Natural Language Engineering **1** (1): 1–38. Printed in the United Kingdom

From Image to Language and Back Again

A. BELZ

Computing, Engineering and Mathematics, University of Brighton
Lewes Road, Brighton BN2 4GJ, UK

T. BERG and L. YU

Computer Science, UNC Chapel Hill
Chapel Hill, NC 27599-3175, USA

(*Received 21 February 2018*)

1 Overview

Work in computer vision and natural language processing involving images and text has been experiencing explosive growth over the past decade, with a particular boost coming from the neural network revolution. The present volume brings together five research articles from several different corners of the area: multilingual multimodal image description (Frank et al.), multimodal machine translation (Madhyastha et al., Frank et al.), image caption generation (Madhyastha et al., Tanti et al.), visual scene understanding (Silberer, et al.), and multimodal learning of high-level attributes (Sorodoc et al.). In this article, we touch upon all of these topics as we review work involving images and text under the three main headings of image description (Section 2), visually grounded referring expression generation and comprehension (Section 3), and visual question answering (Section 4).

2 Image Description

Descriptive text is associated with images in a variety of different ways in the computer vision and NLP fields, in particular (i) individual lexical items associated with images or image regions (typical of image labelling), and (ii) phrases or sentences associated with regions or the image as a whole (typical of image description). Image labelling (or tagging, or indexing) goes back at least to the 1960s (Rosenfeld, 1978); its aim is to attach labels to regions that are meaningful to a human observer such that the labels capture the meaning. Image description aims to produce a summarising description, in structured natural language, of a whole image (or region), typically involving the prioritization of more important elements and relationships. This is the focus of this section, which is divided into three main subsections, on datasets (Section 2.1), models (Section 2.2), and evaluation (Section 2.3). We use the term *image description* as the name of the field, but understand it to cover the automatic generation of any structured text intended to convey the content of an image. We argue below that different image text types can most meaningfully be defined relative to a real-world application context.

Table 1. *Image description datasets. EL=elicited; NO=naturally occurring; YT=there is a clear application task; NT=no task; G=German; E=English; S=Spanish; C=Mandarin; T=Turkish; F=French; D=Dutch; J=Japanese.*

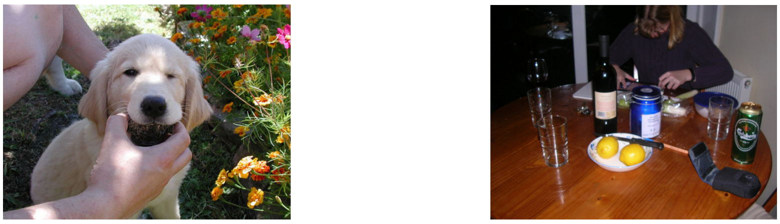
Name	Attribution	Images	Notes	Language(s)
IAPR-TC12	Grubinger et al., 2006	20,000	EL, YT	G, E, S
BBC News	Feng & Lapata, 2008	3,361	NO, YT	E
Pascal1K	Rashtchian et al., 2010	1,000	EL, NT	E
SBU1M Captions	Ordonez et al., 2011	1,000,000	NO, YT	E
VLT2K	Elliott & Keller, 2013	2,424	EL, NT	E
Abstract Scenes	Zitnick & Parikh, 2013	10,020	EL, NT	E
Sentences3D	Kong et al., 2014	1,449	EL, YT	E
Flickr8K	Hodosh & Hockenmaier, 2013	8,092	EL, NT	E
→	Li et al., 2016	=	=	E, C
→	Unal et al., 2016	=	=	E, T
Flickr30K	Young et al., 2014	31,783	EL, NT	E
→	Elliott et al., 2016	=	=	E, G
→	Elliott et al., 2017	=	=	E, G, F
→	van Miltenburg et al., 2017	2,014	=	E, D
Déjà Captions	Chen et al., 2015	4,000,000	NO, NT	E
MSCOCO	Lin et al., 2015	164,062	EL, NT	E
→	Yoshikawa et al., 2017	=	=	E, J
→	Miyazaki & Shimizu, 2016	26,500	=	E, J
→ MMT17-Test2	Elliott et al., 2017	461	=	E, G, F
MS SIND	Huang et al., 2016	81,743	EL, NT	E
Visual Genome	Krishna et al., 2017	108,077	EL, NT	E
MMT17-Test1	Elliott et al., 2017	1,071	EL, NT	E, G

2.1 Data for Image Description Tasks

2.1.1 Datasets

Table 1 provides an overview of image description datasets in terms of number of images, language(s) the descriptions are in, whether there is an explicit or implied real-world application task (e.g. news article image captioning), and whether they were elicited from contributors, or collected from sources where they occur naturally.

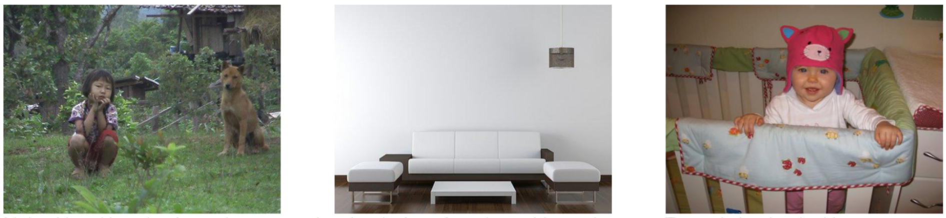
The **IAPR-TC12** benchmark (Grubinger et al., 2006a) has 20,000 images from a travel company's photo collection each with text captions in German, English, and Spanish. The dataset was intended for benchmarking retrieval systems in ImageCLEF 2006. Images depict a wide range of travel-related topics, including sport, landmarks, animals, group shots, landscapes, etc. In contrast to other datasets reviewed here, the collection contains sets of images that depict very similar content (e.g. the same cathedral), but from different angles, dates, etc. Original annotations by the travel company were quality-checked, corrected and completed by direct contributors (not crowdsourced). E.g. *a photo of a brown sandy beach; the dark blue sea with small breaking waves behind it; a dark green palm tree in the foreground on the left; a blue sky with clouds on the horizon in the background.*



A man holds a ball in a puppies mouth.
A puppy bites a ball.
Someone is putting something in the white dog's mouth.
A tan puppy with a hand holding something in his mouth.
A small puppy being fed a chocolate treat.

Woman at table busy with something
A woman by the table preparing drinks.
A woman at the dining table with wine, beer, and lemons.
a woman at a dinner table writing on her notebook
A woman sits with her head down at a table that has alcohol beverages and accessories on it.

Fig. 1. Two images from Pascal1K; original spelling errors (Rashtchian et al., 2010).



Little girl and her dog in northern Thailand. They both seemed interested in what we were doing

Interior design of modern white and brown living room furniture against white wall with a lamp hanging.

Emma in her hat looking super cute

Fig. 2. Image and caption examples from SBU1M.

The **BBC News Database** (Feng and Lapata, 2008) contains 3,361 image-caption-document tuples collected from the BBC News website. Captions are often non-descriptive, e.g. *Breastfed babies tend to be brighter* for an image showing a baby being breastfed. The implicit image description task is news image caption generation, but Feng & Lapata use the data for image labelling.

For **Pascal1K**, Rashtchian et al. (2010) used Mechanical Turk to collect five descriptions each for 1,000 VOC'08 images (50 selected randomly from each of the 20 VOC'08 classes). Contributors had to have high HIT rates and pass a language competence test, leading to relatively high text quality with few grammatical or spelling mistakes. Two example images and their descriptions are shown in Figure 1.

The **SBU1M** team collected one million Flickr images with naturally occurring captions (Ordonez et al., 2011), filtering initial search results to retain only images with captions containing at least 2 words from the original query, and at least one preposition (indicating visible spatial relationships). For examples see Figure 2.

For **VLT2K**, Elliott & Keller (2013) used the images from the VOC'11 action recognition taster competition (Everingham et al., 2011), and collected three descriptions per image via Mechanical Turk. Subsequent annotation steps added visual dependency relations, and associated image regions with descriptions.

