep comprehension of the context and intricate relationships that exist between various entities within both the textual description and the image. The ultimate aim is for the model to produce narratives that are not only coherent and contextually accurate but also richly descriptive. These narratives should encompass references to the objects, scenes, and the complex interplay between entities within the image.

However, when it comes to generating extensive descriptions, standard image captioning models face significant challenges (Meng et al., 2021). The descriptions they generate tend to exhibit repetitive patterns, often lacking a nuanced understanding of contextual references, which if addressed can significantly enhance the quality and depth of the generated narratives. To address these limitations, a potential solution involves incorporating the simultaneous learning of coreferences and narrative grounding as an auxiliary task. In this approach, narrative grounding will serve as a robust signal to guide the model in focusing on important image regions for description, while coreference resolution will aid in capturing the intricate relationships and connections between various entities within the narrative. By using this dual-task approach, we can significantly improve the comprehensiveness and contextual richness of the narratives generated for the images.

Using scene graphs for resolving coreferences and narrative grounding. In Chapter 3 of this thesis, we propose to generate image-based scene graphs which can also be a valuable tool for resolving coreferences and narrative grounding, especially in the context of visual scenes and complex narratives as in the CIN dataset. Scene graphs can help with this by providing a structured representation of objects and their attributes in a given scene or image. When analyzing the text, one approach could be to extract triplets from the text and then learn a matching between the image-based scene graph and the text-based scene graph to determine what a pronoun or noun phrase is referring to. For example, consider an image paired with the sentence, "A cat is sleeping on the couch next to a dog. She has black and white fur." From the scene graph of the image, we can extract relevant triplets such as *cat - sleeping on - couch*, *cat - next to - dog*, *cat - has - white and black fur*. Simultaneously, we can parse triplets from the text using a text-based scene graph parser (Schuster et al., 2015) which will result in triplets of the form *cat - sleeping on - couch* and *she - has - white and black fur*. From matching the image scene graph with the text scene graph, we can determine that *she* refers to the cat and not the dog.

However, due to the complexity of narrations the triplets extracted from the text-based parser might not be complete, making it challenging to learn a perfect matching between visual scene graphs and text scene graphs. Hence, it is crucial to devise methods that could address this noise and effectively use scene graphs as an intermediate representation for resolving coreferences.

Single-stage model for grounding and reasoning. In our models, detailed in Chapter 4 and Chapter 5, our approach involves a two-step process: first, we identify candidate object bounding boxes by utilizing a pre-trained object detector, and second, we then train a join vison-language model to align phrases to object regions and reason about the linked entities in text. It's important to note that due to this two-stage process, the model's performance (especially for grounding) is limited by the quality of object detection, as we can only match the phrases to regions that have been correctly detected in the first step. To address this issue, following previous work such as Li et al. (2022b), we can learn a single-stage method that learns to detect object regions for the phrases directly from the image using a regression objective during training. In this process, the model will automatically learn to disambiguate relevant regions for a phrase and also learn to link phrases based on the surrounding context. The main challenge is the unavailability of annotated training data for paired narration regions and coreferences. One possible solution could be to use distant supervision by aligning the image region to a labeled dataset for object detection by using the mouse traces in the Localized

Narratives dataset. This could provide noisy supervision for phrase-region matching for the narrations which can then be used in an end-to-end learning framework.

6.2.2 Beyond reasoning on images using vision and language

Conditional image generation. Recent advances in generative image modeling have enabled remarkable progress in conditional image generation. Two popular approaches are Generative Adversarial Networks (GANs) (Creswell et al., 2018) and diffusion models (Croitoru et al., 2023), which can synthesize highly realistic images given a text description (Li et al., 2019a; Qiao et al., 2019b). This has sparked great interest in text-to-image generation, where a textual query or caption is transformed into a corresponding image (Nichol et al., 2021; Qiao et al., 2019a,b; Xu et al., 2018). Such methods hold promise for creative applications like image editing (Brooks et al., 2023; Ling et al., 2021), image super-resolution (Saharia et al., 2022), inpainting (Wang et al., 2023; Yu et al., 2018a) and so-on. An alternative technique is a grounded-language-to-image generation (Li et al., 2023c,d), which leverages explicit spatial layout information and grounded text to enable more controlled image synthesis.

A promising research direction is to combine these complementary strengths - utilizing automatic scene graph parsing and captioning of a reference image to extract rich semantic representations as discussed in Chapter 2 and Chapter 3, then leveraging these to assist image generation from a text prompt and conditioned on the automatically generated layout. By automatically constructing detailed scene graphs and captions from reference images, the amount of human input needed for controlled image synthesis could be greatly reduced. Rather than manually specifying object bounding boxes and relationships, a user could simply provide an example image and a text description of the desired edit. This could significantly lower the barrier to utilizing advanced generative image models in creative workflows.

Extending the Coreferenced Image Narratives (CIN) dataset to videos. The multimodal coreference resolution framework proposed in this thesis in Chapter 4 and Chapter 5 focuses on static images. However, extending these coreference capabilities to video could enable an even richer grounding of language in visual data. Video introduces a temporal dimension, where models must leverage cues across frames to track entities and resolve references. For instance, consider a video where a person is sitting on a sofa in one frame, then cooking in the kitchen in the next frame. This requires understanding temporally that the same person is involved in two different tasks across the frames. This connects to challenges in multiple object tracking (Sun et al., 2020) and attaching textual descriptions to tracked identities (Sadhu et al., 2021; Yang et al., 2023).

Enabling more robust video-grounded coreference could unlock new possibilities in tasks like automatic video captioning (Yang et al., 2023), vision-and-language navigation (Anderson et al., 2018c), and action recognition (Jhuang et al., 2013; Yang et al., 2022a). For video captioning, correctly binding pronouns to visually tracked people/objects could improve caption coherence and factual consistency. In navigation scenarios, resolving references like “the person you just saw” can help agents interpret instructions containing temporal dependencies. And for action recognition, linking textual descriptions to tracked entities could provide useful supervision and context.

Overall, by extending the static image coreference models proposed in this thesis to the video domain, there is an exciting opportunity to enable a richer understanding of events, actions, and relationships as they unfold over time. This could lead to AI systems that can not only resolve coreference in images but also reason about the interactions and narratives depicted in videos or even live environments. Tackling the additional challenges introduced by visual dynamics represents a promising research direction.

Evaluation of large vision and language models. Recent work has demonstrated impressive capabilities of large-scale vision-language models like UNITER(Chen et al., 2020), ALBEF (Li et al., 2021), Flamingo (Alayrac et al., 2022), BLIP (Li et al., 2022a), BLIP-2 (Li et al., 2023a), and PaLI (Chen et al., 2022) across a variety of multimodal tasks. Leveraging massive datasets (Penedo et al., 2023; Schuhmann et al., 2022) and model sizes (Hoffmann et al., 2022), these models have achieved strong performance even with zero-shot transfer or simple fine-tuning, particularly on high-level tasks like image captioning (Lin et al., 2014) and visual question answering (Antol et al., 2015). However, evaluating these powerful models on fine-grained vision-language tasks could provide unique insights into their learned representations.

Two promising directions are extending assessments to scene graph generation (Chapter 3) and multimodal coreference resolution (Chapter 4 and Chapter 5). As discussed in this thesis, these tasks require grounding textual concepts and relationships to specific image regions, going beyond caption-level understanding. This grounding could be approached by formulating the tasks as question answering, with scene graph links or coreferent expressions encoded as queries. However, this would mainly assess the alignment of modalities rather than structured representation. Alternately, models could be adapted to directly output connected graphs aligning detected entities and phrases. Although this latter approach requires architectural changes to support structured outputs, it could enable more nuanced probing of learned cross-modal reasoning.

Overall, evaluating emerging multimodal models on structured scene-level tasks beyond captioning presents exciting research for the future. Success would demonstrate deeper semantic visual understanding, including inferring relationships and resolving ambiguity. By innovating new model architectures and training techniques tailored for these challenging tasks, we can continue pushing towards human-like language grounding abilities. Benchmarking performance on scene graphs and coreference thus provides informative next steps beyond standard vision-language benchmarks.

