5-14-2023

Visual-Semantic Learning

Chengxiang Yin
Syracuse University

ABSTRACT

Visual-semantic learning is an attractive and challenging research direction aiming to understand complex semantics of heterogeneous data from two domains, i.e., *visual* signals (i.e., images and videos) and natural *language* (i.e., captions and questions). It requires memorizing the rich information in a single modality and a joint comprehension of multiple modalities. Artificial intelligence (AI) systems with human-level intelligence are claimed to learn like humans, such as efficiently leveraging brain memory for better comprehension, rationally incorporating common-sense knowledge into reasoning, quickly gaining in-depth understanding given a few samples, and analyzing relationships among abundant and informative events. However, these intelligence capacities are effortless for humans but challenging for machines. To bridge the discrepancy between human-level intelligence and present-day visual-semantic learning, we start from its basic understanding ability by studying the visual question answering (e.g., Image-QA and Video-QA) tasks from the perspectives of memory augmentation and common-sense knowledge incorporation. Furthermore, we stretch it to a more challenging situation with limited and partially unlabeled training data (i.e., Few-shot Visual-Semantic Learning) to imitate the fast learning ability of humans. Finally, to further enhance visual-semantic performance in natural videos with numerous spatio-temporal dynamics, we investigate exploiting event-correlated information for a comprehensive understanding of cross-modal semantics.

To study the essential visual-semantic understanding ability of the human brain with memory, we first propose a novel **M**emory **A**ugmented **D**eep **R**ecurrent **N**eural **N**etwork (i.e., **MA-DRNN**) model for Video-QA, which features a new method for encoding videos and questions, and memory augmentation using the emerging Differentiable Neural Computer (i.e., DNC). Specifically, we encode semantic (i.e., questions) information before visual (i.e., videos) information, which leads to better visual-semantic representations. Moreover, we leverage Differentiable Neural Computer (with external memory) to store and retrieve valuable information in questions and videos and

model the long-term visual-semantic dependency.

In addition to basic understanding, to tackle visual-semantic reasoning that requires external knowledge beyond visible contents (e.g., KB-Image-QA), we propose a novel framework that endows the model with capabilities of answering more general questions and achieves better exploitation of external knowledge through generating **M**ultiple **C**lues for **R**easoning with **Mem**ory **N**eural **N**etworks (i.e., **MCR-MemNN**). Specifically, a well-defined detector is adopted to predict image-question-related relation phrases, each delivering two complementary clues to retrieve the supporting facts from an external knowledge base (i.e., KB). These facts are encoded into a continuous embedding space using a content-addressable memory. Afterward, mutual interactions between visual-semantic representation and the supporting facts stored in memory are captured to distill the most relevant information in three modalities (i.e., image, question, and KB). Finally, the optimal answer is predicted by choosing the supporting fact with the highest score.

Furthermore, to enable a fast, in-depth understanding given a small number of samples, especially with heterogeneity in the multi-modal scenarios such as image question answering (i.e., Image-QA) and image captioning (i.e., IC), we study the few-shot visual-semantic learning and present the **H**ierarchical **G**raph **AT**tention Network (i.e., **HGAT**). This two-stage network models the intra- and inter-modal relationships with limited image-text samples. The main contributions of HGAT can be summarized as follows: 1) it sheds light on tackling few-shot multi-modal learning problems, which focuses primarily, but not exclusively, on visual and semantic modalities, through better exploitation of the intra-relationship of each modality and an attention-based co-learning framework between modalities using a hierarchical graph-based architecture; 2) it achieves superior performance on both visual question answering and image captioning in the few-shot setting; 3) it can be easily extended to the semi-supervised setting where image-text samples are partially unlabeled.

Although various attention mechanisms have been utilized to manage contextualized representations by modeling intra- and inter-modal relationships of the two modalities, one limitation of the predominant visual-semantic methods is the lack of reasoning with event correlation, sensing, and

analyzing relationships among abundant and informative events contained in the video. To this end, we introduce the dense caption modality as a new auxiliary and distill event-correlated information to infer the correct answer. We propose a novel end-to-end trainable model, **E**vent-**C**orrelated **G**raph **N**eural **N**etwork**s** (**EC-GNNs**), to perform cross-modal reasoning over information from the three modalities (i.e., caption, video, and question). Besides exploiting a new modality, we employ cross-modal reasoning modules to explicitly model inter-modal relationships and aggregate relevant information across different modalities. We propose a question-guided self-adaptive multi-modal fusion module to collect the question-oriented and event-correlated evidence through multi-step reasoning.

To evaluate our proposed models, we conduct extensive experiments on VTW, MSVD-QA, and TGIF-QA datasets for Video-QA task, Toronto COCO-QA, Visual Genome-QA datasets for few-shot Image-QA task, COCO-FITB dataset for few-shot IC task, and FVQA, Visual7W + ConceptNet datasets for KB-Image-QA task. The experimental results justify these models' effectiveness and superiority over baseline methods.

VISUAL-SEMANTIC LEARNING

By

Chengxiang Yin
B.S., Beijing Institute of Technology, 2016

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical and Computer Engineering

Syracuse University
May 2023

ACKNOWLEDGMENTS

The work of this thesis would not have been possible without the support, encouragement, and guidance of my committee members, friends, and family. I want to acknowledge those who helped me throughout this process.

First and foremost, I would like to take this opportunity to express my most profound appreciation to my supervisor, Dr. Jian Tang, for his help during my doctoral study. I appreciate his decision to accept me, an undergraduate with no research experience, into his group. Not only did he teach me how to accomplish excellent research work, but he also provided me with consistent support and valuable advice on how to overcome difficulties in research and life. The completion of my Ph.D. would not have been possible without his guidance.

In addition, I would also like to express my gratitude to many other faculty members who helped me a lot in my research and studies. In particular, I would like to thank the other supervisor of mine, Dr. Qinru Qiu, who gave me a lot of indispensable help in my course study, proposal, colloquium, and dissertation. I want to extend my deepest gratitude to Dr. Guoliang Xue from Arizona State University, Dr. Dejun Yang from Colorado School of Mines, and Dr. Yanzhi Wang from Northeastern University. Without their inspiration and suggestions, I could not have completed this thesis.

Furthermore, my labmates and many other friends helped me immensely in my research and life. Many thanks for their collaboration, guidance, and encouragement. Specifically, Zhiyuan, Kun, Jing, and Xiang have given me a lot of effective and practical suggestions, both in my research and career path. Tongtong, Zhengping, and Bo helped me

a lot with my thesis. Rui, Chunxu, Qunfang, and Yingya encouraged and cheered me up during my darkest and most depressing times.

Last but not least, I would like to give my everlasting gratitude to my mother, Xiushuang Zhou, and my father, Huanxin Yin, for their unconditional love, care, and encouragement over the past six years of my doctoral career. Most importantly, I would like to thank my fiancée Yanling Yang for her love, care, support, and companionship at every pivotal moment.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Motivation

Recently, significant progress has been made in various applications of a single modality, such as object detection and machine translation, thanks to the application of deep learning technologies [1], such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to a large amount of data. However, to enable an artificial intelligence (AI) system to perform human-level intelligence, such as leveraging brain memory for better comprehension, rationally incorporating common-sense knowledge into reasoning, quickly gaining in-depth understanding with a few samples, and analyzing relationships among numerous events, it requires not only the ability to memorize the rich information contained in a single modality, such as *visual* signals (i.e., images and videos) and natural *language* (i.e., captions and questions), but also a joint comprehension in multiple modalities. This is generally called multi-modal learning [2], and a typical example is visual-semantic learning. For instance, an intelligent cooking robot in the kitchen is expected to make delicious dishes by understanding the recipe instructions and detecting and selecting the right ingredients on the table. The robot can hardly perform this task without the ability to either read the texts or recognize the objects. These intelligence capacities are effortless for humans but challenging for machines.

To bridge the discrepancy between human-level intelligence and present-day visual-semantic learning, we start from its essential visual-semantic understanding ability by studying the video question answering (i.e., Video-QA) task, which requires a deeper understanding of heterogeneous data from two domains, i.e., spatio-temporal video content and question word sequence. Furthermore, combining visual and semantic observations with common-sense knowledge is crucial to answering questions that require external information. Therefore, we investigate how to incorporate external knowledge into visual-semantic learning by tackling knowledge-based image question answering (i.e., KB-Image-QA). Moreover, considering the characteristics of human learning, we extend visual-semantic learning to situations with limited and even partially unlabeled training data (i.e., Few-shot Visual-Semantic Learning) to imitate the fast learning ability of humans. Besides, most natural videos contain numerous events that provide valuable clues for a comprehensive understanding of cross-modal semantics. To further enhance visual-semantic performance, we illustrate how to incorporate event-correlated information into the reasoning process of Video-QA by introducing dense video captions as a new modality for cross-modal reasoning.

1.1.1 Basic Understanding

The ultimate goal of visual-semantic learning tasks is to enable a computer program to fully understand the content of an image/video as a human. However, we are currently still far from achieving this goal. For example, deep learning has made tremendous successes in visual captioning [3] (e.g., Image Captioning [4, 5, 6, 7, 8] and Video Captioning [3, 9, 10, 11]), which generates a natural language sentence to describe the content of an image or video. However, a sentence (or even multiple sentences) cannot fully and accurately describe an image or video (especially a video), which usually contains rich information. We believe it is better to have a machine-learning model that can answer any question related to an image or video, which is undoubtedly challenging. To answer questions as a human, the model needs to fully understand visual observations at a much finer level and generate an accurate answer for each question from a huge answer space. In order to study the basic visual-semantic understanding ability, we start by dealing with the Video-QA

task.

It is well-known that a gated Recurrent Neural Networks (i.e., RNNs), Long Short-Term Memory (i.e., LSTM) [12], is capable of modeling the long-term dependency. Hence, it has been widely used in video processing tasks, including video captioning [9, 3, 13, 14] and Video-QA [15, 16, 17, 18, 19]. However, a seminal work [20] has pointed out that LSTM does not work well when the sequence is long enough. A Video-QA task, however, may involve very long-term visual-textual dependency. A differentiable architecture called Neural Turing Machine (i.e., NTM) was proposed in [20] by Graves *et al.*, which has a neural network controller coupled with external memory. It has been shown that NTM outperforms the standard LSTM in several typical tasks. A more general and advanced memory-augmented neural network model called Differentiable Neural Computer (DNC) was proposed recently in [21], which is capable of solving complex and structured tasks that are inaccessible to neural networks without external memory, such as data storage over long timescales and synthetic question answering. We believe these emerging models (e.g., NTM and DNC) are particularly suitable for the Video-QA task we aim to tackle here.

As far as we know, many works [15, 22, 16] on Video-QA models the visual-textual dependency by encoding visual information before textual information, which leads to little improvement or even degradation training with visual observations. We suspect that the question-related video features are forgotten due to long-term sequence encoding, and a large amount of irrelevant visual information fills up the internal memory of LSTM without proper guidance to select useful information that is closely related to the corresponding question. Therefore, our model encodes a question and then a video to strengthen the influence of video features and select and preserve useful and question-related visual information to effectively utilize internal and external memory.

1.1.2 External Knowledge

Although basic understanding ability has been well-studied by leveraging memory augmentation on Video-QA to select and preserve key information on visual and textual modalities, general cases with more than two modalities are not fully exploited, especially when common-sense knowledge

is available for answer reasoning. Most of the existing VQA models [23, 24, 25, 26, 27] and datasets [28, 29, 30, 31] have focused on simple questions, which are answerable by solely analyzing the question and image, *i.e., no external knowledge is required*. However, a truly 'AI-complete' VQA agent must combine visual and semantic observations with external knowledge for reasoning, which is effortless for humans but challenging for machines. Therefore, to bridge the gap between human-level intelligence and machines algorithm, we address knowledge-based image question answering (i.e., KB-Image-QA) [32] in this thesis to perform visual-semantic reasoning incorporating a third modality of external knowledge.

Some efforts have been made in this direction. Wang *et al.* [32] presented a Fact-based VQA (FVQA) dataset to support much deeper reasoning. FVQA consists of questions that require external knowledge to answer. Several classical solutions [32, 33] have been proposed to solve FVQA by mapping each question to a query and retrieving supporting facts in the knowledge base (i.e., KB) through a keyword-matching manner. These supporting facts are processed to form the final answer. However, these query-mapping-based approaches with solely question parsing suffer from serious performance degradation when the information hint is not captured in the external KB, or the visual concept is not exactly mentioned in the question. Moreover, special information (i.e., visual concept type and answer source) should be determined in advance during the querying and answering phases, which makes it hard to generalize to other datasets. To address these issues, we introduce Multiple Clues for Reasoning, a new KB retrieval method, where a relation phrase detector is proposed to predict multiple complementary clues for supporting facts retrieval.

More recently, Out of the Box (i.e., OB) [34], and Straight to Facts (i.e., STTF) [35] adopt cosine similarity technique to get the highest scoring fact for answer prediction, where the whole visual information (i.e., object, scene, and action) and the whole semantic information (i.e., all question words) are indiscriminately applied to infer the final answer by implicit reasoning. However, for an image-question pair, only part of the visual content in the image and several specific words in the question are relevant to a given supporting fact. Analogously, only part of the supporting fact is needed for visual-semantic reasoning. Additionally, the direct concatenation of

image-question-entity embedding makes capturing information adaptively among different modalities much more difficult. To exploit the inter-relationships among three modalities (i.e., image, question, and KB), we propose a two-way attention mechanism with memory augmentation to model the interactions between visual-semantic representation and the supporting facts. Thus, the most relevant information from the three modalities can be distilled.

1.1.3 Fast Learning

Many visual-semantic learning methods, including our previous efforts on basic understanding and integration of external knowledge, are only capable of multi-modal modeling and reasoning when large amounts of human-annotated data and extensive training time are available. However, a true AI system should be able to quickly deliver an in-depth understanding with a small number of learning samples. Therefore, to provide human-level intelligent agents capable of performing visual-semantic reasoning with limited heterogeneous samples, we investigate few-shot visual-semantic learning in multi-modal scenarios, including visual question answering and image captioning.

Nowadays, for general few-shot learning problems, meta-learning [36, 37, 38] has become a standard methodology. Based on it, a few extensions have been recently made for few-shot visual-semantic learning. Fast Parameter Adaptation for Image-Text Modeling (FPAIT) [39] directly applied Model-Agnostic Meta-Learning (MAML) [40], a well-known meta-learning algorithm to the few-shot visual question answering and image captioning. Analogously, another work [41] adopted a question answering model with two meta-learning techniques, prototypical networks [42] and meta networks [43]. Nonetheless, these attempts left much to be desired in terms of their scope and performance. Firstly and fundamentally, all these methods merely applied existing meta-learning algorithms without explicitly considering the multi-modal nature, to which we paid careful attention in this work. For example, Teney *et al.* [41] obtained their model input through a simple element-wise production between visual and semantic representations. Additionally, neither of them deals with cases where labels are partially unlabeled, which is categorized

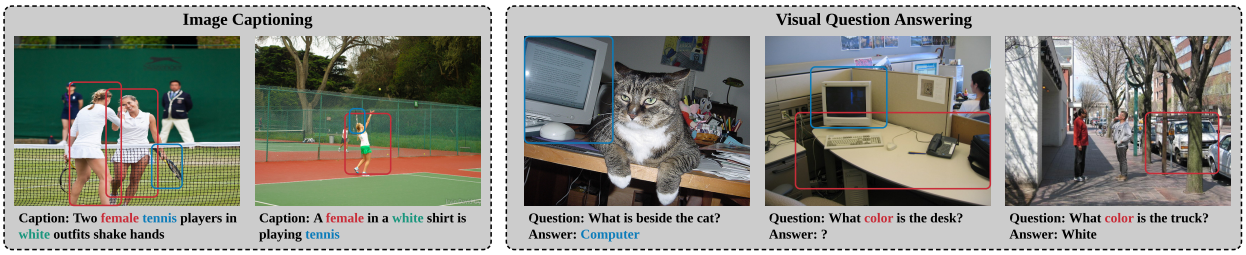

Fig. 1.1: Image-text samples for image captioning (left) and image question answering (right).

as a semi-supervised learning setting. As labeling data can be expensive or even infeasible, semi-supervised learning is very common and of great value in practice, and it becomes more severe when together with the few-shot setting.

For visual-semantic learning involving multiple modalities, especially the cases with limited and partially unlabeled image-text samples, it is vital to fully exploit the potential visual-semantic relationships, such as the intra-relationship of each modality (i.e., intra-modal relationships) and the inter-relationship between different modalities (i.e., inter-modal relationships). While intra-modal learning has been examined meticulously, such as Multi-modal DBMs [44], inter-modal learning leaves more space to explore and endows the model capacities to attentively capture complementary information.

We take two examples in Figure 1.1 for illustration. For the left-hand side image captioning example, both images contain female tennis players in white outfits. Correspondingly, the two ground-truth captions share the words “female”, “white”, and “tennis”. In this example, even if some words in the query caption are missing, the potential inter-modal relationship and the information captured from the visual modality can be used to supplement and strengthen the semantic modality and complete the caption. For the right-hand side visual question answering example, the two left images contain a computer on the desk, which is quite different from the third image. Therefore, the predicted answer of the middle image is likely to be fooled (e.g., computer) by the visual similarity solely. Only if inter-relationship is captured through exploiting modal mutual interactions the right visual clue can be distinguished from the distractors with the help of the semantic information and lead to the correct answer (i.e., white).

1.1.4 Event Correlation

All our previous studies on human-level visual-semantic learning (i.e., basic understanding, integration of external knowledge, and fast learning ability) have focused on images with simple appearance features or short video clips involving only a single event. Nonetheless, most natural videos in real-world scenarios contain a large number of ubiquitous events that are temporally localized and occur concurrently or successively. For example, in Figure 1.2, *a man in the red shirt is playing the violin* and then *the man jumps off the ladder*, and *begins to walk on the side of the roof* are three consecutive events depicting the same person. A mature Video-QA agent with human intelligence should be able to provide an accurate and explainable answer to a given question by perceiving such events in the video and analyzing their relationships to video content and questions by associating them with the actual queried subjects. Although various attention mechanisms have been utilized to manage contextualized representations by modeling intra- and inter-modal relationships of visual and semantic modalities in spatial-temporal dimension, one limitation of the predominant Video-QA methods [45, 46, 16, 17, 47] is the lack of reasoning with event correlation, that is, sensing and analyzing relationships among abundant and informative events contained in the video Therefore, we model Video-QA on long-term videos with more complicated spatial-temporal dynamics by incorporating event-correlated information into answer reasoning.

Events can be properly understood by modeling the complex video dynamics into a natural language description, namely *video captioning*. Massive works [9, 13, 10, 48] explain video with a single sentence. For example, they would most likely focus on *a man in the red shirt is playing the violin* in Figure 1.2, which provides some details about the man and his outfits but fails to articulate other events in the video. Instead, incorporating dense video captions [49, 50, 51] as a new auxiliary modality helps to identify and understand more events involved, which provides valuable clues to infer correct answers.

With multiple modalities available in Video-QA, integrating intra- and inter-modal relationships may further benefit the answer inference. Taking Figure 1.2 as an instance, the model needs to establish subject-predicate relationships between the word *man* and the words *playing* and *walk-*

Fig. 1.2: An illustration of how EC-GNNs handle Video-QA tasks. EC-GNNs aim to associate semantic contents in the generated dense captions and relevant visual contents in frame sequence with the actual subject queried in question. For a complex question such as "What is the man playing while walking on the roof?", EC-GNNs first localize when the man is walking on the roof and then focus on what the man is playing (i.e., the violin).

ing in the question. RNNs are commonly adopted [52, 15, 53] but often fail to model the true complexity of language structure, especially for semantic-complex questions. Alternatively, Graph Convolutional Networks (GCNs) [54] provide a more flexible way to utilize context-aware neural mechanisms to learn the complex intra-modal relationships, and the requirement of intra-modal modeling and the efficacy of GCNs also hold for videos and dense captions. On the other hand, the model needs to establish a solid semantic relationship between *the man playing* in the question, the caption *the man in the red shirt is playing the violin*, and the few frames highlighted in the middle of the video, to correctly answer the question in Figure 1.2. Modeling only one-way interactions [8, 55, 16, 17] is not enough, and methods aggregating relevant information across different modalities and learning long-term semantic dependencies is preferred. Self-attention [56] with different cross-modal key-query pairs provides a promising way of modeling inter-modal relationships.

At the final stage of Video-QA, to aggregate multi-modal features for answer prediction, the predominant approaches [52, 15, 53] generate a monolithic representation using simple vector operations, including concatenation, element-wise addition and/or element-wise multiplication.

However, the monolithic representation cannot fully understand cross-modal semantics, often leading to incorrect answers. As [57] found, the sub-optimal performance may come from distraction from the question. It is promising that multi-step reasoning with question guidance can progressively gather question-oriented and event-correlated evidence and provide better answers.

1.2 Literature Survey and Gap Analysis

1.2.1 Memory Augmented Deep Recurrent Neural Network

Image Question Answering

Due to the emergence of deep-learning architectures (e.g., CNNs [58, 59, 60, 61, 62, 63], RNNs [64, 12, 65]), Image Question Answering (Image-QA) has received increasing attention from both computer vision and natural language processing communities and becomes a typical and popular multimedia application. A vanilla Image-QA model [66] was proposed and considered as the Image-QA benchmark. It leveraged VGGNet [60] to encode images and a two-layer LSTM to encode questions. The question and images features were transformed into a common space and fused via element-wise multiplication for answer prediction. Following the idea of [66], lots of earlier approaches [26, 30, 67, 68, 25, 69, 70, 71, 72, 73] were proposed to learn a joint representation that allows modeling interactions and performing inference over question and image content.

To address the common issue of joint embedding, attention mechanisms were customized and widely adopted to enable the Image-QA models to identify *where to look* in an image and learn soft weights for features from different image regions. Some early works [22, 31, 22, 31, 74] described how to incorporate spatial attention mechanisms to standard LSTM models. They assumed that the answers to image-related questions usually correspond with specific image regions. Furthermore, some contemporaneous works [23, 75] developed schemes by dealing with CNN models. Moreover, to bring *where to look* and *what words to listen to* together, some customized co-attention mechanisms [24, 56] were presented. Concurrently, to obtain more representative and fine-grained

spatially-attended visual features, some works [76, 77, 8, 55, 78] directly leveraged some off-the-shelf object detectors (e.g., Faster R-CNN [79]) to identify regions with higher quality.

To explicitly capture inter-modal and intra-modal interactions, some more recent methods [80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91] represented images and questions using graph-based structures to seamlessly transfer information between graph items (i.e., objects in the image and words in the question) through some graph processing techniques, such as the emerging paradigms proposed for GCNs [54, 92, 93]. In addition to joint embedding, attention-based, and graph-based methods, other works introduced multiple new approaches [94, 95, 96, 97, 98, 99, 100, 101] to address the Image-QA problem from different perspectives.

Although Image-QA has been well studied by many works from different aspects, the problem becomes much more difficult when going from image to video (i.e., a sequence of correlated images). The methods proposed for Image-QA cannot be directly applied to Video-QA because they lack careful consideration of complex spatial-temporal dependencies hidden in videos. In this thesis, we consider a general Video-QA task, which differs from the Image-QA task that only focuses on a single image with simple appearance features being modeled in visual-semantic reasoning.

Video Question Answering

Recently, significant progress has been made to Video-QA, which extends Image-QA to a video domain. The first work [102] on Video-QA was proposed by Tu *et al.*. They introduced a joint parsing system to process video and text jointly for events understanding and query-answering. The Video-QA task was formally introduced by Zhu *et al.* [52]. They presented a RNNs-based encoder-decoder approach for finer understanding of video content in temporal domain (i.e., infer the past, describe present and predict the future) using a question form of "fill-in-the-blank" with multiple choices. Furthermore, Zeng *et al.* [15] collected a large-scale Video-QA dataset (i.e., 18K videos and 175K QA pairs with free-form questions and open-ended answers) by generating Video-QA pairs from video descriptions. In addition, they presented four baseline models, E-MN, E-VQA, E-SA and E-SS, which are extensions from four previous works [103, 66, 10, 13]

respectively.

Compared with Image-QA that focuses primarily on spatial understanding of region-level information from image, Video-QA requires a joint reasoning of spatial and temporal structures of a video to predict an accurate answer. The spatial structure provides in-frame appearance features, while the temporal structure analyzes the actions that occur across the frames. Predominant works [104, 16, 17, 18, 19] addressed Video-QA from the spatial-temporal view point.

Furthermore, to excavate question-specific evidence for Video-QA reasoning, many model-specific attention mechanisms [46, 45, 104, 105, 47, 106, 107, 108, 109, 110] were employed to explicitly model long-term temporal dynamics and incorporate frame-level details, where soft weights are learned under the guidance of question. Spatial-temporal approaches largely overlap with the attention-based approaches because both focus on temporal structure modeling given certain relevant words from questions.

Moreover, it's super difficult for a Video-QA agent to infer the correct answer by tracking a huge amount of information from a extremely long sequence of video frames and question words. Even if the RNN-based networks are able to model temporal correlations among the video frames and question words, they are not capable of representing and storing all these visual-textual contents, especially for the untrimmed long-term video and question. Thus, to address this issue, some Video-QA works [111, 112, 57, 113] leveraged an external memory combined with a neural networks to process, represent and store both visual and textual information more effectively, while trying to locate the salient video and question contents to remove the noise. In addition to spatial-temporal, attention-based, memory-based Video-QA approaches, other works [114, 115, 116, 117, 118, 119, 120, 121] studied the Video-QA problem from different perspectives.

As far as we know, similar to the pioneering learning-based approaches [52, 15] on Video-QA, many spatial-temporal methods [16, 17, 18] and memory-based methods [111, 112, 57] model the visual-textual dependency by encoding visual information before textual information, where a large amount of irrelevant and useless information contained in the video may fill up the limited memory. Therefore, for the purpose of selecting and preserving useful and question-related

visual information to utilize the memory more effectively, our model encodes a question and then encode a video. In addition, unlike the memory-based Video-QA approaches [111, 112, 57, 113] that design and customize learnable memory networks or blocks (e.g., visual memory, question memory) specifically for visual-semantic models, we introduce memory augmentation into Video-QA reasoning with a unified manner for different modalities, by leveraging Differentiable Neural Computer for both question and video encoding, which appends an external memory to LSTM to provide additional storage space for modeling the long-term visual-textual dependency.

1.2.2 Multi-Clue Reasoning with Memory Augmentation

Knowledge-based Image-QA

In real situations, accurately answering a question given an image requires a combination of visual observation and general knowledge. This is an extremely challenging task for current AI agents, but effortless for humans. To bridge the gap between machine learning algorithms and human behaviors, Knowledge-based Image Question Answering is introduced. A pioneering work [122] from Wu *et al.*, named Ask Me Anything, first defined the Knowledge-based VQA as a free-form visual question answering task based on knowledge from external sources, and proposed a flexible approach to develop a deeper understanding of the scene by combining textual representation of image contents with information extracted from a general knowledge base. Similarly, Wang *et al.* [33] introduced a model called Ahab, that not only answers complicated questions that require external information, but also provides explanations for reasoning.

To advocate research in knowledge-based Image-QA, Wang *et al.* [32] introduced a widely-used benchmark called Fact-baesed VQA (FVQA), a VQA dataset that primarily contains complex questions which require much deeper reasoning on common sense or basic factual knowledge to answer. Analogously, Outside Knowledge VQA (OK-VQA) [123] and World Knowledge-Aware VQA (KVQA) [124] were introduced. For both datasets, image contents are insufficient to answer the questions that require reasoning over external knowledge resources.

The aforementioned methods [122, 33, 32] retrieved supporting facts from KBs according to

extracted question keywords or pre-defined query templates; however, synonyms and homographs in questions make it difficult to focus on the most obvious visual concept, which severely hinder obtaining truly relevant facts and lead to incorrect answer predictions. To address this issue, some learning-based approaches were developed. Narasimhan *et al.* [35] proposed a novel framework that goes Straight To The Facts (STTF) via a learned embedding space. Instead of retrieving supporting facts based on question keywords or pre-defined query templates, STTF first employs an LSTM to predict the fact relation that is used to filter the candidate facts from KBs. However, only the top-ranked fact is used to answer to question, which leads to a sub-optimal local decision, since the top-ranked fact is not always the ground truth for question answering. Therefore, they developed a new method called Out of the Box (OB) [34], which creates an entity graph and employs a graph convolutional network (GCN) to jointly consider multiple facts before arriving at an answer. More recently, Zhu *et al.* [125, 126] introduced Multi-Layer Cross-Modal Knowledge Reasoning (Mucko) and comprehensively depict an image by a multi-modal heterogeneous graph, which contains multiple layers of information from visual, semantic and factual modalities, and demonstrates a good interpretability and is able to automatically tells which modality and entity have more contributions to question answering.

Unlike the query-mapping based approaches [122, 33, 32] retrieving supporting facts through a keyword-matching manner, which suffer from serious performance degradation when the information hint is not captured or visual concept is not mentioned, we introduce a learning-based KB retrieval method, where a relation phrase detector is proposed to predict multiple complementary clues for supporting facts retrieval. Even if the learning-based approaches [35, 34] achieve a noticeable improvement over the query-mapping based approaches, they indiscriminately employ the whole visual information (i.e., object, scene, and action) and semantic information (i.e., question words) for answer reasoning through a direct concatenation that makes it hard to adaptively capture information among different modalities. We, however, propose a two-way attention mechanism to model the interactions between visual-semantic representation and supporting facts to distill the most relevant information from each modality. Besides, [35, 34] only utilize the semantic evi-

dence from question space for KB retrieval, while we jointly leverage the visual and semantic information to obtain the supporting facts from KB.

1.2.3 Hierarchical Graph Attention Network

Few-shot Learning

Meta-learning, a standard methodology to tackle few-shot learning problems, has recently attracted extensive attention due to its important roles in achieving human-level intelligence. Several specialized models [127, 128, 42, 129, 130] have been proposed for meta-learning, particularly for few-shot classification, by comparing similarity among data samples using representation learning. Specifically, [128] presented a neural network model, Matching Networks, which learn an embedding function and use the cosine distance in an attention kernel to measure similarity. Similarly, [42] leveraged a similar approach to few-shot classification but used the Euclidean distance with their embedding function.

Another predominant approach to meta-learning is to develop a meta-learner to optimize key hyper-parameters (e.g., initialization) of the learning model. A seminal work [40] presented a model-agnostic meta-learner, MAML, to optimize the initialization of a learning model with the objective of maximizing its performance on a new task after updating its parameters with a small number of samples. Several other methods [131, 132, 133, 134, 135, 136] utilized an additional neural network, such as LSTM, to serve as the meta-learner. In particular, [133] proposed another LSTM-based meta-learner to learn a proper parameter update and a general initialization for the learning model allowing for quick convergence of training for a new task.

Even if the few-shot learning has made tremendous successes, it focuses only on the few-shot classification tasks without a careful consideration for more complicated visual-semantic learning tasks (e.g., Image-QA and Image Captioning), which involve multiple modalities. However, we presented a Hierarchical Graph Attention Network to study the few-shot visual-semantic learning with better exploitation of both intra- and inter-modal relationships.

Visual-semantic Learning

Visual-semantic learning aims to build models that can process and relate the information for both visual and semantic modalities. Generally speaking, visual-semantic learning focuses on multimedia description tasks, such as Image-QA and image captioning. Various methods [66, 26, 67, 23, 75, 74, 24, 71, 80, 8, 55, 137, 87, 89] have been proposed for Image-QA, which are discussed before. Image captioning [138, 139, 140, 141, 142] aims to describe the visual contents of an image in meaningful and syntactically correct language. The most common approach for image captioning is to feed global features extracted from the last layer of a CNN into a language model. This simple recipe is first adopted in [4], where the output features of GoogleNet [61] are used as the initial hidden state of the language model. Other works dealing with image captioning through employing global CNN features includes [143, 5, 3, 144, 145, 146, 147, 148, 6, 149, 150, 151].