The **Abstract Scenes** dataset (Zitnick and Parikh, 2013) consists of 1,002 sets of 10 similar abstract scenes and one associated description. Mechanical Turk contributors created individual scenes of children playing using clip art. Other contributors described the scenes using 1–2 sentences. Finally, contributors created 9 more scenes to match each description. This dataset differs from the others in its use of cartoon-like scenes in which physical properties can be unrealistic.

The **Sentences3D** team (Kong et al., 2014) collected descriptions and annotations for the 1,449 photoss of indoor scenes in the NYU-RGBD v2 dataset via Mechanical Turk. Descriptions vary from one to ten sentences, and tend to be complex with multiple mentions of visual objects. Additional annotations (by direct contributors) link nouns and pronouns to the visual objects they describe.

Flickr8K has 8,092 images of people/animals performing some action from six Flickr groups (Hodosh et al., 2013). Five descriptions per image were collected via Mechanical Turk; QA measures were e.g. a spelling/grammar test, and location in the US. Contributors were asked to write single sentences describing the depicted scenes, situations, events and entities. This dataset was extended in **Flickr30K** (Young et al., 2014) to 31,783 images. As a further extension, **Multi30K** (Elliott et al., 2016) added 31,014 German translations of the original English descriptions (one per image), and 155,070 German original image descriptions (five per image).

Extensions of Flickr30K to other languages exist. Van Miltenburg et al. annotated 2,014 images from the validation and test parts of Flickr30K with five Dutch descriptions each via Crowdflower, using the same collection regime (van Miltenburg et al., 2017). Unal et al. collected Turkish descriptions for Flickr8K, again using the same regime (Unal et al., 2016). Li et al. extended the dataset to Chinese, creating Mandarin captions by (i) machine translating the original descriptions with Google and Baidu, and (ii) crowdsourcing new descriptions (Li et al., 2016).

Lin et al. collected two sets of image descriptions for the **MS COCO** corpus of 2.5 million labeled objects in 328,000 images, one containing 5 descriptions for every image in the training, validation and test sets; the other having 40 descriptions each for a random subset of 5,000 test set images (Lin et al., 2014a). The latter were collected with the aim of achieving higher correlation with human judgments in automatic evaluation via a large number of reference descriptions.

MMT-Test2 (which the MMT team call the Ambiguous COCO test data) is a collection of 461 MS COCO images selected for containing an ambiguous verb (56 verbs in total), in a complex process (Elliott et al., 2017) that involved information from the VerSe dataset of ambiguous-verb captions (Gella et al., 2016).

The **STAIR Captions** dataset (Yoshikawa et al., 2017) is an extension of MS COCO to Japanese, with 5 descriptions for each MS COCO image, obtained with slightly different instructions, using crowdsourcing and direct contributions. An earlier Japanese MS COCO extension for a subset of 26,500 images crowdsourced 3–5 Japanese descriptions per image, again using a slightly different collection regime, including a caption quality filtering step at the end (Miyazaki and Shimizu, 2016).

The **Déjà Captions** team collected 760 million image/text pairs from Flickr, using 693 frequent nouns for queries (Chen et al., 2015a). They segmented texts into sentences and filtered out those that did not contain the query term. Only captions which very closely resembled at least one other caption for a different image were then retained. The result was a collection of 180K unique captions for

4 million images. As with the Abstract Scenes dataset, there are multiple images per caption, whereas with other datasets in this section it is the other way round.

MS SIND (Huang et al., 2016) is a dataset of story-like image sequences paired with: (1) descriptions for each image in isolation, (2) descriptions for each image when seen in a sequence, and (3) descriptions that form a narrative over an image sequence (images/sentences aligned). Image sequences were obtained from Flickr albums, only retaining 'storyable' albums with 10–50 photos, taken within 48 hours.

The **Visual Genome** dataset (Krishna et al., 2017b) has region descriptions (in addition to six other annotation components) for 108,077 images, e.g. for an image with three regions: *man jumping over a fire hydrant, yellow fire hydrant,* and *woman in shorts is standing behind the man.*

MMT-Test1 (a.k.a. Multi30K 2017 test data) is a new dataset of images/texts collected from some of the same Flickr groups as Flickr30K, and some new groups (Elliott et al., 2017) in a multi-step process, resulting in a final set of 1,071 images/texts, each supplemented by one professional German translation, and five crowdsourced German descriptions.

The datasets reviewed in this section differ on many dimensions, including size, ranging from a few thousand images (Pascal1K, BBC News, VLT2K) to a million and more (SBU1M, Déjà Captions). English remains the most frequent language, but other languages are being seen more frequently, mostly as extensions of English datasets. The images in all but one dataset (Abstract Scenes) are photos, mostly user-generated (except BBC News). In some cases, labelled object bounding boxes or region masks (VLT2K, MS COCO, Sentences3D, Visual Genome) around objects are available. Most datasets have image texts elicited from contributors for the specific purpose of creating the corpus, but some, including the very large datasets, have naturally occurring image texts (BBC News, SBU1M, Déjà Captions).

2.1.2 Collecting Human-generated Image Descriptions

Quality assurance measures, instructions and guidelines to contributors when eliciting image descriptions can vary substantively between datasets. The IAPR TC-12 descriptions were intended to describe "what can be recognized in an image without any prior information or extra knowledge" (p. 6). The creators decided not to ask for full sentences, or for descriptions of the entire image, specifically to thwart people's natural storytelling tendencies. They did not constrain the number of phrases that could be used or their order, and considerable variation can be seen in both. A typical example is: *a brown cathedral with two towers and three green doors; a square with street lamps, green spaces, flowers, a tree, benches and people in front of it; grey cobblestones in the foreground; a hill and clouds in the background.*

For VLT2K, Elliot & Keller placed similar restrictions on contributors, asking them to describe an image in two sentences, the first describing the action in the image, the person performing the action and the region involved in the action; the second describing any other regions in the image not directly involved in the action; e.g. *A man is riding a bike down the road. A car and trees are in the background.*

For most datasets, however, the only structural restriction is that descriptions should have one or two sentences describing the whole image. This allows a wide variety of style and focus which researchers seek to control by lists of DOs and DON'Ts which can be detailed. For example, for MS COCO:

- Please describe the image
- Describe all the important parts of the scene.
- Do not start the sentences with "There is.
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentences should contain at least 8 words.

Looking at image descriptions in datasets reveals that contributors do not always follow such instructions, producing descriptions such as: *An empty boat begs to be used*; *The happy lady enjoys her surroundings*; *Take a solitude horse ride in the beautiful country*; and *The curious dog looks to do some damage to the pots*. It appears that more rigorous control, as e.g. for IAPR-TC12 and VLT2K, is needed to constrain people to producing descriptions that describe only what can be seen.

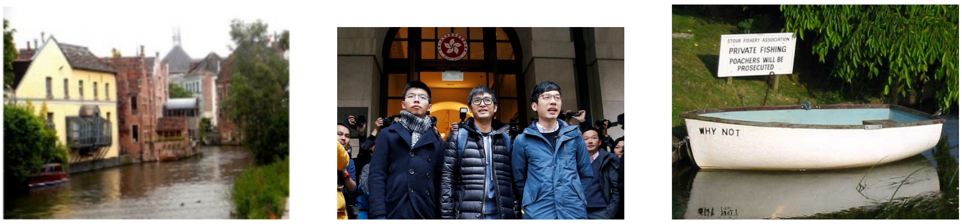
2.1.3 How Humans Describe Images

Human-authored image descriptions tend to prioritize mention of foregrounded and/or large entities, their attributes (color, size, etc.), and relationships linking them, to each other and to their surroundings. However, human authors have strong tendencies to add many different kinds of conjectured content, attributing emotions and intent to people and animals, placing the image in the context of a story, or ascribing subjective properties to image elements. The examples in Figure 1 exhibit several forms of conjecture. For the picture on the left, it is unclear whether the object in the dog's mouth is a chocolate treat, a ball, or something else. Is the object being put into, held in, or in fact retrieved from, the dog's mouth? Is it a puppy or a grown dog? Is its colour white or tan? For the picture on the right, is the woman working on her notebook, preparing drinks, or is she busy with something unidentifiable? In the (naturally occurring) captions of the images in Figure 2 proper names, subjective attributes, and attribution of state of mind are all used.