Appendix A

Injecting prior knowledge into image caption generation

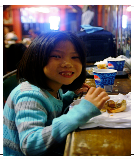
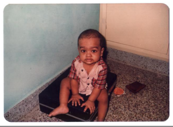
Image	Captions	Ground Truths
	**AoANet:** a woman sitting at a table with a cup of coffee. **+CLTA+SAE-Reg:** a woman sitting at a table eating a hot dog. **Top-20 Topic Words:** woman, table, holding, eating, girl, sitting, standing, group, hand, food, sandwich, plate, young, front, little, picture, lady, hold, next, man.	GT1: a girl eating a hotdog at a wooden table. GT2: an asian child eating a hot dog sitting at a table. GT3: a little girl posing for a picture while eating food. GT4: a young girl smiles while enjoying her meal. GT5: little girl smiles for the camera as she eats her sandwich.
	**AoANet:** a little boy sitting on a chair with a laptop. **+CLTA+SAE-Reg:** a baby sitting on a suitcase on the floor. **Top-20 Topic Words:** luggage, boy, suitcase, bag, child, young, sitting, shoe, elephant, kid, next, little, small, piece, bench, clothes, floor. airport, case, standing.	GT1: young boy sitting on top of a briefcase. GT2: a little boy sitting on a suitcase on the floor. GT3: a toddler boy is sitting on a brief case. GT4: a young baby sits on top of a briefcase. GT5: a small child sitting on top of a briefcase.
	AoANet: a row of motorcycles parked next to each other. **+CLTA+SAE-Reg:** a row of motorcycles parked on the side of a street. **Top-20 Topic Words:** wall, street, motorcycle, brick, city, building, row, meter, group, next, parking. side, clock, two, lined, front, sidewalk, parked, road, graffiti.	GT1: a row of motorcycles parked in front of a building. GT2: a number of motorbikes parked on an alley. GT3: a bunch of motorcycles parked along the side of the street. GT4: a back ally neighborhood with motor bikes in a row. GT5: a bunch of motorcycles parked on the side of the road.
	AoANet: a vase filled with white and white flowers. **+CLTA+SAE-Reg:** a glass vase with some flowers and a doll. **Top-20 Topic Words:** flower, vase, purple, group, vas, white, table, sitting, filled, glass, rose, watch, watching, together, around, small, large, yellow, next, green.	GT1: a doll standing next to a vase filled with flowers and plants. GT2: a clear vase contains some white pretty flowers. GT3: white flowers in a vase with arranged leaves GT4: flowers are placed in a vase with large leaves. GT5: a flower arrangement in a clear glass vase next to a doll.
	AoANet: a person on a surfboard in the water. **+CLTA+SAE-Reg:** a man is para sailing in the ocean. **Top-20 Topic Words:** kite, man, surfboard, water, flying, wave, board, person, ocean, riding, surfer, surf. surfing, group, fly. beach, sky. large, top, red.	GT1: a surfer wrangles a parachute with a scenic mountain background. GT2: a paraglider who has just landed in the ocean GT3: a person riding a surf board with a parachute in a body of water GT4: a man in the water kite surfing on a board. GT5: a man kite boarding over a large body of water.
	AoANet: a display case filled with lots of different flavored donuts. **+CLTA+SAE-Reg:** a box of donuts with different types of sprinkles. **Top-20 Topic Words:** group, many, different, several, various, two, together, type. colorful, donut, bunch, watching, around, color, plate, crowd. table, filled, others, gathered.	GT1: a display of yummy looking chocolate covered donuts with sprinkles. GT2: donuts on a tray in a display case GT3: a bunch of doughnuts with sprinkles on them GT4: tray of chocolate donuts with sprinkles in a display case. GT5: a tray full of sprinkle covered, chocolate glazed doughnuts.
	AoANet: a couple of people sitting on a bench. **+CLTA+SAE-Reg:** two people sitting on a bench looking out at the ocean **Top-20 Topic Words:** people, two, group, sitting, bench, couple, standing, wooden, park, woman, one, water, next, around. near, several, another, together, sits, crowd.	GT1: a couple is sitting on a bench in front of the water. GT2: two people sitting on a bench silhouetted against the sea. GT3: two people are sitting on a bench together in front of water. GT4: the silhouette of two people sitting on a bench in front of the water. GT5: a couple sits on a park bench and watches the water

Figure A.1 Qualitative examples of generated captions from the baseline AoANet model and AoANet with our proposed CLTA and SAE Regularizer. We also visualize the Top-20 Topic words from the learned latent topic space in our CLTA module.

Appendix B

Not all relations are equal: Mining informative labels for scene graph generation

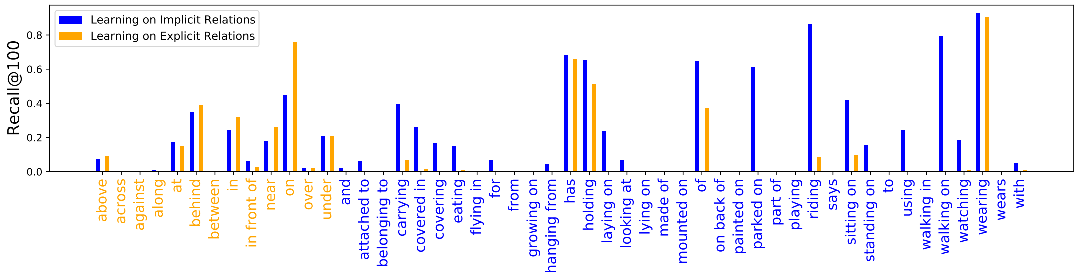


Figure B.1 Class Wise Recall@100 for the set of Implicit and Explicit Relations by training on a subset of relations. All models are trained with Motif-TDE-Sum (Tang et al., 2020; Zellers et al., 2018).

Explicit Relations	above	across	against	along	at	behind	between	in	in front of	near	on	over	under
# of Instances	47341	1996	3092	3624	9903	41356	3411	251756	13715	96589	712409	9317	22596

Table B.1 Explicit Relations for the Visual Genome Dataset (Krishna et al., 2017).

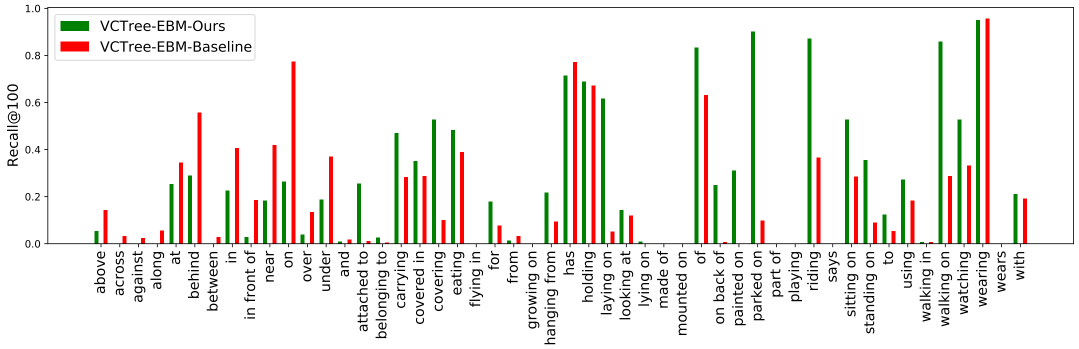


Figure B.2 Relation-wise Recall using the VCTree-EBM (Suhail et al., 2021) as the backbone SGG model trained with our proposed method (Ours) vs. the method proposed in (Suhail et al., 2021) (Baseline).

B.1 Dataset statistics

For the Visual Genome dataset (Krishna et al., 2017), Xu et al. (2017) released a version of the dataset with 50 relations and 150 object categories. These 50 relations are: *above, across, against, along, and, at, attached to, behind, belonging to, between, carrying, covered in, covering, eating, flying in, for, from, growing on, hanging from, has, holding, in, in front of, laying on, looking at, lying on, made of, mounted on, near,*

Implicit Relations	attached to	and	belonging to	carrying	covered in	covering	eating	flying in	for	from	growing on	hanging from	has	holding	laying on	looking at
# of Instances	10190	3477	3288	5213	2312	3806	4688	1973	9145	2945	1853	9894	277936	42722	3739	3083

Table B.2 Implicit Relations for the Visual Genome Dataset (Krishna et al., 2017).

Implicit Relations	lying on	made of	mounted on	of	on back of	painted on	parked on	part of	playing	riding	says	sitting on	standing on	to	using	walking in	walking on	watching	wearing	wears	with
# of Instances	1869	2380	2253	146339.	1914	3095	2721	2065	3810	8856	2241	18643	14185	2517	1925	1740	4613	3490	136099	15457	66425

Table B.3 Implicit Relations for the Visual Genome Dataset (Krishna et al., 2017).

of, on, on back of, over, painted on, parked on, part of, playing, riding, says, sitting on, standing on, to, under, using, walking in, walking on, watching, wearing, wears, with.

In Table B.1, we present the explicit set of relations with the number of training instances for each relation in the Visual Genome dataset. Similarly, in Table B.2 and Table B.3, we define the implicit set of relations[1] and their frequency in the Visual Genome dataset.

B.2 Additional results

Quantitative studies. In Figure B.1, we present the class-wise recall for all the relation classes in the Visual Genome dataset (Krishna et al., 2017) for training on a subset of relations *i.e.,* either learning only on *explicit* relations or only on *implicit* relations. All the models are trained with the MOTIF-TDE-Sum SGG model (Tang et al., 2020; Zellers et al., 2018). The class-wise performances clearly indicate the generalizability of training only on implicit relations as it achieves at-par/similar performances on the explicit relations, whereas the model only trained on explicit relations performs poorly on implicit relations.

Figure B.2 compares the class-wise performances for the VCTree model trained with only Energy Based Modeling (EBM) (Suhail et al., 2021) and also with our proposed method. The performance gains of our model over the baseline in the implicit relations such as “carrying”, “eating”, “covering”, walking” *etc.* shows the importance of mining these informative relations from less informative samples while still maintaining recall of the explicit relations hence, improving generalization.

Regular recall results. In Table B.4, we show the Regular Recall@k results for different SGG backbone architectures when trained with our proposed method compared to

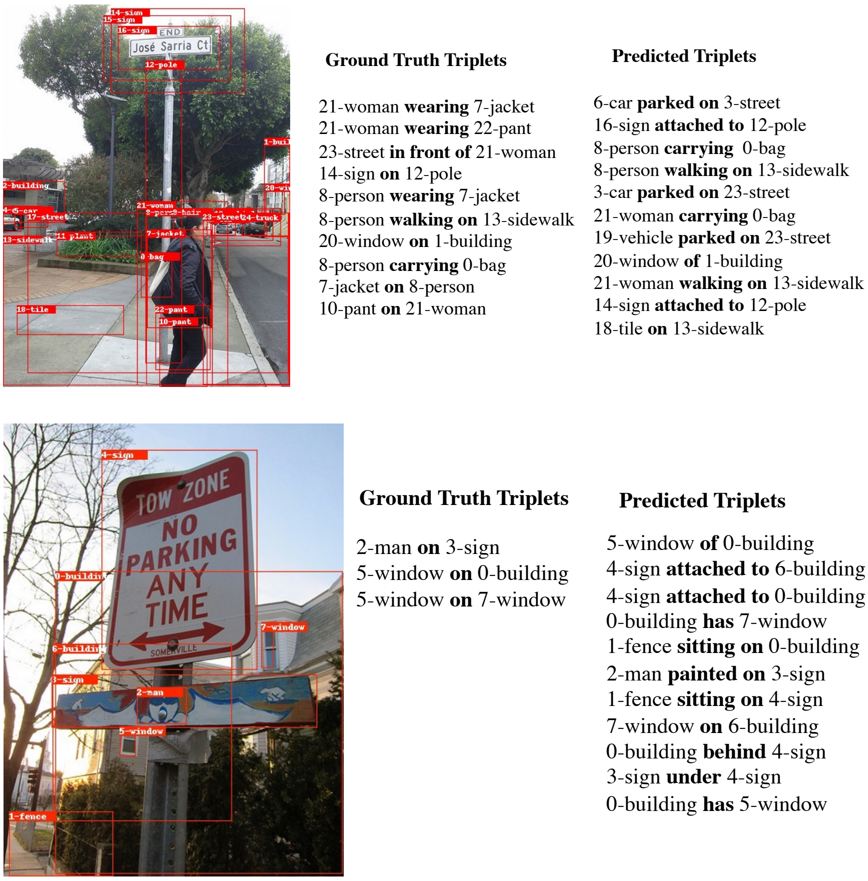
[1] We break down the implicit relations into two tables for better visualization.

the baseline. Although there is no significant improvement in Regular Recall (when compared to the improvements obtained from mean recall), the at-par performance with the baseline shows that our method maintains the performance on frequent relations while improving significantly on the more informative/infrequent relation classes (as measured by mean recall).