In addition, many subsequent works exploit attention mechanisms to achieve greater flexibility and finer granularity in image captioning. Xu [7] leverages additive attention to compute a weight for each grid element of the spatial feature extracted from the last convolutional layer of VGG network [60], so that the model is able to selectively focus on different set of elements to generate different words of an image description. Subsequent works following this line with minor improvements includes [152, 153, 154, 155, 156, 157, 158]. Notably, to imitate how the humans pay attention for image captioning task, some works [159, 160, 161, 162, 163] employed saliency map to achieve stimulus-based attention computation and critically attend to image regions based on whether they are fixated or not. Anderson *et al.* [8] proposed a bottom-up path and employed Faster R-CNN [79] to obtain feature maps of image regions. A top-down attention mechanism is further adopted to weigh the feature map of each region when generating the next word. Many subsequent solutions [164, 165, 166, 167] follow this strategy for image captioning.

To fully exploit and model the semantic and spatial relationships between detected objects or image regions, graph-based encoding is employed for image captioning. Yao *et al.* [168] first proposed to use a GCN-LSTM architecture to integrate semantic and spatial object relationships

into an image encoder. Similarly, Guo *et al.* [169] constructed two structured graphs of visual semantic units, one semantic and one geometry, to exploit the alignment properties between caption words and visual semantic units for image captioning. Other attempts to use graph-based encoding variants for image captioning include [170, 142, 171].

Self-attention mechanism [56] is proposed to enable a complete graph representation for a more flexible image encoding, where each unit is connected with all the others and the relationships between each pair of units can be successfully exploited. The first attempt to leverage self-attention for image captioning is due to Yang *et al.* [172]. They employed a self-attentive module to fuse the detected visual clues (i.e., objects, attributes, and relations) as well as the language context knowledge. At the same time, Li *et al.* [173] introduced an Entangled Attention (ETA) Transformer to encode the detected visual and semantic features respectively with self-attention and feed-forward layers. Self-attention has been widely adopted by several following works [140, 174, 175, 176, 177, 178] with minor modifications tailored for image captioning.

However, unlike most of the existing Image-QA and image captioning methods that rely on a vast amount of human-annotated training data, we propose a model that is able to deal with the cases with limited or even partially unlabeled data samples.

Few-shot Visual-semantic Learning

As aforementioned, meta-learning has become a standard methodology for few-shot learning. Based on it, a few extensions have been recently made for few-shot visual-semantic learning. In particular, Dong *et al.* [39] proposed Fast Parameter Adaptation for Image-Text Modeling (FPAIT) that directly applied Model-Agnostic Meta-Learning (MAML) [40], a well-known meta-learning algorithm, to the few-shot visual question answering and image captioning. Analogously, Teney *et al.* [41] adopted a question answering model with two meta-learning techniques, prototypical networks [42] and meta networks [43]. Nonetheless, these attempts left much to be desired in terms of their scope and performance. In particular, they merely applied existing meta-learning algorithms without explicitly considering the multi-modal nature of visual-semantic learning. For example,

Teney *et al.* [41] obtained the model input through a simple element-wise production between visual and semantic representations. Besides, none of them deals with the cases where labels are partially unlabeled, which is categorized as the semi-supervised learning setting

Unlike [39, 41], we present a model for few-shot visual-semantic learning that considers the multi-modal nature through a fully exploitation of potential visual-semantic relationships (i.e., intra- and inter-modal relationships), and it is easily to be extended to semi-supervised cases with the limited data samples partially unlabeled.

1.2.4 Cross-Modal Reasoning with Event Correlation

Dense Image Captioning

Dense image captioning task is firstly introduced by Johnson *et al.* [179], which concurrently localizes and describes salient image regions with short natural language sentences by spatially extending image captioning. This task can be conceived as a generalization of object detection, where syntactically correct captions replace single-word labels, or image captioning, where multiple salient regions replace a full image. Furthermore, to address this task, they proposed a Fully Convolutional Localization Network (FCLN), consisting of a VGG-16 [60] for apperance feature extraction, a fully differentiable dense localization layer for regions of interest (ROI) prediction, and an LSTM for generating word sequences.

Several subsequent works improve dense image captioning by jointly reasoning with more contextual information involved. Yang *et al.* [180] presented a dense captioning model with two key components, joint inference and context fusion. In particular, for joint inference, the localization bounding boxes are jointly predicted from pooled features of ROI combined with the predicted descriptions; for context fusion, the pooled features are combined with context features to predict more accurate region descriptions. Likewise, Li *et al.* [181] argued that the caption of each region depicts both object properties and its interactions with objects in the image. Therefore, they proposed a novel scheme to learn complementary object context and transfer knowledge from related objects to caption regions.

Furthermore, on the basis of contextual knowledge augmentation, some works proposed to predict dense captions based on linguistic attributes. Notably, Yin *et al.* [182] designed a context and attribute grounded dense captioning model with an attribute-grounded caption generator that not only exploits multi-scale contextual information sharing and message passing, but also incorporates hierarchical linguistic attribute supervision in a coarse-to-fine manner to enhance the distinctiveness of generated captions. To provide more coherent dense descriptions of visual contents, a series of works [183, 184, 185, 186, 187, 188, 189, 190] following these two concepts (i.e., contextual feature mining and attribute augmentation) are proposed for dense image captioning.

Dense Video Captioning

Dense video captioning task is firstly introduced by Krishna *et al.* [49] by extending video caption in event dimension, which simultaneously detects multiple events that occur in the video and describes each event using natural language. The seminal work on dense video captioning [49] employed an existing proposal module [191] to capture both short and long events and caption each event with a newly-designed captioning module that is able to utilize the context from surrounding events, even for streaming videos. Nonetheless, the proposal module and captioning module are trained in an alternative manner, where the caption supervision has no direct benefit for the event proposal prediction.

To overcome this limitation, several subsequent works presented simple yet effective frameworks for end-to-end dense video captioning. Zhou *et al.* [50] proposed an encoder-decoder based end-to-end trainable model consisting of a video encoder, a proposal decoder and a captioning decoder. The captioning decoder employs a masking scheme to convert the predicted event proposals into differentiable masks, which ensures the consistency between proposal and captioning decoders during training, so that the linguistic information from captions is able to guide the proposal module to generate more plausible proposals. Similarly, Wang *et al.* [192] presented another pure end-to-end dense Video Captioning framework with Parallel Decoding (PDVC), which directly feeds the video features into a captioning head parallel to the localization head. Therefore,

the inter-task associations at the feature level are captured and the mutual benefits between two tasks (i.e., proposal localization and caption prediction) can be exploited to improve their performance together.

Additionally, Deep Reinforcement Learning (DRL) is also employed for dense video captioning [193, 194], to alleviate the redundancy and inconsistency caused by duplicate proposal localization and independent caption prediction. Other works on dense video captioning includes [195, 196, 197, 198, 51, 199, 200, 201].

Unlike [125, 126] that leverage dense image captions as a new semantic modality for Knowledge-based Image-QA with cross-modal reasoning on a heterogeneous graph, we introduce dense video captions to deal with Video-QA by incorporating event-correlated information into visual-semantic reasoning. We assume that the event correlation could be successfully exploited to benefit Video-QA inference through modeling intra-modal relationships among dense video captions and inter-modal relationships among video, question and dense captions.

Graph-based Video-QA

As aforementioned, the state-of-the-art Video-QA methods are based on fine-grained representation and model-specific attention, which usually process video and question separately and fuse the contextualized representations of visual and textual modalities for answer prediction. Although they process important information of one modality based on relevant cues from another modality with attention mechanisms, they fail to integrate both inter-modal and intra-modal relationships into a unified module.

Therefore, to explicitly model inter-modal and intra-modal relationships, some works attend to represent Video-QA as a spatial-temporal graph and perform reasoning over the graph representation via GCNs. Wang *et al.* [202] first proposed to perform long-range temporal modeling of human-object and object-object relationships via a graph-based reasoning framework, which clarifies how graph structures can be used for Video reasoning tasks, such as Video-QA. Each input video is represented as a space-time region graph where each node represents a region of inter-

est and each edge denotes appearance similarity or spatial-temporal proximity. Inspired by [202], Huang *et al.* [203] introduced *Location-aware Graph Convolution Networks* (L-GCN) to model the interactions between objects regarding to a question. In particular, they first leverage an off-the-shelf object detector to detect objects of interest to explicitly exclude irrelevant background contents, and construct a fully-connected graph where each node represents an object and each edge between two nodes represents their relationship. Similarly, to explicitly exploit both inter-modal and intra-modal relationships in a more flexible way for Video-QA inference, Jiang *et al.* [204] built an uniform *Heterogeneous Graph Alignment* (HGA) network over video shots and question words. Furthermore, to obtain an enhanced vision-language representation for Video-QA, Jin *et al.* [120] introduced an adaptive spatial-temporal graph module, which refines the spatial-temporal connections for dynamic object representation learning according to their spatial-temporal relations.

Even though [202, 203, 204, 120] provide a better exploitation of inter- and intra-modal relationships by representing Video-QA as a spatial-temporal graph, they only consider the two basic modalities (i.e., visual and semantic) for Video-QA reasoning without any event correlation considered. However, we propose a novel Event-Correlated Graph Neural Networks (EC-GNNs) that introduce dense video captions as a new modality for Video-QA and clarify how to exploit event-correlated information through cross-modal reasoning. To the best of our knowledge, we are the first to integrate dense video captions as a new modality for Video-QA task.

1.3 Contributions

This thesis aims to leverage emerging Deep Learning technologies to investigate human intelligent visual-semantic learning from four important aspects of human-level cognition, including basic understanding, external knowledge integration, faster learning, and event correlation. More specifically, we make the following contributions.

- To study the basic visual-semantic understanding ability of the human brain with memory,

we propose a novel MA-DRNN model for the Video-QA task. We argue that the widely-used visual-textual encoding method doesn't conform to human cognitive habits and often leads to inaccurate answers due to memorizing irrelevant and redundant information. Furthermore, we discuss and show that the commonly-used RNNs (i.e., LSTM, GRU) do not model long-term visual-textual dependencies well. Based on the analysis, we present a novel video and question encoding approach to learn better visual-semantic representations, and we utilize a DNC-based memory augmentation mechanism to store and retrieve valuable and question-related visual information with the incorporated external memory.

- To investigate visual-semantic reasoning capacities with external knowledge integration, we propose a novel MCR-MemNN framework for the KB-Image-QA task. We argue that visual-semantic reasoning using commonsense knowledge is effortless for humans but challenging for current algorithms. In addition, we discuss and show that the predominant methods fail to extract evidence related to image-question and supporting facts across different modalities and degrade severely in extreme cases due to missing key supporting facts. Based on the analysis, we present a bidirectional attention mechanism with memory augmentation to adaptively capture the most relevant cross-modal information by modeling the interactions among different modalities, and we introduce a multi-clue KB retrieval method to obtain ground-truth supporting facts with multiple predicted complementary cues.

- To enable fast and in-depth visual-semantic understanding using a few or even partially unlabeled samples, we investigate few-shot visual-semantic learning and propose a new HGAT framework for Image-QA and Image-Captioning in few-shot and semi-supervised few-shot settings. We claim that the ability to solve visual-semantic tasks with limited samples is critical for a human intelligent system, while current AI agents either fail to consider the multi-modal nature in few-shot scenarios or are incapable of handling limited and partially unlabeled image-text samples. In addition, we discuss and show through two examples that fully exploiting intra- and inter-modal relations is crucial for few-shot visual-semantic

learning. Based on the analysis, we design visual-specific and semantic-specific GNNs to capture the intra-relationships of visual and semantic modalities, respectively, and introduce an attention-based co-learning framework to model the inter-relationships between the two modalities. Furthermore, we employ relation-aware GNNs to jointly learn visual-semantic representations and relations.

- To exploit event correlation for visual-semantic reasoning with complex spatial-temporal dynamics, we propose a novel EC-GNNs framework for Video-QA of long-term videos containing a large number of ubiquitous events. We argue that a mature Video-QA agent with human intelligence is able to provide accurate and explainable answers by perceiving events and analyzing their relations to video content and questions. We show that traditional video captioning with only one sentence cannot articulate all contained events. Furthermore, we illustrate the importance of intra- and inter-modal relationships for learning long-term visual-semantic dependencies through a representative Video-QA example. Based on the analysis, we introduce dense video captions as a new auxiliary modality to identify each involved event and incorporate event-correlated information into the reasoning process, and we propose a cross-modal attention mechanism to explicitly model inter-modal relationships and aggregate relevant information across different modalities.

1.4 Outline

The rest of this thesis is organized as follows. To study the basic visual-semantic understanding ability, we first propose Memory Augmented Deep Recurrent Neural Network (MA-DRNN) for Video-QA in Chapter 2. In addition, in Chapter 3, to investigate visual-semantic reasoning capacities with commonsense knowledge, we present a novel MCR-MemNN framework to deal with KB-Image-QA. Furthermore, we argue that a fast and in-depth visual-semantic understanding using a few or even partially unlabeled samples is crucial to achieve human intelligent visual-semantic learning, and introduce the Hierarchical Graph Attention Network (HGAT) for few-shot

visual semantic learning in Chapter 4. Moreover, in Chapter 5, we clarify how to employ dense video captions as a new auxiliary modality to incorporate event-correlated information into Video-QA reasoning by presenting Event-Correlated Graph Neural Networks (EC-GNNs). Finally, we conclude this thesis and show our future research plan in Chapter 6.

CHAPTER 2
MEMORY AUGMENTED DEEP RECURRENT NEURAL NETWORK FOR VIDEO QUESTION ANSWERING

2.1 Overview

Image Question Answering (i.e., Image-QA) represents a typical and popular multimedia application, which has been studied by quite a few works [66, 26, 30, 70, 75, 74, 8, 80, 89, 90]. However, the problem becomes much more complicated when going from image to video (i.e., a sequence of correlated photos). The methods proposed for Image-QA obviously cannot be directly applied to Video Question Answering since they lack careful consideration for complicated temporal correlations hidden in videos. Even though video captioning shares some similarities with Video-QA, these two tasks are still mathematically and fundamentally different. Hence, existing video captioning models and methods [205, 10, 13, 14, 206] cannot be directly applied or easily extended to solve the Video-QA problem. Among works on Video-QA, some of them [207, 52, 95] focused on multiple choice questions, and another work [16] targeted only three specific kinds of questions. We, however, consider Video-QA in a general case, i.e., general free-form questions that can have

open-ended answers.

In this chapter, we propose a novel **M**emory **A**ugmented **D**eep **RNN** (**MA-DRNN**) based on DNC for Video-QA, which features a new visual-textual encoding method (which is different from many other related works), and DNC-based memory augmentation. We validate and evaluate our model using the VTW dataset [15] and MSVD-QA dataset [46], which are two widely-used large-scale video benchmarks for language-level understanding. It has been shown by experimental results that the proposed model can generate meaningful answers for questions, and more importantly, it outperforms the state-of-the-art methods presented in a recent work [15] in terms of various accuracy-related metrics. Moreover, a comprehensive ablation study has been performed to show the effectiveness and superiority of the new method for video and question encoding as well as memory augmentation based on DNC. To the best of our knowledge, we are the first to propose a memory-augmented deep neural network for Video-QA and show promising results.

2.2 Memory Augmented Deep Recurrent Neural Network

2.2.1 Differential Neural Computer

In this section, we give a necessary background introduction to Differentiable Neural Computer (DNC) [21]. As illustrated in Figure 2.1, a DNC is consisting of two basic components: a neural network controller and an external memory. The controller can interact with the external memory via read and write operations, which is analogous to random-access memory in a conventional computer. However, unlike the conventional computer with discrete read and write operations, each component in DNC is differentiable, allowing end-to-end training with the gradient decent as a regular neural network. This is achieved by defining blurry read and write operations which can access each memory location in a greater or less degree at a time. The degree of blurriness, also called weighting, is determined by the memory addressing mechanism.

Controller. At each time-step t, the recurrent controller receives an input vector $\mathbf{x}_t$ and emits an output vector $\mathbf{y}_t$. Additionally, the controller receives a set of read vectors $\mathbf{r}_{t-1}^1, \mathbf{r}_{t-1}^2, ..., \mathbf{r}_{t-1}^R$

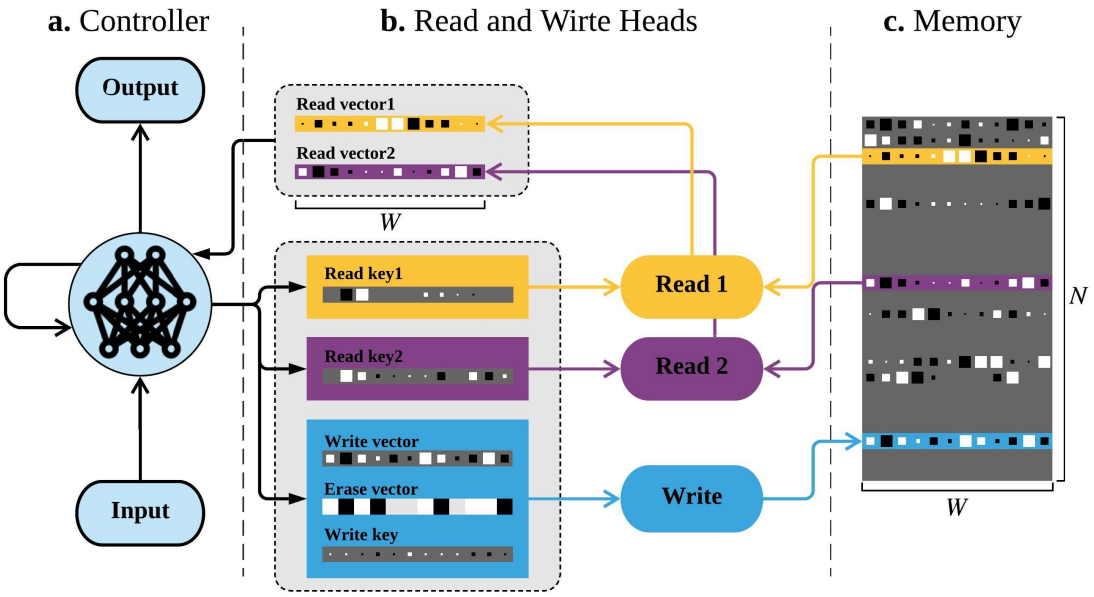

Fig. 2.1: Differentiable Neural Computer (DNC) [21]

from the $N \times W$ memory matrix $\mathbf{M}_{t-1}$ of previous time-step. So the input vector to controller can be expressed as $\mathbf{X}_t = [\mathbf{x}_t; \mathbf{r}^1_{t-1}; \mathbf{r}^2_{t-1}; ...; \mathbf{r}^R_{t-1}]$. Even though any neural network can be used as the controller, it is quite common to use LSTM for the purpose. The output vector from controller can be determined by:

$$\mathbf{y}_t = \mathbf{W}_h[\mathbf{h}^1_t; \mathbf{h}^2_t; ...; \mathbf{h}^L_t] + \mathbf{W}_r[\mathbf{r}^1_t; \mathbf{r}^2_t; ...; \mathbf{r}^R_t]; \tag{2.1}$$

where $\mathbf{h}^1_t, ..., \mathbf{h}^L_t$ are the hidden states of L-layer controller. And for each layer, the hidden state is used to parameterize one write head (blue) and multiple read heads (yellow and purple) (Equation (2.5) and (2.6)).

Write and Read Operations. The write operation of a single write head is mediated by the write weighting vector $\mathbf{w}^w_t$, which is used with an erase vector $\mathbf{e}_t$ and a write vector $\mathbf{v}_t$ to modify the memory as follows:

$$\mathbf{M}_t = \mathbf{M}_{t-1} \circ (\mathbf{E} - \mathbf{w}^w_t \mathbf{e}^\top_t) + \mathbf{w}^w_t \mathbf{v}^\top_t; \tag{2.2}$$

where $\mathbf{e}_t$ controls the erase operation, $\mathbf{v}_t$ controls the add operation, and both of them are determined by the hidden state of the controller (Equation (2.5)). The write weighting vector $\mathbf{w}^w_t$ is calculated based on the write key $\mathbf{k}^w_t$, which is also determined by the hidden state of the controller

(Equation (2.5)). $\mathbf{M}_t$ denotes the memory of the current time-step t, $\mathbf{E}$ is an $N \times W$ matrix of ones and $\circ$ denotes the element-wise multiplication. Then the memory is updated through an erase operation followed by an add operation on $\mathbf{M}_{t-1}$

For the read operation, R read weighting vectors $\mathbf{w}_t^{r,1}, \mathbf{w}_t^{r,2}, ..., \mathbf{w}_t^{r,R}$ are used to compute weighted averages of the contents of the locations to get R read vectors $\mathbf{r}_t^1, \mathbf{r}_t^2, ..., \mathbf{r}_t^R$:

$$\mathbf{r}_t^k = \mathbf{M}_t^\top \mathbf{w}_t^{r,k}. \tag{2.3}$$

where each read weighting vector $\mathbf{w}_t^{r,k}$ is calculated based on the read key $\mathbf{k}_t^{r,k}$, which is determined by the hidden state of the controller (Equation (2.6)).

2.2.2 Memory Augmented LSTMs

In this section, we present the proposed Memory Augmented Deep Recurrent Neural Network (MA-DRNN) model. MA-DRNN introduces two new techniques to improve performance, including 1) a new method for encoding videos and questions (which is different from those in many other related works), and 2) DNC-based memory augmentation.

As illustrated in Figure 2.2, the proposed model works as follows: 1) In the question encoding phase, the embeddings of a sequences of words (in the question) are fed into the DNC (with an LSTM unit as the controller) one by one (one at a time-step). 2) In the video encoding phase, feature maps of a sequence of frames (sampled from the video) are fed into the DNC one by one (one at a timestep). 3) At last, a softmax function is used to produce the probability of each possible answer according to the final visual-textual representation. Basically our model is not restricted to any particular method for generating word embeddings or image feature maps. We will discuss these implementation details later. Moreover, similar to LSTM, multiple DNCs can be stacked together to form a deep model, which usually leads to better performance for those applications related to visual and textual information processing. In our implementation, we employed a deep DNC with two layers, each of which has a distinct external memory.

Our encoding method is different from that in [15]. Even though the models presented in [15]

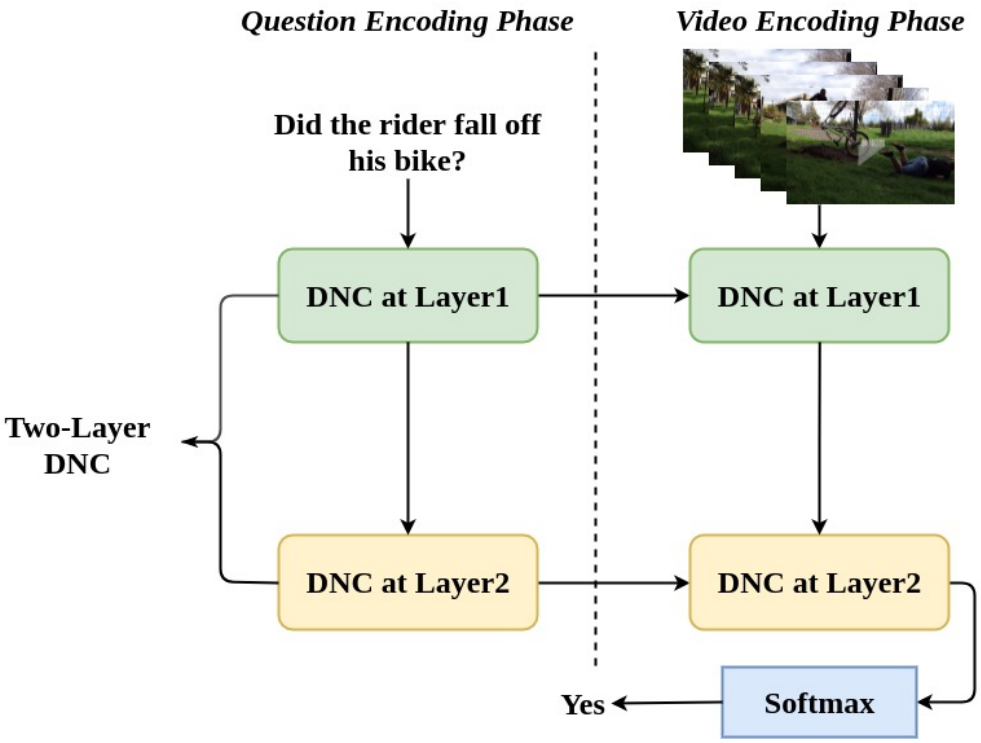

Fig. 2.2: The architecture of Memory Augmented Deep Recurrent Neural Network (MA-DRNN).

represents the state-of-the-art for general Video-QA, there is obvious room for improvement. For example, when testing their models for *Yes/No* questions, the accuracies are around 50%, which is the correct rate of random guessing. An even more important observation from their experimental results [15] is that no significant performance gain has been brought by integrating visual information into the proposed models because there is just a little improvement for each model trained with all data ("Train-all") compared to its counterpart trained without visual observations ("Non-visual") on average classification accuracy; and even a little drop appears on the performance for *Yes/No* questions when using E-SS, E-VQA or E-MN [15]. This leads us to believe that the video features have not play their role well in answering questions, especially those *Yes/No* questions. We suspect that it is the long-term sequence encoding that causes the oblivion of video features, which are encoded before question words; and moreover the final representation lacks useful and question-related visual information, which leads to non-ideal results for Video-QA. In addition, encoding a video before a question [15] may also result in a bad consequence that a large amount of irrelevant and useless information contained in the video may fill up the internal memory of the

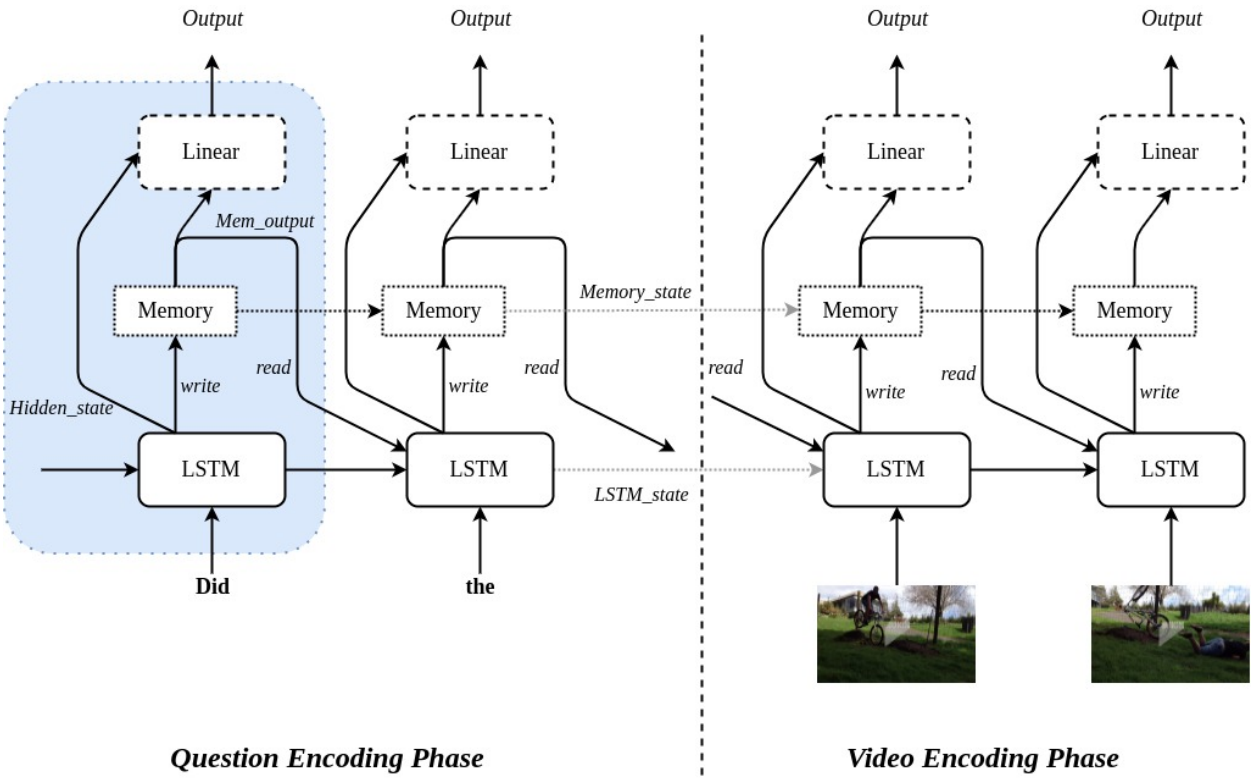

Fig. 2.3: The internal data flow of DNC at each layer of MA-DRNN.

LSTM because of lack of proper guidance for selection and preservation of useful visual information that is closely related to the corresponding question. As far as we know, similar to [15], many other works [22, 16] on modeling the visual-textual dependency encode visual information before textual information. For the purpose of strengthening the influence of visual features on the final representation as well as selecting and preserving useful and question-related visual information through the encoding phase to utilize the memory effectively, our model encodes a question and then encodes a video. The advantage of doing so can be easily understood by thinking about how a human does Video-QA. A video usually includes very rich information, far beyond the normal amount of information received by a person. If he/she first watches a video (probably a long one) and then gets a question (E-SS model [15]), then he/she may not be able to answer it well because he/she may remember a lot of video information that is irrelevant to the question heard later. On the contrary, if he/she gets the question first and then watches the video (our model), then he/she should answer it well since he/she can pay special attention to those relevant and useful information

in the video. The effectiveness of such an encoding method has been justified by our experimental results presented later.

In addition, to enhance the capacity of LSTM on modeling long-term dependency, we introduce memory augmentation into our model by leveraging DNC for question and video encoding, which adds an external memory to LSTM to provide additional storage space for modeling the long-term visual-textual dependency and guiding the generation of final representation for answer prediction. Moreover, the external memory allows more useful textual information to be retained after the question encoding for the better selection of relevant visual information in the video encoding phase. Memory augmentation has also been shown to be effective for Video-QA by our experimental results presented later.

Next, we explain how the LSTM controller interacts with the external memory during the encoding phase in details, which is illustrated in Figure 2.3. Since this process in every layer of the deep DNC is almost the same, we only show it in the first layer. Simply speaking, at each time-step, the content read from the memory of the previous time-step serves as input to LSTM to determine the content written into the memory by the write head at the current time-step. The output at each time-step is determined by both LSTM's hidden state and the content read from the external memory. After the encoding phase, the important textual and visual information as well as the long-term visual-textual dependency will be retained in the external memory and the internal memory of LSTM, which are used to generate the final representation.

Specifically, at each time-step t, the LSTM controller receives two vectors, one is the input $\mathbf{x}_t$ at current time-step, another one is the content/read vector $\mathbf{r}_{t-1}$ read from the memory of the previous time-step, and the two vectors are concatenated to form the input vector $\mathbf{X}_t = [\mathbf{x}_t; \mathbf{r}_{t-1}]$ for the LSTM controller. The computation process of the first layer LSTM controller is shown as

follow:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_\mathbf{i}[\mathbf{X}_t; \mathbf{h}_{t-1}] + \mathbf{b}_\mathbf{i}); \\
\mathbf{f}_t &= \sigma(\mathbf{W}_\mathbf{f}[\mathbf{X}_t; \mathbf{h}_{t-1}] + \mathbf{b}_\mathbf{f}); \\
\mathbf{o}_t &= \sigma(\mathbf{W}_\mathbf{o}[\mathbf{X}_t; \mathbf{h}_{t-1}] + \mathbf{b}_\mathbf{o}); \\
\widetilde{\mathbf{c}}_t &= \tanh(\mathbf{W}_\mathbf{c}[\mathbf{X}_t; \mathbf{h}_{t-1}] + \mathbf{b}_\mathbf{c}); \\
\mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \widetilde{\mathbf{c}}_t; \\
\mathbf{h}_t &= \mathbf{o}_t \circ \tanh(\mathbf{c}_t);
\end{aligned} \tag{2.4}$$

where σ is the sigmoid function, $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$, $\mathbf{c}_t$ are input gate, forget gate, output gate and cell state of LSTM controller of current time-step respectively, $\mathbf{c}_t$ is passed to the LSTM controller of next time-step, $\mathbf{h}_t$ denotes the hidden state, and $\mathbf{W}$ terms and $\mathbf{b}$ terms are learnable weight matrices and biases.