From guessing emotional states to being more precise than the information in an image permits, people have a tendency to fill in the missing bits, to tell a story. Moreover, they do this in a myriad of different ways. On the one hand, humans have these tendencies, on the other hand researchers try to quell them and elicit descriptions that only talk about what can be seen in an image, moreover only what is 'important'. This strong pull between what people come up with when asked to describe an image, and what researchers try to get them to do, raises questions about whether this is a good way to collect training and evaluation data.

2.1.4 Human-generated Image Descriptions as Training and Evaluation Data

The datasets above are used to train the methods in Section 2.2, and as reference data by many of the evaluation methods in Section 2.3. Systems are trained



Sail on by.

The pro-democracy activists Joshua Wong, Alex Chow and Nathan Law outside the Court of Final Appeal in Hong Kong on Tuesday.

An empty boat begs to be used.

Fig. 3. From left: Image and caption from Déjà Captions; news image from New York Times, Feb 6 2018, 11:55 (https://www.nytimes.com/2018/02/06/world/asia/hong-kong-joshua-wong-appeal.html); image and elicited description from Pascal1K.

to produce similar image descriptions to those in these datasets, and the image descriptions they generate are considered good if they are similar to those in the datasets, yet there is a lack of clarity in the field regarding both (i) what these image texts are, and (ii) what they are meant to be *for*. Regarding the former, the main distinction drawn is between *descriptions* and *captions*. For example, Bernardi et al. (2016) distinguish descriptions which “verbalize visual and conceptual information depicted in the image, i.e., descriptions that refer to the depicted entities, their attributes and relations, and the actions they are involved in” (p. 4), and captions which “typically [...] verbalize information that cannot be seen in the image [providing] personal, cultural, or historical context for the image” (p. 18). Similarly, Frank et al. in this volume “define descriptions as sentences that are solely and literally about an image, whereas captions are more naturalistic sentences associated with, but not necessarily descriptions of, an image” (p. 3).

A text accompanying an image in a real-world context (e.g. a caption, article, title, alt text) can normally be unambiguously assigned to a category. Take away the context however, and it is far less clear what category a text belongs to. In Figure 3 for example, the Flickr caption on the left makes no reference to anything visible in the image; the text in the middle is a caption from a news website, and is highly descriptive; the text on the right was elicited for Pascal1K as a description, but is very ‘caption-like’. All examples in Figure 2 are naturally occurring captions, but the first sentence on the left, and the whole caption in the middle, neatly fit both of the definitions of descriptions above.

The question is, does it make sense to say that a text that naturally occurs as a caption is not a caption because it does not fit some definition of captions? It seems more practical to say that a text is a caption because it appears in a particular place alongside an image, regardless of its textual properties, i.e. to tie the definition to application context. Systems trained on naturally occurring captions have this real-world grounding by default, and an implied application task: to generate the kinds of texts normally seen as captions in the particular context data was collected from.

Image description generation systems do not have this real world grounding:

there is no standard definition of what a description is, and there are no naturally occurring image texts unambiguously identifiable as descriptions. This has two implications: (i) for data collection: there is no obvious way to constrain the kinds of texts that should be elicited from contributors; and (ii) for evaluation: because elicited descriptions are used for both training and evaluation where systems are deemed good in proportion to the similarity of their output to the elicited descriptions, the result is a closed system in which questions of what collected tests are meant to be good *for*, and whether they are in fact good for it, are not directly addressed at all (Belz, 2009). This is why real-world grounding is needed: an explicitly stated application context would address both of these questions, an issue which we will pick up again in the section on extrinsic evaluation below (Section 2.3.3).

2.2 Image Description Methods

A basic division in image description is between (i) methods that create descriptions for new images from scratch, and (ii) methods that retrieve similar image/description pairs from the training data, and use those to create a description for a new image. The latter are a form of memory-based learning, known as **retrieval-based** methods in image description. These subdivide into methods that assess the similarity of new cases with known cases in visual space, and generate descriptions in textual space (Vinyals et al., 2015; Chen and Zitnick, 2015; Karpathy et al., 2014; Hodosh et al., 2013); and those, now the more common, which involve some form of joint modelling of the visual and textual spaces (Yagcioglu et al., 2015; Mason and Charniak, 2014; Gupta et al., 2012; Ordonez et al., 2011).

Methods that create a new description for a given image from scratch, often called **generative** methods (Ortiz et al., 2015; Lin et al., 2014a; Fang et al., 2015; Elliott and de Vries, 2015), tend to have the following component steps: (1) Image analysis, sometimes broken down into (a) identification of type and, optionally, location of, objects and background/scene in the image, and (b) detection of attributes, relations and activities involving objects from Step 1; and (2) generation of a word string from a representation of the output from Step 1. Sometimes, a third, re-ranking step is added. The distinguishing difference between the two types of approaches is that retrieval-based approaches must consult a memory bank of training instances during application, whereas generative approaches create models that abstract away from the individual instances seen during training, generalize over them, and are usually in some respect more effective and/or efficient than consulting training instances individually during application.

The above division is into two contrasting paradigms, broad-strokes outlines of general approach, which do not imply specific techniques to implement them. In the next section, we select a small number of reference papers, provide a detailed description of the methods presented in them, and describe a set of paradigmatically similar methods in relation to them. In Section 2.2.3, we briefly highlight some current trends in the field. Given that a very recent survey reviews a large cross-section of image description methods in detail (Bernardi et al., 2016), we do not aim to provide an exhaustive survey of image description papers here.

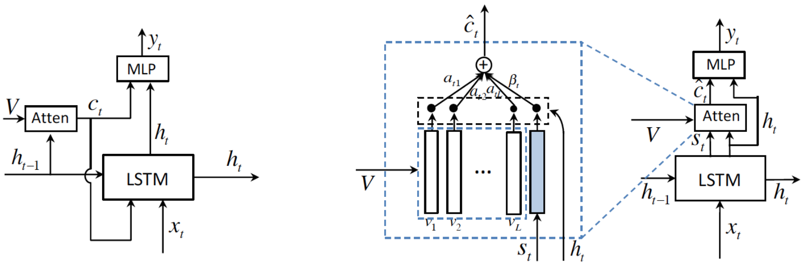
2.2.1 Generative approaches

As laid out in more detail above, generative methods start with some form of image analysis, mapping images to representations that encode information intended to be more useful or efficient for generating descriptions than the raw pixel-grid values. These may be readily interpretable by humans (symbolic representations of objects, attributes, relations, 'stuff', etc.), or not (vectors of real numbers). For Step 1a, some systems identify labelled regions (Farhadi et al., 2010; Yatskar et al., 2014; Kulkarni et al., 2011), others directly map images to words (Fang et al., 2015). Step 1b determines object attributes (Yatskar et al., 2014; Kulkarni et al., 2011), spatial relationships (Yang et al., 2011; Elliott and Keller, 2013; Muscat and Belz, 2017), activities (Yatskar et al., 2014; Elliott and Keller, 2013), etc. In Step 2, systems differ in linguistic knowledge brought to bear on the generation process. Some view the task as linearising labels, relations and attributes from Step 1 (Fang et al., 2015; Li et al., 2011); others slot them into templates (Yang et al., 2011; Elliott and Keller, 2013; Kulkarni et al., 2011), yet others use grammar to construct descriptions (Mitchell et al., 2012; Kuznetsova et al., 2014). Some approaches (Fang et al., 2015; Wang et al., 2017) add a final re-ranking step, e.g. the latter uses CIDEr (see Section 2.3.2) to calculate a 'consensus evaluation score' between candidate captions and their nearest neighbours retrieved via a cross-modal embedding space.