		Predicate Classification			Scene Graph Classification			Scene Graph Detection		
Models	Method	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
Motif-TDE-Sum (Tang et al., 2020; Zellers et al., 2018)	Baseline	33.38	45.88	51.25	20.47	26.31	28.79	11.92	16.56	20.15
	Ours	33.36	43.53	47.44	24.31	29.91	31.75	14.59	17.96	19.70
VCTree (Tang et al., 2019)	Baseline	59.82	65.93	67.57	41.49	45.16	46.10	24.90	32.02	36.30
	Ours	58.66	64.69	67.05	35.49	38.71	39.51	24.63	31.52	36.42
VCTree-EBM (Suhail et al., 2021)	Baseline	57.31	63.99	65.84	40.31	44.72	45.84	24.21	31.36	35.87
	Ours	57.42	64.37	66.43	35.42	38.79	39.66	23.70	30.74	35.62
VCTree-TDE (Tang et al., 2020)	Baseline	40.12	50.83	54.91	26.00	33.03	35.97	13.97	19.43	23.34
	Ours	36.90	47.62	52.03	25.67	32.83	35.76	15.20	19.00	20.98

Table B.4 Scene Graph Generation performance comparison on Regular Recall@K (Tang et al., 2020) under all three experimental settings. We compare the results of our proposed framework (Ours) with the original model (Baseline) using different SGG architectures.

Qualitative studies. We present additional qualitative visualizations in Figure B.3 and Figure B.4. Our proposed method predicts informative relations for both the set of pairs present in the ground truth and the new set of object pairs that further help to define a scene comprehensively. In the quantitative evaluation, we only reward object pairs that have corresponding ground truth relations, hence, the relations for the remaining set of object pairs can only be visualized qualitatively.



Ground Truth Triplets

21-woman **wearing** 7-jacket
21-woman **wearing** 22-pant
23-street **in front of** 21-woman
14-sign **on** 12-pole
8-person **wearing** 7-jacket
8-person **walking on** 13-sidewalk
20-window **on** 1-building
8-person **carrying** 0-bag
7-jacket **on** 8-person
10-pant **on** 21-woman

Predicted Triplets

6-car **parked on** 3-street
16-sign **attached to** 12-pole
8-person **carrying** 0-bag
8-person **walking on** 13-sidewalk
3-car **parked on** 23-street
21-woman **carrying** 0-bag
19-vehicle **parked on** 23-street
20-window **of** 1-building
21-woman **walking on** 13-sidewalk
14-sign **attached to** 12-pole
18-tile **on** 13-sidewalk

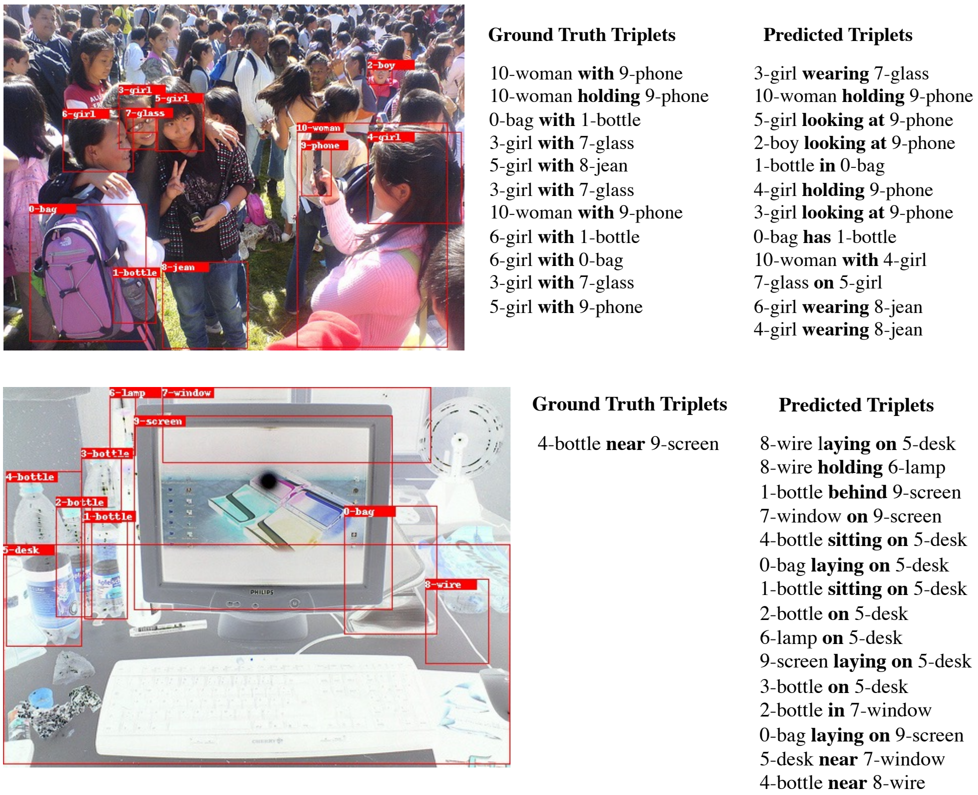
Ground Truth Triplets

2-man **on** 3-sign
5-window **on** 0-building
5-window **on** 7-window

Predicted Triplets

5-window **of** 0-building
4-sign **attached to** 6-building
4-sign **attached to** 0-building
0-building **has** 7-window
1-fence **sitting on** 0-building
2-man **painted on** 3-sign
1-fence **sitting on** 4-sign
7-window **on** 6-building
0-building **behind** 4-sign
3-sign **under** 4-sign
0-building **has** 5-window

Figure B.3 Additional Qualitative Results with the ground truth triplets and the predicted triplets from the VCTree-EBM model trained with our proposed training framework. The predicted triplets are from the SGCls setting.



Ground Truth Triplets

10-woman **with** 9-phone
10-woman **holding** 9-phone
0-bag **with** 1-bottle
3-girl **with** 7-glass
5-girl **with** 8-jean
3-girl **with** 7-glass
10-woman **with** 9-phone
6-girl **with** 1-bottle
6-girl **with** 0-bag
3-girl **with** 7-glass
5-girl **with** 9-phone

Predicted Triplets

3-girl **wearing** 7-glass
10-woman **holding** 9-phone
5-girl **looking at** 9-phone
2-boy **looking at** 9-phone
1-bottle **in** 0-bag
4-girl **holding** 9-phone
3-girl **looking at** 9-phone
0-bag **has** 1-bottle
10-woman **with** 4-girl
7-glass **on** 5-girl
6-girl **wearing** 8-jean
4-girl **wearing** 8-jean

Ground Truth Triplets

4-bottle **near** 9-screen

Predicted Triplets

8-wire **laying on** 5-desk
8-wire **holding** 6-lamp
1-bottle **behind** 9-screen
7-window **on** 9-screen
4-bottle **sitting on** 5-desk
0-bag **laying on** 5-desk
1-bottle **sitting on** 5-desk
2-bottle **on** 5-desk
6-lamp **on** 5-desk
9-screen **laying on** 5-desk
3-bottle **on** 5-desk
2-bottle **in** 7-window
0-bag **laying on** 9-screen
5-desk **near** 7-window
4-bottle **near** 8-wire

Figure B.4 Additional Qualitative Results.

Appendix C

Who are you referring to? Coreference resolution in image narrations

C.1 Annotation details

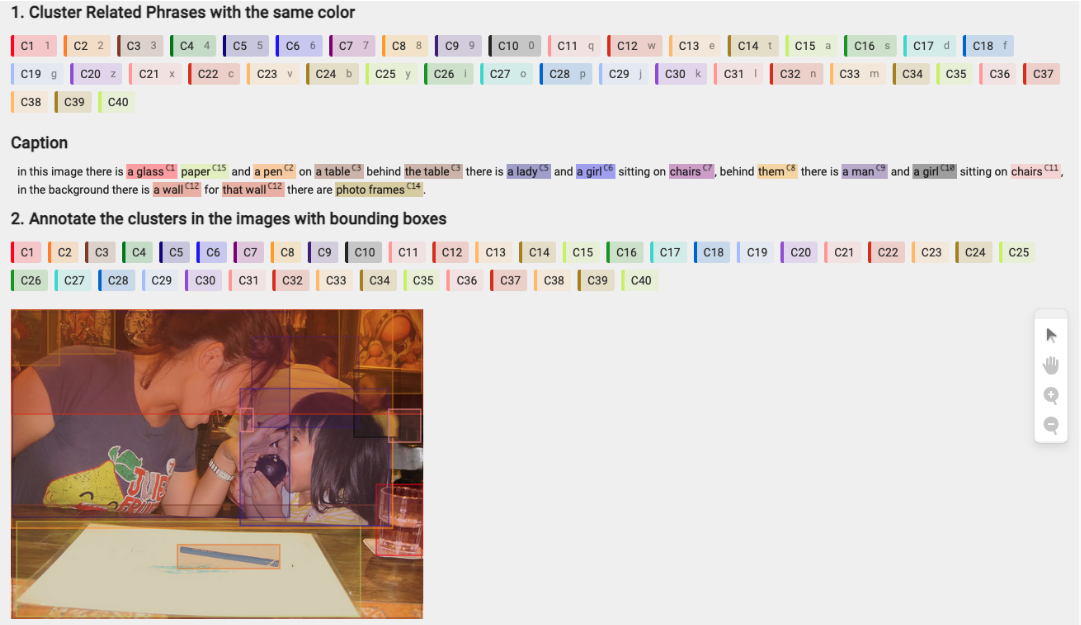
Localized Narratives dataset. Pont-Tuset et al. (2020) proposed the Localized Narratives dataset, a new form of multimodal image annotations connecting vision and language. In particular, the annotators describe an image with their voice while simultaneously hovering their mouse over the region they are describing. Hence, each image is described with a natural language description attending to different regions of the image. In addition to textual descriptions (obtained using speech-to-text conversion), they additionally provide mouse traces for the words.