$\mathbf{h}_t$ is used to control the write and read operations on external memory.

$$\begin{aligned}
\mathbf{k}_t^w &= \mathbf{W}_{key}^w \mathbf{h}_t; \\
\mathbf{e}_t &= \mathbf{W}_e \mathbf{h}_t; \\
\mathbf{v}_t &= \mathbf{W}_v \mathbf{h}_t.
\end{aligned} \tag{2.5}$$

Specifically, it determines the erase vector $\mathbf{e}_t$, write vector $\mathbf{v}_t$ and write key $\mathbf{k}_t^w$ (introduced in *Preliminaries*) for the write head to edit the external memory. The write key $\mathbf{k}_t^w$, erase vector $\mathbf{e}_t \in [0, 1]^W$ and write vector $\mathbf{v}_t$ are computed according to Equation (2.5), where $\mathbf{W}_{key}^w$, $\mathbf{W}_e$ and $\mathbf{W}_v$ are all learnable weight matrices. Then the write wighting vector $\mathbf{w}_t^w$ can be calculated based on $\mathbf{k}_t^w$, to further update the external memory as shown in Equation (2.2).

$$\mathbf{k}_t^{r,k} = \mathbf{W}_{key}^{r,k} \mathbf{h}_t. \tag{2.6}$$

$\mathbf{h}_t$ is also used to calculate the read key $\mathbf{k}_t^{r,k}$ for each read head according to Equation (2.6), where $\mathbf{W}_{key}^{r,k}$ is a learnable weight matrix. Then the read weighting vector $\mathbf{w}_t^{r,k}$ can be calculated based on $\mathbf{k}_t^{r,k}$, to further obtain read vector $\mathbf{r}_t^k$ as shown in Equation (2.3), where k denotes the index of each read head.

Description: As a line of bikers attempted the same jump, almost the entire row successfully landed and continued riding. But when the last rider tried to follow their lead, he crashed hard on the grass.
Question: Did the last biker crash on the grass?
Answer: **Our Model:** Yes; **Ground Truth:** Yes

Description: This BMX rider tried to perform a wild flip off a dirt incline and fell down hard.
Question: Does the BMX rider successfully complete the front flip off the dirt incline?
Answer: **Our Model:** No; **Ground Truth:** No

Description: Some great POV footage of a fellow skier going off a jump. If only he made it!
Question: Who is in the footage tumbling off a jump?
Answer: **Our Model:** Skier; **Ground Truth:** Skier

Description: When a snorkeler dove down to get a closer look at a cool shell, he was surprised when a perfectly camouflaged octopus emerged from the reef.
Question: Who emerges from the reef camouflage?
Answer: **Our Model:** Octopus; **Ground Truth:** Octopus

Fig. 2.4: The examples from the testing results given by the proposed MA-DRNN model

Then the read vectors $\mathbf{r}_t^1, \mathbf{r}_t^2, ..., \mathbf{r}_t^R$ read from the memory can be concatenated together to obtain the final read vector $\mathbf{r}_t = [\mathbf{r}_t^1; \mathbf{r}_t^2; ...; \mathbf{r}_t^R] \in \mathbb{R}^{R \times W}$ at current time-step. The final output $\mathbf{y}_t$ at each time-step t is given according to Equation (2.7), where $\mathbf{W}_h$ and $\mathbf{W}_r$ are both learnable weight matrices.

$$\mathbf{y}_t = \mathbf{W}_h \mathbf{h}_t + \mathbf{W}_r \mathbf{r}_t. \tag{2.7}$$

2.3 Performance Evaluation

In this section, we first introduce the dataset, some implementation details and evaluation metrics; and then present and analyze the experimental results.

2.3.1 Dataset Preparation

Two large-scale datasets for VideoQA were used in our experiments.

VTW: The dataset was first introduced in [15] based on a state-of-the-art question generation method [208]. It contains 151,263 QA pairs from 14,100 videos for training, 21,352 QA pairs from 2,000 videos for validation and 2,000 videos for testing. The QA pairs for training and validation were automatically generated from user-curated descriptions, while the QA pairs for testing were human-generated. In our experiments, 146,700 QA pairs from 14,091 videos were chosen for training and 3,600 QA pairs from 1,999 videos were chosen for testing.

MSVD-QA: The dataset was generated based on Microsoft Research Video Description Corpus [209], a widely used dataset for video captioning, using the generation method introduced in [210]. It contains 1,970 video clips and 50,505 QA pairs. For fair comparison, we used the same dataset split method as [46]: 30,933 QA pairs from 1,200 videos for training, 6,415 QA pairs from 250 videos for validation and 13157 QAs from 520 videos for testing.

Note that the VTW dataset contains two types of QA pairs, *Others* and *Yes/No*, while the MSVD-QA dataset only has *Others* QA pairs.

2.3.2 Experimental Setup

We used TensorFlow to implement the proposed model. In our implementation, we sampled 50 frames and 20 frames for each video in VTW and MSVD-QA datasets respectively in an evenly-paced manner, then resized and normalized each frame to zero mean and unit norm with a size of $224 \times 224 \times 3$. We extracted the appearance features of each frame using a widely-used CNN, VGG network [60] pre-trained on ImageNet 2012 classification dataset [211]. Moreover, 20 evenly

distributed clips (each with 16 frames) were sampled for each video, then the motion features of each clip were also extracted using C3D network [212], pre-trained on Sports-1M dataset [213]. Both appearance features and motion features have a dimension of $4,096$.

There are two types of text inputs: question and answer. The questions in VTW dataset with length larger than 15 were trimmed to a maximum of 15 words, while the questions with length smaller than 15 were padded with zeros to the length of 15 (10 for the MSVD-QA dataset), and each word in the questions was represented as a $300D$ vectors using the GloVe word embedding [214] pre-trained on the Common Crawl dataset. For answers, following the same rule as [15, 46], we used a simple but widely-used method to generate word embeddings. Specifically, we first calculated the occurrence frequency of each answer in training and validation set, then retained the top-1000 most frequent answers to create the answer space, then we represented each answer using a one-hot vector with length of 1000 based on the answer space. This set covers 81% and 96% of the training and validation answers in the VTW and MSVD-QA datasets respectively. In the training process, we set the batch size to 100 and used the RMSProp algorithm [215] with a learning rate of 0.0001.

We used the same metrics for performance evaluation as [15, 46]. For QA pairs whose answers are *Yes/No*, in order to penalize false-positive answers, we used

$$Acc^{\dagger} = \frac{TP}{TP + FP + FN} \tag{2.8}$$

to evaluate both *Yes* $Acc^{\dagger}$ and *No* $Acc^{\dagger}$, where TP is true-positive, FP is false-positive and FN is false-negative. $Acc^{\dagger}$ becomes higher when the number of true *Yes* or true *No* answers gets larger; while it becomes lower when the number of false *Yes* or false *No* answer gets larger. Meanwhile, we also used the standard classification accuracy: $Acc = N_c/N_s$ to evaluate the performance for answering the *Yes/No* questions, where N_c is the number of testing examples which have been classified correctly and N_s gives the total number of testing examples.

For *Others* QA pairs whose questions are general “what/how" kind of questions, we used both the standard classification accuracy Acc and the widely-used metric $WUPS$ [26] to evaluate their

performance. This is because the classification accuracy Acc is too strict, while $WUPS$ can be considered as a relaxed version that deals with the word-level ambiguities. Similar as in [15], we show $WUPS$ scores with thresholds of 0.0 and 0.9 respectively. In addition, we also calculated the average classification accuracy $Avg.Acc$ of *Yes/No* and *Others* to show the overall performance for each method for VTW dataset. For the MSVD-QA dataset, since five types of questions were included, besides the overall standard classification accuracy Acc and $WUPS$ scores, we also reported the accuracy of each question type.

Table 2.1: Performance comparisons with the VTW dataset in terms of various accuracy-related metrics

Models	Others WUPS 0.0 (%)	Others WUPS 0.9 (%)	Others Acc (%)	Yes $Acc^{\dagger}$ (%)	No $Acc^{\dagger}$ (%)	Yes/No Acc (%)	Avg.Acc (%)
ST	32.9	5.51	2.1	11.9	26.7	49.3	25.7
E-MN	47.9	10.1	2.9	40.0	13.0	49.5	26.2
E-VQA	49.3	13.2	5.0	38.8	22.3	46.7	25.9
E-SA	51.4	15.5	8.4	36.4	28.8	52.4	30.4
E-SS	48.7	14.2	7.3	34.5	25.8	49.5	28.4
MA-DRNN	**56.3**	**20.7**	**11.4**	**54.9**	**38.6**	**67.4**	**39.4**

2.3.3 Analysis of Results

First, we show some examples from the VTW dataset given by MA-DRNN during our testing in Figure 2.4, which include four sets of frames from some testing videos and the corresponding questions and answers. For instance, the first example shows that the last rider failed to jump and land on the ground in the last part of the video, and when the question “Did the last biker crash in the grass?" is asked, the correct answer “Yes" was given by our model. In the fourth example, the question is “Who emerges from reef camouflage?". As shown in the video, it is a camouflaged octopus gradually emerged from the reef. The correct answer “Octopus" was given by our model.

In our experiments with the VTW dataset, we compared the proposed MA-DRNN model with the four models presented in [15], including E-MN, E-VQA, E-SA and E-SS, which represent the state-of-the-art on Video-QA. In addition, we included the results related to another method, Skip-Thought (ST) [216], which was used as the baseline for comparisons in [15]. Note that a self-paced method was used in [15] to identify and get rid of those bad QA pairs during training,

Table 2.2: Performance comparisons and ablation study with the MSVD-QA dataset

Models	what (%)	who (%)	how (%)	when (%)	where (%)	Acc (%)	WUPS 0.0 (%)	WUPS 0.9 (%)
E-MN	18.7	46.8	56.8	64.7	73.7	30.4	72.9	38.4
E-VQA	11.8	40.7	84.2	84.3	47.4	24.8	70.7	33.3
E-SA	16.5	45.2	83.1	82.4	73.7	29.3	73.7	38.1
GRA	22.8	45.0	73.4	82.4	**78.9**	32.7	73.9	41.0
DRNN-v1/E-SS	21.9	47.5	79.2	66.7	15.8	33.1	72.1	41.0
DRNN-v2	22.4	50.0	69.0	76.5	31.6	34.1	73.2	41.8
MA-DRNN-v1	23.3	49.6	**83.1**	74.5	10.5	34.9	73.2	42.6
MA-DRNN-v2	**24.3**	**51.6**	82.0	**86.3**	26.3	**36.2**	**74.1**	**43.7**

which has been shown to lead to minor performance improvement. Since this method is just a data pre-processing method and we are not sure which QA pairs were eliminated by them, we used all the data for training and included their “Train-all" results from [15] for fair comparisons. We presented all the results with the VTW dataset in TABLE 2.1.

In our experiments with the MSVD-QA dataset, we compared the proposed MA-DRNN model with the five models presented in [15, 46] including E-MN, E-VQA, E-SA, E-SS and GRA, whose results were obtained by running the implementation described in [217]. We conducted a comprehensive ablation study with the MSVD-QA dataset to evaluate the effectiveness and superiority of the proposed encoding method as well as memory augmentation. We implemented the DRNN-v1/E-SS, DRNN-v2 and MA-DRNN-v1 to serve as the baseline for comparisons. Here, compared to MA-DRNN, DRNN does not use memory augmentation; and v1 denotes encoding videos before questions, v2 denotes encoding question before videos (the proposed encoding method). Note that DRNN-v1 is the same as E-SS so we put them together; and MA-DRNN-v2 here is the same as MA-DRNN in TABLE 2.1. We presented all the results on the MSVD-QA dataset in TABLE 2.2. We can make the following observations from these experimental results:

1) For the *Yes/No* questions of VTW dataset, MA-DRNN significantly outperforms all the other methods in terms of classification accuracies. E-SA turns out to perform best among all the four models in [15]. MA-DRNN achieves classification accuracies of 67.4%, 54.9% and 38.6% in terms of the three metrics respectively, which represent 15.0%, 18.5% and 9.8% improvements over those of E-SA. It is worth mentioning that compared to the baseline ST, all the four models in [15] only lead to minor improvement (or even worse performance) in terms of *Yes/No* Acc; and moreover, the

corresponding results are all around 50%, which is the correct rate of random guessing. However, MA-DRNN offers a much better accuracy of 67.4%.

2) For the *Others* questions, MA-DRNN outperforms all the other methods in terms of both standard classification accuracy and $WUPS$. On the VTW datast, MA-DRNN achieves a standard classification accuracy of 11.4%, representing 3.0% improvement over that of E-SA, which performs best among all the four models in [15]. MA-DRNN achieves $WUPS$ scores of 56.3% and 20.7% with thresholds of 0.0 and 0.9, which represent 4.9% and 5.2% improvements over those of E-SA. On the MSVD-QA dataset, MA-DRNN-v2 (i.e., the proposed model) achieves a standard classification accuracy of 36.2% and $WUPS$ scores of 74.1% and 43.7%, which represents the best performance among all the models mentioned above. In addition, MA-DRNN achieves the highest accuracies in most question types.

3) In term of average accuracy (i.e., $Avg.ACC$), MA-DRNN still achieves better performance than all the other baselines. Particularly, on the VTW datast (TABLE 2.1), compared to ST, E-MN, E-VQA, E-SA and E-SS, MA-DRNN improves the average accuracy by $13.7\%, 13.2\%, 13.5\%, 9.0\%$ and 11.0% respectively.

4) For the ablation study with the MSVD-QA dataset, first of all, we can see that MA-DRNN-v1 and MA-DRNN-v2 perform consistently better than DRNN-v1 and DRNN-v2 respectively, which do not use memory augmentation. For example, compared to DRNN-v2, MA-DRNN-v2 improves the standard classification accuracy from 34.1% to 36.2%. These results well justify the effectiveness of DNC-based memory augmentation. Furthermore, DRNN-v2 and MA-DRNN-v2 perform consistently better than DRNN-v1 and MA-DRNN-v1 respectively, which uses a commonly-used method to encode videos before questions. For example, compared to MA-DRNN-v1, MA-DRNN-v2 improves the standard classification accuracy from 34.9% to 36.2%. These results provide a convincing evidence that the textual information (i.e. questions) can provide a proper guidance for selection and preservation of useful visual information. Thus, the proposed encoding method leads to better visual-textual representations, which offer superior performance on the Video-QA task.

2.4 Summary

In this chapter, we presented a novel MA-DRNN model for Video-QA, which features a new question and video encoding method and memory augmentation using the emerging DNC. Extensive experiments and ablation studies were conducted with the widely-used VTW dataset and MSVD-QA dataset for performance evaluation. The ablation study justified the effectiveness and superiority of the proposed encoding method and the use of memory augmentation. Moreover, the experimental results have shown that MA-DRNN delivers state-of-the-art performance in terms of various accuracy-related metrics. The proposed encoding method, which has been proven effective for Video-QA, can be extended to other tasks which need both visual and textual encoding. In the future, we will investigate how to improve the memory addressing mechanism of augmented memory and make it more suitable to the temporal structure for encoding visual and textual information over long timescales.

CHAPTER 3
MULTI-CLUE REASONING WITH MEMORY AUGMENTATION FOR KNOWLEDGE-BASED IMAGE QUESTION ANSWERING

3.1 Overview

In this chapter, to deal with the knowledge-based image question answering (KB-Image-QA) task and address the aforementioned issues, we present a novel framework that focuses on achieving a better exploitation of external knowledge through generating **M**ultiple **C**lues for **R**easoning with **Mem**ory **N**eural **N**etworks. We name it as **MCR-MemNN** for short. Specifically, as illustrated by Figure 3.1, the image-question pair is encoded into a visual-semantic representation, which serves as the input to a well-defined detector to obtain a three-term relation phrase (i.e., $\langle$ *Subject, Relation, Object* $\rangle$). In this case, either the subject (i.e., cat) or the object (i.e., tiger) acts a clue to retrieve supporting facts in the external KB, and both deliver a fact set including the ground-truth fact (i.e., cat RelatedTo tiger). In this manner, the ground-truth fact can be success-

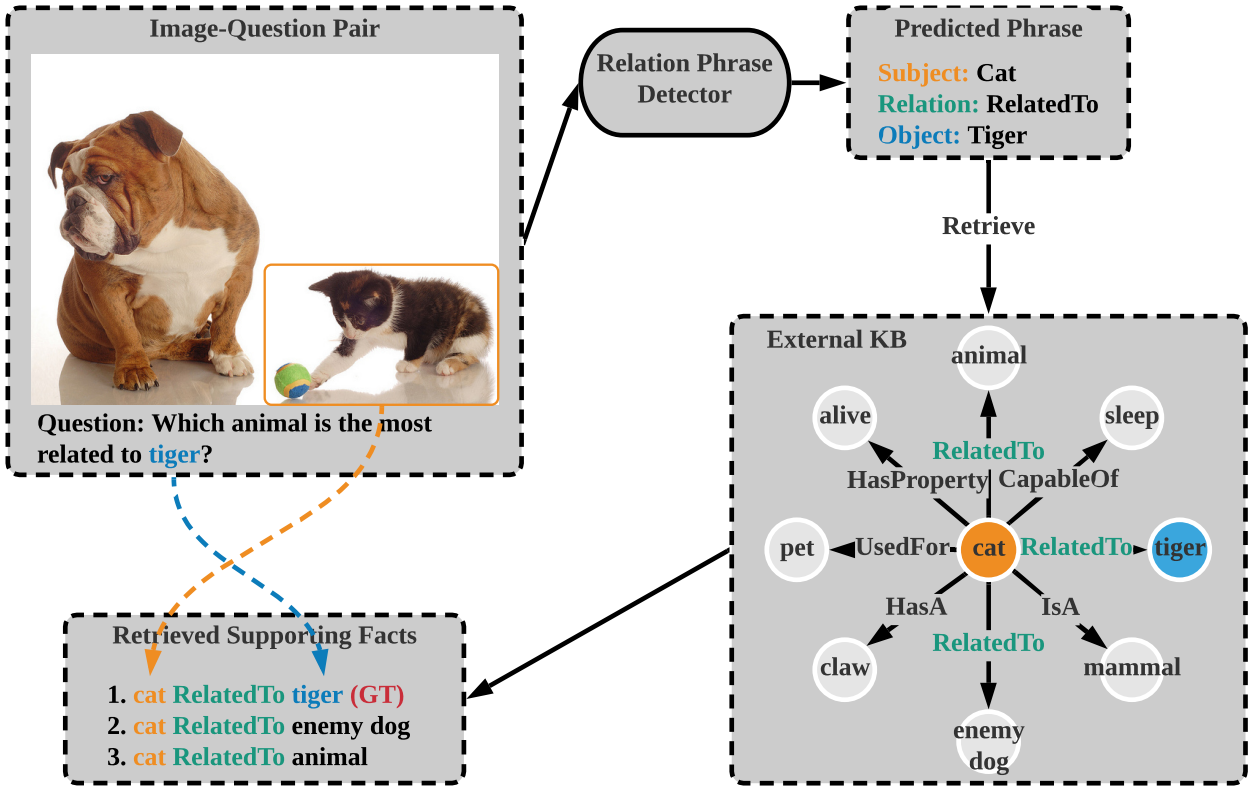

Fig. 3.1: An illustration of Multi-Clue Reasoning.

fully fetched as long as one of the two complementary clues is predicted. Note that the predicted relation (i.e., RelatedTo) is adopted to filter out some facts with different relations. The retained supporting facts, including the ground-truth, are further encoded into a continuous embedding space and stored in a content-addressable memory, where each memory slot corresponds to a supporting fact. Afterwards, for reasoning procedure, we assume that the visual and semantic content in the image-question pair can contribute to a better exploitation of the external knowledge (i.e., supporting facts). Analogously, the external knowledge is helpful for a better understanding of the image-question pair. Therefore, we employ a two-way attention mechanism, which is not only intended to focus on the important aspects of the memory in light of the image-question pair, but also the important parts of image and question in light of the memory.

The main contributions of this chapter are summarized as follows: 1) We demonstrate a new KB retrieval method with two complementary clues (i.e., subject and object), which makes it a lot easier to fetch the ground-truth supporting fact. 2) A two-way attention mechanism with memory

augmentation is employed to model the inter-relationships among image, question and KB to distill the most relevant information in the three modalities. 3) We perform extensive experiments on two widely-used benchmarks, which shows that the proposed MCR-MemNN is an effective framework customized for KB-Image-QA.

3.2 Problem Statement

Given an image **I** and a related question **Q**, the KB-Image-QA task aims to predict an answer **A** from the external knowledge base (KB), which consists of facts in the form of triplet (i.e., $\langle$ *Subject, Relation, Object* $\rangle$), where subject represents a visual concept, object represents an attribute or a visual concept and relation denotes the relationship between the subject and the object. Note that the answer **A** can be either the subject or object in the triplet. The key of KB-Image-QA is to select the correct supporting fact and then determine the answer. For convenience, in our notations, the fact $\langle$ Subject, Relation, Object $\rangle$ corresponds to the answer subject; while its reversed form $\langle$ Object, Relation, Subject $\rangle$ corresponds to the answer object. For instance, as for the ground-truth supporting fact (i.e., cat *RelatedTo* tiger) in Figure 3.3, the corresponding answer is cat, and the answer of its reversed form (i.e., tiger *RelatedTo* cat) is tiger. In this manner, during inference stage, the optimal answer is the first term of the predicted fact.

3.3 Multi-Clue Reasoning with Memory Augmentation

3.3.1 Relation Phrase Detector

Same as the fact in KB, the relation phrase is depicted in a form of triplet (i.e., $\langle$ *Subject, Relation, Object* $\rangle$), which contains two complementary clues (i.e., subject and object) for KB retrieval. As mentioned, the two complementary clues make it a lot easier to fetch the ground-truth supporting fact, and the relation predicted can be leveraged to filter out the facts with wrong relations. Therefore, we develop a detector to obtain a relation phrase based on the visual and semantic content of the

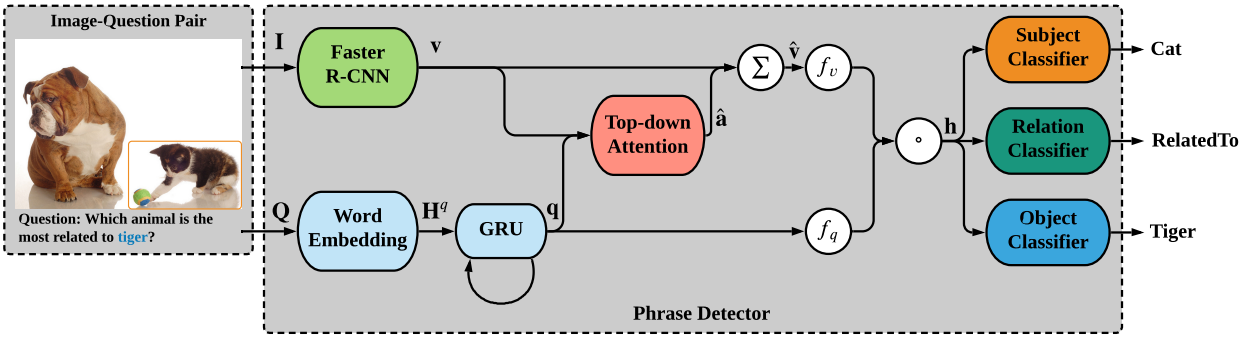

Fig. 3.2: Relation Phrase Detector.

image-question pair.

As shown in Figure 3.2, given an image-question pair $\{\mathbf{I}, \mathbf{Q}\}$, the relation phrase prediction is formulated as a multi-task classification problem following [218]. For image embedding, we use a Faster R-CNN [79] in conjunction with the ResNet-101 [62] pretrained on Visual Genome dataset [28] to generate an output set of image features $\mathbf{v}$ as the visual representation of the input image.

$$\mathbf{v} = \text{Faster R-CNN}(\mathbf{I}) \tag{3.1}$$

where $\mathbf{v} \in \mathcal{R}^{2048 \times K}$ is based on the bottom-up attention [8], which represents the ResNet features centered on the top-K objects in the image. K is set to 36 in our experiments.

For question embedding, we first transfer each word in $\mathbf{Q}$ into a feature vector using the pre-trained 300D GloVe [214] vectors, and use randomly initialized vectors for the words which are out of GloVe's vocabulary. Here we denote the resulting sequence of word embeddings as $\mathbf{H}^q$. Then the Gated Recurrent Unit (GRU) [219] with hidden state dimension 512 is adopted to encode the $\mathbf{H}^q$ as semantic representation $\mathbf{q} \in \mathcal{R}^{512}$.

$$\mathbf{q} = \text{GRU}(\mathbf{H}^q) \tag{3.2}$$

To encode the image and question in a shared embedding space, top-down attention [8] is employed to fuse the visual representation $\mathbf{v}$ and semantic representation $\mathbf{q}$. Specifically, for each

object from $i = 1...K$, its feature $\mathbf{v}_i$ is concatenated with the semantic representation $\mathbf{q}$, and then passed through a non-linear layer f_a and a linear layer to obtain its corresponding attention weight a_i.

$$a_i = \mathbf{W}_a f_a([\mathbf{v}_i, \mathbf{q}]) \tag{3.3}$$

$$\hat{\mathbf{a}} = \text{softmax}(\mathbf{a}) \tag{3.4}$$

$$\hat{\mathbf{v}} = \sum_{i=1}^{K} \hat{a}_i \mathbf{v}_i \tag{3.5}$$

where $\mathbf{W}_a$ is a learnable weight vector and the attention weights $\mathbf{a}$ are normalized with a softmax function to $\hat{\mathbf{a}}$. The image features are weighted by the normalized attention values to get the weighed visual representation $\hat{\mathbf{v}} \in \mathcal{R}^{2048}$. Following [55], the non-linear layer $f_a : \mathbf{x} \in \mathcal{R}^m \rightarrow \mathbf{y} \in \mathcal{R}^n$ with parameters a is defined as follows:

$$\widetilde{\mathbf{y}} = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{3.6}$$

$$\mathbf{g} = \sigma(\mathbf{W}^{'}\mathbf{x} + \mathbf{b}^{'}) \tag{3.7}$$

$$\mathbf{y} = \widetilde{\mathbf{y}} \circ \mathbf{g} \tag{3.8}$$

where σ is the sigmoid activation function, $\mathbf{W}, \mathbf{W}^{'} \in \mathcal{R}^{n \times m}$ and $\mathbf{b}, \mathbf{b}^{'} \in \mathcal{R}^n$ are learnable parameters, and $\circ$ denotes the element-wise multiplication.

A joint embedding $\mathbf{h}$ of the question and the image is obtained by the fusion of the weighed visual representation $\hat{\mathbf{v}}$ and the semantic representation $\mathbf{q}$.

$$\mathbf{h} = f_v(\hat{\mathbf{v}}) \circ f_q(\mathbf{q}) \tag{3.9}$$

where $\mathbf{h} \in \mathcal{R}^{512}$. Both f_v and f_q are the non-linear layers with the same form as f_a. The joint embedding $\mathbf{h}$ is then fed into a group of linear classifiers for the prediction of subject, relation and

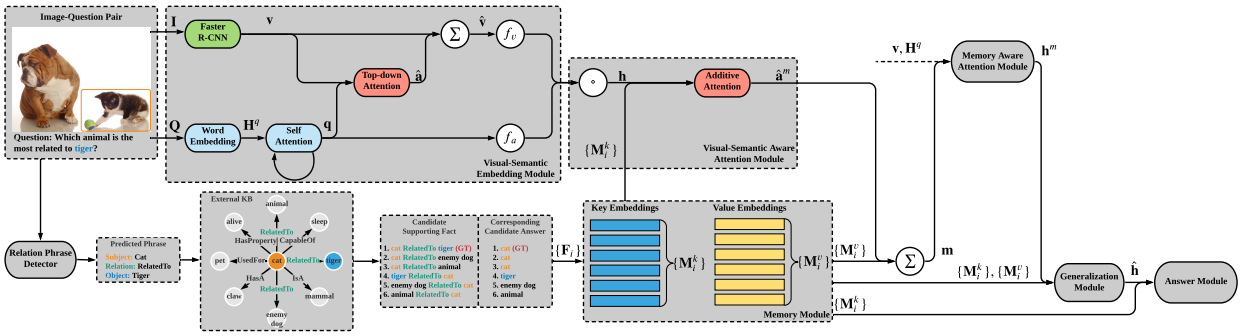

Fig. 3.3: Architecture of MCR-MemNN.

object in a relation phrase.

$$\hat{\mathbf{s}} = \text{softmax}(\mathbf{W}_s\mathbf{h} + \mathbf{b}_s) \tag{3.10}$$

$$\hat{\mathbf{r}} = \text{softmax}(\mathbf{W}_r\mathbf{h} + \mathbf{b}_r) \tag{3.11}$$

$$\hat{\mathbf{o}} = \text{softmax}(\mathbf{W}_o\mathbf{h} + \mathbf{b}_o) \tag{3.12}$$

where $\mathbf{W}_s, \mathbf{W}_r, \mathbf{W}_o$ and $\mathbf{b}_s, \mathbf{b}_r, \mathbf{b}_o$ are learnable parameters, and $\hat{\mathbf{s}}, \hat{\mathbf{r}}, \hat{\mathbf{o}}$ denote the predicted probability for subject, relation and object over each candidate, respectively. The loss function for the relation phrase detector is defined as

$$\mathcal{L}_d = \lambda_s\mathcal{L}_c(\mathbf{s}, \hat{\mathbf{s}}) + \lambda_r\mathcal{L}_c(\mathbf{r}, \hat{\mathbf{r}}) + \lambda_o\mathcal{L}_c(\mathbf{o}, \hat{\mathbf{o}}) \tag{3.13}$$

where $\mathbf{s}, \mathbf{r}, \mathbf{o}$ are the ground-truth labels for subject, object and relation, respectively. $\mathcal{L}_c$ represents the cross-entropy loss, and the weights of the loss terms (i.e., $\lambda_s, \lambda_r, \lambda_o$) are set to 1.0 in our experiments.

3.3.2 MCR-MemNN

As shown in Figure 3.3, besides the Relation Phrase Detector, our proposed MCR-MemNN framework consists of six components, which are Memory Module, Visual-Semantic Embedding Module, Visual-Semantic Aware Attention Module, Memory Aware Attention Module, Generalization

Module and Answer Module.

Memory Module For an image-question pair $\{\mathbf{I}, \mathbf{Q}\}$, after the three-term phrase (i.e., Subject, Relation, Object) is predicted by the relation phrase detector, all the facts in the external KB with entities pointed by the subject or pointed to the object within h hops are collected as candidate supporting facts. h is set to 1 in our experiments. The relation predicted is leveraged to further filter out the facts with wrong relations. Then a set of candidate supporting facts $\{\mathbf{F}_i\}_{i=1}^{N}$ is collected, where N denotes the number of facts in the fact set, and each fact $\mathbf{F}_i$ consists of a sequence of words.