The standard architecture that has emerged for generative image description comprises an *encoder*, usually a CNN (convolutional neural network), which maps images to more efficient and/or more task-suitable representations of themselves, and a *decoder*, an RNN (recurrent neural network) or LSTM (an RNN with long short-term memory), which maps the new representations to descriptions. In a typical example of this approach, Lu et al. (2017) use the last convolutional layer of a ResNet with dimensionality $2048 \times 7 \times 7$ to produce encodings, obtaining a global image feature vector as the normalised sum over the spatial CNN feature vectors at each of the k grid locations. The decoder is a single layer LSTM with hidden vector size 512 which takes as input the global image feature vector from the CNN stage concatenated with the current word embedding vector, and produces a prediction of the next word as output. During training CIDEr is used to assess progress.

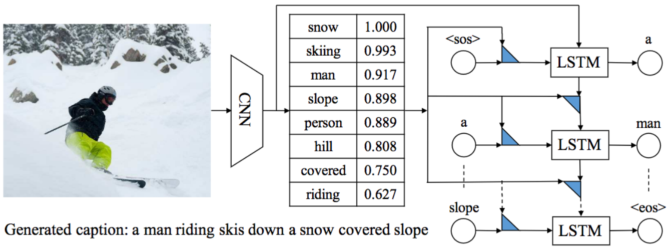
An increasingly common addition to this basic architecture is a visual attention mechanism which typically produces a spatial map that identifies the specific image region(s) most relevant to the current word prediction task (Karpathy and Fei-Fei, 2015; Xu et al., 2015). Lu et al.'s (2017) contribution is a version that only switches on when needed, based on the insight that non-visual words such as determiners, as well as other words in contexts where the predictive power of the preceding word(s) is particularly strong, do not benefit from visual attention. The key idea is that the model learns to extract a 'visual sentinel' vector from the decoder's memory of visual and linguistic information; an adaptive context vector is modeled as a mixture of the spatially attended image features and the visual sentinel vector, the latter controlled by a weight called the 'sentinel gate'. The diagrams below show the standard attention architecture (left) in comparison with Lu et al.'s adaptive extension (right), where $V = [v_1, ..., v_L]$ are the spatial image features at time t,

$a_{t,1}, ...a_{t,L}$ the attention weights, h_t the hidden state, s_t the visual sentinel vector, β_t the sentinel gate, c_t the context vector, and $\hat{c}_t$ the new adaptive context vector):



Karpathy & Fei-Fei (2015), in the image analysis step, detect objects with a Region CNN, pre-trained on ImageNet and finetuned on the 200 ImageNet classes. They use the top 19 detected locations as well as the whole image, and compute representations (sets of vectors) based on the 19 bounding boxes (region-based embedding). They obtain a word-based embedding in the same space with a bi-RNN, and compute pairwise similarities between individual region and word vectors as their inner products. They then obtain an alignment that pairs multiple words to single regions with a Markov Random Field. The resulting single-region/multi-word alignments are used in Step 2 which outputs a list of snippets for identified regions.

Gan et al. (2017b) also use a standard CNN-LSTM set-up, but extend each weight matrix of the conventional LSTM to an ensemble of tag-specific weight matrices (blue triangles below). The degree to which each member of the ensemble is used to generate a caption is tied to the image-dependent probability of the corresponding tag. The following diagram presents the generation process in outline:



2.2.2 *Retrieval-based approaches*

Gupta et al.'s (2012) description generator is an archetypal example of a retrieval-based approach, and comprises the following five steps:

1. Extract image features: RGB and HSV histograms for colour; Gabor and Haar descriptors for texture; GIST for scene; SIFT for shape. Feature extraction is repeated (except GIST) for 3 vertical and 3 horizontal image slices. Finally, vectors are concatenated into a single feature vector for each feature type.
2. Retrieve k nearest images: compute distance between image feature vectors, using L_1 distance for colour vectors, L_2 for texture and scene, χ^2 for shape. Image distance is then the dot product of distance weights and feature vectors.

3. Parse the descriptions of the k most similar images, using the Stanford dependency parser; extract object 1-tuples (subjects and objects), attribute/object 2-tuples (attribute+subject, attribute+object), action 2-tuples (e.g. verb+subject), and relation 3-tuples (e.g. verb+preposition+object) from the dependency parse.
4. Compute a probability score for each candidate tuple (any tuple derived from one of the k retrieved descriptions) on the basis of relative image similarity (compared to the other $k - 1$ most similar images) and relative Google frequency (compared to the other candidate tuples). Tuples are 'integrated' by slotting them into a predefined tripartite syntactic template.
5. Score the resulting 'triples' with the joint probability of their component tuples. Depending on the dataset, the top-scoring triple or the syntactically aggregated top 3 triples are passed to SimpleNLG for surface realisation.

One of two seminal papers in the retrieval-based area, Ordonez et al. (2011) present a simpler method that uses GIST and tiny-image for Step 1, and the sum of GIST similarity and tiny-image colour similarity for Step 2. Following re-ranking of the most similar images, Steps 3–5 are trivial as the description of the top image is simply transferred as the output description. Kulkarni et al. (2011) and Yang et al. (2011) use approaches similar to Gupta et al. for Step 3, but apply different syntactic templates in Step 4. Some techniques are familiar from generative approaches, e.g. Yagcioglu et al. (2015) use encodings produced by a CNN trained on ImageNet for Step 1. Mason and Charniak (2014) construe Step 4 as multi-document extractive summarisation over the retrieved descriptions.

The above methods do not involve representations in a shared visual-textual space. Other retrieval-based methods, in addition to image similarity, also assess the match between possible descriptions and the input image. For example, Farhadi et al. (2010), in the original retrieval-based method, map both images and descriptions to $< object, action, scene >$ triples, using small multi-label Markov random fields. They consider the top k triples predicted for images and descriptions, and compute a rank-based similarity measure to select the description to be transferred.

Hodosh et al. (2013) construe image description explicitly as a matter of ranking candidate descriptions, and the natural inverse of image retrieval, best implemented by a uniform approach. They focus on the problem of learning an appropriate mapping between images and descriptions for which they use Kernel Canonical Correlation Analysis with a wide range of different image and text kernels. Learned projection weights map KCCA image and description vectors to an induced shared space in which images are expected to appear nearer sentences they are more strongly associated with (i.e. that describe them well). Candidate descriptions are ranked in order of their cosine similarity in this space with the new image to be described.

2.2.3 Some recent trends

Attention mechanisms have been garnering increasing interest as additions to encoder-decoder architectures for image description (Xu et al., 2015; You et al., 2016; Lu et al., 2017), with extensions to the basic mechanism emerging. For example, You et al. selectively attend to candidate semantic concepts, fusing them into hidden states and outputs. Lu et al. (see above) introduce a selective visual attention mechanism that switches off when not needed.

Another trend is region-based image description (Karpathy and Fei-Fei, 2015; Krishna et al., 2017a; Kinghorn et al., 2018). E.g. the latter use a regional object detector and RNN-based attribute prediction in addition to encoder-decoder language generation, e.g. performing well at cross-domain generalisation.

There is growing interest in incorporating high-level concepts into neural architectures, rather than relying on lower-level image features alone. One approach trains a CNN classifier for each attribute (word) in the training descriptions (Wu et al., 2017); the resulting set of attribute likelihoods for an image is viewed as a high-level representation of its content. An RNN then generates captions on the basis of the attribute likelihoods. Similarly, Gan et al. (2017) compute tags (words) from images, and use the probability of each tag to compose the parameters in an LSTM (see Section 2.2.1).

More generally, bringing linguistic knowledge into neural-based image description is being explored. One approach uses dependency trees to embed sentences for image retrieval (Socher et al., 2014); another (Venugopalan et al., 2016) integrates a neural LM and distributional semantics obtained from large text corpora into an LSTM for video description. The ACL 2018 Workshop on Relevance of Linguistic Structure in Neural Architectures for NLP is a sign of growing interest.

Other recent developments are generating captions with creativity (Chen et al., 2015a), sentiment (Mathews et al., 2016), and humorous/romantic/plain styles (Gan et al., 2017a); unsupervised learning of image-to-text mappings (Hendricks et al., 2016); and generating paragraph-long descriptions (Krause et al., 2016).