The Localized Narratives dataset is built on top of COCO (Lin et al., 2014), Flickr30k (Plummer et al., 2015), ADE20k (Zhou et al., 2017) and Open Images (Kuznetsova et al., 2020). The statistics of the individual datasets are shown in Table C.1.

Localized Narratives Subsets (Pont-Tuset et al., 2020)	#images	#captions	#words/capt.
COCO	123,287	142,845	41.8
Flickr30k	31,783	32,578	57.1
ADE20k	22,210	22,529	43.0
Open Images	671,469	675,155	34.2

Table C.1 Statistics of Localized Narratives for COCO, Flickr30k, ADE20k, and Open Images.

Annotation tool and analysis. We develop an HTML-based interface on the Label Studio annotation tool (lab). Figure C.1 shows the annotation interface from Label Studio. We hired 6 high-quality annotators (all from computer science background) for an average of 54 hours of annotation time. The annotators were trained with the exact description of the task and given a pilot study before proceeding with the complete annotations. The pilot study was useful to correct and retrain annotators if needed. As shown in Figure C.1, the annotators had to select a mention in the caption with a given label (C1, C2, etc.) in Step 1 and draw a bounding box in the image for the selected mention in Step 2 (with the same label).



1. Cluster Related Phrases with the same color

Caption

in this image there is a glass^{C1} paper^{C15} and a pen^{C2} on a table^{C3} behind the table^{C3} there is a lady^{C5} and a girl^{C6} sitting on chairs^{C7}, behind them^{C8} there is a man^{C9} and a girl^{C10} sitting on chairs^{C11}, in the background there is a wall^{C12} for that wall^{C12} there are photo frames^{C14}.

2. Annotate the clusters in the images with bounding boxes

Figure C.1 Annotation interface from Label Studio.

For Step 1, if the mention is coreferring then it is selected with the same label to define coreference chains. It is important to note that the captions are pre-marked with noun phrases parsed from (spa). The annotators are instructed to correct the phrases if they are wrong (*e.g.,* for a mention glass windows, the parser parses *glass* and *windows* as two different mentions rather than belonging to the same label/cluster) and remove the phrases that do not correspond to a region in the image.

In Step 2, if there are plural mentions such as *two men*, we ask the annotators to draw two separate bounding boxes for this. In the case of mentions such as *several* *people* if the people are less than five, they are instructed to draw separate bounding boxes otherwise a group bounding box (covering all the people).

Given the challenging nature of the task, we doubly annotate 30 images with coreference chains and bounding boxes to compute the inter-annotator agreement. More specifically, for the coreference chain we compute *Exact Match* which denotes whether the coreference chains annotated by the two annotators are the same. We get an exact match of 79.9% in the coreference chains, which is a high agreement given the complexity of the task. For the bounding box localization, we compute the Intersection over Union (IoU) to compute the overlap between the two annotations. It is considered to be correct/matching if the IoU is above 0.6. We achieve bounding box accuracy of 81% on this subset of images. This analysis shows good agreement between the annotators given the subjective nature and complexity of the task.

Coreferenced Image Narratives dataset. In total, we annotate all the 1000 test images and 880 validation images (out of 1000) in the Flickr30k dataset. The text descriptions from the Localized Narratives dataset are very noisy with a lot of words/sequence of words. We manually filter phrases such as - *in this image, in the front, in the background, we can see, i can see, in this picture*. If there are some other mentions that are pre-marked and not filtered, we ask the annotators explicitly to filter them out. By doing this, we make sure that the dataset is clear of any unnecessary and noisy mentions.

All the words that are marked as mentions and are not noun phrases (as detected by the part of speech tagger (spa)) are considered as pronouns *e.g., them, they, their, this, that, which, those, it, who, he, she, her, him, its*.

Statistics for the Coreferenced Image Narratives . In Figure C.2, we show the statistics for the frequency of pronouns in the dataset. Few pronouns (*e.g.,* he, it, them) are more frequent than the others. Overall, the occurrence of pronouns is frequent to conduct a fair evaluation of the coreference based models. Similarly in Figure C.3, we evaluate how many mentions occur in the coreference chains. Coreference chains with 2 and 3 mentions have a very high frequency in the dataset. There are few chains that

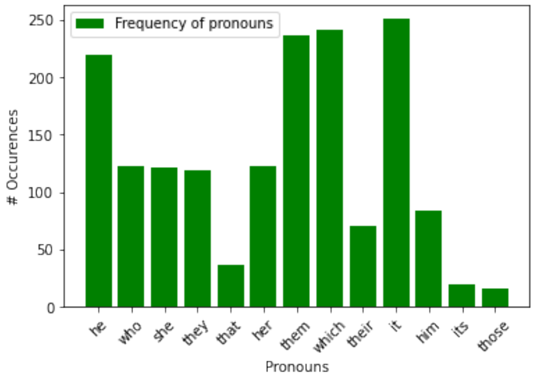


Figure C.2 Total number of occurrences of pronouns in Coreferenced Image Narratives .

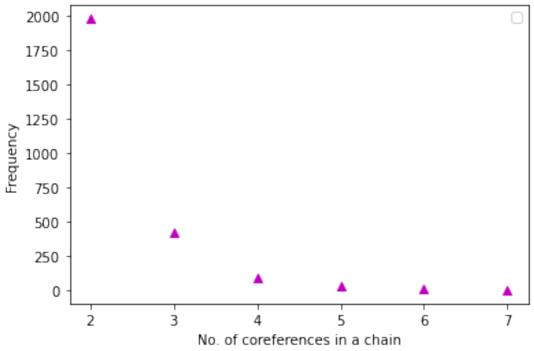


Figure C.3 Number of coreference chains with 2 or more than 2 mentions in a chain in Coreferenced Image Narratives .

have longer mentions (*e.g.,* 6 and 7). Hence, we can safely conclude that the dataset is a powerful tool to evaluate coreference chains and learn complex coreferencing and grounding models. Moreover, the average length of the mentions (excluding pronouns) is 1.93.

C.2 Evaluation metrics

In this section, we discuss in detail the evaluation metrics used for CR and narrative grounding. For CR, we use the MUC and the BLANC metrics, which are discussed below.

(a) MUC F-measure. It measures the number of coreference links (pairs of mentions) common to the predicted R and ground-truth chains K. It involves computing the partitions with respect to the two chains:

$$\text{MUC-R} = \frac{\sum_{i=1}^{N_k}(|K_i| - |p(K_i)|)}{\sum_{i=1}^{N_k}(|K_i| - 1)}, \tag{C.1}$$

$$\text{MUC-P} = \frac{\sum_{i=1}^{N_r}(|R_i| - |p^{'}(R_i)|)}{\sum_{i=1}^{N_r}(|R_i| - 1)} \tag{C.2}$$

where K_i is the i^{th} ground-truth chain and $p(K_i)$ is the set of partitions created by intersecting K_i with the output chains; R_i is the i^{th} output chain and $p^{'}(R_i)$ is the set of partitions created by intersecting R_i with the ground-truth chains; and N_k and N_r are the total number of ground-truth and output chains, respectively.

(b) BLANC. Let C_k and C_r be the pairs of coreference links respectively, and N_k and N_r be the set of non-coreference links in the ground-truth and output respectively. The BLANC Precision and Recall for coreference links is calculated as follows:

$R_c = \frac{|C_k \cup C_r|}{|C_k|}$ and $P_c = \frac{|C_k \cup C_r|}{|C_r|}$, where R_c and P_c are the recall and precision respectively.

Similarly, recall R_n and precision P_n for non-coreference links (N_k and N_r) are computed. The overall precision and recall are:

BLANC-R $= \frac{(R_c + R_n)}{2}$ and BLANC-P $= \frac{(P_c + P_n)}{2}$, respectively.

For evaluating narrative grounding in images, we consider a prediction to be correct if the IoU (Intersection over Union) score between the predicted bounding box and the ground truth box is larger than 0.5 (Gupta et al., 2020; Wang et al., 2020a). Following (Kamath et al., 2021), if there are phrases with multiple ground truth boxes (*e.g.,* several people), we use the any-box protocol *i.e.,* if any ground truth bounding box overlaps the predicted bounding box, it is a correct prediction. We report percentage accuracy for evaluating narrative grounding.

C.3 Implementation details

Inputs and modules. For the image modeling, we extract bounding box regions, visual features, and object class labels using the Faster-RCNN object detector (Ren et al., 2015). We use Glove embeddings (Pennington et al., 2014) to encode the object class labels and the mentions from the textual branch. For the mouse traces, we follow (Pont-Tuset et al., 2020) and extract the trace for each word in the sentence and then convert it into bounding box coordinates for the initial representation. All the modules *i.e.,* image encoder, text encoder, trace encoder, and joint text-trace encoder are a stack of two transformer encoder layers. Each transformer encoder layer includes a multi-head self-attention layer and an FFN. There are two heads in the multi-head attention layer, and two FC layers followed by ReLU activation layers in the FFN. The output channel dimensions of these two FC layers are 2048 and 1024, respectively. The input to the joint text-trace encoder comes from the separate text and trace encoder branches. We add a special embedding to the learned embeddings following (Chen et al., 2020) to distinguish between the two modalities (text and trace) in the transformer encoder.