Afterwards, similar to question embedding, each fact $\mathbf{F}_i$ is transformed to a sequence of word embeddings $\mathbf{H}_i^f$ based on GloVe's vocabulary, and encoded using a BiLSTM to get its representation $\mathbf{f}_i \in \mathcal{R}^{128}$.

$$\mathbf{f}_i = \text{BiLSTM}(\mathbf{H}_i^f) \tag{3.14}$$

To store the candidate supporting facts, a key-value structured memory network [220] is leveraged. The key memory is used in the addressing stage, while the value memory is used in the reading stage. The representation $\mathbf{f}_i$ of each fact is passed through two separate linear layers to obtain its key embedding $\mathbf{M}_i^k \in \mathcal{R}^{128}$ and value embedding $\mathbf{M}_i^v \in \mathcal{R}^{128}$ respectively.

$$\mathbf{M}_i^k = \mathbf{W}_k \mathbf{f}_i \tag{3.15}$$

$$\mathbf{M}_i^v = \mathbf{W}_v \mathbf{f}_i \tag{3.16}$$

where $\mathbf{W}_k \in \mathcal{R}^{128\times128}$ and $\mathbf{W}_v \in \mathcal{R}^{128\times128}$ are learnable parameters. For the set of candidate supporting facts $\{\mathbf{F}_i\}_{i=1}^{N}$, we have a set of key embeddings $\mathbf{M}^k = \{\mathbf{M}_i^k\}_{i=1}^{N}$ and a set of value embeddings $\mathbf{M}^v = \{\mathbf{M}_i^v\}_{i=1}^{N}$. Note that one memory slot is defined as a pair of key embedding and value embedding (i.e., $\{\mathbf{M}_i^k, \mathbf{M}_i^v\}$) of one candidate supporting fact.

Visual-Semantic Embedding Module The visual-semantic embedding module has basically the same architecture as the relation phrase detector and the only difference is about the question

embedding before top-down attention, where self-attention are applied over the sequence of word embeddings $\mathbf{H}^q$ to get the semantic representation $\mathbf{q} \in \mathcal{R}^{128}$.

$$\hat{\mathbf{a}}^{qq} = \text{softmax}((\mathbf{H}^q)^\intercal \mathbf{H}^q) \tag{3.17}$$

$$\mathbf{q} = \text{BiLSTM}([\mathbf{H}^q(\hat{\mathbf{a}}^{qq})^\intercal, \mathbf{H}^q]) \tag{3.18}$$

where the $\mathbf{H}^q$ are weighted by the normalized values, concatenated with itself and fed into a BiLSTM. Afterwards, same procedures are conducted to obtain the visual-semantic representation $\mathbf{h} \in \mathcal{R}^{128}$.

Visual-Semantic Aware Attention Module Given visual-semantic representation $\mathbf{h}$, we apply an attention over all the memory slots $\{\mathbf{M}_i^k, \mathbf{M}_i^v\}_{i=1}^N$ to obtain the memory summary $\mathbf{m} \in \mathcal{R}^{128}$ in light of the visual and semantic content of the image-question pair.

$$a_i^m = \mathbf{W}_3 \text{tanh}(\mathbf{W}_1 \mathbf{h} + \mathbf{W}_2 \mathbf{M}_i^k) \tag{3.19}$$

$$\hat{\mathbf{a}}^m = \text{softmax}(\mathbf{a}^m) \tag{3.20}$$

$$\mathbf{m} = \sum\nolimits_{i=1}^{N} \hat{a}_i^m \mathbf{M}_i^v \tag{3.21}$$

where $\mathbf{W}_1 \in \mathcal{R}^{128\times128}, \mathbf{W}_2 \in \mathcal{R}^{128\times128}, \mathbf{W}_3 \in \mathcal{R}^{1\times128}$ are learnable parameters. The attention for each memory slot is calculated and normalized based on $\mathbf{h}$ and the corresponding key embedding $\mathbf{M}_i^k$. Then the set of value embeddings $\mathbf{M}^v = \{\mathbf{M}_i^v\}_{i=1}^N$ are weighted to get the memory summary $\mathbf{m}$.

Memory Aware Attention Module As we have obtained the memory summary $\mathbf{m}$, we proceed to compute the attentions over all the question words and all the image features in light of the memory.

Given memory summary $\mathbf{m}$, the sequence of word embeddings $\mathbf{H}^q$ and the set of image features $\mathbf{v}$, the memory-aware question embedding $\mathbf{q}^m \in \mathcal{R}^{128}$ and the memory-aware image embedding

$\mathbf{v}^m \in \mathcal{R}^{2048}$ are derived as follows:

$$\hat{\mathbf{a}}^q = \text{softmax}((\mathbf{H}^q)^\intercal \mathbf{m}) \tag{3.22}$$

$$\mathbf{q}^m = \mathbf{H}^q \hat{\mathbf{a}}^q \tag{3.23}$$

$$\hat{\mathbf{a}}^v = \text{softmax}((\mathbf{W}_v \mathbf{v})^\intercal \mathbf{m}) \tag{3.24}$$

$$\mathbf{v}^m = \mathbf{v}\hat{\mathbf{a}}^v \tag{3.25}$$

where $\hat{\mathbf{a}}^q$ represents the normalized memory aware attention over all the question words, $\hat{\mathbf{a}}^v$ represents the normalized memory aware attention over all the image features and $\mathbf{W}_v \in \mathcal{R}^{128\times2048}$ are learnable parameters.

The visual-semantic representation $\mathbf{h}^m$ in light of the memory is obtained by the fusion of the memory-aware question embedding $\mathbf{q}^m$ and the memory-aware image embedding $\mathbf{v}^m$.

$$\mathbf{h}^m = f_v^m(\mathbf{v}^m) \circ f_q^m(\mathbf{q}^m) \tag{3.26}$$

where $\mathbf{h}^m \in \mathcal{R}^{128}$. Both f_v^m and f_q^m are the non-linear layers with the same form as f_a. Note that the aforementioned two-way mechanism corresponds to Sections 3.3.2 and 3.3.2.

Generalization Module Inspired by [221], another hop of the attention process is conducted over the memory before answer prediction. Attention mechanism of Section 3.3.2 is leveraged here, given memory aware visual-semantic representation $\mathbf{h}^m$, to fetch the most relevant information $\mathbf{m}^h \in \mathcal{R}^{128}$ from the memory to obtain the final visual-semantic representation $\hat{\mathbf{h}} \in \mathcal{R}^{128}$. To be more specific, the fetched information $\mathbf{m}^h$ is concatenated with $\mathbf{h}^m$ and fed into a GRU to

update the visual-semantic representation. To the end, a batch normalization (BN) layer is used.

$$a_i^h = \mathbf{W}_6 \text{tanh}(\mathbf{W}_4 \mathbf{h}^m + \mathbf{W}_5 \mathbf{M}_i^k) \tag{3.27}$$

$$\hat{\mathbf{a}}^h = \text{softmax}(\mathbf{a}^h) \tag{3.28}$$

$$\mathbf{m}^h = \sum_{i=1}^{N} \hat{a}_i^h \mathbf{M}_i^v \tag{3.29}$$

$$\hat{\mathbf{h}} = \text{BN}(\mathbf{h}^m + \text{GRU}([\mathbf{h}^m, \mathbf{m}^h])) \tag{3.30}$$

where $\mathbf{W}_4 \in \mathcal{R}^{128\times128}$, $\mathbf{W}_5 \in \mathcal{R}^{128\times128}$, $\mathbf{W}_6 \in \mathcal{R}^{1\times128}$ are learnable parameters.

Answer Module Given the final visual-semantic representation $\hat{\mathbf{h}}$, and the set of key embeddings $\mathbf{M}^k \in \mathcal{R}^{128\times N}$, the key embedding $\mathbf{M}_i^k$ of each candidate supporting fact is concatenated with the $\hat{\mathbf{h}}$ to compute the probability of whether the fact is correct. The predicted supporting fact is

$$\text{argmax}_i \text{softmax}(\mathbf{W}_7[\hat{\mathbf{h}}, \mathbf{M}^k] + \mathbf{b}^7) \tag{3.31}$$

where $i = 1, ..., N$, $\mathbf{W}_7 \in \mathcal{R}^{1\times256}$ and $\mathbf{b}^7 \in \mathcal{R}^N$ are learnable parameters. As we have stated in Section 3.2, the optimal answer is the first term of the predicted fact.

Loss Function Once we have the candidate supporting facts retrieved from the external KB, all the learnable parameters of the proposed MCR-MemNN (besides the Relation Phrase Detector) are trained in an end-to-end manner by minimizing the following loss function over the training set.

$$\mathcal{L} = \frac{1}{D} \sum_{k=1}^{D} \mathcal{L}_c(\mathbf{Y}_k, \hat{\mathbf{Y}}_k) \tag{3.32}$$

where $\mathcal{L}_c$ is defined as the cross-entopy loss, D is the number of training samples, $\mathbf{Y}_k$ and $\hat{\mathbf{Y}}_k$ represent the ground-truth label and the predicted probability over each candidate supporting fact.

3.4 Performance Evaluation

We employed two compelling benchmarks, FVQA [32] and Visual7W+ConceptNet [222] to evaluate the proposed MCR-MemNN on Knowledge-based Image Question Answering (KB-Image-QA) task.

3.4.1 Benchmark Datasets

FVQA. The Factual Visual Question Answering (FVQA) dataset consists of $2,190$ images, $5,286$ questions and $4,126$ unique facts corresponding to the questions. The external KB of FVQA, consisting of $193,449$ facts, are constructed based on three structured KBs, including DBpedia [223], ConceptNet [224] and WebChild [225]. Following [32], the top-1 and top-3 accuracies are averaged over five test splits.

Visual7W+ConceptNet. The Visual7W+ConceptNet dataset, built by [222], is a collection of knowledge-based questions with images sampled from the test split of Visual7W [31] dataset. It consists of $16,850$ open domain question-answer pairs based on $8,425$ images. Note that the supporting facts of each question-answer pair are retrieved directly from the ConceptNet, which serves as the external KB. Following [32], the top-1 and top-3 accuracies are calculated over the test set.

3.4.2 Experimental Setup

Implementation Details For the training of the relation phrase detector, the model was trained with Adam optimizer [226] with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-6}, and the batch size is set to 32. For the training of the proposed MCR-MemNN (besides the relation phrase detector), the model was trained with Adam optimizer with an initial learning rate of 1×10^{-3} and weight decay of 1×10^{-6}, and the batch size is set to 64. Top 40 predicted subjects and objects were adopted as clues to retrieve candidate supporting facts from the external KB. The memory size N_{mem} was set to 96. Note that no matter the number of candidate supporting

Table 3.1: Results for the relation phrase detector on FVQA.

Case	Classification Acc.			Recall		QA Acc.	
	Sub	Obj	Union	w/ Rel	w/o Rel	w/ Rel	w/o Rel
Top-20	64.31	39.43	69.80	78.56	82.08	60.36	62.76
Top-30	69.46	43.30	74.24	83.61	86.90	66.34	65.28
Top-40	72.65	45.71	76.22	85.76	88.85	**70.92**	**68.85**
Top-60	73.59	46.23	77.10	86.58	89.81	68.52	67.30
Top-100	75.60	47.61	79.37	88.21	90.33	68.02	65.57

facts N is larger than, equal to or smaller than N_{mem}, the ground-truth fact is preserved and the negative facts are randomly selected to fill up the remaining memory. Our code was implemented in PyTorch [227] and run with NVIDIA Tesla P100 GPUs.

Baselines For the FVQA dataset, CNN-RNN based approaches including LSTM-Quest-ion+Image+Pre-VQA [32] and Hie-Question+Image+Pre-VQA [32], semantic parsing bas-ed approaches including FVQA (top-3-QQmaping) [32] and FVQA (Ensemble) [32], learni-ng-based approaches including Straight to the Facts (STTF) [35], Out of the Box (OB) [34], Reading Comprehension [228], and Multi-Layer Cross-Modal Knowledge Reasoning (Muc-ko) [125] are compared with the proposed MCR-MemNN.

For the Visual7W+ConceptNet dataset, learning based approaches including KDMN [222], Out of the Box (OB) are compared with the proposed MCR-MemNN. Note that both KDMN and MCR-MemNN adopted memory augmentation technique for storing the retrieved external knowledge.

3.4.3 Experimental Results

Relation Phrase Detector As shown in Table 3.1, to evaluate the performance of the relation phrase detector, five different cases are considered, including Top-20, Top-30, Top-40, Top-60 and Top-100. For each case, the classification accuracies of subject (i.e., 'Sub'), object (i.e, 'Obj') and both union (i.e., 'Union') are calculated. In addition, the answer recall is reported with the

Table 3.2: Accuracy comparison on FVQA.

Method	Overall Accuracy	
	top-1	top-3
LSTM-Question+Image+Pre-VQA [32]	24.98	40.40
Hie-Question+Image+Pre-VQA [32]	43.14	59.44
FVQA (top-3-QQmaping) [32]	56.91	64.65
FVQA (Ensemble) [32]	58.76	-
Straight to the Facts (STTF) [35]	62.20	75.60
Reading Comprehension [228]	62.94	70.08
Out of the Box (OB) [34]	69.35	80.25
Mucko (w/o Semantic Graph) [125]	**71.28**	**82.76**
MCR-MemNN (Ours)	70.92	81.83

top-3 relation limitation (i.e., 'w/ Relation') or not (i.e., 'w/o Relation'). Finally, the downstream question answering accuracy ('QA Acc.') is also calculated.

For more clarity, the 'Sub' in the case of Top-40 represents the fraction of test samples that the ground-truth subject is included in the top 40 predicted subjects, and these subjects are further adopted as the clues for KB retrieval. The Recall w/ Relation in the case of Top-40 represents the fraction of test samples that the correct answer is included in the candidate answer set corresponding to the supporting facts retrieved by the top-40 subject or object clues and filtered by the top 3 predicted relations. Note that the top-3 classification accuracy for relation prediction using the relation phrase detector is 93.20%.

Results in Table 3.1 show that both union accuracy and answer recall of the Top-40 case are higher than those of the Top-20 case, and improvements of 10.56% and 6.09% on downstream QA accuracies (with and without relation filtering) are caused. Even though much more relevant supporting facts are retrieved from the external KB in the Top-100 case, which delivers higher both union accuracy and answer recall, the downstream QA accuracies are lower than those of the Top-40 case. This observation clearly shows that excessive retrieved facts can lead to more redundant information, which is harmful to the reasoning process. We choose the top 40 subjects and objects as clues for KB retrieval as this gives the best downstream QA accuracies.

MCR-MemNN Tables 3.2 and 3.3 show the comparison of the proposed MCR-MemNN with the above-mentioned baselines on FVQA and Visual7W+ConceptNet, respectively, and the following observations can be made:

1) On FVQA dataset, MCR-MemNN outperforms almost all the baselines in terms of top-1 and top-3 accuracies. To be specific, MCR-MemNN outperforms the semantic parsing based approaches including FVQA (top-3-QQmaping) and FVQA (Ensemble), and achieves more than 12% boost on top-1 accuracy and more than 15% boost on top-3 accuracy. In addition, compared with the typical learning based approaches including STTF and OB, which wholly introduce the visual and semantic information without selection, MCR-MemNN gains an improvement by leveraging a two-way attention mechanism (i.e., Sections 3.3.2 and 3.3.2) to exploit the inter-relationships among image, question and KB and distill the most relevant information in each of the three modalities.

2) Even though a heterogeneous graph neural network with high complexity is employed in Mucko, the proposed MCR-MemNN can still deliver a comparable performance. Note that the full model of Mucko leveraged dense captions [179] as input for performance improvement. For fair comparison, the semantic graph for dense caption parsing is removed.

3) On Visual7W+ConceptNet dataset, MCR-MemNN consistently outperforms a series of models based on KDMN, which leverages a dynamic memory network to preserve the retrieved external knowledge. Since both MCR-MemNN and KDMN adopted the memory augmentation technique, this observation further evidences the effectiveness of modeling inter-relationships among image, question and KB for KB-Image-QA task. Note that Mucko does not provide the result without Semantic Graph.

Ablation Studies To validate the superiority of the proposed MCR-MemNN on KB-Image-QA, several ablation experiments were conducted based on FVQA dataset, and we can make the following observations based on Table 3.4:

1) MCR-MemNN adopts both subject and object as clues for KB retrieval, which leads to better

Table 3.3: Accuracy comparison on Visual7W+ConceptNet.

Method	Overall Accuracy	
	top-1	top-3
KDMN-NoKnowledge [222]	45.1	-
KDMN-NoMemory [222]	51.9	-
KDMN [222]	57.9	-
KDMN-Ensemble [222]	60.9	-
Out of the Box (OB) [34]	57.32	71.61
MCR-MemNN (Ours)	**64.23**	**79.18**

performance. Specifically, compared with Case-1, where only subject is adopted as clues for KB retrieval, there exists a jump on top-1 (i.e., 3.44%) and top-3 (i.e., 5.40%) accuracies when both of the subject and object are leveraged as clues in Case-3. A similar gain can be observed when Case-3 is compared with Case-2.

2) To validate the effectiveness of the relation filtering, the experiment is conducted in Case-4, which achieves 2.48% improvement over Case-3 on top-1 accuracy. This observation clearly implies that the predicted relation can successfully remove the redundant supporting facts retrieved from the external KB. Similarly, compared with Case-5 without relation filtering, Case-6 brings up an improvement of 2.07%.

3) The two-way attention mechanism (i.e., Sections 3.3.2 and 3.3.2) can deliver an additional performance gain on KB-Image-QA task. For instance, compared with Case-4, where the inter-relationships are not exploited among image, question and KB, Case-6 brings up improvements of 3.40% on top-1 accuracy. This indicates that the redundant information of the three modalities (i.e., image, question and KB) is removed during the reasoning process, and the most relevant information is collected by modeling inter-relationships among the three modalities.

Qualitative Results Figure 3.4 shows several success and failure cases using MCR-MemNN. Two steps are indispensable to predict the correct answer: (1) Either the correct subject or object is included in the top-40 predicted subject set or object set. (2) The correct relation is predicted

Table 3.4: Ablation studies on FVQA. (Sub: Subject as Clue; Obj: Obj as Clue; Rel: Relation Filtering; Att: Two-way Attention)

Case	Sub	Obj	Rel	Att	Overall Accuracy top-1	Overall Accuracy top-3
1	✓				61.60	68.18
2		✓			44.07	52.72
3	✓	✓			65.04	73.58
4	✓	✓	✓		67.52	76.48
5	✓	✓		✓	68.85	78.37
6	✓	✓	✓	✓	**70.92**	**81.83**

as the top-3 relations. The ground-truth fact will be retrieved as one of the candidate supporting facts, if both the two steps are successfully accomplished. For instance, all five samples in the first row have their corresponding ground-truth facts successfully predicted and the correct answers are delivered. Specifically, the first two samples have both of their subjects and objects correctly predicted. The 3rd sample have its subject correctly predicted while the last two have their objects correctly predicted. This clearly verifies the advantage of multiple clues reasoning, retrieval using two complementary clues (i.e., subject and object) makes it a lot easier to fetch the ground-truth fact and deliver the correct answer. Some other cases are presented in the second row. Generally, if a wrong fact is predicted (e.g., the 2nd, 3rd and 5th samples), the correct answer cannot be given. However, in some special cases, even if the ground-truth fact is not successfully predicted (e.g., the 1st sample), the correct answer can still be delivered. If the correct relation is not included in the top-3 (e.g., the 3rd sample), the correct answer cannot be given even if the subject or object is correctly predicted. The MCR-MemNN always fails if the question is intend to know the relationship between subject and object (e.g., 4th sample).

3.5 Summary

This chapter, by introducing Multiple Clues for Reasoning with Memory Neural Network (MCR-MemNN), presents a novel framework for knowledge-based image question answering (KB-Image-

Fig. 3.4: Success and failure cases. (Terms in yellow/blue indicates the correctly predicted subjects/objects (in top-40), terms in green denotes the correctly predicted relations (in top-3) and terms underlined represents the predicted answer.)

QA). Comprehensive experiments have been conducted on two widely-used benchmarks and the extensive experimental results have shown that 1) Retrieval using two complementary clues (i.e., subject and object) makes it a lot easier to fetch the ground-truth fact and deliver the correct answer; 2) Two-way attention mechanism with memory augmentation can successfully model the inter-relationships among the three modalities including image, question and KB, and brings up a remarkable performance gain on KB-Image-QA; 3) MCR-MemNN outperforms most of the KB-Image-QA methods and achieves a comparable performance with the state-of-the-art.

CHAPTER 4
HIERARCHICAL GRAPH ATTENTION NETWORK FOR FEW-SHOT VISUAL-SEMANTIC LEARNING

4.1 Overview

In this chapter, to deal with the few-shot visual-semantic learning tasks, we propose the **H**ierarchical **G**raph **AT**tention network (**H-GAT**). This two-stage network is able to model the intra-modal relationships and the inter-modal relationships with a few image-text samples and can be extended to a semi-supervised setting. In the first stage, visual-specific and semantic-specific GNNs are leveraged to model the intra-relationship of images and texts (i.e., visual-specific relationships and semantic-specific relationships), respectively. To model the inter-relationship between the visual and semantic modalities, an attention-based co-learning framework is presented to guide the node feature update of these GNNs. In the second stage, relation-aware GNNs are used to predict the result of the query sample by jointly learning visual representations, semantic representations, visual-specific relationships and semantic-specific relationships. We perform extensive experiments on three widely-used benchmarks, Toronto COCO-QA [229], Visual Genome-QA [28] and

COCO-FITB [39], which showed that HGAT is a strong and effective model customized for few-shot visual-semantic learning.

The superiority of our proposed method can be summarized as follows: 1) It sheds light on tackling few-shot multi-modal learning problems, especially for few-shot visual-semantic learning, a fairly new but critical setting for human-level intelligence, through taking advantage of the intra- and inter-modal relationships. 2) Compared with FPAIT and several few-shot learning methods, it delivers state-of-the-art performance in terms of accuracy on both visual question answering and image captioning in the few-shot setting. 3) Several ablation experiments show the benefits of modeling of the visual-specific and semantic-specific relationships, the attention-based co-learning framework and the hierarchical graph-based architecture. 4) It can be easily extended to the semi-supervised setting and delivers better performance compared with the other two graph-based methods.

4.2 Problem Statement

The general visual-semantic learning aims to build models that process and relate information from multiple modalities [2]. We focus primarily, but not exclusively, on the visual and semantic modalities, and study the visual-semantic learning problem by tackling the image question answering (Image-QA) and image captioning (IC) tasks. For Image-QA, given an image **I** and a related question **Q**, we need to generate a corresponding answer **A**. For IC, we follow the fill-in-the-blank setting [39], attempting to fill in the blank **A** of a given description **Q** for an image **I**. Note that both the question/description **Q** and the answer/blank **A** are represented in a natural language format. Regularly, **A** is picked from a pre-defined set of different answers/labels. The traditional Image-QA and IC tasks seek a model **F**, which can be a neural network, to map the observations $\mathbf{I}, \mathbf{Q}$ to the output **A**.

In few-shot learning, given only a few training samples, the model is expected to be able to adapt to a new task quickly. N-way K-shot problem settings are usually used to measure few-

shot learning methods. Take an N-way K-shot VQA/IC task $\mathcal{T}$ with M queries as an example: $\mathcal{T}$ consists of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$, on which the model is learnt and evaluated respectively. $\mathcal{S}$ is a set of $N \times K$ samples, containing K labeled image-text pairs for each of N unique answers. $\mathcal{Q}$ contains another M samples with the same answers as those in $\mathcal{S}$. Formally speaking, $\mathcal{T} = \mathcal{S} \cup \mathcal{Q}$, where $\mathcal{S} = \{(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)\}_{i=1}^{N\times K}$ and $\mathcal{Q} = \{(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)\}_{i=N\times K+1}^{N\times K+M}$; The label space of task $\mathcal{T}$ is defined as $\mathcal{C}_\mathcal{T} = \{\mathbf{A}_i\}_{i=1}^{N\times K}$. We have $\mathbf{A}_{(n-1)\times K+i} = \mathbf{A}_{(n-1)\times K+j}$ for $n = 1, \cdots, N$ and $1 \leq i, j \leq K$ (i.e., $|\mathcal{C}_\mathcal{T}| = N$), with $\{\mathbf{A}_i\}_{i=N\times K+1}^{N\times K+M} \subset \mathcal{C}_\mathcal{T}$.

In this work, we use meta-learning [40] to define few-shot visual-semantic learning problems, and it generally consists of two phases, meta-training and meta-testing. During the meta-training, a set of T tasks $\{\mathcal{T}_t\}_{t=1}^{T}$ are generated from a meta-training dataset $\mathcal{D}_{\text{mtr}}$, and we develop a method that takes as input the support sets $\{\mathcal{S}_t\}_{t=1}^{T}$ and returns a model which minimizes the loss over the corresponding query sets $\{\mathcal{Q}_t\}_{t=1}^{T}$. During the meta-testing, another set of T' tasks $\{\mathcal{T}_{T+t}\}_{t=1}^{T'}$ are generated from a meta-testing dataset $\mathcal{D}_{\text{mte}}$, and for $t = 1, \cdots, T'$, we expect the trained model can learn quickly from the $N \times K$ labeled image-text samples in the support set $\mathcal{S}_{T+t}$ and deliver highly-accurate labels for samples from the query set $\mathcal{Q}_{T+t}$. Note that the labels used in meta-training and meta-testing are mutually exclusive, i.e., $\mathcal{C}_{\text{mtr}} \cap \mathcal{C}_{\text{mte}} = \emptyset$ where $\mathcal{C}_{\text{mtr}} = \bigcup_{1\leq t\leq T} \mathcal{C}_{\mathcal{T}_t}$ and $\mathcal{C}_{\text{mte}} = \bigcup_{1\leq t\leq T'} \mathcal{C}_{\mathcal{T}_{T+t}}$.

Additionally, the problem can be extended to semi-supervised learning if a portion of labels in all support sets $\{\mathcal{S}_t\}_{t=1}^{T+T'}$ are unknown. In Section 4.4.3, the effectiveness of our model on semi-supervised setting will be presented.

4.3 Hierarchical Graph Attention Neural Network

4.3.1 Image and Text Embedding

To capture and preserve useful visual and semantic representations, we resort to modality-specific deep networks on image and text inputs. For image embedding of an image $\mathbf{I}_i$, we build a neural network ϕ to output visual representation $\phi(\mathbf{I}_i; \boldsymbol{\theta}_\phi)$ by following the architecture used by

FPAIT [39], which contains four 3×3 convolutional blocks with batch normalizations and ReLU activations. The numbers of feature channels in these four blocks are 64, 96, 128 and 256, respectively. There is a 2×2 max-pooling layer after each of the first three blocks, and a global-pooling layer after the last block. The output image embedding has a dimension of 256. For text embedding of a question/description $\mathbf{Q}_i$, we build a neural network ψ to output the semantic representation $\psi(\mathbf{Q}_i; \boldsymbol{\theta}_\psi)$ as follows. We first transfer each word in $\mathbf{Q}_i$ into a feature vector using the pre-trained $100D$ GloVe [214] vectors, and use randomly initialized weights for all words which are out of GloVe's vocabulary, and then use the temporal convolutional network (TCN) [230] to obtain the embedding of the word sequence. Same as FPAIT [39], the output text embedding has a dimension of 512. The neural networks for both image embedding and text embedding, i.e., $\phi(\cdot; \boldsymbol{\theta}_\phi)$ and $\psi(\cdot; \boldsymbol{\theta}_\psi)$, are jointly trained with other modules of HGAT.

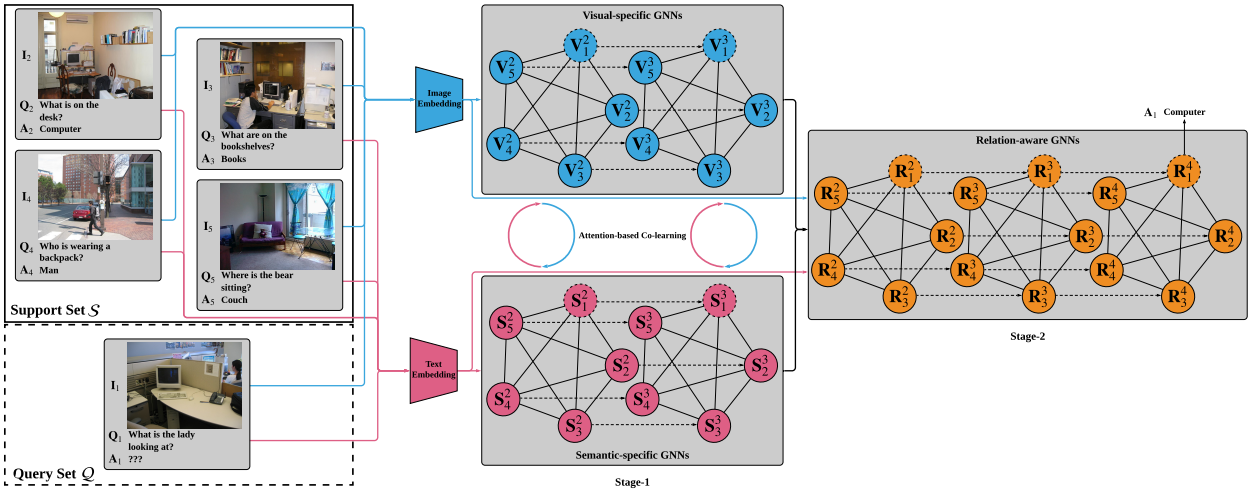

Fig. 4.1: The architecture of Hierarchical Graph Attention Network. A 4-way 1-shot problem with one query sample ($N = 4$, $K = 1$, $M = 1$) is presented for simplicity. GNN nodes with solid line (Nodes 2-5) correspond to the samples from the support set $\mathcal{S}$, and nodes with dashed line (Node 1) correspond to the samples from the query set $\mathcal{Q}$. Dotted arrows between GNN layers represent node inheritances.

4.3.2 Graph Construction

For each task $\mathcal{T}$, given the visual and semantic representations (extracted from the image and text embedding neural networks respectively) of all the image-text samples, we construct two graphs,

the visual-specific GNNs (with blue nodes in Figures 4.1 and 4.2) and the semantic-specific GNNs (with red nodes), respectively. As shown in Figure 4.1, in Stage-1 of HGAT, both of the visual-specific and semantic-specific GNNs are two-layer GNNs. Each GNN layer contains $N \times K + M$ fully-connected nodes, and each node corresponds to an image-text sample from either the support set or the query set.

For each sample $(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)$, the feature vector of its corresponding node in the first layer of GNNs ($\mathbf{V}_i^1$ and $\mathbf{S}_i^1$) is initialized as the concatenation of its visual or semantic representation and the one-hot encoding of its label.

$$\mathbf{V}_i^1 = [\phi(\mathbf{I}_i; \boldsymbol{\theta}_\phi)||h(\mathbf{A}_i)] \tag{4.1}$$

$$\mathbf{S}_i^1 = [\psi(\mathbf{Q}_i; \boldsymbol{\theta}_\psi)||h(\mathbf{A}_i)] \tag{4.2}$$

where $||$ denotes vector concatenation operation, and $h(\mathbf{A}_i) \in [0, 1]^N$ represents the one-hot encoding of the label $\mathbf{A}_i$. For any image-text sample $(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)$ from the query set $\mathcal{Q}$ or with unknown labels in the semi-supervised setting, we set $h(\mathbf{A}_i)$ to be a zero vector $\mathbf{0}^N$ instead.

For each node in the lth layer ($l > 1$), its feature vector is a concatenation of features inherited from its corresponding node in previous layer ($\mathbf{V}_i^{l-1}$ or $\mathbf{S}_i^{l-1}$) and an updated feature vector ($\mathbf{V}_i^l$ or $\mathbf{S}_i^l$) computed via the attention-based co-learning described in the following section.