2.3 Evaluation of Image Description Methods

A range of evaluation methods have been used in image description. Using the taxonomy developed in previous work (Belz and Hastie, 2014), we distinguish the following method categories. *Intrinsic* measures assess properties of systems or components in their own right, for example comparing their outputs to model outputs in a corpus, whereas *extrinsic* measures assess the effect of a system on something that is external to it, for example human performance at a given task or the value added to an application. One subcategory of intrinsic methods are *output quality measures* which can be either *automatically assessed* or *human-assessed.* Subcategories of extrinsic measures are *user task success measures* which assess impact on users' ability to perform a given task, and *system purpose success measures which* assess impact on a system's achievement of (an aspect of) its stated purpose.

By far the most common evaluation measures in image description are intrinsic assessments of output quality. Both automatic and human-assessed measures have been used, and we assess each of those in turn below (Sections 2.3.2 and 2.3.1). In Section 2.3.3 we briefly review the few extrinsic measures in the field.

2.3.1 Intrinsic human-assessed output-quality measures

Human assessment of the quality of generated outputs in image description tends to take the form of asking participants, mainly on crowdsourcing platforms, to answer questions about aspects of the texts, by selecting a score on a verbal descriptor scale of 1–3 or 1–5 where each number is accompanied by an explanatory bit of text. For example, Elliot & Keller crowdsourced five judgments each for 101 image/description pairs, using three criteria assessed on scales of 1–5:

1. *Grammaticality*: give high scores if the description is correct English and doesn't contain any grammatical mistakes.
2. *Action:* give high scores if the description correctly describes what people are doing in the image.
3. *Scene*: give high scores if the description correctly describes the rest of the image (background, other objects, etc).

Gupta et al. (2012) collected human judgements on 100 and 500 images from the Pascal and IAPR TC-12 datasets, respectively, using rating criteria of Readability and Relevance, and scales from 1–3, adopted from Li et al. (2011).

The Readability and Grammaticality criteria above seek to assess if a text is the kind of text a native speaker would produce (most commonly called 'Grammaticality'); the other criteria address aspects of what is called Adequacy in MT, in this context the appropriateness of the text for the image. **Grammaticality** (e.g. Kulkarni et al., 2011; Li et al., 2011; Yang et al., 2011; Gupta et al., 2012; Kuznetsova et al., 2012; Mitchell et al., 2012; Elliott and Keller, 2013; Hodosh et al., 2013) and **Adequacy** (e.g. Li et al., 2011; Yang et al., 2011; Gupta et al., 2012; Kuznetsova et al., 2012; Mitchell et al., 2012; Elliott and Keller, 2013) are the two most common criteria used in the field. Other criteria have been used, for example **Creativity** (Li et al., 2011), and **Human-likeness** (Mitchell et al., 2012).

The 2015 COCO Image Captioning Challenge took a different approach. Here, texts generated by all 15 competing systems, plus human and random texts, were assessed on five criteria; scores were derived either from verbal descriptor scale judgments, or the assessors' response was converted to a percentage, as follows:

1. *Overall caption quality*:
   (a) Percentage of captions evaluated as better or equal to human caption.
   (b) Percentage of captions that pass the Turing Test.
2. *Correctness*: Average correctness of the captions on a scale 1–5 (incorrect–correct).
3. *Detailedness*: Average detail of the captions from 1–5 (lacking details–very detailed).
4. *Saliency*: Percentage of captions that are similar to human description.

Two criteria are assessed on verbal descriptor scales as above, with Correctness a form of Adequacy. However, with the other criteria the organisers made an attempt to reduce subjectivity and variability in judgments by making them comparative.

Reporting of human-assessed evaluation experiments in non-competition contexts in the field is frequently patchy, omitting crucial details such as how many evaluators were used, who they were, or reporting statistical significance assessments without giving the method used for the assessment. Human assessment is notoriously hard to reproduce and compare across experiments even where those involve the same data; an established standard framework of assessment criteria, experimental design and contributor recruitment would go some way towards addressing this.

2.3.2 Intrinsic automatic output-quality measures

The main automatic metrics for assessing output quality that have been used in image description are BLEU and Meteor from machine translation, ROUGE from summarisation, and CIDEr and SPICE which were specifically developed for evaluation of image descriptions. Figure 2.3.2 presents an overview of metrics, the field they originated in, when they were introduced, and a sample of papers they have been used in. Below, we briefly summarise the metrics developed for image description (assuming the other three are well enough known).

CIDEr (Vedantam et al., 2014) differs from other n-gram metrics such as BLEU mainly in that it assigns lower weights to n-grams that are common to reference image descriptions (using tf-idf).

SPICE (Anderson et al., 2016) starts by dependency-parsing the generated sentence and the reference sentences, then maps the result to a 'scene graph' of objects, relations and object attributes. It constructs the union of scene graphs for the reference sentences, then turns both the graph for the generated sentence, and the union-graph for the reference sentences into tuple sets comprising 1-tuples for objects, 2-tuples for attributes, and 3-tuples for relations. Finally, Recall, Precision and F-score are computed on the two tuple sets.

Most recently, Kilickaya et al. have proposed the use of the word mover distance (WMD) document similarity metric for image description (Kilickaya et al., 2017). WMD is similar in spirit to edit-distance metrics and computes the distance between generated text and reference text on the basis of the Euclidean distance between word2vec embeddings of words used as the cost of replacing one word with another.

Other metrics have been used, e.g. where a system produces ranked outputs, model performance can be measured by the rank of the original image or caption in the ranked list of outputs, e.g. R@k (Recall at k) is the percentage of queries for which the correct response was among the first k results; median rank of the correct response in the ranked list of results is also used (Hodosh et al., 2013).

Some research has shown Meteor to correlate well with human judgments in this field (Huang et al., 2016). The paper that introduced CIDEr (Vedantam et al., 2014) found that the latter outperformed Meteor in most cases, but by a small margin. Evaluated on the 2015 COCO Challenge test data and human judgments for all 5 assessment criteria (see previous section for details), SPICE was shown (Anderson et al., 2016) to correlate far better with the human judgments than any of the other metrics discussed above in terms of Pearson's r, with extremely high values for r except for detailedness which it clearly is not suitable for. WMD has not been shown to clearly outperform SPICE (Kilickaya et al., 2017).

The aim of meta-evaluation is often presented as determining which metric is best at predicting human judgment, not which metric is best at assessing *a specific criterion* (best = strongest correlation with human assessments of the same criterion). Clearly, the metrics in this section are not suitable for assessing how detailed

Table 2. *Intrinsic output-quality metrics that have been used in image description.*

Metric	Origin	Examples of use
BLEU-n	2002, MT	(Farhadi et al., 2010; Kulkarni et al., 2011; Yang et al., 2011; Li et al., 2011; Ordonez et al., 2011; Gupta et al., 2012; Elliott and Keller, 2013; Hodosh et al., 2013; Karpathy et al., 2014; Kuznetsova et al., 2014; Devlin et al., 2015; Huang et al., 2016; Wu et al., 2017; Lu et al., 2017; Gan et al., 2017b; Dai et al., 2017; Kinghorn et al., 2018)
ROUGE	2004, Sum	(Yang et al., 2011; Gupta et al., 2012; Hodosh et al., 2013; Fang et al., 2015; Gan et al., 2017b; Wu et al., 2017; Dai et al., 2017; Kinghorn et al., 2018)
Meteor	2005, MT	(Yang et al., 2011; Karpathy et al., 2014; Kuznetsova et al., 2014; Chen and Zitnick, 2015; Devlin et al., 2015; Elliott and de Vries, 2015; Fang et al., 2015; Jia et al., 2015; Karpathy and Fei-Fei, 2015; Ortiz et al., 2015; Vinyals et al., 2015; Xu et al., 2015; Yagcioglu et al., 2015; Huang et al., 2016; Gan et al., 2017b; Wu et al., 2017; Dai et al., 2017; Kinghorn et al., 2018)
CIDEr	2014, ID	(Vedantam et al., 2014; Karpathy et al., 2014; Chen and Zitnick, 2015; Fang et al., 2015; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Yagcioglu et al., 2015; Lu et al., 2017; Gan et al., 2017b; Wu et al., 2017; Dai et al., 2017)
SPICE	2016, ID	(Anderson et al., 2016; Lu et al., 2017; Dai et al., 2017)
WMD	2017, ID	(Kilickaya et al., 2017)

a description is (only if a description is as detailed as the average human one); SPICE is not suitable for Fluency, BLEU is, etc. Which metric is best depends on the assessment criterion. The evidence currently is that SPICE, CIDEr and Meteor, in this order, predict human Adequacy and Grammaticality assessments well.