Training details. The whole architecture is trained end-to-end with the AdamW (Loshchilov and Hutter, 2017) optimizer. We train the transformer encoders with the learning rate of 3e-5, batch size of eight, weight decay of 0.01, and the loss coefficient λ of 0.001. We train the model for 60 epochs and choose the best-performing model based on the validation set.

C.4 Zero-shot results on Flickr30k dataset

Method	zs-MUC-R	zs-MUC-P	zs-MUC-F1	zs-Grounding Acc. (%)
VinVL	59.16	60.78	57.24	-
MAF†	61.97	68.46	63.91	57.1
Ours (w/o MT)	70.11	68.67	68.48	59.4

Table C.2 Zero-shot performance on the Flickr30k entities dataset.

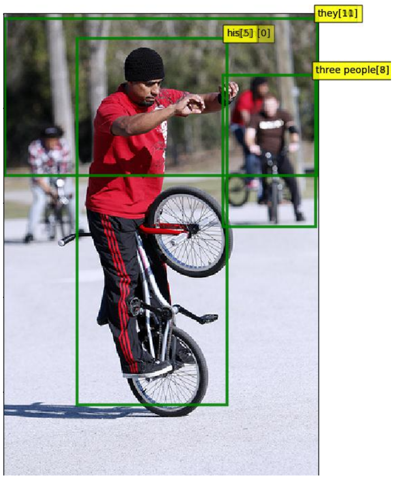
In Table C.2, we evaluate our model and baselines using the zero-shot setting on the Flickr30k entities dataset Plummer et al. (2015) for CR and grounding. These results indicate that our method better generalizes to unseen CR chains and narrative grounding than the baselines.

C.5 Additional qualitative results

In Fig. C.4, we show additional qualitative results from our proposed method. The model correctly chains mentions and grounds them to the correct entities in the image even for complex and ambiguous cases. Our model finds coreferences for people (*e.g., [a man, his]*) or for objects (*e.g., [a barbecue grill, it]*). Moreover, it also finds links for plurals such as *[two men, them]*. There is a huge potential in learning to disambiguate the mentions in the descriptions and this work paves the way for future research.

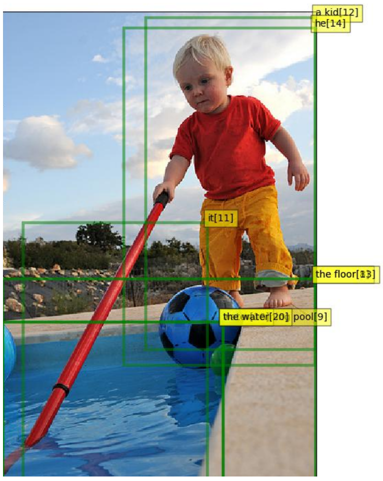
Narration: in this picture i can see **a man[0]** doing stunts with **a bicycle[1]**, **he[2]** is wearing **a cap[3]** on **his[5] head[4]**. i can see **three people[6-8]** in the back, **they[9-11]** are riding **bicycles[12]**. i can see **the ground[13]** at the bottom and **the trees[14]** in the background and **it[15]** looks like **grass[16]** on **the ground[17]** in the back.

Predicted Coreference Chains: [**a man[0], he[2], his[5]**],
[**three people[6-8], they[9-11]**]



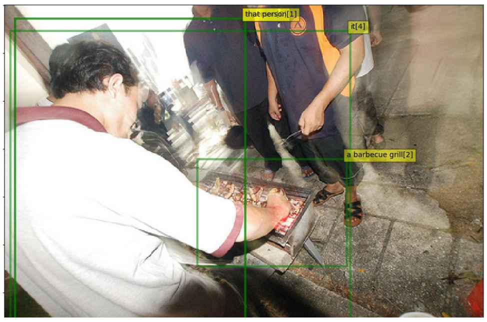
Narration: this image is taken outdoors. at the top of the image there is **sky with clouds[1]**. in the background we can see there are **many plants[2]** and **trees[3]**. we can see **the mesh[4]**. there are **many rocks[5]**. at the bottom of the image there is **the floor[6]**. we can see **the swimming pool[9]** with **water[10]** in **it[11]**. in the middle of the image **a kid[12]** is standing on **the floor[13]** and **he[14]** is holding **a stick[15]** in **the hand[16]** and playing. we can see **the balls[17]** in **the water[20]**.

Predicted Coreference Chains: [**a kid[12], he[14]**],
[**the swimming pool[9], water[10], it[11], the water[20]**] ,
[**the floor[6], the floor[13]**]



Narration: on the left side of the image there is **a person[0]**. in front of **that person[1]** there is **a barbecue grill[2]** with a **food item[3]** on **it[4]**. and there are **few people[5]** standing. this is an edited image. and there is a blur background. and there are few other things in the background.

Predicted Coreference Chains: [**a person[0], that person[1]**],
[**a barbecue grill[2], it[4]**]



Narration: in front of the picture, we see **two men[0]**. **the man[2]** on the left side is wearing the **spectacles[3]** and **he[4]** is trying to talk something. **the man[5]** on the right side is wearing the **goggles[6]** and **an orange cap[7]**. **it[8]** looks like **a man[9]** is holding **a wooden stick[10]**. behind **them[11-12]**, we see **the people[13]** and **some of them[14]** are wearing **the orange color caps[15]**. this picture is blurred in the background.

Predicted Coreference Chains: [**the man[2], he[4], a man[9]**],
[**two men[0], them[11-12]**]

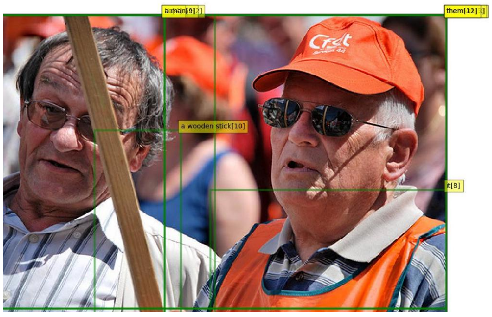


Figure C.4 Additional qualitative results for coreference chains. For each image, we show the predicted coreference chain (mentions more than 2) and the grounding results for the corresponding mentions in the chain. The colored mentions in the descriptions are the ground-truth coreference chains.

Appendix D

Semi-supervised coreference resolution in image narrations

D.1 Baselines

Method	MUC			B^3			CEAF$_{\phi 4}$			CoNLL
	R	P	F1	R	P	F1	R	P	F1	F1
Weights from ALBEF (Li et al., 2021)	**31.70**	25.97	26.47	70.78	78.35	73.77	54.58	89.02	66.94	55.73
Ours (Weights from BERT (Devlin et al., 2018))	31.11	**35.25**	**31.86**	**70.63**	**87.85**	**78.06**	**63.99**	**93.44**	**75.47**	**61.79**

Table D.1 CR performance with ALBEF (Li et al., 2021) as pre-trained weights.

Grounding	MUC			B^3			CEAF$_{\phi 4}$			CoNLL
threshold	R	P	F1	R	P	F1	R	P	F1	F1
0.0	28.20	22.47	23.08	70.10	76.29	72.40	52.32	87.22	64.71	53.40
0.5	31.18	30.04	28.92	70.80	82.67	75.76	59.68	92.18	71.77	58.82
0.7	30.48	33.34	30.58	70.63	86.74	77.58	63.30	93.22	74.85	61.01
0.9	**31.11**	**35.25**	**31.86**	**70.63**	**87.85**	**78.06**	**63.99**	**93.44**	**75.47**	**61.79**

Table D.2 Performance of our proposed method by varying the grounding threshold t to include samples above this threshold in Eq. (5.9).

We consider the following baselines to fairly compare and evaluate our proposed method:

(a) Text-only CR: For all these methods, we directly evaluate the coreference chains using the narration only without the image. (1) *Neural-Coref (Lee et al., 2017):* This method is trained end-to-end using a neural network on a large corpus of Wikipedia data to detect mentions and coreferences. For this baseline, we use the predicted mentions instead of the gold mentions. (2) *longdoc (Toshniwal et al., 2021):* This is a strong transformer-based method using Longformer-Large as the backbone for coreference resolution, trained on multiple datasets. We use the gold mentions to predict coreference chains for this model.

(b) Multi-modal CR: We evaluate strong vision and language models for the task on coreference resolution on the CIN dataset. (1) *VisualBERT (Su et al., 2019), UNITER (Chen et al., 2020), VinVL (Zhang et al., 2021b):* All these three baselines are strong vision language models trained on image-caption data and shows improvements on a variety of downstream tasks such as VQA, NLVR etc. To test it for CR, we compute the cosine similarity for the multi-modal mention embeddings in a zero-shot way. (2) *MAF (Wang et al., 2020a):* MAF is a weakly supervised phrase grounding method, originally trained on the Flickr30k-Entities (Plummer et al., 2015). We train this model on narrations data and evaluate it for CR as elaborated in Chapter 4. (3) *WS-MCR:* As discussed in Chapter 4, this is a weakly supervised method for multimodal coreference resolution trained on the CIN dataset. We train a vision and text encoder with an image-text contrastive loss and weak prior linguistic rules as a regularizer.

D.2 Bounding box regression loss (BBR)

We define the smmoth-L1 loss (Ren et al., 2015) for the bounding box transformation as follows:

$$\mathscr{L}_{bbr} = \begin{cases} 0.5(b_m - b'_m)^2/\beta, & \text{if } |b_m - b'_m| < \beta \\ |b_m - b'_m| - 0.5 * \beta, & \text{otherwise} \end{cases} \tag{D.1}$$

where β is set to 1 following previous work (Girshick, 2015; Ren et al., 2015), b'_m are the transformed bounding box coordinates after applying the delta transformation δ_m on the maximum region proposal r_m similar to (Girshick, 2015).