4.3.3 Attention-based Co-learning Framework

Each layer of the two modal-specific GNNs conducts associated node feature update in the proposed attention-based co-learning framework. For the node feature update in the lth layer ($l = 1, 2$), the inputs are two sets of nodes $\{\mathbf{V}_i^l\}_{i=1}^{N\times K+M}$, $\mathbf{V}_i^l \in \mathcal{R}^{F_V^l}$ and $\{\mathbf{S}_i^l\}_{i=1}^{N\times K+M}$, $\mathbf{S}_i^l \in \mathcal{R}^{F_S^l}$, and the outputs are two updated sets of nodes $\{\mathbf{V}_i^{l+1}\}_{i=1}^{N\times K+M}$, $\mathbf{V}_i^{l+1} \in \mathcal{R}^{2F_V^{l'}}$ and $\{\mathbf{S}_i^{l+1}\}_{i=1}^{N\times K+M}$, $\mathbf{S}_i^{l+1} \in \mathcal{R}^{2F_S^{l'}}$, where F_V^l, $2F_V^{l'}$, F_S^l, and $2F_S^{l'}$ represent the number of input and output feature channels of each node in the two modal-specific GNNs, respectively.

As an initial step, two shared learnable linear transformations, parametrized by $\mathbf{W}_V^l \in \mathcal{R}^{F_V^{l'}\times F_V^l}$

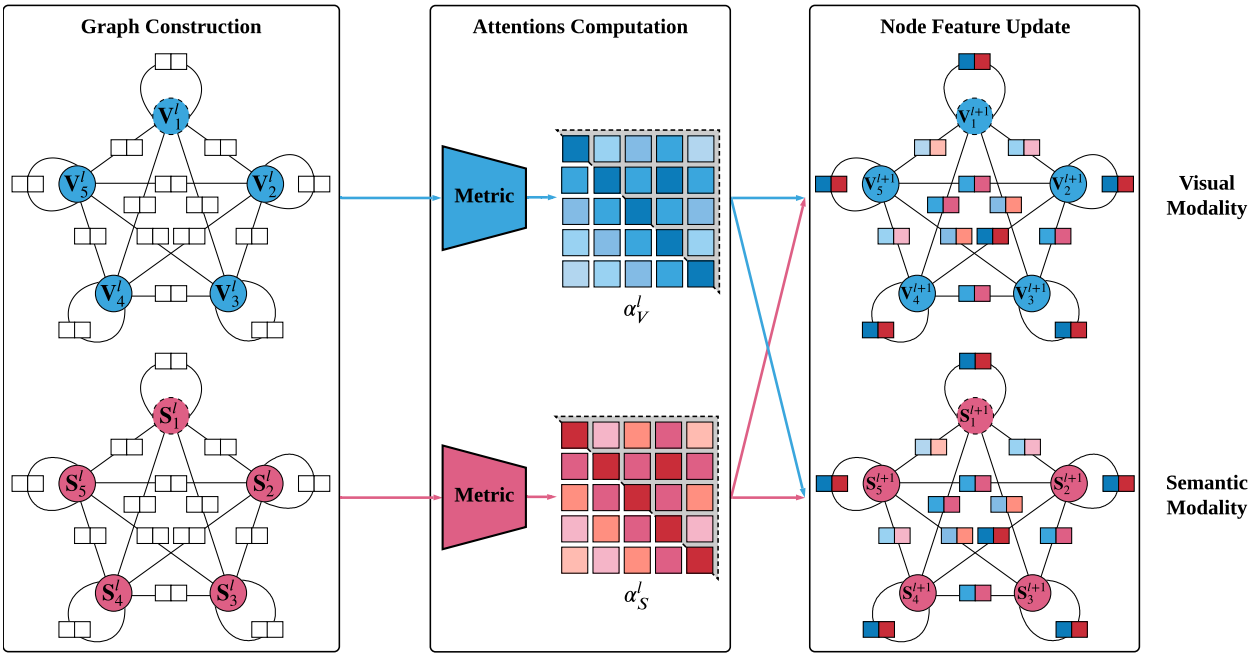

Fig. 4.2: An illustration of the attention-based co-learning framework for one GNN layer. For simplicity, a 4-way 1-shot problem with one query sample ($N = 4$, $K = 1$, $M = 1$) is presented as an example. Nodes with solid line (Nodes 2-5) represent the samples from the support set $\mathcal{S}$, and nodes with dashed line (Node 1) represent the samples from the query set $\mathcal{Q}$. A two-dimensional attention is computed for each pair of nodes, to capture their relationship of the visual and semantic modalities, respectively. For simplicity, only half of the attentions (within the dotted triangles) are depicted in the node feature update.

and $\mathbf{W}_S^l \in \mathcal{R}^{F_S^{l'} \times F_S^l}$, are applied to the two sets of nodes. Then, for each modal-specific GNN layer, a shared attentional mechanism a is performed for each pair of nodes to compute the attention coefficients $e_{V_{ij}}^l \in \mathcal{R}$ and $e_{S_{ij}}^l \in \mathcal{R}$.

$$e_{V_{ij}}^l = a(\mathbf{W}_V^l \mathbf{V}_i^l, \mathbf{W}_V^l \mathbf{V}_j^l) = \text{LReLU}\left(\mathbf{a}_V^{l\top}\left[\mathbf{W}_V^l \mathbf{V}_i^l || \mathbf{W}_V^l \mathbf{V}_j^l\right]\right) \tag{4.3}$$

$$e_{S_{ij}}^l = a(\mathbf{W}_S^l \mathbf{S}_i^l, \mathbf{W}_S^l \mathbf{S}_j^l) = \text{LReLU}\left(\mathbf{a}_S^{l\top}\left[\mathbf{W}_S^l \mathbf{S}_i^l || \mathbf{W}_S^l \mathbf{S}_j^l\right]\right) \tag{4.4}$$

where $e_{V_{ij}}^l$ and $e_{S_{ij}}^l$ indicate the importance of node $\mathbf{V}_j^l$ to node $\mathbf{V}_i^l$ in the visual-specific GNN and that of node $\mathbf{S}_j^l$ to node $\mathbf{S}_i^l$ in the semantic-specific GNN, respectively. LReLU denotes the Leaky Rectified Linear Unit [231] function. Both $\mathbf{a}_V^l \in \mathcal{R}^{2F_V^{l'}}$ and $\mathbf{a}_S^l \in \mathcal{R}^{2F_S^{l'}}$ serve as learnable weight vectors, and $\mathbf{a}^\top$ represents the transpose of $\mathbf{a}$. The attentions $\alpha_{V_{ij}}^l \in \mathcal{R}$ and $\alpha_{S_{ij}}^l \in \mathcal{R}$ are obtained

by normalizing the attention coefficients using the softmax function.

$$\alpha^l_{V_{ij}} = \text{softmax}(e^l_{V_{ij}}) = \frac{\exp(e^l_{V_{ij}})}{\sum_{k=1}^{N\times K+M} \exp(e^l_{V_{ik}})} \tag{4.5}$$

$$\alpha^l_{S_{ij}} = \text{softmax}(e^l_{S_{ij}}) = \frac{\exp(e^l_{S_{ij}})}{\sum_{k=1}^{N\times K+M} \exp(e^l_{S_{ik}})} \tag{4.6}$$

where $\alpha^l_{V_{ij}}$ and $\alpha^l_{S_{ij}}$ represent the attentions of the visual modality (blue-hue square matrix in Figure 4.2) and those of the semantic modality (red-hue square matrix), respectively. The values of $\alpha^l_{V_{ij}}$ and $\alpha^l_{S_{ij}}$ are expected to be positively correlated.

Once obtained, both attentions are shared by the associated node feature update of the two modal-specific GNNs. For example, the attentions of visual modality not only serve for the node feature update of the visual-specific GNNs but also utilize the relationship on visual modality to refine the semantic-specific GNNs. Analogously, the attentions of semantic modality are also helpful to both semantic-specific and visual-specific GNNs.

$$\mathbf{V}^{l+1}_i = \text{ELU}\left(\sum_{j=1}^{N\times K+M} \alpha^l_{V_{ij}} \mathbf{W}^l_V \mathbf{V}^l_j || \sum_{j=1}^{N\times K+M} \alpha^l_{S_{ij}} \mathbf{W}^l_V \mathbf{V}^l_j\right) \tag{4.7}$$

$$\mathbf{S}^{l+1}_i = \text{ELU}\left(\sum_{j=1}^{N\times K+M} \alpha^l_{V_{ij}} \mathbf{W}^l_S \mathbf{S}^l_j || \sum_{j=1}^{N\times K+M} \alpha^l_{S_{ij}} \mathbf{W}^l_S \mathbf{S}^l_j\right) \tag{4.8}$$

where $\alpha^l_{V_{ij}}$ and $\alpha^l_{S_{ij}}$ form the two-dimensional attention between each pair of nodes in both modal-specific GNNs and ELU denotes the Exponential Linear Unit [232] function.

Based on the attentions shared by the visual and semantic modalities, the associated node feature update is conducted, and the inter-modal relationships are modeled under the attention-based co-learning framework. Note that while the basic attention mechanism used here follows the Graph Attention Network [92], our proposed attention-based co-learning framework is agnostic to the particular choice of attention mechanism.

4.3.4 Relation-aware GNNs

The layerwise outputs of the modal-specific GNNs, i.e., $\{\mathbf{V}_i^{l+1}\}_{i=1}^{N\times K+M}$ and $\{\mathbf{S}_i^{l+1}\}_{i=1}^{N\times K+M}$ for $l = 1, 2$, extracting the hierarchical features with intra-relationship and inter-relationship from both the visual and semantic modalities, are further exploited by the relation-aware GNNs in Stage-2 of HGAT.

We construct the relation-aware GNNs with $N \times K + M$ nodes in each layer, which share similar structure with the modal-specific GNNs but also take the relationships obtained in Stage-1 for the feature initialization of each node. To be more specific, for the node feature update in the lth layer ($l = 1, 2, 3$), the inputs is a set of nodes $\{\mathbf{R}_i^l\}_{i=1}^{N\times K+M}$, $\mathbf{R}_i^l \in \mathcal{R}^{F_R^l}$ and the outputs are an updated set of nodes $\{\mathbf{R}_i^{l+1}\}_{i=1}^{N\times K+M}$, $\mathbf{R}_i^{l+1} \in \mathcal{R}^{F_R^{l'}}$ where F_R^l, $F_R^{l'}$ represent the number of input and output feature channels of each node in the relation-aware GNNs, respectively. The input to the first layer $\mathbf{R}_i^1$ is the concatenation of the visual and semantic embeddings, the one-hot encoding of the label, and the multi-modal features obtained in Stage-1.

$$\mathbf{R}_i^1 = \left[\phi(\mathbf{I}_i;\boldsymbol{\theta}_\phi)||\psi(\mathbf{Q}_i;\boldsymbol{\theta}_\psi)||h(\mathbf{A}_i)||\mathbf{V}_i^2||\mathbf{V}_i^3||\mathbf{S}_i^2||\mathbf{S}_i^3\right] \tag{4.9}$$

The input to the lth layer ($l > 1$) is a concatenation of features inherited from its corresponding node in previous layer $\mathbf{R}_i^{l-1}$ and an updated feature vector $\mathbf{R}_i^l$, which is computed in a similar way to the modal-specific GNNs. First, the attention coefficient $e_{R_{ij}}^l \in \mathcal{R}$ indicating the importance of node $\mathbf{R}_j^l$ to node $\mathbf{R}_i^l$ is calculated.

$$e_{R_{ij}^l} = a(\mathbf{W}_R^l\mathbf{R}_i^l, \mathbf{W}_R^l\mathbf{R}_j^l) = \text{LReLU}\left(\mathbf{a}_R^{l\top}\left[\mathbf{W}_R^l\mathbf{R}_i^l||\mathbf{W}_R^l\mathbf{R}_j^l\right]\right) \tag{4.10}$$

where $\mathbf{W}_R^l \in \mathcal{R}^{F_R^{l'}\times F_R^l}$ and $\mathbf{a}_R^l \in \mathcal{R}^{2F_R^{l'}}$ are learnable parameters. Then the attentions are computed by normalizing the attention coefficients using softmax function.

$$\alpha_{R_{ij}}^l = \text{softmax}(e_{R_{ij}}^l) = \frac{\exp(e_{R_{ij}}^l)}{\sum_{k=1}^{N\times K+M}\exp(e_{R_{ik}}^l)} \tag{4.11}$$

Afterwards, the attentions are used to compute the updated node features through a linear combination of the corresponding features, followed by a non-linearity activation.

$$\mathbf{R}_i^{l+1} = \text{ELU}\left(\sum_{j=1}^{N\times K+M} \alpha_{R_{ij}}^l \mathbf{W}_R^l \mathbf{R}_j^l\right) \tag{4.12}$$

Finally, to get the final prediction of the ith sample from HGAT, we set the last output dimension $F_R^{3'}$ to N, and use softmax$(\mathbf{R}_i^4) \in [0,1]^N$ as the confidence score vector over the N answers. The predicted label is $\hat{\mathbf{A}}_i = \text{argmax}_n \mathbf{R}_{i,n}^4$, where $\mathbf{R}_{i,n}^4$ is the nth element of $\mathbf{R}_i^4$ and $1 \leq n \leq N$.

4.3.5 Meta-Training & Meta-Testing

Given a set of T tasks in the meta-training phase, the learnable parameters of the proposed HGAT, $\boldsymbol{\theta}_\phi \cup \boldsymbol{\theta}_\psi \cup \{\mathbf{W}_V^l, \mathbf{W}_S^l, \mathbf{a}_V^l, \mathbf{a}_S^l\}_{l=1}^2 \cup \{\mathbf{W}_R^l, \mathbf{a}_R^l\}_{l=1}^3$, are trained in an end-to-end manner by minimizing the following loss function over the task set.

$$\mathcal{L} = \frac{1}{TM} \sum_{\mathcal{T}\in\{\mathcal{T}_t\}_{t=1}^T} \sum_{i=N\times K+1}^{N\times K+M} \mathcal{L}_c(\mathbf{A}_i, \hat{\mathbf{A}}_i) \tag{4.13}$$

where $\mathcal{L}_c$ is defined as the cross-entropy loss, $\mathbf{A}_i$ and $\hat{\mathbf{A}}_i$ represent the ground truth answer and the predicted answer of the image-text samples from the query set $\mathcal{Q}$.

The overall meta-training and meta-testing algorithms for HGAT are summarized in Algorithms 1 and 2.

Algorithm 1: The process of meta-training for HGAT

Input: A set of T tasks $\{\mathcal{T}_t\}_{t=1}^{T}$ generated from a meta-training dataset $\mathcal{D}_{\text{mtr}}$, where $\{\mathcal{T}_t\}_{t=1}^{T} = \{\mathcal{S}_t\}_{t=1}^{T} \cup \{\mathcal{Q}_t\}_{t=1}^{T}$
Initialize $\boldsymbol{\theta}_\phi \cup \boldsymbol{\theta}_\psi \cup \{\mathbf{W}_V^l, \mathbf{W}_S^l, \mathbf{a}_V^l, \mathbf{a}_S^l\}_{l=1}^{2} \cup \{\mathbf{W}_R^l, \mathbf{a}_R^l\}_{l=1}^{3}$
while not done **do**
  Sample batch of tasks $< \mathcal{T}_t >$ from task set $\{\mathcal{T}_t\}_{t=1}^{T}$
  for all $\mathcal{T}_t$ **do**
    $\{\mathbf{V}_i^1\}, \{\mathbf{S}_i^1\} \leftarrow$ GraphConstruct$(\{(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)\}; \{\boldsymbol{\theta}_\phi, \boldsymbol{\theta}_\psi\})$, $\forall i$
    for all $l = 1, 2$ **do**
      $\{\mathbf{V}_i^{l+1}\}, \{\mathbf{S}_i^{l+1}\} \leftarrow$ NodeFeatureUpdate$(\{\mathbf{V}_i^l\}, \{\mathbf{S}_i^l\}; \{\mathbf{W}_V^l, \mathbf{W}_S^l, \mathbf{a}_V^l, \mathbf{a}_S^l\})$, $\forall i$
    end for
    Initialize $\mathbf{R}_i^1 \leftarrow$ GraphConstruct$(\{(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)\}, \{\mathbf{V}_i^2\}, \{\mathbf{S}_i^2\}, \{\mathbf{V}_i^3\}, \{\mathbf{S}_i^3\}; \{\boldsymbol{\theta}_\phi, \boldsymbol{\theta}_\psi\})$, $\forall i$
    for all $l = 1, 2, 3$ **do**
      $\{\mathbf{R}_i^{l+1}\} \leftarrow$ NodeFeatureUpdate$(\{\mathbf{R}_i^l\}; \{\mathbf{W}_R^l, \mathbf{a}_R^l\})$, $\forall i$
    end for
    Compute the loss and update $\boldsymbol{\theta}_\phi \cup \boldsymbol{\theta}_\psi \cup \{\mathbf{W}_V^l, \mathbf{W}_S^l, \mathbf{a}_V^l, \mathbf{a}_S^l\}_{l=1}^{2} \cup \{\mathbf{W}_R^l, \mathbf{a}_R^l\}_{l=1}^{3}$
  end for
end while

Algorithm 2: The process of meta-testing for HGAT

Input: A task $\mathcal{T}$ sampled from $\{\mathcal{T}_{T+t}\}_{t=1}^{T'}$, where $\mathcal{T} = \mathcal{S} \cup \mathcal{Q}$, $\mathcal{S} = \{(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)\}_{i=1}^{N\times K}$, $\mathcal{Q} = \{(\mathbf{I}_i, \mathbf{Q}_i)\}_{i=N\times K+1}^{N\times K+M}$
Parameters: $\boldsymbol{\theta}_\phi \cup \boldsymbol{\theta}_\psi \cup \{\mathbf{W}_V^l, \mathbf{W}_S^l, \mathbf{a}_V^l, \mathbf{a}_S^l\}_{l=1}^2 \cup \{\mathbf{W}_R^l, \mathbf{a}_R^l\}_{l=1}^3$
Output: $\{\hat{\mathbf{A}}_i\}_{i=N\times K+1}^{N\times K+M}$
$\{\mathbf{V}_i^1\}, \{\mathbf{S}_i^1\} \leftarrow$ GraphConstruct$(\{(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)\}; \{\boldsymbol{\theta}_\phi, \boldsymbol{\theta}_\psi\}), \forall i$
for all $l = 1, 2$ **do**
    $\{\mathbf{V}_i^{l+1}\}, \{\mathbf{S}_i^{l+1}\} \leftarrow$ NodeFeatureUpdate$(\{\mathbf{V}_i^l\}, \{\mathbf{S}_i^l\}; \{\mathbf{W}_V^l, \mathbf{W}_S^l, \mathbf{a}_V^l, \mathbf{a}_S^l\}), \forall i$
end for
Initialize $\mathbf{R}_i^1 \leftarrow$ GraphConstruct$(\{(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)\}, \{\mathbf{V}_i^2\}, \{\mathbf{S}_i^2\}, \{\mathbf{V}_i^3\}, \{\mathbf{S}_i^3\}; \{\boldsymbol{\theta}_\phi, \boldsymbol{\theta}_\psi\}), \forall i$
for all $l = 1, 2, 3$ **do**
    $\{\mathbf{R}_i^{l+1}\} \leftarrow$ NodeFeatureUpdate$(\{\mathbf{R}_i^l\}; \{\mathbf{W}_R^l, \mathbf{a}_R^l\}), \forall i$
end for
$\{\hat{\mathbf{A}}_i\}_{i=N\times K+1}^{N\times K+M} \leftarrow$ QueryLabelPredict$(\{\mathbf{R}_i^4\}_{i=N\times K+1}^{N\times K+M})$

4.4 Performance Evaluation

We employed three compelling benchmarks, Toronto COCO-QA [229], Visual Genome-QA [28] and COCO-FITB [39] to evaluate the proposed HGAT on two typical visual-semantic learning tasks, image question answering (Image-QA) and image captioning (IC).

Benchmark		Toronto COCO-QA	Visual Genome-QA	COCO-FITB
Task		Image-QA	Image-QA	IC
#Pair	Meta-training	57,834	554,795	181,844
	Meta-testing	13,965	136,473	34,919
#Class	Meta-training	256	244	159
	Meta-testing	65	82	43

Table 4.1: Statistics on the three benchmark datasets.

4.4.1 Benchmark Datasets

Table 4.1 shows the statistics on the three benchmark datasets for few-shot Image-QA and few-shot IC tasks.

Toronto COCO-QA consists of $78,736$ question-answer (QA) pairs for training and $38,948$ QA pairs for testing. Each QA pair associates with one image from MSCOCO [233] and is labeled with one of the four QA types (i.e., object, number, color and location). Following the same pre-processing steps of FPAIT [39], the Toronto COCO-QA is transformed into the format which can be used for the few-shot Image-QA. Consequently, $57,834$ QA pairs with a set of 256 unique answers are used in the meta-training phase, and $13,965$ QA pairs with a set of 65 unique answers are used in the meta-testing phase. The two answer sets are mutually exclusive.

Visual Genome-QA, the largest dataset for Image-QA, contains over 1.7 million QA pairs with more than $100,000$ images sampled from MSCOCO [233]. Compared with Toronto COCO-QA, more categories of questions, which may start with the "who", "what", "where", "when", "why", "how" and "which", are provided. The Visual Genome-QA is transformed into the format for few-shot Image-QA with similar pre-processing steps for Toronto COCO-QA. Finally, there are $554,795$ QA pairs with a set of 244 unique answers for meta-training, and $136,473$ QA pairs for meta-testing with a set of 82 unique answers. The two answer sets are mutually exclusive.

COCO-FITB, proposed and used by FPAIT [39] for few-shot IC, is transformed from MSCOCO [233] by processing MSCOCO Captions [234] to generate image-caption pairs in the fill-in-the-blank format. $181,844$ image-caption pairs with a set of 159 unique blank words are used in meta-training, and $34,919$ image-caption pairs with a set of 43 unique blank words are used in meta-testing. The two sets of blank words are mutually exclusive.

4.4.2 Experimental Setup

Few-shot setup Following the common setup in few-shot learning [40, 42], for each task $\mathcal{T}$ of N-way K-shot learning, we set $N \in \{5, 10\}$, $K \in \{1, 5\}$ and $M = 1$. Take a 10-way 5-shot Image-QA task for example: given 10 different answers, each answer has 5 labeled image-question pairs, and these 50 samples serve as the support set to predict the result out of the 10 answers for the 1 unlabeled image-question pair from the query set. Therefore, we can evaluate both Image-QA and IC tasks in terms of the standard classification accuracy.

Implementation details In the meta-training phase, the proposed model was trained with Adam optimizer [226] with an initial learning rate of 1×10^{-3} and weight decay of 1×10^{-6}. The task mini-batch sizes were set to 128, 32, 64, and 16 for 5-way 1-shot, 5-way 5-shot, 10-way 1-shot, and 10-way 5-shot, respectively. Our code was implemented in PyTorch [227] and run with NVIDIA Tesla P100 GPUs.

Baselines FPAIT [39] directly leverages MAML [40] to deal with few-shot Image-QA and IC tasks; Prototypical Net [42], Relation Net [129], R2D2 [136], and DN4 [130] focus on few-shot classification. GNN [235] and EGNN [236] are two GNN-based few-shot classification models. None of these algorithms, including MAML, has paid any attention to the few-shot visual-semantic learning, but it is noting that all of them can be extended to tackle few-shot Image-QA and IC as few-shot classification tasks.

Baselines Implementation We re-implemented the Prototypical Net, Relation Net, R2D2, DN4, GNN, and EGNN and extended these algorithms from few-shot classification to few-shot visual-semantic learning. For Prototypical Net, Relation Net, R2D2 and DN4, we adopted the concatenation of the corresponding visual and semantic representations as input feature for each sample in the support/training sets and query/test sets. For GNN, each sample from the support or query sets is represented as a node in the first layer of GNNs, and each node is initialized as the concatenation of its visual and semantic representations as well as the one-hot encoding of its label. Note that for the unknown labels (e.g., query samples), unlike Garcia *et al.* [235], the one-hot encoding is initialized to a zero vector instead of a uniform vector. For EGNN, both of the node features and the edge features need to be initialized. The node features are initialized as the concatenation of its visual and semantic representations. Following the definition in Kim *et al.* [236], each edge feature is a 2-dimensional vector with a value representing the relationship between its two connected nodes. If the two connected nodes belong to one class, the edge feature is set to $[1\ 0]$; otherwise, it is set to $[0\ 1]$. In addition, all the edges connected to the samples with unknown labels are set to $[0.5\ 0.5]$. It is worth noting that for both of the GNN and EGNN, only

Method	Toronto COCO-QA				Visual Genome-QA				COCO-FITB			
	5-way accuracy		10-way accuracy		5-way accuracy		10-way accuracy		5-way accuracy		10-way accuracy	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
FPAIT	59.38	71.92	45.11	60.20	75.49	79.12	61.66	67.62	60.13	70.88	47.10	59.31
FPAIT+CLT	60.61	72.17	46.37	60.92	75.05	79.28	60.82	67.48	61.01	71.13	47.79	60.91
Prototypical Net	60.12	71.72	45.31	59.67	75.43	80.33	62.32	67.23	60.56	71.16	47.52	59.38
Relation Net	61.75	71.89	45.60	60.13	77.21	80.72	63.14	68.10	61.35	71.68	47.92	59.55
R2D2	61.83	72.60	47.13	59.36	77.44	81.08	64.71	71.55	60.87	71.60	47.73	59.33
DN4	62.60	74.12	47.68	60.44	78.33	84.25	64.92	71.20	62.09	73.62	48.57	60.82
GNN	61.42	72.55	46.35	58.95	76.72	81.43	63.19	68.65	61.85	72.70	48.14	59.86
EGNN	62.21	73.41	46.99	60.01	77.67	83.26	64.07	70.87	62.67	72.98	48.22	60.13
HGAT	**63.13**	**75.41**	**48.10**	**61.50**	**79.56**	**86.10**	**66.62**	**72.13**	**63.36**	**74.14**	**49.26**	**61.31**

Table 4.2: Comparison of accuracy on Toronto COCO-QA, Visual Genome-QA, and COCO-FITB.

one query node exists in each layer of GNNs.

4.4.3 Experimental Results

Results on the three benchmarks are shown in Table 4.2, and we can make the following observations:

1) HGAT outperforms all baselines in terms of classification accuracy in all settings. Concretely, in the case of 5-way 5-shot Image-QA on Toronto COCO-QA, HGAT gives an accuracy of 75.41%, excelling the second best by 1.29 percentage points, which indicates that the modeling of the intra- and inter-modal relationships using the hierarchical graph-based structure can lead to consistent advantages on few-shot visual-semantic learning. Similar trends can be observed for other test cases and benchmark datasets.

2) Generally, the graph-based methods (e.g., GNN, EGNN and HGAT) perform better than non-graph-based few-shot learning algorithms, such as FPAIT, Prototypical Net, Relation Net, R2D2 and DN4. For instance, in the case of 5-way 1-shot IC on COCO-FITB, GNN, EGNN and HGAT achieve higher classification accuracies of 61.85%, 62.67% and 63.36%, respectively, than those of FPAIT (60.13%), Prototypical Net (60.56%), Relation Net (61.35%), R2D2 (60.87%) and DN4 (62.09%). This observation clearly evidences that relationship modeling using the graph-based structure is crucial for cases with limited data on visual-semantic learning.

3) Among the graph-based methods, our HGAT brings noticeable improvements over GNN

and EGNN. For example, when few-shot Image-QA is conducted on Visual Genome-QA, HGAT obtains classification accuracies of 79.56%, 86.10%, 66.62% and 72.13% in the four test cases, respectively, which are 1.89%, 2.84%, 2.55% and 1.26% higher than those of EGNN. Similar improvements can be observed regarding GNN. Although GNN and EGNN utilize the pairwise relationships of nodes, the intra-relationship of each modality, as well as the inter-relationship between different modalities, have not been fully exploited.

To further justify the superiority of HGAT for the few-shot visual-semantic learning tasks, experimental comparisons have been expanded to standard Image-QA and IC methods that are not specifically designed for few-shot learning, including HCA [24], SAAA [237], and CNN+TCN [39]. Since Dong *et al.* [39] compared on COCO-QA and COCO-FITB, we followed its setting and strategies for a fair comparison. Results in Table 4.3 show that the HGAT is more suitable and outperforms standard methods by a significant margin.

Method	COCO-QA		COCO-FITB	
	5-way accuracy		5-way accuracy	
	1-shot	5-shot	1-shot	5-shot
HCA	55.40	66.78	54.33	62.91
SAAA	56.72	67.23	55.67	64.16
CNN+TCN	57.19	71.82	59.95	70.32
HGAT	**63.13**	**75.41**	**63.36**	**74.14**

Table 4.3: Comparison with standard Image-QA and IC methods.

4.4.4 Ablation Study

To validate the superiority of the proposed HGAT, several ablation experiments were conducted based on Visual Genome-QA for few-shot Image-QA. The following observations are made based on Table 4.4:

1) HGAT conducts separate exploitation of the intra-relationship of each modality, which can lead to better performance. Compared with Case-1, where no intra-modal relationships are exploited, there exists a jump on accuracy when the visual-specific relationships are modeled in

Case	St1	St2	Vis	Sem	Att	5-way accuracy		10-way accuracy	
						1-shot	5-shot	1-shot	5-shot
1		✓				76.10	82.14	63.99	66.30
2	✓		✓			75.33	81.67	62.08	66.42
3	✓			✓		75.84	80.25	62.64	65.76
4	✓		✓	✓		76.78	82.32	64.16	68.22
5	✓		✓	✓	✓	77.47	83.26	64.90	70.03
6	✓	✓	✓			77.55	84.01	63.77	69.26
7	✓	✓		✓		78.14	83.88	64.23	68.61
8	✓	✓	✓	✓		78.86	84.55	65.21	70.06
9	✓	✓	✓	✓	✓	**79.56**	**86.10**	**66.62**	**72.13**

Table 4.4: Ablation study on Visual Genome-QA for few-shot Image-QA. (St1: Stage-1; St2: Stage-2; Vis: Visual Relations; Sem: Semantic Relations; Att: Attention-based Co-learning) Based on the model only with the three-layer relation-aware GNNs in Stage-2 (Case-1), we gradually add visual-specific GNNs (Case-6), semantic-specific GNNs (Case-7), both visual-specific GNNs and semantic-specific GNNs (Case-8) and attention-based co-learning framework (Case-9) in Stage-1 for ablation study. Corresponding cases (i.e., Case-2 to Case-5) without the 3-layer relation-aware GNNs in Stage-2 are also exploited.

Case-6. A similar improvement can be observed in Case-7, where semantic-specific relationships are modeled. Moreover, an additional gain can be noticed if both visual- and semantic-specific relationships are exploited in Case-8.

2) To validate the effectiveness of the attention-based co-learning framework, the experiment is conducted in Case-9, which achieves 0.70%, 1.55%, 1.41% and 2.07% improvements over Case-8. Note that the attention-based co-learning can only be achieved when both visual- and semantic-specific GNNs are leveraged in Stage-1.

3) The relation-aware GNNs in Stage-2 can deliver an additional performance gain on few-shot visual-semantic learning. For instance, compared with Case-5, where the relation-aware GNNs are replaced by fully-connected neural networks for label prediction, Case-9 brings up improvements of 2.09%, 2.84%, 1.72% and 2.10% on accuracies.

4) It should be noted that the Case-1 with only Stage-2 represents a 3-layer GNNs, and the initial feature of each node is the concatenation of the corresponding visual and semantic represen-

tations as well as the one-hot encoding of label. Case-1 performs comparably with the graph-based methods, GNN and EGNN as expected.