2.3.3 Extrinsic evaluation measures

An extrinsic form of evaluation for image description, more specifically a user-task-success measure, was proposed by Ordonez et al. (2011) who presented contributors on Mechanical Turk with two images and one caption, and asked them to assign the caption to the 'more relevant' image. One of the two images was a system-generated one, whereas the other was selected randomly from the dataset. One of the 'systems' evaluated was the set of original human descriptions. The evaluation involved 100 images and showed that contributors were able to identify the correct picture from an original human description 96% of the time. For the best system, contributors were able to select the correct image 66.7% of the time.

Huang et al. (2016) used crowdsourcing to ask five contributors per story to rate how strongly they agreed with the statement *If these were my photos, I would like using a story like this to share my experience with my friends* (on a Likert-type scale of 'strongly disagree' to 'strongly agree'). This measure can be seen as assessing system purpose success (see above), in terms of the likelihood that end users will actually use the image-series descriptions generated by systems. However, rather than evaluate actual use rates in a real-world context such as Flickr, contributors are asked to judge how likely they would be to use the texts in a real-world context. This is a surrogate measure reminiscent of the 'pseudo-extrinsic' measure of Overall Responsiveness used in the TAC'08 summarisation competition where the question was *What would I pay for this summary of the answers to my questions?*

In many situations, real-world extrinsic evaluation is not feasible, simply because it is expensive and time-consuming to set up and run. However, extrinsic grounding, where an application task is explicitly defined, data is collected within the context of the application task, and evaluations can be carried out by comparing against extrinsically grounded reference data, should be feasible in many situations, and would help begin to address the vexed questions from Section 2.1.4.

3 Referring Expression Generation and Comprehension

Much of everyday language and discourse concerns the visual world around us; this makes understanding the relationship between objects in the physical world and language describing the objects an important challenge for AI. While image description strives to construct broad descriptions of image content, referring expressions, REs, are a more focused form of language, used to identify a particular object or temporal event in an image or video. People use such expressions all the time, especially in dialogue to indicate a particular object or event to a co-observer, e.g. *the woman in the blue shirt*, or *when she took a bite of the apple*. Computational models that generate and comprehend such expressions have broad applicability to human-computer interaction, especially for agents such as robots, interacting with people in the real world. Successful models need to connect visual interpretations of objects in the world to natural language that describes an object or event.

In the RE problem there is a pragmatic interaction between agents that involves two main tasks: (a) a speaker task where one must generate a natural language expression given a target and its surrounding world context; and (b) a listener task where one must interpret and comprehend the expression and map it to the correct target. We refer to these two tasks as referring expression generation and comprehension, respectively. In this section we review work on REs, including datasets and methods for generation and comprehension in images and videos.

3.1 Referring Expression Datasets

Some initial datasets in referring expression generation (REG) used graphics engines to produce images of objects (van Deemter et al., 2006; Viethen and Dale, 2008) with corresponding shared evaluation challenges (Gatt and Belz, 2010). Recently more realistic datasets have been introduced, consisting of craft objects like pipecleaners, and ribbons (Mitchell et al., 2010), or everyday home and office objects such as staplers or combs (Mitchell et al., 2013a), arrayed on a simple background. These datasets helped move REG research into the domain of real world objects.

In the past few years, datasets have become even larger and more realistic and expanded to include video REs. The ReferIt Dataset (Kazemzadeh et al., 2014) was perhaps the first large-scale RE dataset to be based on complex real world scenes. The images used to construct this dataset were originally sampled from the ImageCLEF IAPR image retrieval dataset (Grubinger et al., 2006b), a large collection of scene images with associated object segmentations. The ReferIt dataset was collected via a simple two-player online game (the ReferItGame) to crowdsource REs. In this game, Player 1 is shown an image with a highlighted target object and asked to write a natural language expression referring to the target. Player 2 is shown only the image and RE and asked to click on the corresponding object. If the players do their job correctly, they receive points and the expression is added to the dataset. This allows both data collection and verification within the game.

Based on this game, Yu et al. (2016a) further collected the RefCOCO and RefCOCO+ datasets, building on the MS COCO image collection (Lin et al., 2014b). In the RefCOCO dataset, no restrictions are placed on the type of language used in the REs, while in the RefCOCO+ dataset players are stopped from using location words in their REs by adding 'taboo' words to the ReferItGame. Thus, RefCOCO+ tends to focus more on appearance based descriptions. Another dataset based on MS COCO images has been collected, called the Google Refexp dataset (Mao et al., 2016). During collection of this dataset, one set of workers on Mechanical Turk were asked to write REs for objects. Another set of workers were asked to click on the indicated object given an RE. In Table 3, we show the statistics of each of the above-mentioned 4 datasets. REs in RefCOCO and RefCOCO+ tend to contain fewer words than those in Refexp since the competitive and time-based nature of games encourages players to write only the amount of information necessary to convey the correct object to the other player. Refexp contains more caption-like REs with many details about each referred object since labelers were encouraged to do so. Fig. 5 shows example images and expressions.

More recently, inspired by the two-player game GuessWhat, a task for localizing an unknown object by comprehending a sequence of questions and answers was introduced (De Vries et al., 2017). An example sequence is ("Is it a vase?", "Yes"), ("Is it in the left corner?", "No"), ("Is it the purple one?", "Yes"), etc.

In addition to image-based RE datasets, in the past year several video-based RE datasets and related tasks have been proposed. One example is the task of RE-guided tracking where a natural language specification indicates what object to track in a video (Li et al., 2017). Other work (Hendricks et al., 2017) considers retrieving a specific temporal video segment (a moment rather than an object) given a natural language text description. They introduce a dataset called Distinct Describable Moments (DiDeMo) with language annotations of video segments. We show an example of a video-expression pair in Fig. 4. The whole dataset consists of 40,000 pairs of localized video moments and corresponding expressions.

3.2 Referring Expressions for Images

Research on understanding how people generate REs has a long history, dating back to the 1970s (Winograd, 1972). Early work in REG (Dale and Reiter, 1995; Dale and

Table 3. *4 referring expression datasets that use realistic images.*

Dataset	#images	#expressions	collection way	expression style
Referit	19.894	130,525	Referit Game	Free style
RefCOCO	19,994	142,210	Referit Game	Free style
RefCOCO+	19,992	141,564	Referit Game	Abs. Loc forbidden
Google Refexp	104,560	26,711	Two rounds	COCO-caption style

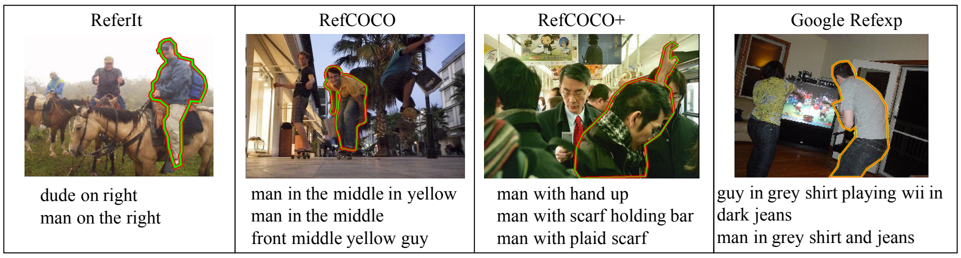


Fig. 4. Example images and referring expressions from RE datasets.

Reiter, 2000) explored research related to the Gricean maxims (Grice, 1975) which provide principles for how people will behave in conversation, including quality, , quantity, relevance, and manner. More recently, there has been progress examining other aspects of the RE problem such as types of attributes used (Mitchell et al., 2013a), modeling variations between speakers (Viethen and Dale, 2010; Viethen et al., 2013; Van Deemter et al., 2012; Mitchell et al., 2013b), incorporating visual classifiers (Mitchell et al., 2011), producing algorithms to refer to object sets (Ren et al., 2010; FitzGerald et al., 2013), or examining impoverished perception REG (Fang et al., 2013). There have been REG shared-task competitions since 2007 (Gatt and Belz, 2010). Krahmer and van Deemter provide a good survey of work in this area (Krahmer and van Deemter, 2012).