D.3 Training details

The whole architecture is trained end-to-end with the AdamW (Loshchilov and Hutter, 2017) optimizer. The initial learning rate of the model is 1e-5. The learning rate is gradually warmed up for 2 epochs with a unit multiplier and then decayed following a step scheduler with a step size of 10 epochs and gamma of 0.95. We use a batch size of eight and a weight decay of 0.01. The model is trained for 30 epochs and we choose the best performing model based on the test set. The model is trained on 4 V100 GPUs with data parallelism. All code and models will be made available at https://github.com/VICO-UoE/CIN-SSL.

D.4 Further ablations

Pre-trained weights from ALBEF (Li et al., 2021). In Table D.1, we show CR results by replacing the text-encoder and the multi-modal encoder in our model from the ALBEF (Li et al., 2021) framework with 6 transformer blocks in each encoder. Moreover, we use the pre-trained weights from ALBEF (Li et al., 2021) and then fine-tune them using our semi-supervised training strategies. Despite strong pre-training from ALBEF, the model does not fine-tune and transfer well to the task of MCR compared to our method which is initialized with BERT weights.

Varying the threshold for pseudo-grounding loss. In Table D.2, we compare the results with different grounding thresholds for the pseudo grounding loss. Including all the pseudo predictions (threshold of 0.0) leads to a significant drop in performance showing the importance of thresholding-based training strategy. Furthermore, as presented in the results the threshold of 0.9 works best compared to 0.5 and 0.7.

Varying the amount of labeled and unlabeled data. In Table D.3 and Table D.4, we present detailed results from the discussion in Section 5.5.3 on different CR metrics by varying the amount of labeled and unlabeled data.

% $\mathscr{D}_s$	MUC			B^3			CEAF$_{\phi 4}$			CoNLL
	R	P	F1	R	P	F1	R	P	F1	F1
20%	26.40	31.18	27.26	69.83	88.26	77.75	64.25	92.02	75.11	60.04
50%	28.65	34.13	29.91	70.26	88.83	78.27	64.55	92.56	75.53	61.24
100%	**31.11**	**35.25**	**31.86**	**70.63**	**87.85**	**78.06**	**63.99**	**93.44**	**75.47**	**61.79**

Table D.3 CR performance by varying the number of labels in the labeled dataset.

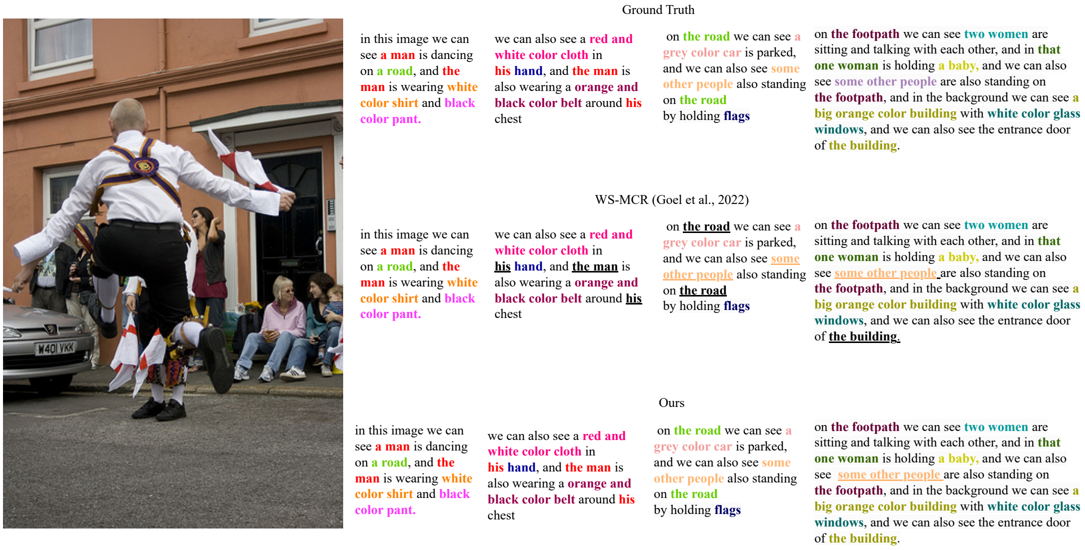
% $\mathscr{D}_u$	MUC			B^3			CEAF$_{\phi 4}$			CoNLL
	R	P	F1	R	P	F1	R	P	F1	F1
20%	30.96	27.09	27.20	70.84	79.36	74.23	56.49	90.79	69.04	56.82
50%	30.87	29.97	28.83	70.75	83.58	76.23	60.24	92.02	72.27	59.11
100%	**31.11**	**35.25**	**31.86**	**70.63**	**87.85**	**78.06**	**63.99**	**93.44**	**75.47**	**61.79**

Table D.4 CR performance by varying the number of labels in the unlabeled dataset.

D.5 Qualitative results

In Fig. D.1 and Fig. D.2, we present detailed qualitative visualizations. Figure D.1 shows the coreference chains predicted by our proposed method and the baseline, WS-MCR in Chapter 4. The baseline model misses the instances of *his* to relate it to the *the man* (row 2, column 2) which is correctly clustered by our method. Moreover, WS-MCR cannot relate *big orange color building* to *the building* (row2, column 4) unlike our method. This highlights that our method is able to learn fine-grained correlations between the image regions and text to effectively resolve such ambiguities. Despite significant advantages, our method still fails to resolve cases like *some other people* by clustering them into the same chain. We believe that the model needs more complex visual understanding (localize different instance of *some other people*) and contextual knowledge from text (*people standing on road* vs *people standing on footpath*) for these specific cases.

In Fig. D.2, we show grounding of the corresponding mentions on the image from our proposed method in Chapter 5 and the baseline WS-MCR. Compared to the baseline, our method successfully grounds *entrance door*, *some other people* and



Ground Truth

in this image we can see a man is dancing on a road, and the man is wearing white color shirt and black color pant.

we can also see a red and white color cloth in his hand, and the man is also wearing a orange and black color belt around his chest

on the road we can see a grey color car is parked, and we can also see some other people also standing on the road by holding flags

on the footpath we can see two women are sitting and talking with each other, and in that one woman is holding a baby, and we can also see some other people are also standing on the footpath, and in the background we can see a big orange color building with white color glass windows, and we can also see the entrance door of the building.

WS-MCR (Goel et al., 2022)

in this image we can see a man is dancing on a road, and the man is wearing white color shirt and black color pant.

we can also see a red and white color cloth in his hand, and the man is also wearing a orange and black color belt around his chest

on the road we can see a grey color car is parked, and we can also see some other people also standing on the road by holding flags

on the footpath we can see two women are sitting and talking with each other, and in that one woman is holding a baby, and we can also see some other people are also standing on the footpath, and in the background we can see a big orange color building with white color glass windows, and we can also see the entrance door of the building.

Ours

in this image we can see a man is dancing on a road, and the man is wearing white color shirt and black color pant.

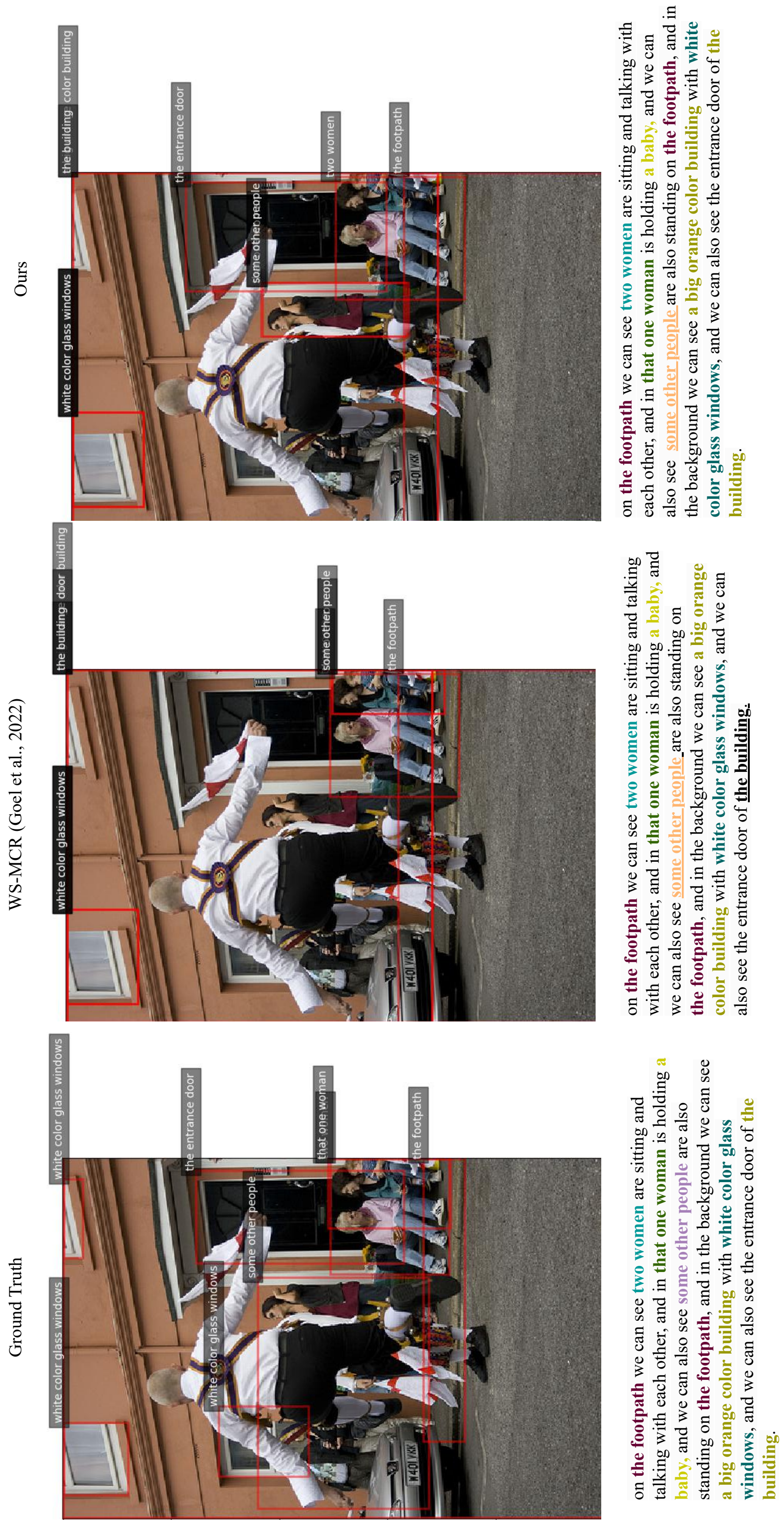
we can also see a red and white color cloth in his hand, and the man is also wearing a orange and black color belt around his chest

on the road we can see a grey color car is parked, and we can also see some other people also standing on the road by holding flags

on the footpath we can see two women are sitting and talking with each other, and in that one woman is holding a baby, and we can also see some other people are also standing on the footpath, and in the background we can see a big orange color building with white color glass windows, and we can also see the entrance door of the building.