Besides the $\mathbf{V}_i$ and $\mathbf{S}_i$ terms (which contains the intra- and inter-relationship from both visual and semantic features) in Eq. 3.9, whose effects have been justified, we also studied the efficacy of $\phi(\mathbf{I}_i)$, $\psi(\mathbf{Q}_i)$, and $h(\mathbf{A}_i)$ using 5-way 1-shot task on Visual-Genome-QA. As shown in Table 4.5, the absence of any term leads to a performance degradation.

	w/o h	w/o ϕ	w/o ψ	Full HGAT
5-way 1-shot	77.88	78.01	78.53	**79.56**

Table 4.5: Study on the efficacy of $\phi(\mathbf{I}_i)$, $\psi(\mathbf{Q}_i)$, and $h(\mathbf{A}_i)$ in Equation 9.

Experiments have been conducted on Visual Genome-QA to study the number of GNNs layers in Stage-1 (i.e., modal-specific GNNs) and Stage-2 (i.e., relation-aware GNNs). Results on 5-way 1-shot and 5-way 5-shot classifications in terms of the accuracy are shown in Table 4.6. Each row represents the results of the same number of model-specific GNN layers (Stage-1), and each column represents the results of the same number of relation-aware GNN layers (Stage-2). Specifically, when the numbers of GNN layers in Stage-1 and Stage-2 are set to 2 and 3, respectively, the optimal performance is achieved.

# of Layers	2	3	4
2	78.33/84.69	**79.56/86.10**	79.16/85.51
3	77.85/83.97	78.90/85.01	78.77/85.63

Table 4.6: Study on GNNs layers in Stage-1 (row-wise) and Stage-2 (column-wise).

4.4.5 Semi-supervised Few-shot Learning

Table 4.7 presents the comparisons of semi-supervised learning among HGAT, GNN, and EGNN. Experiments are conducted on the 5-way 5-shot VQA on Toronto COCO-QA, and results are presented when 40%, 60%, 80% of the image-text samples are labeled. Note that the labeled samples are balanced among the 5 classes. Take the 40% case for example, for a task, each of

Toronto COCO-QA	5-way 5-shot accuracy			
	40%	60%	80%	100%
GNN-LabeledOnly	64.62	67.30	70.31	72.55
GNN-Semi	66.04	68.44	71.48	72.55
EGNN-LabeledOnly	65.86	69.08	71.57	73.41
EGNN-Semi	**67.18**	69.92	72.61	73.41
HGAT-LabeledOnly	66.09	69.83	73.12	**75.41**
HGAT-Distractor	64.25	68.94	73.01	**75.41**
HGAT-Semi	67.16	**70.78**	**73.95**	**75.41**

Table 4.7: Comparison of semi-supervised learning results on Toronto COCO-QA for few-shot visual question answering.

the class contains 2 labeled samples and 3 unlabeled samples from the support set. 'LabeledOnly' is equivalent to the supervised few-shot setting, where only the labeled support samples are used. For instance, the 5-way 5-shot 40% VQA with 'LabeledOnly' is equivalent to the 5-way 2-shot VQA. 'Semi' denotes the semi-supervised few-shot setting, where all the support samples are used, regardless of whether they are labeled. In addition, 'Distractor' means the unlabeled support samples are randomly sampled from other classes instead of the 5 classes of the labeled support samples.

Besides, each of the three methods can acquire noticeable improvements when semi-supervised learning is performed compared with 'LabeledOnly' which demonstrates the unlabeled support samples can contribute to the learning in a few-shot setting. Moreover, for the proposed HGAT, the 'Distractor' leads to a minor performance degradation for each case compared with 'LabeledOnly'. This observation clearly shows that only the unlabeled samples from the classes of interest can contribute to the few-shot visual-semantic learning. Notably, for semi-supervised few-shot visual-semantic learning, the HGAT consistently outperforms the GNN and EGNN, except the 40% case, where HGAT achieves a comparable accuracy given by EGNN.

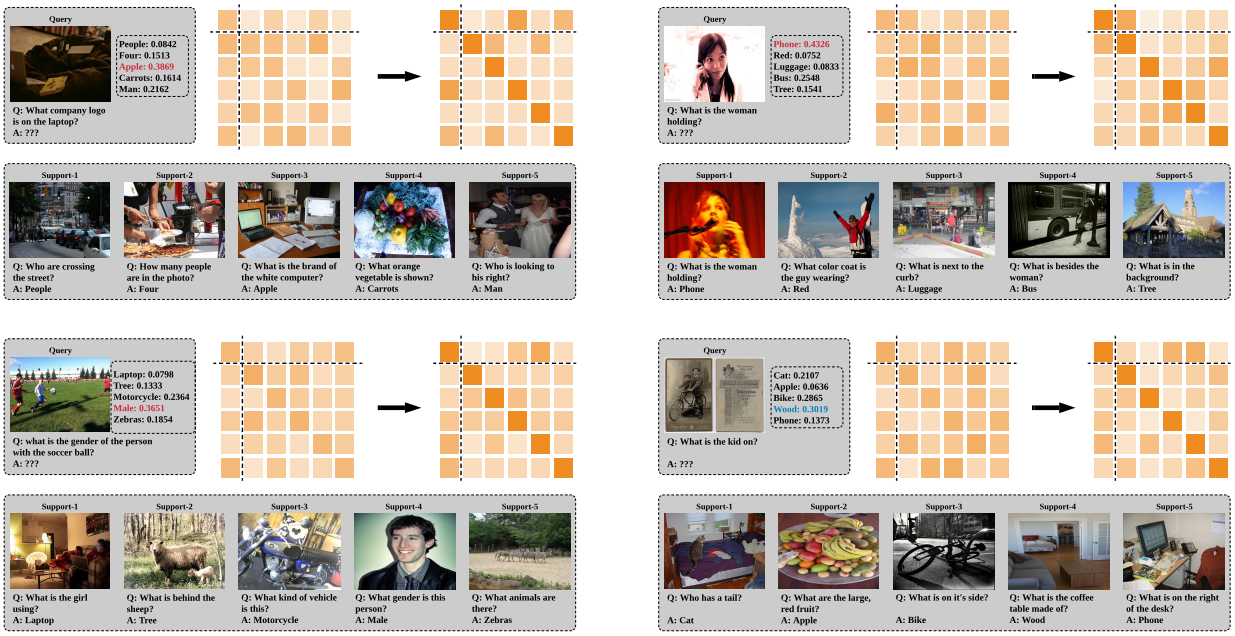

Fig. 4.3: Attention visualizations of the 3rd layer in the relation-aware GNNs for 5-way 1-shot VQA on Visual Genome-QA. Dark/light color denotes higher/lower values. Red/blue color denotes correct/wrong predictions on query samples.

4.4.6 Visualization

Figure 4.3 shows the computed attentions for several 5-way 1-shot VQA tasks sampled from the meta-testing set. The left and right square matrices represent the attentions before and after the meta-training, respectively. Take the top-left task for instance, we can notice that the attention between the query sample and the third support sample is larger than other off-diagonal attentions, which implies a stronger correlation between these two samples. Though the "apple" trademark, which is the decisive clue, occupies only a small portion of both images, HGAT can still associate the query to the support sample within the same class and give the correct answer.

4.5 Summary

This chapter, through introducing Hierarchical Graph Attention network (HGAT), presents a novel method for few-shot visual-semantic learning. Comprehensive experiments have been conducted on the widely-used Toronto COCO-QA, Visual Genome-QA and COCO-FITB benchmarks. The

extensive experimental results have shown that 1) HGAT delivers the state-of-the-art performance in terms of accuracy on both few-shot Image-QA and IC tasks compared with few-shot learning and standard (non-few-shot) methods; 2) It sheds light on tackling the few-shot multi-modal learning problems, especially for the few-shot visual-semantic learning tasks, through hierarchical exploitation and co-learning of the multiple modalities; 3) It can be easily extended to the semi-supervised setting, outperforming other few-shot visual-semantic learning baselines in the semi-supervised setting.

CHAPTER 5
CROSS-MODAL REASONING WITH EVENT CORRELATION FOR VIDEO QUESTION ANSWERING

5.1 Overview

In this paper, to deal with Video-QA, we propose **E**vent-**C**orrelated **G**raph **N**eural **N**etwork**s** (**EC-GNNs**), a novel end-to-end trainable model, which performs reasoning with event correlation, by exploiting the relationships among dense events, video contents and question words. To the best of our knowledge, this is the first work that explicitly incorporates dense video captions to perform the reasoning of Video-QA. Specifically, we develop a dense caption modality and handle the Video-QA task via three modality-aware graphs, where the dense caption features, visual features, and question embedding features are kept in the corresponding graphs. Three procedures are performed afterwards to infer the final answer. First, graph reasoning modules capture the intra-relations of each modality. Then, cross-modal reasoning modules aggregate relevant information across different modalities by leveraging a cross-modal attention mechanism to model the inter-modal relationships. Finally, after repeating the above two procedures several times, a question-guided

self-adaptive multi-modal fusion module conducts multiple inference loops to collect the question-oriented and event-related evidences and refines the final answer prediction.

The contributions are summarized as follows: 1) We propose a novel scheme to perform cross-modal reasoning with event correlation over information from three modalities (i.e., caption, video, and question). 2) We are the first to introduce dense video captions for Video-QA and clarify how to incorporate the event-correlated information into reasoning process. 3) We explicitly model the inter-modal relationships through a cross-modal attention mechanism in cross-modal reasoning. 4) We design a question-guided self-adaptive multi-modal fusion module that adaptively collects question-oriented and event-correlated evidence. 5) The proposed model outperforms most Video-QA methods and performs on par with the state-of-the-art on two Video-QA benchmarks.

5.2 Event-Correlated Graph Neural Networks

This section describes the architecture of the Event-Correlated Graph Neural Networks (EC-GNNs) as shown in Figure 5.1. We first introduce the representation (Sec. 5.2.1) and graph construction (Sec. 5.2.2) of the three modalities (i.e., the caption, the video, and the question). Then we demonstrate how the graph reasoning modules (Sec. 5.2.3) and the cross-modal reasoning modules (Sec. 5.2.4) can model the intra- and inter-relationships. Finally, we elaborate on the design of the question-guided self-adaptive multi-modal fusion module (Sec. 5.2.5) and the answer prediction (Sec. 5.2.6).

5.2.1 Contextual Representation with Generated Modality

Generating caption modality Dense video captions are a set of captions in descriptive natural language, which describe multiple events temporally localized in a video. To generate dense captions of a video, we first extract its appearance features and optical flow features using ResNet-200 [62] and BN-Inception [238] respectively. Given extracted features, we employ the dense video captioning model [50] trained on both ActivityNet Captions [49] and YouCookII [239]

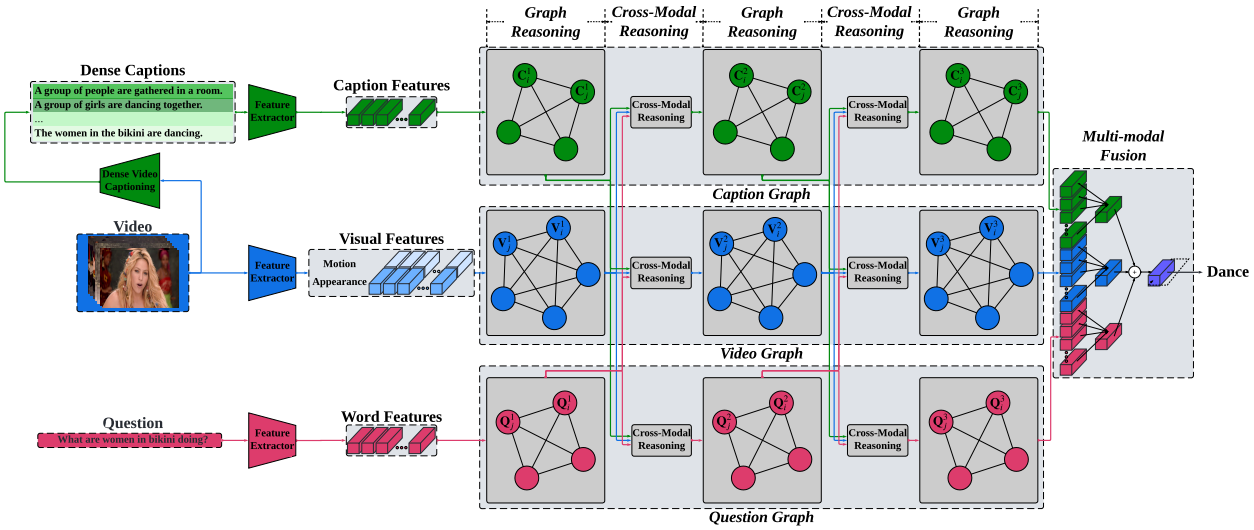

Fig. 5.1: The architecture of Event-Correlated Graph Neural Networks (EC-GNNs), which consists of three components: graph reasoning modules, cross-modal reasoning modules, and a question-guided self-adaptive multi-modal fusion module. Caption features, visual features, and word features are forwarded to *caption graph*, *video graph*, and *word graph* respectively. Each of the graphs contains three graph reasoning procedures where two cross-modal reasoning procedures are interleaved. Finally, multi-modal fusion is performed to infer the answer. Green, blue, and red arrows between modules corresponds to the information flow of caption, video, and question modalities respectively.

datasets to obtain the corresponding dense captions. Each caption is first represented as a sequence of word embeddings based on GloVe 300-D [214], and then encoded with a trainable GRU [219] to obtain its final hidden state as a monolithic representation. Therefore, the dense captions are denoted as a sequence of monolithic representations $\mathbf{F}^c = \{\mathbf{f}_i^c : i \leq N_c, \mathbf{f}_i^c \in \mathcal{R}^{d_c}\}$, where N_c represents the number of captions and d_c is the dimensionality of the caption features. In Figure 5.1, we highlight the caption features *in green*. In order to match the words in captions with the words in questions, same dictionary are used when dealing with dense captions and questions.

Representing video modality In addition to the generated caption modality described in the paper, the video modality and question modality are also properly represented and utilized in the proposed model. In order to take into account the appearance and motion information involved in video contents, we represent a video at both frame-level and shot-level. Specifically, to obtain appearance features, we use the 2D ConvNets (i.e., ResNet-152 [62] and VGG [60]) pre-trained

on the ImageNet 2012 classification dataset [211]. For shot-level motion features, we leverage 3D ConvNets (i.e., C3D [240]) pre-trained on the Sport1M dataset [213]. Then, the video is represented as two feature views, appearance features $\mathbf{F}^a = \{\mathbf{f}_i^a : i \leq N_v, \mathbf{f}_i^a \in \mathcal{R}^{d_a}\}$, and motion features $\mathbf{F}^m = \{\mathbf{f}_i^m : i \leq N_v, \mathbf{f}_i^m \in \mathcal{R}^{d_m}\}$, where N_v is number of frames and d_a, d_m are dimensionalities of the two feature views. We project the concatenation of the two features into a common visual space by two fully-connected layers, to obtain a joint representation of the video $\mathbf{F}^v = \{\mathbf{f}_i^v : i \leq N_v, \mathbf{f}_i^v \in \mathcal{R}^{d_v}\}$, where d_v represents the dimensionality of visual features. In Figure 5.1, we highlight the appearance/motion/visual features in *light/normal/dark blue*.

Representing question modality For a given question, we represent each word as a fixed-length vector initialized with the GloVe $300D$ word embedding [214] pre-trained on the Common Crawl dataset. The question is then denoted as a sequence of word embeddings $\mathbf{F}^q = \{\mathbf{f}_i^q : i \leq N_q, \mathbf{f}_i^q \in \mathcal{R}^{d_q}\}$, in which N_q represents the number of words and d_q equals 300. In Figure 5.1, we highlight the word features in red.

Obtaining contextualized representations All the features of the three modalities (i.e., caption features, visual features, and word features) are time series. To exploit the dynamic temporal information and obtain contextual representations for each modality, we leverage three independent GRUs to encode these features separately.

$$\mathbf{C}^1, \mathbf{c}_{N_c} = \text{GRU}(\mathbf{F}^c; \boldsymbol{\theta}^c_{GRU}) \tag{5.1}$$

$$\mathbf{V}^1, \mathbf{v}_{N_v} = \text{GRU}(\mathbf{F}^v; \boldsymbol{\theta}^v_{GRU}) \tag{5.2}$$

$$\mathbf{Q}^1, \mathbf{q}_{N_q} = \text{GRU}(\mathbf{F}^q; \boldsymbol{\theta}^q_{GRU}) \tag{5.3}$$

The contextualized caption features are denoted as $\mathbf{C}^1 = \{\mathbf{C}_i^1 : i \leq N_c, \mathbf{C}_i^1 \in \mathcal{R}^{d_C}\}$, the contextualized visual features are denoted as $\mathbf{V}^1 = \{\mathbf{V}_i^1 : i \leq N_v, \mathbf{V}_i^1 \in \mathcal{R}^{d_V}\}$, and the contextualized word features are denoted as $\mathbf{Q}^1 = \{\mathbf{Q}_i^1 : i \leq N_q, \mathbf{Q}_i^1 \in \mathcal{R}^{d_Q}\}$, where d_C, d_V, and d_Q are dimensionalities of the contextualized features; $\mathbf{c}_{N_c} \in \mathcal{R}^{d_C}$, $\mathbf{v}_{N_v} \in \mathcal{R}^{d_V}$ and $\mathbf{q}_{N_q} \in \mathcal{R}^{d_Q}$ are the last hidden

states, which represent the global features of the three modalities.

5.2.2 Graph Construction

Given the contextualized features of three modalities, we construct three modality-aware graphs: *caption graph*, *video graph*, and *question graph*. As shown in Figure 5.1, each modality-aware graph is a three-layer fully-connected GNNs (i.e., $l = 1, 2, 3$), where the caption features $\mathbf{C}^l$, visual features $\mathbf{V}^l$, and word features $\mathbf{Q}^l$ are kept in the *caption*, *video*, *question graphs*, with green, blue and red nodes, respectively. The numbers of nodes in each layer of the three graphs are N_c, N_v, and N_q, respectively, which equals to the numbers of dense captions, frames, and words in questions.

5.2.3 Intra-Modal Graph Reasoning

As each layer of the modality-aware GCNs contains modality-specific knowledge, we first capture the intra-modal relationships from each graph by performing graph reasoning independently. The intra-modal graph reasoning procedures in these three graphs share the common operations but differ in their node representations. Thus we take the *video graph* as an instance to illustrate the operations of graph reasoning.

We first project the node features $\mathbf{V}^l$ into an interaction space by a non-linear transformation operation $\phi(.)$. Then we calculate the dot-product similarity [202], which measures the semantic correlations among nodes, and employ softmax function on each row of the matrix to obtain the normalized adjacency matrix.

$$\mathbf{G}_V^l = \text{softmax}(\phi(\mathbf{V}^l)\phi(\mathbf{V}^l)^T) \tag{5.4}$$

where $\mathbf{G}_V^l \in \mathcal{R}^{N_v \times Nv}$ represents the adjacency matrix and ${\mathbf{G}_V}_{i,j}^l$ indicates the weight between $\mathbf{V}_i^l$ and $\mathbf{V}_j^l$.

To perform intra-modal reasoning on the graph, we apply the Graph Convolutional Networks (GCNs) proposed in [54], which allows us to update each node based on its neighbors and the

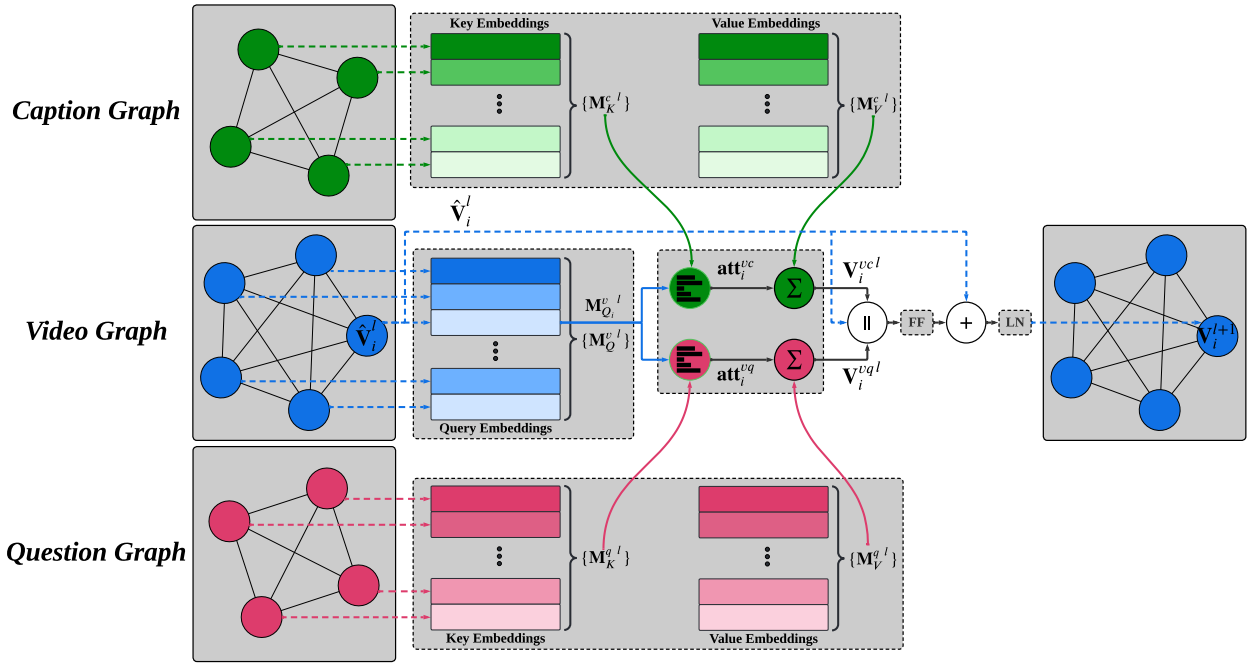

Fig. 5.2: An illustration of the cross-modal reasoning module. For simplicity, the cross-modal reasoning module in *video graph* between layer l and $l+1$ is presented, and the query node $\mathbf{V}_i^l$ is taken as an example for demonstration. Green, blue, and red arrows between modules corresponds to the information flow of caption, video, and question modalities respectively. Dotted arrows represent the transmission of node features. Different shades of color indicate representations of different nodes. FF and LN denote feed-forward network and layer normalization.

corresponding weights. Given the node features $\mathbf{V}^l$ and weights specified by the adjacency matrix $\mathbf{G}_V^l$, the nodes are updated by a linear transformation of aggregated excitation of its neighbors and itself. Formally, we represent the graph convolution on layer l of *video graph* as

$$\hat{\mathbf{V}}^l = \text{Relu}(\text{LayerNorm}(\mathbf{G}_V^l \mathbf{V}^l \mathbf{W})) \tag{5.5}$$

where $\mathbf{W} \in \mathcal{R}^{d_V \times d_V}$ is a learnable weight matrix and $\hat{\mathbf{V}}^l = \{\hat{\mathbf{V}}_i^l : i \leq N_v, \hat{\mathbf{V}}_i^l \in \mathcal{R}^{d_V}\}$ represents the updated node features. Two non-linear functions including Layer Normalization [241] and ReLU are applied after the graph convolution.

We conduct the above intra-modal graph reasoning on *caption graph*, *video graph*, and *question graph* independently and obtain the updated node features $\hat{\mathbf{C}}^l$, $\hat{\mathbf{V}}^l$ and $\hat{\mathbf{Q}}^l$ accordingly, which are forwarded to cross-modal reasoning ($l = 1, 2$) and multi-modal fusion ($l = 3$) respectively.

5.2.4 Cross-Modal Reasoning with Cross-Modal Attention Mechanism

A video footage tends to focus on part of the question rather than the whole, and tends to pay more attention on a few captions instead of assigning equal weight to all captions. Same rule applies for a word in the question and a caption in the set of dense captions. Therefore, to infer the correct answer, we need to fully understand the dense interactions between factors of different modalities.

As shown in Figure 5.1, between two intra-modal graph reasoning procedures, we propose a cross-modal attention mechanism (CAM) in cross-module reasoning procedure to explicitly exploit the inter-modal relationships with two other modalities. For instance, in the cross-modal reasoning of *video graph*, CAM allows a factor from the video modality as a clue to determine the weights of factors from question and caption modalities to aggregate the relevant information.

Given a query and a set of key-value pairs, CAM computes the weighted sum of values based on the dot-product similarity of the query and keys. With query, key and value denoted as a set of vectors (i.e., $\mathcal{M}_Q$, $\mathcal{M}_K$, and $\mathcal{M}_V$), we formulate the cross-modal attention mechanism as

$$\text{CAM}(\mathcal{M}_Q, \mathcal{M}_K, \mathcal{M}_V) = \text{softmax}(\frac{\mathcal{M}_Q {\mathcal{M}_K}^T}{\sqrt{d}})\mathcal{M}_V \tag{5.6}$$

in which $\mathcal{M}_K$ and $\mathcal{M}_V$ belong to the same modality which is different from that of $\mathcal{M}_Q$, and d indicates the dimensionality of the vectors.

With CAM, which is designed to better embed and capture information across multiple modalities, as the core module, we build the cross-modal reasoning procedure, which is illustrated in Figure 5.2. The cross-modal attention mechanism (CAM) is used to explicitly exploit the inter-modal relationships of one modality with two other modalities. As same operations shared across the cross-modal reasoning procedures in three graphs, in Figure 5.2, we give a description of cross-modal reasoning using *video graph* as an example, in which the node features of *video graph* are employed as query vectors, and node features of *question graph* and *caption graph* serve as key-value pairs.

We first linearly project the updated node features $\hat{\mathbf{V}}^l$, $\hat{\mathbf{Q}}^l$ and $\hat{\mathbf{C}}^l$ into a transformed space.

$$\mathbf{M}_Q^{v\,l} = \mathbf{W}_Q^v \hat{\mathbf{V}}^l \tag{5.7}$$

$$\mathbf{M}_K^{q\,l}, \mathbf{M}_V^{q\,l} = \mathbf{W}_K^q \hat{\mathbf{Q}}^l, \mathbf{W}_V^q \hat{\mathbf{Q}}^l \tag{5.8}$$

$$\mathbf{M}_K^{c\,l}, \mathbf{M}_V^{c\,l} = \mathbf{W}_K^c \hat{\mathbf{C}}^l, \mathbf{W}_V^c \hat{\mathbf{C}}^l \tag{5.9}$$

where $\mathbf{W}_Q^v$, $\mathbf{W}_K^q$, $\mathbf{W}_V^q$, $\mathbf{W}_K^c$ and $\mathbf{W}_V^c$ are the learnable weight matrices. $\mathbf{M}_Q^{v\,l} = \{\mathbf{M}_{Q_i}^{v\,l} : i \leq N_v\}$ represents the transformed visual features in the query space. $\mathbf{M}_K^{q\,l} = \{\mathbf{M}_{K_i}^{q\,l} : i \leq N_q\}$ and $\mathbf{M}_V^{q\,l} = \{\mathbf{M}_{V_i}^{q\,l} : i \leq N_q\}$ represent the transformed word features in the key and value spaces. $\mathbf{M}_K^{c\,l} = \{\mathbf{M}_{K_i}^{c\,l} : i \leq N_c\}$ and $\mathbf{M}_V^{c\,l} = \{\mathbf{M}_{V_i}^{c\,l} : i \leq N_c\}$ represent the transformed caption features in the key and value spaces. Then we apply the CAM to embed the visual information into the feature spaces of question words and dense captions respectively to collect the relevant information from question and caption modalities.

$$\mathbf{V}^{vql} = \text{CAM}(\mathbf{M}_Q^{v\,l}, \mathbf{M}_K^{q\,l}, \mathbf{M}_V^{q\,l}) \tag{5.10}$$

$$\mathbf{V}^{vcl} = \text{CAM}(\mathbf{M}_Q^{v\,l}, \mathbf{M}_K^{c\,l}, \mathbf{M}_V^{c\,l}) \tag{5.11}$$

where $\mathbf{V}^{vql} = \{\mathbf{V}_i^{vql} : i \leq N_v\}$ and $\mathbf{V}^{vcl} = \{\mathbf{V}_i^{vcl} : i \leq N_v\}$ indicate the word features and caption features attended by the visual features (i.e., CAM-based features), which are combined with the node features $\hat{\mathbf{V}}^l$ to obtain the updated visual features $\mathbf{V}^{l+1} = \{\mathbf{V}_i^{l+1} : i \leq N_v, \mathbf{V}_i^{l+1}\}$ after cross-modal reasoning.

$$\mathbf{V}^{l+1} = \text{LayerNorm}(\text{FF}([\mathbf{V}^{vql}||\mathbf{V}^{vcl}||\hat{\mathbf{V}}^l]) + \hat{\mathbf{V}}^l) \tag{5.12}$$

where $||$ denotes the concatenation operation and FF represents a feed-forward network. To achieve a better performance, a residual connection [62] is adopted for multiple layers stacking.

The above cross-modal reasoning is performed on *video graph*, *question graph* and *caption graph* and obtains the updated node features $\mathbf{V}^{l+1}$, $\mathbf{Q}^{l+1}$ and $\mathbf{C}^{l+1}$, which are forwarded to the next

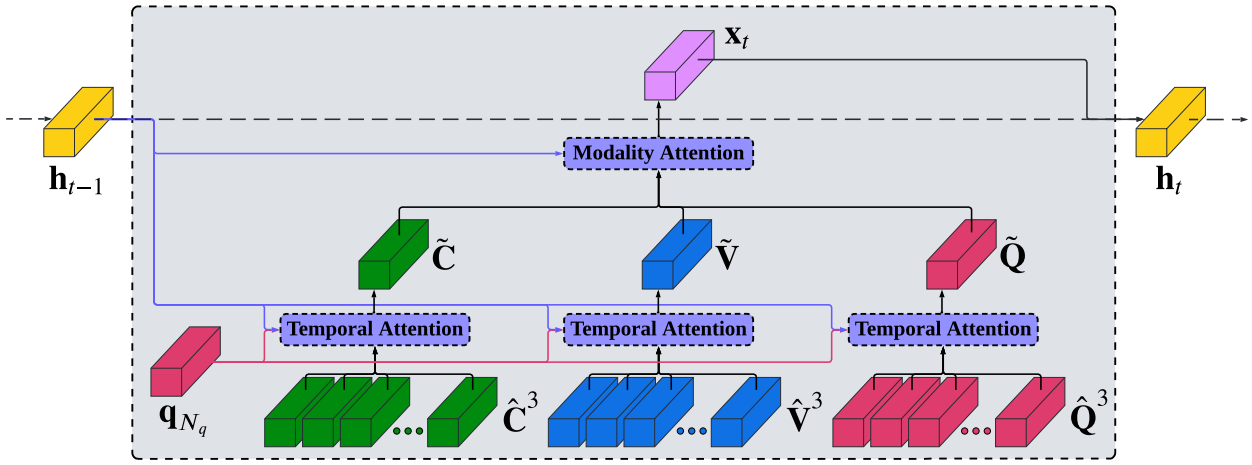

Fig. 5.3: An illustration of the question-guided self-adaptive multi-modal fusion module at the t-th step. An LSTM controller with previous state $\mathbf{h}_{t-1}$ attends to relevant caption features, visual features, and word features with question guidance, and combines them to obtain current state $\mathbf{h}_t$.

GNN layer for the next round of intra-modal graph reasoning.

5.2.5 Question-Guided Self-Adaptive Multi-Modal Fusion

In multi-modal fusion procedure, to collect the question-oriented and event-correlated evidence from the three modalities, we design a question-guided self-adaptive multi-modal fusion module, which performs multiple cycles of reasoning to selectively attend to multi-modal features through gradually refining the soft attention weights with the guidance of question.