In the past few years, deep learning techniques have been widely applied in RE research. In the following, we denote r as the RE and o as the target object. As described above, there are typically two tasks explored in the literature. The first task is **referring expression comprehension**, requiring a system to select the region described by a given RE. To address this problem, some work (Hu et al., 2016; Mao et al., 2016; Nagaraja et al., 2016; Yu et al., 2016a) models $P(r|o)$, selecting the object o from the image that maximizes this probability. Alternatively, some

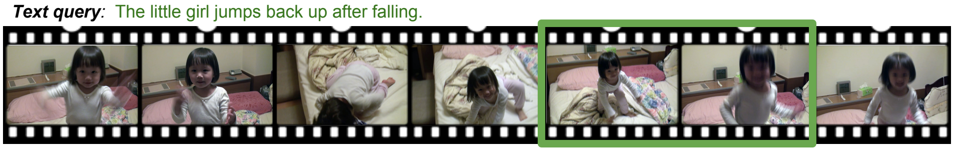


Fig. 5. Example video and temporal RE in DiDeMo (Hendricks et al., 2017).

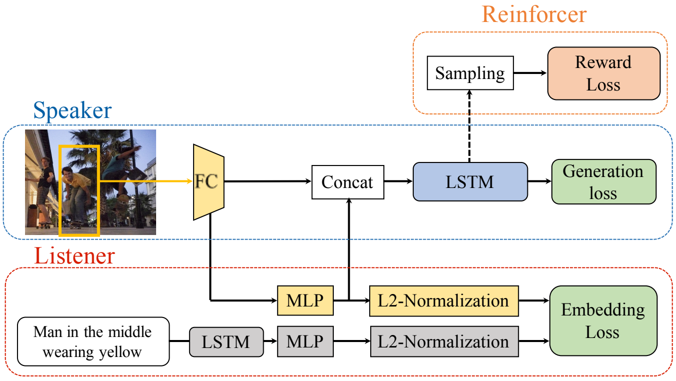


Fig. 6. Joint speaker-listener-reinforcer model for RE generation/comprehension (Yu et al., 2017).

works model $P(o, r)$ directly (Rohrbach et al., 2016; Wang et al., 2016a; Wang et al., 2018; Liu et al., 2017; Yu et al., 2018), by learning an embedding that minimizes the distance between object-expression pairs. The second task is **referring expression generation**, which asks a system to compose a natural language expression for a specified object within an image, i.e., $P(r|o)$. Many recent works (Mao et al., 2016; Yu et al., 2016a; Liu et al., 2017) use CNN-LSTM structures to generate expressions.

One current state-of-art model is the speaker-listener-reinforcer model (Yu et al., 2017), a unified framework for comprehension and generation tasks. The speaker module generates REs, the listener comprehends REs, and the reinforcer uses a reward function to guide sampling of more discriminative expressions. The speaker is modeled using a CNN-LSTM framework. VGGNet (Simonyan and Zisserman, 2014) is used to extract a visual representation for the target object and other visual context. Then, an LSTM (Hochreiter and Schmidhuber, 1997) is used to generate the most likely expression given the visual representation. Given a target object o_i, its VGG-fc7 feature v_i is first extracted. Its global context g_i, is modeled as features extracted from the VGG-fc7 layer for the entire image. Finally, its location/size is modeled as a 5-dimensional vector, l_i, encoding the top-left and bottom-right corners of o_i, as well as its relative size with respect to the image, i.e., $l_i = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}]$. The speaker model also considers visual comparisons to produce expressions contrasting the target object from other related objects. The comparison features are composed of: (a) appearance similarity δv_i, and (b) location and size similarity δl_i. The final visual representation for the target object is a concatenation of the above features followed by a fully connected layer fusing them together, $r_i = W_m[v_i, g_i, l_i, \delta v_i, \delta l_i] + b_m$. This joint feature is then fed into the LSTM for RE generation. During training the negative log-likelihood is minimised:

$$\begin{aligned} L_1^s(\theta) &= -\sum_i \log P(r_i|o_i;\theta) \\ &= -\sum_i \sum_t \log P(r_i^t|r_i^{t-1}, \ldots, r_i^1, o_i;\theta) \end{aligned} \tag{1}$$

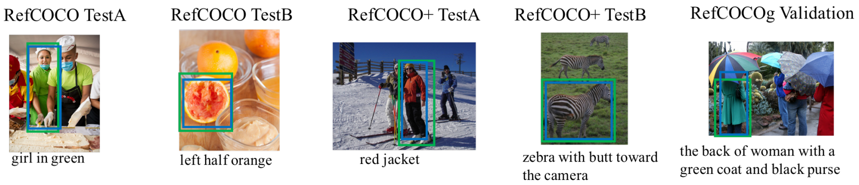


Fig. 7. Comprehension examples in (Yu et al., 2017). Green box shows the ground-truth region, blue box shows correct comprehension using the proposed model.

A joint-embedding model is used for the listener which merges visual information from the target object and semantic information of the corresponding RE into a joint embedding space such that their embedded vectors are close to each other. An LSTM encodes the input RE and the same visual representation as the speaker is used to encode the target object. Visual representation and word-embedding are shared with the speaker so that speaker and listener are aware of each other's behaviour. In the embedding part, two MLPs and two L2 normalization layers are applied on top of each view. The inner product of the two normalized representations is computed as their similarity score $S(r, o)$. In training, two contrastive triplets are sampled for enforcing a higher similarity between a positive match than the negative matches, which constructs a ranking loss:

$$\begin{aligned} L^l(\theta) = \sum_i [&\lambda_1^l \max(0, M + S(r_i, o_k) - S(r_i, o_i)) \\ &+ \lambda_2^l \max(0, M + S(r_j, o_i) - S(r_i, o_i))] \end{aligned} \tag{2}$$

where the negative matches are randomly chosen from the other objects and expressions in the same image. The reinforcer guides the speaker to generate less ambiguous expressions. It is composed of a discriminative reward function and performs a non-differentiable policy gradient update to the speaker. During training, the reinforcer takes the sampled expression $w_{1:T}$ from the speaker and feeds it to a pre-trained reward function. The goal is to maximize the reward expectation $F(w_{1:T})$ under the distribution of $p(w_{1:T}; \theta)$ parameterized by the speaker, i.e., $J = E_{p(w_{1:T})}[F]$. This reward function is another listener trained with 1-d Logistic Regression loss to produce a score between 0 and 1. At inference time, the speaker output $P(r|o)$ and listener output $P(r, o)$ are used together for both the comprehension and generation tasks. Fig. 7 and Fig. 8 show example results on these two tasks using the joint speaker-listener-reinforcer model.

3.3 Referring Expressions for Video

To address the temporal localization task (Fig. 4) the Moment Context Network (MCN) was proposed (Hendricks et al., 2017). Given input video frames $v = v_t$ where $t \in 0, ..., T-1$ indexes time, a proposed temporal interval $\hat{\tau} = \tau_{\text{start}} : \tau_{\text{end}}$, and an expression r, the goal is to find the moment described by r:

$$\hat{\tau} = \text{argmin}_{\tau} D_{\theta}(s, v, \tau) \tag{3}$$

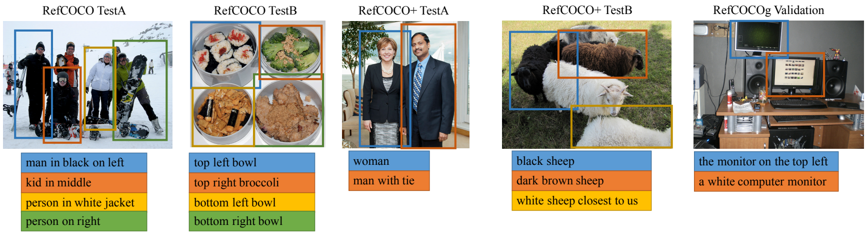


Fig. 8. Generation examples in (Yu et al., 2017). Each sentence shows the generated expression for one of the depicted objects (color coded to indicate correspondence).

where $D_\theta(r, v, \tau)$ measures the distance between a temporal interval τ and RE r. The MCN network is shown in Fig. 9. Video moments are encoded into visual temporal context features: video features reflecting what is occurring within each moment, global video features providing broader context for each moment, and temporal endpoint features indicating when a moment occurs within a longer video.