Figure D.1 Coreference resolution for an image-narration pair. We break down the full narration as individual sentences in the columns for simplicity. The rows from top to bottom show ground-truth annotations, predictions from the WS-MCR method in Chapter 4 and Ours in Chapter 5. The mentions in the same color form a part of the coreference chain.

two women. Hence our method clearly exhibits strong coreference resolution and grounding capabilities compared to previous work.


Ground Truth

white color glass windows
the entrance door
that one woman
the footpath
some other people
white color glass windows
white color glass windows

on **the footpath** we can see **two women** are sitting and talking with each other, and in **that one woman** is holding **a baby**, and we can also see **some other people** are also standing on **the footpath**, and in the background we can see **a big orange color building** with **white color glass windows**, and we can also see the entrance door of **the building**.

WS-MCR (Goel et al., 2022)

on **the footpath** we can see **two women** are sitting and talking with each other, and in **that one woman** is holding **a baby**, and we can also see **some other people** are also standing on **the footpath**, and in the background we can see **a big orange color building** with **white color glass windows**, and we can also see the entrance door of **the building**.

Ours

on **the footpath** we can see **two women** are sitting and talking with each other, and in **that one woman** is holding **a baby**, and we can also see **some other people** are also standing on **the footpath**, and in the background we can see **a big orange color building** with **white color glass windows**, and we can also see the entrance door of **the building**.

Figure D.2 Visualization of Grounding and CR performance. Zoom in for better visualization of bounding boxes.

Bibliography

Labelstudio. https://labelstud.io/.

Pytorch. https://pytorch.org/.

Spacy. https://spacy.io/.

Abulikemu Abuduweili, Xingjian Li, Humphrey Shi, Cheng-Zhong Xu, and Dejing Dou. Adaptive consistency regularization for semi-supervised transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6923–6932, 2021.

George Adaimi, David Mizrahi, and Alexandre Alahi. Composite relationship fields with transformers for scene graph generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 52–64, 2023.

Berfin Aktaş, Veronika Solopova, Annalena Kohnert, and Manfred Stede. Adapting coreference resolution to twitter conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2454–2460, 2020.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Hyogyun An, Huiju Kim, Hyuntae Park, and Victoire Cyrot. Perceptual-iq: Visual commonsense reasoning about perceptual imagination. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 6581–6583. IEEE, 2022.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018a.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018b.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018c.

Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. Sim-to-real transfer for vision-and-language navigation. In *Conference on Robot Learning*, pages 671–681. PMLR, 2021.

Stephen R Anderson. Where's morphology? *Linguistic inquiry*, 13(4):571–612, 1982.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer, 2007.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Junwei Bao, Nan Duan, Ming Zhou, and Tiejun Zhao. Knowledge-based question answering as machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 967–976, 2014.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. Generating sentences from disentangled syntactic and semantic spaces. *arXiv preprint arXiv:1907.05789*, 2019.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.

Yoshua Bengio et al. Markovian models for sequential data. *Neural computing surveys*, 2(199):129–162, 1999.

Eric Bengtson and Dan Roth. Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, 2008.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

Yi Bin, Yang Yang, Chaofan Tao, Zi Huang, Jingjing Li, and Heng Tao Shen. Mr-net: exploiting mutual relation for visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8110–8117, 2019.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

Serhat Selcuk Bucak, Rong Jin, and Anil K Jain. Multi-label learning with incomplete class assignments. In *CVPR 2011*, pages 2801–2808. IEEE, 2011.

Ricardo S Cabral, Fernando Torre, Joao P Costeira, and Alexandre Bernardino. Matrix completion for multi-label image classification. In *Advances in neural information processing systems*, pages 190–198, 2011.

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, Antonio Feraco, et al. A practical guide to sentiment analysis. 2017.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

Soravit Changpinyo, Jordi Pont-Tuset, Vittorio Ferrari, and Radu Soricut. Telling the what while pointing to the where: Multimodal queries for image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12136–12146, 2021a.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021b.

Hong Chen, Zhenhua Fan, Hao Lu, Alan L Yuille, and Shu Rong. Preco: A large-scale dataset in preschool vocabulary for coreference resolution. *arXiv preprint arXiv:1810.09807*, 2018.

Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. Cross-modal image-text retrieval with semantic consistency. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1749–1757, 2019a.

Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 824–832, 2017.

Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019b.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581—-1590, 2021.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014a.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014b.

Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 2014.

Jun Chu and Gui-Hua Zhao. Scene classification based on sift combined with gist. In *2014 International Conference on Information Science, Electronics and Electrical Engineering*, volume 1, pages 331–336. IEEE, 2014.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Kevin Clark and Christopher D Manning. Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*, 2016.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. Acquiring common sense spatial knowledge through implicit spatial templates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384, 2021.

Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087, 2015.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2601–2610, 2019.

Joe Davison, Joshua Feldman, and Alexander M Rush. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1173–1178, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017.

Nai Ding, Lucia Melloni, Xing Tian, and David Poeppel. Rule-based and word-level statistics-based processing of language: insights from neuroscience. *Language, Cognition and Neuroscience*, 32(5):570–575, 2017.

Apoorva Dornadula, Austin Narcomey, Ranjay Krishna, Michael Bernstein, and Fei-Fei Li. Visual relationships as functions: Enabling few-shot scene graph prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 647–657, 2019.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C Lawrence Zitnick, et al. Visual storytelling. *arXiv preprint arXiv:1604.03968*, 2016.

Julie A Fiez and Marcus E Raichle. Linguistic processing. In *International Review of Neurobiology*, volume 41, pages 233–254. Elsevier, 1997.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org, 2017.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Arushi Goel, Basura Fernando, Thanh-Son Nguyen, and Hakan Bilen. Injecting prior knowledge into image caption generation. In *European Conference on Computer Vision*, pages 369–385. Springer, 2020.

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Not all relations are equal: Mining informative labels for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15596–15606, 2022a.

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Who are you referring to? weakly supervised coreference resolution with multimodal grounding. *arXiv preprint arXiv:2211.14563*, 2022b.

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Semi-supervised multimodal coreference resolution in image narrations. *arXiv preprint arXiv:2310.13619*, 2023a.

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Who are you referring to? coreference resolution in image narrations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15247–15258, 2023b.

Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2019.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021.

Danfeng Guo, Arpit Gupta, Sanchit Agarwal, Jiun-Yu Kao, Shuyang Gao, Arijit Biswas, Chien-Wei Lin, Tagyoung Chung, and Mohit Bansal. Gravl-bert: Graphical visual-linguistic representations for multimodal coreference resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 285–297, 2022.

Rahul Gupta, Nikolaos Malandrakis, Bo Xiao, Tanaya Guha, Maarten Van Segbroeck, Matthew Black, Alexandros Potamianos, and Shrikanth Narayanan. Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 33–40, 2014.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020.

Vishal Gupta and Gurpreet Singh Lehal. A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence*, 2(3):258–268, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Ulf Hermjakob, Eduard H Hovy, and C Lin. Knowledge-based question answering. In *Proceedings of the Sixth World Multiconference on Systems, Cybernetics, and Informatics (SCI-2002)*, volume 1, pages 17–26. Citeseer, 2000.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Jerry R Hobbs. Resolving pronoun references. *Lingua*, 44(4):311–338, 1978.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030, 2022.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. Visual writing prompts: Character-grounded story generation with curated image sequences. *arXiv preprint arXiv:2301.08571*, 2023.

Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao Huang, and Lun-Wei Ku. Knowledge-enriched visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7952–7960, 2020.

De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding" it": Weakly-supervised reference-aware visual grounding in instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5948–5957, 2018.

Jun Huang, Linchuan Xu, Kun Qian, Jing Wang, and Kenji Yamanishi. Multi-label learning with missing and completely unobserved labels. *Data Mining and Knowledge Discovery*, 35(3):1061–1086, 2021.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019a.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019b.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016.

Zanming Huang, Zhongkai Shangguan, Jimuyang Zhang, Gilad Bar, Matthew Boyd, and Eshed Ohn-Bar. Assister: Assistive navigation via conditional instruction generation. In *European Conference on Computer Vision*, pages 271–289. Springer, 2022.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

Zih Siou Hung, Arun Mallya, and Svetlana Lazebnik. Union visual translation embedding for visual relationship detection and scene graph generation. *Unknown Journal*, 2019.

Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Contextual translation embedding for visual relationship detection and scene graph generation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9423–9432, 2020.

Filip Ilievski, Pedro Szekely, and Bin Zhang. Cskg: The commonsense knowledge graph. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, pages 680–696. Springer, 2021.

Sho Inayoshi, Keita Otani, Antonio Tejero-de Pablos, and Tatsuya Harada. Bounding-box channels for visual relationship detection. In *European Conference on Computer Vision*, pages 682–697. Springer, 2020.

Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2018.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Person recognition in personal photo collections. In *Proceedings of the IEEE international conference on computer vision*, pages 3862–3870, 2015.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh, and Mo Nours Arab. Human-computer interaction: Overview on state of the art. *International journal on smart sensing and intelligent systems*, 1(1):137–159, 2008.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Kolesnikov, Alina Kuznetsova, Christoph Lampert, and Vittorio Ferrari. Detecting visual relationships using box attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3558–3565, 2014.