The multi-modal fusion module takes the updated node features from the last GNN layer (i.e., caption features $\hat{\mathbf{C}}^3$, visual features $\hat{\mathbf{V}}^3$, and word features $\hat{\mathbf{Q}}^3$) and the last hidden state of question (i.e., $\mathbf{q}_{N_q}$) as the inputs. Inspired by the multi-modal fusion module proposed by HME [57], we adopts multi-step reasoning with an LSTM controller.

At t-th step of reasoning, as shown in Figure 5.3, by interacting with the last hidden state of question as well as the previous state of LSTM controller, the module first employs temporal attention mechanism to attend to different parts of the caption features, visual features, and word

features, independently, and collects the question-oriented information from each modality.

$$\mathbf{att}^c = \text{softmax}(\mathbf{W}^c \text{tanh}(\mathbf{W}_q^c \mathbf{q}_{N_q} + \mathbf{W}_h^c \mathbf{h}_{t-1} + \mathbf{W}_c^c \hat{\mathbf{C}}^3 + \mathbf{b}^c)) \tag{5.13}$$

$$\mathbf{att}^v = \text{softmax}(\mathbf{W}^v \text{tanh}(\mathbf{W}_q^v \mathbf{q}_{N_q} + \mathbf{W}_h^v \mathbf{h}_{t-1} + \mathbf{W}_v^v \hat{\mathbf{V}}^3 + \mathbf{b}^v)) \tag{5.14}$$

$$\mathbf{att}^q = \text{softmax}(\mathbf{W}^q \text{tanh}(\mathbf{W}_q^q \mathbf{q}_{N_q} + \mathbf{W}_h^q \mathbf{h}_{t-1} + \mathbf{W}_q^q \hat{\mathbf{Q}}^3 + \mathbf{b}^q)) \tag{5.15}$$

$$\tilde{\mathbf{C}} = \sum_{i=1}^{N_c} \mathbf{att}_i^c \hat{\mathbf{C}}_i^3 \quad \tilde{\mathbf{V}} = \sum_{i=1}^{N_v} \mathbf{att}_i^v \hat{\mathbf{V}}_i^3 \quad \tilde{\mathbf{Q}} = \sum_{i=1}^{N_q} \mathbf{att}_i^q \hat{\mathbf{Q}}_i^3 \tag{5.16}$$

where $\mathbf{att}^c \in \mathcal{R}^{N_c}$, $\mathbf{att}^v \in \mathcal{R}^{N_v}$, and $\mathbf{att}^q \in \mathcal{R}^{N_q}$ indicate the normalized attention weights over the caption features $\hat{\mathbf{C}}^3$, visual features $\hat{\mathbf{V}}^3$, and word features $\hat{\mathbf{Q}}^3$ respectively. $\mathbf{W}^c$, $\mathbf{W}^v$, $\mathbf{W}^q$, $\mathbf{W}_q^c$, $\mathbf{W}_q^v$, $\mathbf{W}_q^q$, $\mathbf{W}_h^c$, $\mathbf{W}_h^v$, $\mathbf{W}_h^q$, $\mathbf{W}_c^c$, $\mathbf{W}_v^v$, $\mathbf{W}_q^q$, $\mathbf{b}^c$, $\mathbf{b}^v$ and $\mathbf{b}^q$ are the learnable weights. $\tilde{\mathbf{C}}$, $\tilde{\mathbf{V}}$, and $\tilde{\mathbf{Q}}$ represent the attended caption features, attended visual features, and attended word features respectively, which are then incorporated together with the learned modality weights $\boldsymbol{\alpha} = \{\alpha^c, \alpha^v, \alpha^q\}$ to gather the event-correlated information across the three modalities.

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{W}^\alpha \text{tanh}([\mathbf{W}_c^\alpha \tilde{\mathbf{C}} || \mathbf{W}_v^\alpha \tilde{\mathbf{V}} || \mathbf{W}_q^\alpha \tilde{\mathbf{Q}}] + \mathbf{W}_h^\alpha \mathbf{h}_{t-1} + \mathbf{b}^\alpha)) \tag{5.17}$$

$$\mathbf{x}_t = \text{tanh}(\alpha^c \mathbf{W}_c^x \tilde{\mathbf{C}} + \alpha^v \mathbf{W}_v^x \tilde{\mathbf{V}} + \alpha^q \mathbf{W}_q^x \tilde{\mathbf{Q}} + \mathbf{W}_h^x \mathbf{h}_{t-1} + \mathbf{b}^x) \tag{5.18}$$

where $\mathbf{W}^\alpha$, $\mathbf{W}_c^\alpha$, $\mathbf{W}_v^\alpha$, $\mathbf{W}_q^\alpha$, $\mathbf{W}_h^\alpha$, $\mathbf{W}_c^x$, $\mathbf{W}_v^x$, $\mathbf{W}_q^x$, $\mathbf{W}_h^x$, $\mathbf{b}^\alpha$ and $\mathbf{b}^x$ are the learnable weights. The output feature $\mathbf{x}_t$ is forwarded to obtain the current state $\mathbf{h}_t$ of the LSTM controller by $\mathbf{h}_t = \text{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1})$.

We iteratively perform the above reasoning step to obtain the final representation of the multi-modal fusion. After N_r steps, the state LSTM controller $\mathbf{h}_{N_r}$ captures the question-oriented and event-correlated information across the three modalities. We set the value of N_r to 3, referring to that in HME [57]. Following [16], we apply the standard temporal attention on the contextualized visual features $\mathbf{V}^1$ and caption features $\mathbf{C}^1$, and concatenate with $\mathbf{h}_{N_r}$ to obtain the final representation $\mathbf{s}_a$ for answer prediction.

5.2.6 Answer Prediction

Based on the final representation $\mathbf{s}_a$ from the fusion module, we predict the final answer of the Video-QA tasks. We aim at solving three different types of Video-QA, same as [16]: *open-ended words*, *open-ended numbers* and *multiple-choice*. Given the input video v and the corresponding question q, the open-ended tasks is to infer an answer $\hat{a}$ from a pre-defined answer set $\mathcal{A}$ of size C that matches the ground-truth $a^\star \in \mathcal{A}$, while the multiple-choice task is to select the correct answer $\hat{a}$ out of the candidate set $\{a_i\}_{i=1}^K$. With the representation $\mathbf{s}_a$, for each of the three tasks, we train and evaluate a separate model with different predicting functions and training losses.

Open-ended word task The *open-ended words* is treated as a classification problem. We apply a linear layer and a softmax function upon the representation $\mathbf{s}_a$ to generate the probabilities $\mathbf{p}$ for all answers in the pre-defined answer set $\mathcal{A}$.

$$\mathbf{p} = \text{softmax}(\mathbf{W}_a^w \mathbf{s}_a + \mathbf{b}^w) \tag{5.19}$$

in which $\mathbf{W}_a^w$ and $\mathbf{b}^w$ are the learnable weights. The cross-entropy loss is adopted to train the model, and the predicted answer is represented as $\hat{a} = \text{argmax}_{\mathcal{A}}(\mathbf{p})$ in testing phase.

Open-ended number task We consider the *open-ended numbers* as a linear regression problem, which takes the representation $\mathbf{s}_a$ and outputs a rounded integer value as the predicted answer.

$$\hat{a} = \text{round}(\mathbf{W}_a^n \mathbf{s}_a + \mathbf{b}^n) \tag{5.20}$$

where $\mathbf{W}_a^n$ and $\mathbf{b}^n$ are the learnable weights. The l_2 loss is leveraged to minimize the gap between the predicted integer and the ground-truth. Note that $\mathcal{A}$ is an integer-valued answer set (i.e., 0-10) for the *open-ended numbers*.

Multiple choice task For *multiple-choice*, we concatenate the question q with each answer in the candidate set $\{a_i\}_{i=1}^{K}$ to obtain K word sequences, which are forward to the model individually to get K final representations $\{\mathbf{s}_a\}_{i=1}^{K}$. Then we use a linear function to transfer the K representations to K scores $\mathbf{s} = \{s^p, s_1^n, s_2^n, ..., s_{K-1}^n\}$. The answer with the highest score is selected as the predicted one. We optimize the model by minimizing the hinge loss L_{hinge} between the score for the correct answer s^p and the scores for the incorrect answers $\{s_i^n\}_{i=1}^{K-1}$.

$$L_{hinge} = \sum_{i=1}^{K-1} \max(0, 1 + s_i^n - s^p) \tag{5.21}$$

5.3 Performance Evaluation

In this section, we evaluate our proposed model on two widely used large-scale Video-QA benchmarks and discuss the experimental results quantitatively and qualitatively.

5.3.1 Benchmark Datasets

TGIF-QA TGIF-QA [16] is the most commonly used benchmark for Video-QA with 165K QA pairs from 72 animated GIFs. It includes four different Video-QA tasks: (1) *FrameQA*: An *open-ended words* task, in which the question can be answered based on a single video frame. (2) *Count*: An *open-ended numbers* task, asking the number of repetition of a given action. (3) *Action*: A 5-option *multiple-choice* task, which aims to recognize a repeated action given its specific times. (4) *Trans*: A 5-option *multiple-choice* task, asking the transition happened between two states.

MSVD-QA MSVD-QA [46] is generated automatically based on Microsoft Research Video Description Corpus [209]. It contains 50K QA pairs associated with 1,970 video clips and consists of five different types of questions, including *What*, *Who*, *How*, *When* and *Where*. All questions are *open-ended words* tasks, where the answer is predicted from a pre-defined answer set of size 1,000.

5.3.2 Experimental Setup

Implementation details We employed the pre-trained ResNet-152 and VGG to obtain the appearance features for TGIF-QA and MSVD-QA respectively and the pre-trained C3D to extract the motion features. d_c, d_a, d_m, d_v, and d_q are set to 512, 2,048, 4,096, 4,096, and 300, respectively. d_V, d_Q, and d_C are all set to 512. The dimensions of $\mathbf{h}_t$ and $\mathbf{s}_a$ are set to 512 and 1,536 respectively. The proposed model was trained with Adam optimizer [226] with an initial learning rate 10^{-4} and a batch size 64. Our code was implemented in PyTorch [227] and run with NVIDIA Tesla P100 GPUs.

Evaluation metrics For both *open-ended words* and *multiple-choice* tasks, we adopted the classification accuracy (i.e., Accuracy) as the evaluation metric. For *open-ended number* task, as all the answers belong to 11 possible integer values ranging from 0 to 10, we evaluated the model in terms of the mean square error (i.e., MSE) between the predicted integer and the ground-truth. Note that the model delivers better performance with higher classification accuracy while lower mean square error.

Baselines Our model was compared with some typical and state-of-the-art methods. For TGIF-QA dataset, we compared EC-GNNs with ST-VQAs [16], ST-VQA$\star$ [17], Co-Mem [112], HME [57], PSAC [47], and HGA [204]. For MSVD-QA dataset, in addition to ST-VQA, Co-Mem, HME and HGA, we also compared against E-VQA [15], E-MN [15] and AMU [46].

5.3.3 Experimental Results

We compare our proposed model with the aforementioned baselines. The results on TGIF-QA and MSVD-QA datasets are shown in Tables 5.1 and 5.2, and we can make the following observations.

State-of-the art performance especially on event-related tasks EC-GNNs outperforms most baselines and performs comparably with state-of-the-art method (i.e., HGA [204]) for both

Table 5.1: Comparison on TGIF-QA in terms of MSE (↓) for *Count* and Accuracy (%, ↑) for others.

Methods	*FrameQA* (↑)	*Count* (↓)	*Action* (↑)	*Trans* (↑)
ST-VQA-Sp [16]	45.5	4.28	57.3	63.7
ST-VQA-Tp [16]	49.3	4.40	60.8	67.1
ST-VQA-SpTp [16]	47.8	4.56	57.0	59.6
ST-VQA⋆ [17]	52.0	4.22	73.5	79.7
Co-Mem [112]	51.5	4.10	68.2	74.3
PSAC [47]	55.7	4.27	70.4	76.9
HME [57]	53.8	4.10	73.9	77.8
HGA [204]	55.1	4.09	75.4	81.0
Ours	**55.3**	**4.18**	**75.8**	**81.2**

Table 5.2: Comparison on MSVD-QA in terms of Accuracy (%, ↑).

Methods	*What* (↑)	*Who* (↑)	*How* (↑)	*When* (↑)	*Where* (↑)	*All* (↑)
E-VQA [15]	9.7	42.2	**83.8**	72.4	**53.6**	23.3
E-MN [15]	12.9	46.5	80.3	70.7	50.0	26.7
AMU [46]	20.6	47.5	83.5	72.4	**53.6**	32.0
ST-VQA [16]	18.1	50.0	**83.8**	72.4	28.6	31.3
Co-Mem [112]	19.6	48.7	81.6	**74.1**	31.7	31.7
HME [57]	22.4	50.1	73.0	70.7	42.9	33.7
HGA [204]	23.5	**50.4**	83.0	72.4	46.4	34.7
Ours	**24.0**	49.9	79.7	70.7	50.0	**34.8**

datasets. Notably, EC-GNNs delivers an accuracy of 75.8% on the *Action* task of TGIF-QA, excelling the second best by 0.4 percentage points; On MSVD-QA, it performs the best on *What* task and the union of all tasks (denoted as *All* in Table 5.2). The results suggest that leveraging event-related information (i.e., dense video captions) is beneficial for the cases where obvious event information (e.g., actions) is involved and multiple events occur concurrently or successively (e.g., state transitions), and clearly demonstrate the effectiveness and feasibility of the proposed EC-GNNs.

Advantages of modality-aware graph networks Comparing to HGA [204], which also uses GCNs for the modeling of intra-modal relationships, EC-GNNs performs better on most tasks and question types. The reason may be that HGA uses one graph, ignoring the pattern differences between visual and linguistic modalities and leading to semantic bias, while EC-GNNs employ

modality-aware graphs that model heterogeneous intra-modal relationships and revealing multi-modal features.

Gains of question guidance EC-GNNs consistently outperforms HME [57], which employs similar multi-modal fusion but lacks question guidance. This suggests that gathering question-oriented information is crucial for a comprehensive understanding of cross-modal semantics. An ablation study is conducted to further justify the superiority of question guidance in multi-modal fusion.

Superiority of GCNs to RNNs EC-GNNs perform consistently better than E-VQA [15], a typical RNNs-based Video-QA method. It implies that for Video-QA tasks, GCNs learn better long-term dependencies than RNNs, especially when complex semantics are involved.

5.3.4 Visualization

To demonstrate the efficacy of the proposed model, we use Figure 5.4 to visualize the learned attention weights $\mathbf{att}^c$, $\mathbf{att}^v$, and $\mathbf{att}^q$ (in Equations (5.13)–(5.15)) over caption, visual, and word features (finalized by three modality-aware GCNs) in the question-guided self-adaptive multi-modal fusion.

The first sample shows that the proposed model can effectively focus on the most relevant frames, words and captions for a given question. This example is from the MSVD-QA dataset. From the point view of the video as a whole, it is difficult to know *who is in the middle of the stage*, especially when close-up shots take up most of the video. Furthermore, the model needs to find out the most important ones from a large number of dense video captions, which helps to infer the correct answer. Our model subtly pays more attention to a few overall shots in the middle of the video, where a woman leads the dance in the center of the stage while others dance around. Correspondingly, in the caption modality, the model assigns a higher weight to the caption *the women in the dress continue dancing and one of the women in the middle of the stage* while reducing other weights.

In the latter two samples, the model is able to directly infer the correct answer by referring to some correlated captions. In addition to providing some visually intuitive information, such as the presence of two women in the video of the second sample, the captions can also incorporate some commonsense knowledge that may not be represented by an end-to-end model for Video-QA datasets with limited sample diversity [90]. For instance, in the third sample, the caption *a man is riding a horse in a field* can naturally associate the *horse riding* with *field*, which provides exactly where *the horse and rider trot across*. Nonetheless, without assistance from dense video captions, most other models have to recognize the environment based on visual features, which is quite challenging due to the ambiguities contained in video frames.

5.3.5 Ablation Study

To validate the contribution of each component in the proposed EC-GNNs, several ablation experiments were conducted on TGIF-QA. The following observations are made based on Table 5.3.

Benifits of generating an event-correlated modality EC-GNNs incorporates dense video captions as a new modality to perform reasoning of Video-QA, which leads to better performance. Compared with Case-ALL, which learns the full model, the accuracy of the *Action* task decreases by 2.4% in Case-1, where the caption modality is not modeled. This suggests that dense video captions can provide valuable evidence for answer reasoning, especially when the videos and questions contain complex semantics. Similar trends can be observed in *FrameQA*, *Count* and *Trans* tasks.

Effectiveness of cross-modal reasoning The model in Case-2, without the cross-modal reasoning module, gets 1.8% accuracy degradation of the *Action* task compared with Case-ALL. Degraded performance of this ablated model is also observed on the other three tasks. This demonstrates the benefits of cross-modal reasoning in the modeling inter-modal relationships, which can aggregate relevant information across different modalities.

Table 5.3: Ablation study on TGIF-QA. MSE (↓) for *Count* and Accuracy (%, ↑) for others. (Vid: Video Modality; Cap: Caption Modality; CMR: Cross-Modal Reasoning Module; MMF: Multi-Modal Fusion module from [57]; Q-MMF: Question-guided self-adaptive Mutli-Modal Fusion module.)

Case #	Vid	Cap	CMR	MMF	Q-MMF	*FrameQA* (↑)	*Count* (↓)	*Action* (↑)	*Trans* (↑)
1	✓		✓		✓	52.1	4.33	73.4	77.5
2	✓	✓			✓	53.3	4.28	74.0	76.8
3	✓	✓	✓			52.6	4.32	74.7	78.5
4	✓	✓	✓	✓		54.1	**4.15**	75.3	80.8
ALL	✓	✓	✓		✓	**55.3**	4.18	**75.8**	**81.2**

Gains of question guidance in multi-modal fusion The leveraging of multi-modal fusion delivers a performance gain over Case-3, where only the self-attention pooling and bi-linear fusion are applied for feature fusion [204]. For instance, for the *Action* task, Case-4 brings up improvements of 0.6% on accuracy, and further adopting question-guided self-adaptive multi-modal fusion in Case-ALL achieves an additional gain of 0.5%. Similar trends can be observed in *FrameQA* and *Trans* tasks. In brief, collecting the question-oriented and event-correlated evidence facilitates a comprehensive understanding of cross-modal semantics and outperforms simple multi-modal fusion.

5.4 Summary

By introducing Event-Correlated Graph Neural Networks (EC-GNNs), we propose a unified end-to-end trainable model for Video Question Answering (Video-QA). Notably, this paper clarifies how event-correlated information can be extracted as a new modality and integrated into Video-QA inference. Furthermore, we solve the Video-QA task through a multi-modal graph consisting of three modality-aware graph convolution networks that perform cross-modal reasoning over three modalities (i.e., the caption, the video, and the question). Comprehensive experiments have been conducted on two widely-used Video-QA benchmarks, and extensive experimental results show the superiority of introducing the event-based caption modality and the effectiveness of the proposed model equipped with the cross-modal attention mechanism and the question guidance.

Fig. 5.4: Visualization of learned attention weights over the caption (in green), visual (in blue), and word (in red) features. Dark/light color denotes higher/lower values.

CHAPTER 6
CONCLUSIONS AND FUTURE PLAN

The major goal of this thesis is to bridge the discrepancy between human-level intelligence and visual-semantic learning from several different aspects: basic understanding, external knowledge integration, fast learning, and event correlation. During our investigation, some high-level conclusions and insights into human intelligent visual-semantic learning can be made based on the experimental results and analysis. Specifically, we find a good practice for designing models, summarize and discuss several common ideas for visual-semantic learning. In addition, despite emerging Deep Learning (DL) delivers promising techniques to mitigate this gap, there are still many barriers and challenges to achieve a powerful AI agent with visual-semantic understanding abilities of human-level. While we have discussed and solved some of them in the thesis, many crucial and interesting research directions still lie ahead.

6.1 Conclusions

In this section, we first summarize a practical approach to design and tailor visual-semantic learning models by imitating human cognition. We then conclude several decent ideas adopted in our model design and validated in experiments.

6.1.1 Human Cognition Imitation

Through our efforts in approaching human-level intelligence, we have learned a good practice in the design of visual-semantic learning models, that is, customizing learning modules in deep models by mimicking human cognition and thinking process when dealing with visual-semantic tasks. For example, considering how a human approaches a Video-QA task, if he/she watches a long video and reads a question, it is difficult or even clueless for him/her to provide the correct answer, because he/she may remember a lot of video information irrelevant to the question. However, he/she is able to pay special attention to the part of the video that is relevant to the question and give an accurate answer by reading the question before watching the video. Therefore, in MA-DRNN, by thinking about a better way for humans to watch videos and read questions, we propose to encode question before video for Video-QA reasoning, and demonstrate its effectiveness through ablation studies. Similarly, by mimicking the way the human brain facilitates cognition, we employ external memory augmentation in MA-DRNN and MCR-MemNN to improve the understanding and reasoning capacities of visual-semantic learning models. Furthermore, by considering how humans recognize and process heterogeneous information from multiple modalities (i.e., video/image, question, knowledge, and dense video captioning) , we introduce variant approaches in MCR-MemNN, HGAT, and EC-GNNs to take full advantage of intra- and inter-modal relationships. At the same time, we shed light on how to incorporate new modalities (e.g., knowledge, dense video captioning) into the model reasoning process to mimic human-level visual-semantic learning.

6.1.2 Differentiable Memory

According to our experimental results and analysis of MA-DRNN and MCR-MemNN to study the basic understanding ability for Video-QA and KB-Image-QA, the external memory augmentation improves visual-semantic reasoning performance for the both cases with and without commonsense knowledge. In particular, MA-DRNN utilizes a differentiable memory block and a customized memory addressing system to provide additional storage space for long-term visual-

textual dependency modeling, while MCR-MemNN stores candidate supporting facts in a key-value format using a learnable structured memory module. Even with different memory formats, both the integrated external memories are differentiable, allowing end-to-end training with the model. This is inline with the human cognitive system, where certain specific areas of human brain help memorize relevant and key information, thereby enabling human-level understanding and reasoning process.

6.1.3 Inter-Modal Relationships

Based on the experimental results and analysis of MCR-MemNN, HGAT, and EC-GNNs to investigate some advanced visual-semantic learning abilities (i.e., external knowledge integration, fast learning, and event correlation), exploiting and modeling inter-modal relationships among different modalities leads to superior cross-modal reasoning ability with a new modality of supporting facts or dense video captions, and enables the model to perform visual-semantic reasoning with limited or even partially unlabeled samples. Specifically, MCR-MemNN employs a self-attention-based mechanism on memory slots to learn the summary of supporting facts in light of visual-semantic representation, and uses dot-product to compute the memory-aware attention over question words and image features to deliver the question and image embeddings in light of memory. Similarly, with a new modality of dense video caption involved, EC-GNNs proposes a cross-modal attention mechanism based on self-attention to exploit the inter-modal relationships among the three modalities (i.e., caption, video, and question). However, HGAT presents a novel attention-based co-learning mechanism to capture the visual-semantic interactions by sharing attentions between associated nodes in two modal-specific GNNs. Even with different mechanisms to model inter-modal relations, they are all trying to imitate the human cognition and thinking process when performing visual-semantic learning, that is to adaptively collect the question-oriented and information-complementary evidence for multi-modal understanding and reasoning by taking advantage of inherent associations across different modalities.

6.1.4 Attention Mechanisms

Variant attention mechanisms are widely-adopted in MCR-MemNN, HGAT, and EC-GNNs for intra-modal and cross-modal reasoning, multi-modal feature embedding and fusion. Specifically, MCR-MemNN employs top-down attention combined with self-attention to obtain the visual-semantic representation, and leverages basic attention mechanisms to model the interactions among question words, image regions, and supporting facts and deliver a memory-aware visual-semantic representation and a visual-semantic aware memory summary. HGAT proposes a shared attentional mechanism to share the attentions of visual modality with semantic modality and vice-versa for the node feature update of visual-specific GNNs and semantic-specific GNNs respectively, so that both intra- and inter-modal relationships are exploited among the samples from support set and query set. EC-GNNs introduces a cross-modal attention mechanism to explicitly exploit the inter-modal relationships among three modalities (i.e., caption, video, and question). Moreover, a novel question-guided self-adaptive multi-modal fusion is presented by employing temporal attention mechanism to attend to caption features, visual features, and word features through multi-step reasoning, and iteratively collects the question-oriented and event-correlated evidence from each modality for the final representation of multi-modal fusion.

When doing visual-semantic learning, with a large amount of heterogeneous information involved, humans always filter out what they need according to their experience (e.g., commonsense knowledge integration) and some correlations among information from different modalities (i.e., intra- and inter-modal relationships modeling), then quickly understand and reason over the digested information to give accurate response. For example, when performing Video-QA reasoning on a long-term video with complicated spatial-temporal dynamics, humans only focus on the informative frames based on the given question, meanwhile optimize the awareness on each question word according to visual observations, and finally deliver the correct answer. To imitate this learning process, variant attention mechanism has been proposed and widely adopted during investigation, and its superiority is well-justified through our efforts to mitigate the gap between visual-semantic learning and human-level intelligence from several different perspectives.

6.2 Future Plan

6.2.1 Explainable Knowledge-based Video-QA

As aforementioned, most of the existing Video-QA models [52, 15, 104, 16, 17, 53] focus on questions that are answerable by solely referring to the visible video content without any commonsense. However, such models are incapable of answering questions that require external knowledge beyond what is contained in the video. In real scenarios, most questions given a video requires a joint analysis of visual content and general knowledge, which is extremely challenging for current AI agents but effortless of humans. That's why knowledge-based Video-QA is introduced as a brand new topic under visual-semantic learning.

Even if significant progress has been made to knowledge-based Image-QA, research on knowledge-based Video-QA is still in its infancy. As far as we know, there are only two work dealing with KB-Video-QA, that created a benchmark KB-Video-QA dataset (i.e., KnowIT VQA) and proposed two baseline models (i.e., ROCK and ROLL). However, these existing efforts have not touched the root of KB-Video-QA. For instance, the proposed dataset and model can only handle multi-choice questions rather than free-form questions with open-ended answers, which are more general and closer to the real situation. In addition, only question and candidate answers are leveraged to retrieve relevant knowledge facts from KB, while the visual content with much more valuable spatial-temporal clues are not incorporated. Furthermore, during video reasoning, the cross-modal relationships among video, text (i.e., subtitles, question, candidate answers), and the retrieved knowledge facts are not exploited. More importantly, neither of the existing models can interpret why a particular answer is given, which is an ability necessary for a mature model, as advanced AI agents should be able to provide reasonable explanations for the decisions they made and the knowledge facts they referred. To address those limitations, we aim to propose a general KB-Video-QA framework that exploits cross-modal relationships during KB retrieval and reasoning, to deliver explainable open-ended answers for free-form questions. More specifically, our plan is summarized as follows:

- Reformat current KB-Video-QA datasets to create a dataset with free-form questions and each corresponds to an open-ended answer.
- Spatial-temporal features should be properly extracted from the given video and employed as clues for both external knowledge retrieval and answer inference.
- Implement a modality-aware GNN-based reasoning framework to capture question-oriented and information-complementary facts from different modalities (i.e., video, text and KB).
- The learned model is able to highlight related visual concepts (e.g., video frames and objects), semantic facts (e.g., question words and subtitles), and knowledge facts, to provide reliable interpretations for the delivered answers.

6.2.2 Video-QA with Spatial-Temporal Dense Captions

Analogous to external knowledge, dense visual captions provide much more detailed information which is hard to perceive during conventional Visual-QA reasoning. Thus, how to employ dense visual captions as another modality to facilitate Visual-QA reasoning becomes a popular research direction. Take Mucko [125] as an example, it handles KB-Image-QA with a multi-modal heterogeneous graph, where the dense image captions [179] are extracted, parsed and kept in the semantic layer to bridge the gap between visual and factual information. Likewise, in this thesis, we proposed Event-Correlated Graph Neural Networks (EC-GNNs) to perform reasoning of Video-QA with event correlation. Notably, dense video captions [49] are incorporated, and the relationships among dense events, video contents and question words are fully exploited. Proceeding with those efforts, it is verified that incorporating dense visual captions as a new auxiliary modality helps to identify and understand visual concepts in spatial dimension and events involved in temporal dimension. Besides, the intra- and inter-modal relationships among video, question and dense captions are better exploited, which provides valuable clues to infer more accurate answers during Visual-QA reasoning.

Nonetheless, dense image captions gives only visual concepts of a single frame as well as their spatial relationships without any considerations of temporal correlations of objects; similarly, dense video captions provides only descriptions of multiple events along video time span without understanding of fine-grained visual elements over each frame. An ideal Video-QA agent requires a comprehensive spatial-temporal understanding during inference, thus, it requires not only employing captions of dense events across temporal axis, but also incorporating captions of dense proposals which are spatially distributed in each video frame. Therefore, instead of exploiting them separately for Video-QA, it's worth investigating how to jointly incorporate dense video captions and dense image captions in reasoning process to deliver a comprehensive spatial-temporal representation, as the information provided on both side is able to complement each other, especially for the Videos with complicated spatial-temporal dynamics.

6.2.3 Visual-QA with Concise Visual Captions

It is verified that dense visual captions are able be incorporated to enhance the reasoning process of both Image-QA and Video-QA; however, a critical issue exists. Rich information is contained in image and video of high complexity. An image may contain more than ten objects with complicated relationships among them, and a long-term video is consisting of more than 1000 frames. During Visual-QA reasoning, there are many object proposals of image with inconsistent clues attended and many frames of video with duplicated and redundant visual appearance information selected. Thus, the incorporated dense visual captions are prone to be redundant and inconsistent. For example, dense video captions of two event proposals that are highly overlapped temporally have exact the same meaning and a lot of repeated words. A more unfavorable case is that the two captions express two meanings which are completely different. Same issue exists for dense image captions of any pairs of object proposals that are overlapping in spatial dimension. In addition, given a question, only a few captions are correlated to answer prediction. At this time, all Visual-QA models with the guidance of dense visual captions are struggling to select the most informative one to infer accurate answer. Even if some works on video captioning [242] and dense

video captioning [193, 194] have devoted to tackle the inconsistent and redundant caption generation from the perspective of informative frame selection or temporal consistency modeling, none of them considers how to deal with the redundancy and inconsistency in the context of Visual-QA reasoning with dense captions incorporated. Therefore, it's worthwhile to develop a Visual-QA agent to process the most informative and question-correlated dense visual captions, and skip captions when redundancy and inconsistency happens. With this effort, Visual-QA model can not only precisely localize the most valuable and concise captions from a large set of dense captions to supplement Visual-QA reasoning, but also save a lot of computation efforts to analyze uninformative and uncorrelated captions.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE TPAMI*, vol. 41, no. 2, pp. 423–443, 2018.

[3] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of IEEE CVPR*, 2015, pp. 2625–2634.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of IEEE CVPR*, 2015, pp. 3156–3164.

[5] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of IEEE CVPR*, 2015, pp. 3128–3137.

[6] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of IEEE CVPR*, 2017, pp. 7008–7024.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of ICML*, 2015, pp. 2048–2057.

[8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of IEEE CVPR*, 2018.

[9] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," *arXiv preprint arXiv:1412.4729*, 2014.

[10] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of IEEE ICCV*, 2015, pp. 4507–4515.

[11] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proceedings of IEEE CVPR*, 2016, pp. 4584–4593.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proceedings of IEEE ICCV*, 2015.

[14] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[15] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Niebles, and M. Sun, "Leveraging video descriptions to learn video question answering," in *Proceedings of AAAI*, 2017.

[16] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *Proceedings of IEEE CVPR*, 2017, pp. 2758–2766.

[17] Y. Jang, Y. Song, C. D. Kim, Y. Yu, Y. Kim, and G. Kim, "Video question answering with spatio-temporal reasoning," *IJCV*, vol. 127, no. 10, pp. 1385–1412, 2019.

[18] L. Gao, P. Zeng, J. Song, Y.-F. Li, W. Liu, T. Mei, and H. T. Shen, "Structured two-stream attention network for video question answering," in *Proceedings of AAAI*, vol. 33, no. 01, 2019, pp. 6391–6398.

[19] J. Jiang, Z. Chen, H. Lin, X. Zhao, and Y. Gao, "Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering," in *Proceedings of AAAI*, vol. 34, no. 07, 2020, pp. 11 101–11 108.

[20] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[21] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou *et al.*, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, 2016.

[22] A. Jiang, F. Wang, F. Porikli, and Y. Li, "Compositional memory for visual question answering," *arXiv preprint arXiv:1511.05676*, 2015.

[23] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "Abc-cnn: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015.

[24] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NeurIPS*, 2016, pp. 289–297.

[25] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," in *Proceedings of AAAI*, vol. 3, no. 7, 2016, p. 16.

[26] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proceedings of IEEE ICCV*, 2015, pp. 1–9.

[27] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proceedings of ICML*, 2016, pp. 2397–2406.

[28] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.

[29] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *NeurIPS*, 2014, pp. 1682–1690.

[30] M. Ren, R. Kiros, and R. Zemel, "Image question answering: A visual semantic embedding model and a new dataset," *NeurIPS*, vol. 1, no. 2, p. 5, 2015.

[31] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of IEEE CVPR*, 2016, pp. 4995–5004.

[32] P. Wang, Q. Wu, C. Shen, A. Dick, and A. Van Den Hengel, "Fvqa: Fact-based visual question answering," *IEEE TPAMI*, vol. 40, no. 10, pp. 2413–2427, 2017.

[33] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," in *IJCAI*, 2017, p. 1290–1296.

[34] M. Narasimhan, S. Lazebnik, and A. Schwing, "Out of the box: Reasoning with graph convolution nets for factual visual question answering," in *NeurIPS*, 2018, pp. 2654–2665.

[35] M. Narasimhan and A. G. Schwing, "Straight to the facts: Learning knowledge base retrieval for factual visual question answering," in *ECCV*, 2018, pp. 451–468.

[36] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few-shot learning," *arXiv preprint arXiv:1707.09835*, 2017.

[37] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[38] C. Yin, J. Tang, Z. Xu, and Y. Wang, "Adversarial meta-learning," *arXiv preprint arXiv:1806.03316*, 2018.

[39] X. Dong, L. Zhu, D. Zhang, Y. Yang, and F. Wu, "Fast parameter adaptation for few-shot image captioning and visual question answering," in *Proceedings of the ACMMM*, 2018, pp. 54–62.

[40] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*. JMLR. org, 2017, pp. 1126–1135.

[41] D. Teney and A. van den Hengel, "Visual question answering as a meta learning task," in *ECCV*, 2018.

[42] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017.

[43] T. Munkhdalai and H. Yu, "Meta networks," in *ICML*. JMLR. org, 2017, pp. 2554–2563.

[44] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *NeurIPS*, 2012, pp. 2222–2230.

[45] Y. Ye, Z. Zhao, Y. Li, L. Chen, J. Xiao, and Y. Zhuang, "Video question answering via attribute-augmented attention network learning," in *Proceedings of International ACM SIGIR*, 2017, pp. 829–832.

[46] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the ACMMM*. ACM, 2017, pp. 1645–1653.

[47] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *Proceedings of AAAI*, vol. 33, no. 01, 2019, pp. 8658–8665.

[48] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of IEEE CVPR*, 2016, pp. 4594–4602.

[49] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of IEEE ICCV*, 2017, pp. 706–715.

[50] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of IEEE CVPR*, 2018, pp. 8739–8748.

[51] T. Wang, H. Zheng, and M. Yu, "Dense-captioning events in videos: Sysu submission to activitynet challenge 2020," *arXiv preprint arXiv:2006.11693*, 2020.

[52] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *IJCV*, vol. 124, no. 3, pp. 409–421, 2017.

[53] C. Yin, J. Tang, Z. Xu, and Y. Wang, "Memory augmented deep recurrent neural network for video question answering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3159–3167, 2019.

[54] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[55] D. Teney, P. Anderson, X. He, and A. Van Den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proceedings of IEEE CVPR*, 2018, pp. 4223–4232.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[57] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proceedings of IEEE CVPR*, 2019, pp. 1999–2007.

[58] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[61] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of IEEE CVPR*, 2015, pp. 1–9.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE CVPR*, 2016, pp. 770–778.

[63] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of IEEE CVPR*, 2017, pp. 4700–4708.

[64] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[65] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[66] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of IEEE ICCV*, 2015, pp. 2425–2433.

[67] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question," *NeurIPS*, vol. 28, 2015.

[68] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *arXiv preprint arXiv:1512.02167*, 2015.

[69] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in *Proceedings of IEEE CVPR*, 2016, pp. 4976–4984.

[70] H. Noh, P. Hongsuck Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proceedings of IEEE CVPR*, 2016, pp. 30–38.

[71] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[72] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual qa," *NeurIPS*, vol. 29, 2016.

[73] K. Saito, A. Shin, Y. Ushiku, and T. Harada, "Dualnet: Domain-invariant network for visual question answering," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 829–834.

[74] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *ECCV*. Springer, 2016, pp. 451–466.

[75] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of IEEE CVPR*, 2016, pp. 21–29.

[76] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proceedings of IEEE CVPR*, 2016, pp. 4613–4621.

[77] I. Ilievski, S. Yan, and J. Feng, "A focused dynamic attention model for visual question answering," *arXiv preprint arXiv:1604.01485*, 2016.

[78] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh, "Pythia v0.1: the winning entry to the vqa challenge 2018," *arXiv preprint arXiv:1807.09956*, 2018.

[79] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, vol. 28, 2015.

[80] D. Teney, L. Liu, and A. van Den Hengel, "Graph-structured representations for visual question answering," in *Proceedings of IEEE CVPR*, 2017, pp. 1–9.

[81] W. Norcliffe-Brown, S. Vafeias, and S. Parisot, "Learning conditioned graph structures for interpretable visual question answering," *NeurIPS*, vol. 31, 2018.

[82] Z. Yang, Z. Qin, J. Yu, and Y. Hu, "Scene graph reasoning with prior visual relationship for visual question answering," *arXiv preprint arXiv:1812.09681*, 2018.

[83] Y. Cheng and C. Yuan, "Reasoning-aware graph convolutional network for visual question answering," 2019.

[84] Z. Zheng, W. Wang, S. Qi, and S.-C. Zhu, "Reasoning visual dialogs with structural and partial observations," in *Proceedings of IEEE CVPR*, 2019, pp. 6669–6678.

[85] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, "Language-conditioned graph networks for relational reasoning," in *Proceedings of IEEE ICCV*, 2019, pp. 10 294–10 303.

[86] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proceedings of IEEE ICCV*, 2019, pp. 10 313–10 322.

[87] Q. Huang, J. Wei, Y. Cai, C. Zheng, J. Chen, H.-f. Leung, and Q. Li, "Aligned dual channel graph convolutional network for visual question answering," in *ACL*, 2020, pp. 7166–7176.

[88] R. Saqur and K. Narasimhan, "Multimodal graph networks for compositional generalization in visual question answering," *NeurIPS*, vol. 33, pp. 3070–3081, 2020.

[89] D. Gao, K. Li, R. Wang, S. Shan, and X. Chen, "Multi-modal graph neural network for joint reasoning on vision and scene text," in *Proceedings of IEEE CVPR*, 2020, pp. 12 746–12 756.

[90] M. Khademi, "Multimodal neural graph memory networks for visual question answering," in *ACL*, 2020, pp. 7177–7188.

[91] W. Liang, Y. Jiang, and Z. Liu, "Graphvqa: Language-guided graph neural networks for scene graph question answering," *NAACL-HLT 2021*, p. 79, 2021.

[92] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *ICLR*, 2018.

[93] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model cnns," in *Proceedings of IEEE CVPR*, 2017, pp. 5115–5124.

[94] C. Ma, C. Shen, A. Dick, Q. Wu, P. Wang, A. van den Hengel, and I. Reid, "Visual question answering with memory-augmented networks," in *Proceedings of IEEE CVPR*, 2018, pp. 6975–6984.

[95] J. Mun, K. Lee, J. Shin, and B. Han, "Learning to specialize with knowledge distillation for visual question answering," *NeurIPS*, vol. 31, 2018.

[96] C. Wu, J. Liu, X. Wang, and R. Li, "Differential networks for visual question answering," in *Proceedings of AAAI*, vol. 33, no. 01, 2019, pp. 8997–9004.

[97] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *Proceedings of IEEE CVPR*, 2020, pp. 10 800–10 809.

[98] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, and A. v. d. Hengel, "Counterfactual vision and language learning," in *Proceedings of IEEE CVPR*, 2020, pp. 10 044–10 054.

[99] Z. Liang, W. Jiang, H. Hu, and J. Zhu, "Learning to contrast the counterfactual samples for robust visual question answering," in *Proceedings of EMNLP*, 2020, pp. 3285–3292.

[100] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual vqa: A cause-effect look at language bias," in *Proceedings of IEEE CVPR*, 2021, pp. 12 700–12 710.

[101] S. Sheng, A. Singh, V. Goswami, J. Magana, T. Thrush, W. Galuba, D. Parikh, and D. Kiela, "Human-adversarial visual question answering," *NeurIPS*, vol. 34, pp. 20 346–20 359, 2021.

[102] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu, "Joint video and text parsing for understanding events and answering queries," *IEEE MultiMedia*, vol. 21, no. 2, pp. 42–70, 2014.

[103] S. Sukhbaatar, J. Weston, R. Fergus *et al.*, "End-to-end memory networks," in *NeurIPS*, 2015, pp. 2440–2448.

[104] Z. Zhao, J. Lin, X. Jiang, D. Cai, X. He, and Y. Zhuang, "Video question answering via hierarchical dual-level attention network learning," in *Proceedings of the ACMMM*. ACM, 2017, pp. 1050–1058.

[105] K.-M. Kim, S.-H. Choi, J.-H. Kim, and B.-T. Zhang, "Multimodal dual attention memory for video story question answering," in *ECCV*, 2018, pp. 673–688.

[106] J. Kim, M. Ma, K. Kim, S. Kim, and C. D. Yoo, "Progressive attention memory network for movie story question answering," in *Proceedings of IEEE CVPR*, 2019, pp. 8337–8346.

[107] T. Yu, J. Yu, Z. Yu, and D. Tao, "Compositional attention networks with two-stream fusion for video question answering," *IEEE Transactions on Image Processing*, vol. 29, pp. 1204–1218, 2019.

[108] W. Zhang, S. Tang, Y. Cao, S. Pu, F. Wu, and Y. Zhuang, "Frame augmented alternating attention network for video question answering," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1032–1041, 2019.

[109] T. Yang, Z.-J. Zha, H. Xie, M. Wang, and H. Zhang, "Question-aware tube-switch network for video question answering," in *Proceedings of the ACMMM*, 2019, pp. 1184–1192.

[110] J. Kim, M. Ma, T. Pham, K. Kim, and C. D. Yoo, "Modality shifting attention network for multi-modal video question answering," in *Proceedings of IEEE CVPR*, 2020, pp. 10 106–10 115.

[111] K.-M. Kim, M.-O. Heo, S.-H. Choi, and B.-T. Zhang, "Deepstory: Video story qa by deep embedded memory networks," *arXiv preprint arXiv:1707.00836*, 2017.

[112] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in *Proceedings of IEEE CVPR*, 2018, pp. 6576–6585.

[113] T. Yu, J. Yu, Z. Yu, Q. Huang, and Q. Tian, "Long-term video question answering via multimodal hierarchical memory attentive networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 931–944, 2020.

[114] Z. Zhao, X. Jiang, D. Cai, J. Xiao, X. He, and S. Pu, "Multi-turn video question answering via multi-stream hierarchical attention context network." in *IJCAI*, vol. 2018, 2018, p. 27th.

[115] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proceedings of IEEE CVPR*, 2020, pp. 9972–9981.

[116] C. Lei, L. Wu, D. Liu, Z. Li, G. Wang, H. Tang, and H. Li, "Multi-question learning for visual question answering," in *Proceedings of AAAI*, vol. 34, no. 07, 2020, pp. 11 328–11 335.

[117] Z. Yang, N. Garcia, C. Chu, M. Otani, Y. Nakashima, and H. Takemura, "Bert representations for video question answering," in *Proceedings of IEEE WACV*, 2020, pp. 1556–1565.

[118] A. U. Khan, A. Mazaheri, N. D. V. Lobo, and M. Shah, "Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering," *arXiv preprint arXiv:2010.14095*, 2020.

[119] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proceedings of IEEE ICCV*, 2021, pp. 1686–1697.

[120] W. Jin, Z. Zhao, X. Cao, J. Zhu, X. He, and Y. Zhuang, "Adaptive spatio-temporal graph enhanced vision-language representation for video qa," *IEEE Transactions on Image Processing*, vol. 30, pp. 5477–5489, 2021.

[121] M. Peng, C. Wang, Y. Gao, Y. Shi, and X.-D. Zhou, "Temporal pyramid transformer with multimodal interaction for video question answering," *arXiv preprint arXiv:2109.04735*, 2021.

[122] Q. Wu, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proceedings of IEEE CVPR*, 2016, pp. 4622–4630.

[123] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of IEEE CVPR*, 2019, pp. 3195–3204.

[124] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar, "Kvqa: Knowledge-aware visual question answering," in *Proceedings of AAAI*, vol. 33, no. 01, 2019, pp. 8876–8884.

[125] Z. Zhu, J. Yu, Y. Wang, Y. Sun, Y. Hu, and Q. Wu, "Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering," *arXiv preprint arXiv:2006.09073*, 2020.

[126] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, and J. Tan, "Cross-modal knowledge reasoning for knowledge-based visual question answering," *Pattern Recognition*, vol. 108, p. 107563, 2020.

[127] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML*, vol. 2. Lille, 2015.

[128] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *NeurIPS*, 2016, pp. 3630–3638.

[129] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of IEEE CVPR*, 2018, pp. 1199–1208.

[130] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of IEEE CVPR*, 2019, pp. 7260–7268.

[131] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas, "Learning to learn by gradient descent by gradient descent," in *NeurIPS*, 2016, pp. 3981–3989.

[132] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*, 2016, pp. 1842–1850.

[133] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2017.

[134] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," in *ICLR*, 2018.

[135] A. Antoniou, H. Edwards, and A. Storkey, "How to train your maml," *ICLR*, 2019.

[136] L. Bertinetto, J. F. Henriques, P. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," in *ICLR*, 2019.

[137] A. Zadeh, M. Chan, P. P. Liang, E. Tong, and L.-P. Morency, "Social-iq: A question answering benchmark for artificial social intelligence," in *Proceedings of IEEE CVPR*, 2019, pp. 8807–8817.

[138] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proceedings of IEEE CVPR*, 2018, pp. 5561–5570.

[139] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *ECCV*, 2018, pp. 499–515.

[140] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *NeurIPS*, 2019, pp. 11 135–11 145.

[141] F. Chen, R. Ji, J. Ji, X. Sun, B. Zhang, X. Ge, Y. Wu, F. Huang, and Y. Wang, "Variational structured semantic inference for diverse image captioning," in *NeurIPS*, 2019, pp. 1929–1939.

[142] T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy parsing for image captioning," in *Proceedings of IEEE ICCV*, 2019, pp. 2621–2629.

[143] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[144] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proceedings of IEEE CVPR*, 2015, pp. 2422–2431.

[145] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *Proceedings of IEEE CVPR*, 2015, pp. 1473–1482.

[146] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of IEEE ICCV*, 2015, pp. 2407–2415.

[147] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of IEEE CVPR*, 2016, pp. 4651–4659.

[148] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proceedings of IEEE CVPR*, 2016, pp. 203–212.

[149] J. Gu, G. Wang, J. Cai, and T. Chen, "An empirical study of language cnn for image captioning," in *Proceedings of IEEE ICCV*, 2017, pp. 1222–1231.

[150] F. Chen, R. Ji, J. Su, Y. Wu, and Y. Wu, "Structcap: Structured semantic embedding for image captioning," in *Proceedings of the ACMMM*, 2017, pp. 46–54.

[151] F. Chen, R. Ji, X. Sun, Y. Wu, and J. Su, "Groupcap: Group-based image captioning with structured relevance and diversity constraints," in *Proceedings of IEEE CVPR*, 2018, pp. 1345–1353.

[152] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of IEEE ICCV*, 2017, pp. 4894–4902.

[153] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of IEEE CVPR*, 2017, pp. 375–383.

[154] X. Chen, L. Ma, W. Jiang, J. Yao, and W. Liu, "Regularizing rnns for caption generation by reconstructing the past with the present," in *Proceedings of IEEE CVPR*, 2018, pp. 7995–8003.

[155] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, "Skeleton key: Image captioning by skeleton-attribute decomposition," in *Proceedings of IEEE CVPR*, 2017, pp. 7272–7281.

[156] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *Proceedings of AAAI*, vol. 32, no. 1, 2018.

[157] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," *NeurIPS*, vol. 29, 2016.

[158] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of IEEE CVPR*, 2017, pp. 5659–5667.

[159] Y. Sugano and A. Bulling, "Seeing with humans: Gaze-assisted neural image captioning," *arXiv preprint arXiv:1608.05203*, 2016.

[160] H. R. Tavakoli, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *Proceedings of IEEE ICCV*, 2017, pp. 2487–2496.

[161] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *Proceedings of IEEE CVPR*, 2017, pp. 7206–7215.

[162] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying more attention to saliency: Image captioning with saliency and context attention," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2, pp. 1–21, 2018.

[163] S. Chen and Q. Zhao, "Boosted attention: Leveraging human attention for image captioning," in *ECCV*, 2018, pp. 68–84.

[164] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective decoding network for image captioning," in *Proceedings of IEEE ICCV*, 2019, pp. 8888–8897.

[165] Y. Qin, J. Du, Y. Zhang, and H. Lu, "Look back and predict forward in image captioning," in *Proceedings of IEEE CVPR*, 2019, pp. 8367–8375.

[166] L. Huang, W. Wang, Y. Xia, and J. Chen, "Adaptively aligned image captioning via adaptive attention time," *NeurIPS*, vol. 32, 2019.

[167] L. Wang, Z. Bai, Y. Zhang, and H. Lu, "Show, recall, and tell: Image captioning with recall mechanism," in *Proceedings of AAAI*, vol. 34, no. 07, 2020, pp. 12 176–12 183.

[168] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *ECCV*, 2018, pp. 684–699.

[169] L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu, "Aligning linguistic words and visual semantic units for image captioning," in *Proceedings of the ACMMM*, 2019, pp. 765–773.

[170] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of IEEE CVPR*, 2019, pp. 10 685–10 694.

[171] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of captions," *arXiv preprint arXiv:2006.11807*, 2020.

[172] X. Yang, H. Zhang, and J. Cai, "Learning to collocate neural modules for image captioning," in *Proceedings of IEEE ICCV*, 2019, pp. 4250–4260.

[173] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proceedings of IEEE ICCV*, 2019, pp. 8928–8937.

[174] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in *Proceedings of IEEE CVPR*, 2020, pp. 10 327–10 336.

[175] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of IEEE ICCV*, 2019, pp. 4634–4643.

[176] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of IEEE CVPR*, 2020, pp. 10 971–10 980.

[177] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of IEEE CVPR*, 2020, pp. 10 578–10 587.

[178] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," in *Proceedings of ACCV*, 2020.

[179] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of IEEE CVPR*, 2016, pp. 4565–4574.

[180] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context," in *Proceedings of IEEE CVPR*, 2017, pp. 2193–2202.

[181] X. Li, S. Jiang, and J. Han, "Learning object context for dense captioning," in *Proceedings of AAAI*, vol. 33, no. 01, 2019, pp. 8650–8657.

[182] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, "Context and attribute grounded dense captioning," in *Proceedings of IEEE CVPR*, 2019, pp. 6241–6250.

[183] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proceedings of IEEE CVPR*, 2017, pp. 317–325.

[184] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition gan for visual paragraph generation," in *Proceedings of IEEE ICCV*, 2017, pp. 3362–3371.

[185] Y. Mao, C. Zhou, X. Wang, and R. Li, "Show and tell more: Topic-oriented multi-sentence image captioning." in *IJCAI*, 2018, pp. 4258–4264.

[186] M. Chatterjee and A. G. Schwing, "Diverse and coherent paragraph generation from images," in *ECCV*, 2018, pp. 729–744.

[187] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Dense relational captioning: Triple-stream networks for relationship-based captioning," in *Proceedings of IEEE CVPR*, 2019, pp. 6271–6280.

[188] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," *IEEE TPAMI*, vol. 44, no. 2, pp. 710–722, 2019.

[189] Y. Luo, Z. Huang, Z. Zhang, Z. Wang, J. Li, and Y. Yang, "Curiosity-driven reinforcement learning for diverse visual paragraph generation," in *Proceedings of the ACMMM*, 2019, pp. 2341–2350.

[190] Z. Chen, A. Gholami, M. Nießner, and A. X. Chang, "Scan2cap: Context-aware dense captioning in rgb-d scans," in *Proceedings of IEEE CVPR*, 2021, pp. 3193–3203.

[191] V. Escorcia, F. Caba Heilbron, J. C. Niebles, and B. Ghanem, "Daps: Deep action proposals for action understanding," in *ECCV*. Springer, 2016, pp. 768–784.

[192] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-end dense video captioning with parallel decoding," in *Proceedings of IEEE ICCV*, 2021, pp. 6847–6857.

[193] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, "Streamlined dense video captioning," in *Proceedings of IEEE CVPR*, 2019, pp. 6588–6597.

[194] M. Suin and A. Rajagopalan, "An efficient framework for dense video captioning," in *Proceedings of AAAI*, vol. 34, no. 07, 2020, pp. 12 039–12 046.

[195] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue, "Weakly supervised dense video captioning," in *Proceedings of IEEE CVPR*, 2017, pp. 1916–1924.

[196] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proceedings of IEEE CVPR*, 2018, pp. 7190–7198.

[197] B. Shi, L. Ji, Y. Liang, N. Duan, P. Chen, Z. Niu, and M. Zhou, "Dense procedure captioning in narrated instructional videos," in *ACL*, 2019, pp. 6382–6391.

[198] T. Rahman, B. Xu, and L. Sigal, "Watch, listen and tell: Multi-modal weakly supervised dense event captioning," in *Proceedings of IEEE ICCV*, 2019, pp. 8908–8917.

[199] S. Fujita, T. Hirao, H. Kamigaito, M. Okumura, and M. Nagata, "Soda: Story oriented dense video captioning evaluation framework," in *ECCV*. Springer, 2020, pp. 517–531.

[200] L. Ji, X. Guo, H. Huang, and X. Chen, "Hierarchical context-aware network for dense video event captioning," in *ACL*, 2021, pp. 2004–2013.

[201] S. Chen and Y.-G. Jiang, "Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning," in *Proceedings of IEEE CVPR*, 2021, pp. 8425–8435.

[202] X. Wang and A. Gupta, "Videos as space-time region graphs," in *ECCV*, 2018, pp. 399–417.

[203] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-aware graph convolutional networks for video question answering," in *Proceedings of AAAI*, vol. 34, no. 07, 2020, pp. 11 021–11 028.

[204] P. Jiang and Y. Han, "Reasoning with heterogeneous graph alignment for video question answering," in *Proceedings of AAAI*, vol. 34, no. 07, 2020, pp. 11 109–11 116.

[205] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[206] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *Proceedings of IEEE CVPR*, 2019, pp. 12 487–12 496.

[207] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of IEEE CVPR*, 2016, pp. 4631–4640.

[208] M. Heilman and N. A. Smith, "Good question! statistical ranking for question generation," in *Proceedings of HLT*. ACL, 2010, pp. 609–617.

[209] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. ACL, 2011, pp. 190–200.

[210] M. Heilman and N. A. Smith, "Question generation via overgenerating transformations and ranking," Carnegie-Mellon Univ Pittsburgh PA Language Technologies Inst, Tech. Rep., 2009.

[211] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[212] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE TPAMI*, vol. 35, no. 1, pp. 221–231, 2013.

[213] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of IEEE CVPR*, 2014, pp. 1725–1732.

[214] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of EMNLP*, 2014, pp. 1532–1543.

[215] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini–batch gradient descent," 2012.

[216] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *NeurIPS*, 2015, pp. 3294–3302.

[217] D. Xu, https://github.com/xudejing/VideoQA, 2017.

[218] P. Lu, L. Ji, W. Zhang, N. Duan, M. Zhou, and J. Wang, "R-vqa: learning visual relation facts with semantic attention for visual question answering," in *Proceedings of ACM SIGKDD*, 2018, pp. 1880–1889.

[219] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[220] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," *arXiv preprint arXiv:1606.03126*, 2016.

[221] Y. Chen, L. Wu, and M. J. Zaki, "Bidirectional attentive memory networks for question answering over knowledge bases," *arXiv preprint arXiv:1903.02188*, 2019.

[222] G. Li, H. Su, and W. Zhu, "Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks," *arXiv preprint arXiv:1712.00733*, 2017.

[223] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.

[224] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[225] N. Tandon, G. De Melo, and G. Weikum, "Acquiring comparative commonsense knowledge from the web," in *Proceedings of AAAI*, 2014, pp. 166–172.

[226] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[227] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," *NeurIPS Autodiff Workshop*, 2017.

[228] H. Li, P. Wang, C. Shen, and A. v. d. Hengel, "Visual question answering as reading comprehension," in *Proceedings of IEEE CVPR*, 2019, pp. 6319–6328.

[229] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *NeurIPS*, 2015, pp. 2953–2961.

[230] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proceedings of IEEE CVPR*, 2017, pp. 156–165.

[231] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML*, vol. 30, 2013, p. 3.

[232] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *ICLR*, 2016.

[233] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[234] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[235] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," in *ICLR*, 2018.

[236] J. Kim, T. Kim, S. Kim, and C. D. Yoo, "Edge-labeling graph neural network for few-shot learning," in *Proceedings of IEEE CVPR*, 2019, pp. 11–20.

[237] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," *arXiv preprint arXiv:1704.03162*, 2017.

[238] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of IEEE ICCV*. PMLR, 2015, pp. 448–456.

[239] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proceedings of AAAI*, 2018.

[240] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of IEEE ICCV*, 2015, pp. 4489–4497.

[241] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[242] Y. Chen, S. Wang, W. Zhang, and Q. Huang, "Less is more: Picking informative frames for video captioning," in *ECCV*, 2018, pp. 358–373.

Chengxiang Yin

1075 Rollins Rd, Burlingame, CA
Contact: 315-243-4938 • cyin02@syr.edu

Experience

Applied Research Scientist — **Burlingame CA, USA**
Reality Labs, Meta Platforms — *04/2022 – Present*

Software Engineer Intern, Perception - CV/DL — **Fremont CA, USA**
Detection-Segmentation Group, PONY.AI — *05/2021 – 08/2021*

Research Assistant — **Syracuse NY, USA**
EECS NetLab, Syracuse University — *08/2016 – 04/2022*

Research Intern — **Beijing, China**
Intelligent-Control Group, DIDI AI Labs — *06/2019 – 01/2020*

Research Intern — **Beijing, China**
Image-Tech Group, DIDI AI Labs — *05/2018 – 08/2018*

Education

Syracuse University, NY, USA — **08/2016**
PhD Candidate of Electrical & Computer Engineering, GPA: 3.82/4.0
Research interests: Deep Learning, Computer Vision, Multi-modal Learning

Beijing Institute of Technology, Beijing, China — **09/2012**
B.S. of Electronic Science & Technology, GPA: 3.91/4.0
Research interests: Cloud Computing, Internet of Things(IoT)

Publications

ICCV 2021 (AR: 25.9%): **Chengxiang Yin**, Kun Wu, Zhengping Che, Bo Jiang, Jian Tang. Hierarchical Graph Attention Network for Few-shot Visual-Semantic Learning. In 2021 International Conference on Computer Vision (ICCV).

TMM 2020 (IF: 8.2): **Chengxiang Yin**, Jian Tang, Tongtong Yuan, Zhiyuan Xu, Yanzhi Wang. Bridging the Gap between Semantic Segmentation and Instance Segmentation. In IEEE Transactions on Multimedia (TMM), 2020.

TNNLS 2019 (IF: 14.3): **Chengxiang Yin**, Jian Tang, Zhiyuan Xu, Yanzhi Wang. Memory Augmented Deep Recurrent Neural Network for Video Question Answering. In IEEE Transactions on Neural Networks and Learning Systems (TNNLS), Vol. 31, No. 9, pp. 3159-3167, 2019.

TMC 2021 (IF: 6.1): Zhiyuan Xu, Dejun Yang, **Chengxiang Yin**, Jian Tang, Yanzhi Wang, Guoliang Xue. A Co-Scheduling Framework for DNN Models on Mobile and Edge Devices with Heterogeneous Hardware. In IEEE Transactions on Mobile Computing (TMC), 2021.

TMC 2020 (IF: 6.1): Zhiyuan Xu, Jian Tang, **Chengxiang Yin**, Yanzhi Wang, Guoliang Xue. ReCARL: Resource Allocation in Cloud RANs with Deep Reinforcement Learning. In IEEE Transactions on Mobile Computing (TMC), 2020.

JSAC 2019 (IF: 13.1): Zhiyuan Xu, Jian Tang, **Chengxiang Yin**, Yanzhi Wang, Guoliang Xue. Experience-Driven Congestion Control: When Multi-Path TCP Meets Deep Reinforcement Learning. In IEEE Journal on Selected Areas in Communications (JSAC), Vol. 37, No.6, pp. 1325-1336, 2019.

IPDPS 2019 (AR: 27.7%): Jielong Xu, Jian Tang, Zhiyuan Xu, **Chengxiang Yin**, Kevin Kwiat, Charles Kamhoua. A Deep Recurrent Neural Network Based Predictive Control Framework for Reliable Distributed Stream Data Processing. In Proceedings of the 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 262-272, 2019.

GRSL 2022 (IF: 5.3): Xiaojie Li, Jian Tang, **Chengxiang Yin**. Sequence-to-Sequence Learning for Prediction of Soil Temperature and Moisture. In IEEE Geoscience and Remote Sensing Letters (GRSL), 2022.

IIH-MSP 2015: **Chengxiang Yin**, Jin Hu, Xuejun Zhang, Xiang Xie. Advertising System Based on Cloud Computing and Audio Watermarking. In Proceedings of the 2015 IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), pp. 150-155, 2015.

Awards & Honors

Second Place of IEEE Low-Power Image Recognition Challenge (LPIRC) **02/2020**
Intelligent-Control Group, DIDI AI Labs
Object Detection Online Track

Student Travel Grant Award (NSF) for 2019 IEEE IPDPS **05/2019**
EECS NetLab, Syracuse University

Outstanding Performance with Athena Certificate Issued by DIDI **08/2018**
Image-Tech Group, DIDI AI Labs

Student Grant Awarded by CSC for Overseas Exchange **03/2016**
France, ESIEE-Amiens

First Prize/Special Award of China College Students IoT Design Contest **09/2015**
Multimedia-Tech Lab, Beijing Institute of Technology
Hosted by Texas Instruments; National Top 12 among 1000+ teams; First place in North China

Second Prize of China College Students Computer Game Contest **09/2015**
Beijing Institute of Technology
Dots and Boxes Track

Student Travel Grant Award for 2015 IEEE IIH-MSP **09/2015**
Multimedia-Tech Lab, Beijing Institute of Technology

Student Travel Grant Award for 2015 JSPS A3 Foresight Seminar **06/2015**
Multimedia-Tech Lab, Beijing Institute of Technology

First Prize of China College Students Mathematical Modeling Contest **12/2013**
Beijing Institute of Technology