Kazuhiro Kosuge and Yasuhisa Hirata. Human-robot interaction. In *2004 IEEE International Conference on Robotics and Biomimetics*, pages 8–11. IEEE, 2004.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169, 2018.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128 (7):1956–1981, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Giulio E Lancioni and Nirbhay N Singh. *Assistive technologies for people with diverse abilities*. Springer, 2014.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the fifteenth conference on computational natural language learning: Shared task*, pages 28–34, 2011.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*, 2017.

Kenton Lee, Luheng He, and Luke Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-2108. URL https://aclanthology.org/N18-2108.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018b.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019a.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4654–4662, 2019b.

Lin Li, Jun Xiao, Hanrong Shi, Wenxiao Wang, Jian Shao, An-An Liu, Yi Yang, and Long Chen. Label semantic knowledge distillation for unbiased scene graph generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023b.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022b.

Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems*, 34: 19652–19664, 2021.

Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. Dreamedit: Subject-driven image editing. *arXiv preprint arXiv:2306.12624*, 2023c.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1347–1356, 2017a.

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1261–1270, 2017b.

Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023d.

Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual relationship detection with deep structural ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020.

Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 34:16331–16345, 2021.

Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019a.

Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2611–2620, 2019b.

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Li Su, and Qingming Huang. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 539–547, 2019c.

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5612–5621, 2021.

Gordon D Logan and Daniel D Sadler. A computational analysis of the apprehension of spatial relations. 1996.

Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40 (3):1–49, 2008.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016a.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29:289–297, 2016b.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. *Advances in Neural Information Processing Systems*, 30, 2017.

Pengcheng Lu and Massimo Poesio. Coreference resolution for the biomedical domain: A survey. *arXiv preprint arXiv:2109.12424*, 2021.

Xiaoqiang Luo and Sameer Pradhan. Evaluation metrics. In *Anaphora Resolution*, pages 141–163. Springer, 2016.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*, 2015a.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015b.

Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.

Diego Marcheggiani, Joost Bastings, and Ivan Titov. Exploiting semantics in neural machine translation with graph convolutional networks. *arXiv preprint arXiv:1804.08313*, 2018.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

Zihang Meng, Licheng Yu, Ning Zhang, Tamara L Berg, Babak Damavandi, Vikas Singh, and Amy Bearman. Connecting what to say with where to look by modeling human attention traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12679–12688, 2021.

Li Mi and Zhenzhong Chen. Hierarchical graph attention network for visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13886–13895, 2020.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2939, 2016.

Marie-Francine Moens, Katerina Pastra, Kate Saenko, and Tinne Tuytelaars. Vision and language integration meets multimedia fusion. *Ieee Multimedia*, 25(2):7–10, 2018.

Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Kien Nguyen, Subarna Tripathi, Bang Du, Tanaya Guha, and Truong Q Nguyen. In defense of scene graphs for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1407–1416, 2021.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Joakim Nivre. Dependency grammar and dependency parsing. *MSI report*, 5133 (1959):1–32, 2005.

Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos. Automatic image captioning. In *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, volume 3, pages 1987–1990. IEEE, 2004.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. *arXiv preprint arXiv:1611.06641*, 2016.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pages 647–664. Springer, 2020.

Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. Investigating failures to generalize for coreference resolution models. *arXiv preprint arXiv:2303.09092*, 2023.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40, 2012.

James Pustejovsky and Branimir Boguraev. Lexical knowledge representation and natural language processing. *Artificial Intelligence*, 63(1-2):193–223, 1993.

James Pustejovsky and Nikhil Krishnaswamy. Embodied human computer interaction. *KI-Künstliche Intelligenz*, 35(3-4):307–327, 2021.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge. *Advances in neural information processing systems*, 32, 2019a.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019b.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 492–501, 2010.

Rajat Raina, Andrew Y Ng, and Daphne Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 713–720, 2006.

Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. Linking people in videos with "their" names using coreference resolution. In *European conference on computer vision*, pages 95–110. Springer, 2014.

Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Charles Rosenberg, and Li Fei-Fei. Learning semantic relationships for better action retrieval in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1100–1109, 2015.

Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. Breakingnews: Article annotation by image and text processing. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1072–1085, 2017.

Marta Recasens and Eduard Hovy. Blanc: Implementing the rand index for coreference evaluation. *Natural language engineering*, 17(4):485–510, 2011.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.

Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. Generating descriptions with grounded and co-referenced people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4979–4989, 2017.

Mohammad Amin Sadeghi and Ali Farhadi. *Recognition using visual phrases*. IEEE, 2011.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5600, 2021.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

Katharine Sanderson. Gpt-4 is here: what scientists think. *Nature*, 615(7954):773, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Karin Kipper Schuler. Verbnet: A broad-coverage, comprehensive verb lexicon. 2005.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884, 2019.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

Thomas B Sheridan. Human–robot interaction: status and challenges. *Human factors*, 58(4):525–532, 2016.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K Reddy. Neural abstractive text summarization with sequence-to-sequence models. *ACM Transactions on Data Science*, 2(1):1–37, 2021.

Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 299–315, 2018.

Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Alessandro Stolfo, Chris Tanner, Vikram Gupta, and Mrinmaya Sachan. A simple unsupervised approach for coreference resolution using rule-based weak supervision. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 79–88, 2022.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13945, 2021.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162, 2020.

Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

Peng Sun, Xuezhen Yang, Xiaobing Zhao, and Zhijuan Wang. An overview of named entity recognition. In *2018 International Conference on Asian Language Processing (IALP)*, pages 273–278. IEEE, 2018.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1017–1024, 2011.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Qiaoyu Tan, Yanming Yu, Guoxian Yu, and Jun Wang. Semi-supervised multi-label classification using incomplete label information. *Neurocomputing*, 260:192–202, 2017.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Conference on Computer Vision and Pattern Recognition*, 2019.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55, 2020.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. On generalization in coreference resolution. *arXiv preprint arXiv:2109.09667*, 2021.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.

Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 2021.

Marc Vilain, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

Jakob Voss. Measuring wikipedia. 2005.

Atro Voutilainen. *Part-of-speech tagging*, volume 219. The Oxford handbook of computational linguistics, 2003.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Morton Wagman. Artificial intelligence and human cognition: a theoretical intercomparison of two realms of intellect. 1991.

Jian Wang, Yonghao He, Cuicui Kang, Shiming Xiang, and Chunhong Pan. Image-text cross-modal retrieval via modality-specific feature learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 347–354, 2015.

Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.

Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14090–14100, 2021.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017.

Qinxin Wang, Hao Tan, Sheng Shen, Michael W Mahoney, and Zhewei Yao. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. *arXiv preprint arXiv:2010.05379*, 2020a.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023.

Tzu-Jui Julius Wang, Selen Pehlivan, and Jorma Laaksonen. Tackling the unannotated: Scene graph generation with bias-reduced models. *arXiv preprint arXiv:2008.07832*, 2020b.

Ryen W White. Skill discovery in virtual assistants. *Communications of the ACM*, 61 (11):106–113, 2018.

Sam Wiseman and Alexander M Rush. Sequence-to-sequence learning as beam-search optimization. *arXiv preprint arXiv:1606.02960*, 2016.

Sam Wiseman, Alexander M Rush, and Stuart M Shieber. Learning global features for coreference resolution. *arXiv preprint arXiv:1604.03035*, 2016.

Sam Joshua Wiseman, Alexander Matthew Rush, Stuart Merrill Shieber, and Jason Weston. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2015.

Baoyuan Wu, Fan Jia, Wei Liu, Bernard Ghanem, and Siwei Lyu. Multi-label learning with missing labels using mixed dependency graphs. *International Journal of Computer Vision*, 126(8):875–896, 2018.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Bei Xie, Jiaohua Qin, Xuyu Xiang, Hao Li, and Lili Pan. An image retrieval algorithm based on gist and sift features. *Int. J. Netw. Secur.*, 20(4):609–616, 2018.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

Linli Xu, Zhen Wang, Zefan Shen, Yubo Wang, and Enhong Chen. Learning low-rank label correlations for multi-label classification with missing labels. In *2014 IEEE international conference on data mining*, pages 1067–1072. IEEE, 2014.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 265–273, 2020.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023.

Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12536, 2021.

Hao Yang, Joey Tianyi Zhou, and Jianfei Cai. Improving multi-label learning with missing labels by structured semantic correlations. In *European conference on computer vision*, pages 835–851. Springer, 2016.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14063–14073, 2022a.

Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. What gpt knows about who is who. *arXiv preprint arXiv:2205.07407*, 2022b.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 1–17. Springer, 2020.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022c.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902, 2017.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018.

Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 322–338, 2018.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018a.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018b.

Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. *arXiv preprint arXiv:2303.13233*, 2023.

Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE international conference on computer vision*, pages 1974–1982, 2017.

Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. *arXiv preprint arXiv:1805.03508*, 2018c.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *European Conference on Computer Vision*, pages 606–623. Springer, 2020a.

Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3736–3745, 2020b.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1476–1485, 2019.

Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. On exploring undetermined relationships for visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5128–5137, 2019.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021a.

Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017a.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017b.

Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5678–5686, 2017c.

Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021b.

Yifeng Zhang, Ming Jiang, and Qi Zhao. Explicit knowledge incorporation for visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1356–1365, 2021c.

Yifeng Zhang, Ming Jiang, and Qi Zhao. New datasets and models for contextual reasoning in visual dialog. In *European Conference on Computer Vision*, pages 434–451. Springer, 2022.

Feipeng Zhao and Yuhong Guo. Semi-supervised multi-label learning with incomplete labels. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

Hao Zhou, Chongyang Zhang, and Chuanping Hu. Visual relationship detection with relative location mining. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 30–38, 2019a.

Yimin Zhou, Yiwei Sun, and Vasant Honavar. Improving image captioning by leveraging knowledge graphs. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 283–293. IEEE, 2019b.

Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *Proceedings of the IEEE international conference on computer vision*, pages 589–598, 2017.