To construct the local and global visual features, fc7 features are extracted for each frame using VGGNet. Then, the local features are constructed by temporally pooling features within each specific moment, and global features are constructed by averaging over all video frames. Temporal endpoint features indicate the start and endpoint of a candidate moment (normalized to the interval [0,1]). The concatenation of these features are fed into a MLP to get the final visual feature for a moment P_θ^V. Additionally, the authors also incorporate optical flow (Wang et al., 2016b) as a motion feature for each moment P_θ^F. The language encoding is similar to (Yu et al., 2017), where an LSTM is used to encode the input expression and its last hidden state is fed into a MLP to yield the embedded feature P_θ^L. Then, the distance between a moment and an RE is computed as:

$$D_\theta(r, v, t) = \|P_\theta^V(v, \tau) - P_\theta^L(r)\| + \eta \|P_\theta^F(f, \tau) - P_\theta^L(r)\| \tag{4}$$

where η is a tunable 'late fusion' scalar. A ranking loss similar to Eqn. 2 is used for training. At inference time, each temporal segment is compared with the input expression, and the nearest one is selected as the referring moment (Eqn. 3).

4 Visual Question Answering

Another language and vision task that has received increasing attention recently is Visual Question Answering. VQA systems take as input an image or video along with relevant natural language questions, and produce an answer to those questions. Questions can be open ended, requiring systems to produce a natural language answer, or a set of multiple choice answers is provided, requiring systems to select the best answer from a list. One driving factor for the introduction of VQA was that despite progress on image and video captioning, automatic evaluation of descriptions is still a challenging open research problem. Multiple choice VQA provides a task that is simple to evaluate automatically. Additionally, VQA provides a nice tool for more fine-grained evaluation of algorithms since different types of questions can be used to probe and evaluate various aspects of visual understanding, ranging

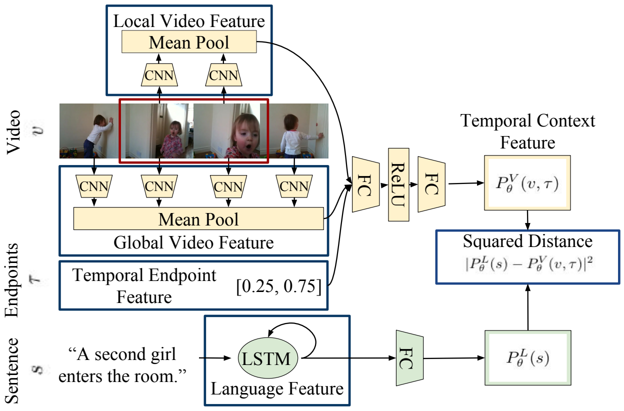


Fig. 9. Moment Context Network (MCN) used in (Hendricks et al., 2017).

from object identification, counting, or appearance, to more complex visual understanding of interactions, and inferences about why or how something is occurring in an image or video. In this section we describe existing VQA datasets and review some efforts toward building VQA systems.

4.1 VQA in Images

4.1.1 Image-based VQA Datasets

Several VQA datasets have recently been constructed. We review some of the prominent efforts here. Statistics about all of the datasets are presented in Table 4.

DAQUAR (Malinowski and Fritz, 2014) was built on the NYU indoor scene RGB-D dataset (Silberman et al., 2012), a collection of indoor environments with associated RGB and depth camera images and annotated object class labels. To construct the dataset, the DAQUAR authors asked 5 in-house participants to provide questions and answers based on these images. Questions generally refer to everyday objects and relationships between objects, e.g. "Q: what is on the right side of the notebook on the desk in image4, A: plastic cup of coffee". Answers are evaluated using the WUPS score to compute how close the produced answer from a system matches the ground truth answer. WUPS is a soft measure based on the Wu and Palmer score (Wu and Palmer, 1994), which calculates the semantic relatedness of terms by considering the depths of their synsets in the WordNet taxonomy, along with the depth of the least common subsumer.

COCO-QA (Ren et al., 2015a) is built on the MS COCO dataset (Section 2.1.1). QA pairs are automatically generated from image descriptions using four question templates: Object Questions, Number Questions, Color Questions, and Location Questions. For example, a description reading "A man is riding a horse" can be automatically transformed into the question "What is the man riding?" Each answer consists of a single word, allowing models to treat the problem as a classification task without considering natural language generation, simplifying evaluation.

FM-IQA (Gao et al., 2015) is also built on MS COCO (FM stands for Freestyle

Multilingual). Annotators provide freestyle question-answer pairs in Chinese, then each question-answer pair is translated into English. Arguing that automatic metrics like WUPS, BLEU, METEOR, or CIDEr cannot accurately evaluate model capacity, the authors conduct a Visual Turing Test (Turing, 1950) instead, where answers are mixed between humans and model, then human judges are asked to distinguish models from humans, and provide a score indicating the answer quality.

Visual Madlibs (Yu et al., 2015) is again built on MS COCO. Questions are designed with 12 fill-in-the-blank templates, to collect targeted descriptions about: people & objects, their appearance, activities, and interactions, as well as inferences about the general scene or its broader context. Collected descriptions are used for two tasks: (a) fill-in-the-blank description generation (similar to image captioning, but more focused on a particular image aspect), and (b) multiple-choice fill-in-the-blank question answering. In the latter, given an image and a partial description such as "The person is [blank] the frisbee", the task is to select the correct choice from 4 answers. This provides an multiple-choice test for evaluation; varying the selection of negative answers can make questions for model testing easier or harder.

VQA (Antol et al., 2015) is built on top of MS COCO. The questions in VQA are free-form and open-ended and the answers are also free-form, both of which were written by humans. For each question, there are 10 answers gathered from humans. Similar to Visual Madlibs, there are also two tasks in VQA: open-ended answering and multiple-choice. For evaluation of the open-ended task, a predicted answer is deemed accurate if at least 3 humans provided that exact answer. As most answers (89.32%) are single word, there is no high-order n-gram matching issue. For the multiple-choice task, each question is associated with 18 candidate answers. Most recent research works on the first open-ended task.

VQA v2 (Goyal et al., 2017) is a second, more balanced version of the VQA dataset, created to address the visual priming bias problem in the original VQA. For example, people tend to raise the question "Is there a clock tower in the picture?" only on images that contain clock towers. This makes blindly answering "Yes" to "Do you see...?" and "Is there ...?" an easy way to achieve high model accuracy. In order to ease this bias issue, the authors collected complementary images for biased questions so that each question has two complementary images that look similar but have different answers. This balanced dataset was constructed to encourage VQA models to focus more on visual understanding than learning dataset biases.

Visual7W (Zhu et al., 2016) is part of the Visual Genome project (Krishna et al., 2017b) and similar to Visual Madlibs. Arguing that many relevant image question pairs relate to local image regions rather than to the entire image, the authors establish a link between text descriptions and regions through object grounding to construct region based visual questions. There are in total six W question types (*what*, *where*, *when*, *who*, *why* and *how*), and a 7th *which* question category. Each question is associated with 4 answers, only one of which is correct. In addition, for each question, the object-level grounding (object being mentioned by the QA pairs) is provided, resolving the co-reference ambiguity between images and questions. At test time, this provides a way to analyze the behavior of attention-based models.

CLEVR (Johnson et al., 2017a) is somewhat different from the above datasets.

Arguing that existing VQA datasets have strong biases that models can exploit to correctly answer questions without reasoning, the authors propose CLEVR which is specifically designed for visual reasoning. Images in CLEVR are computer generated using Blender. Each scene contains three to ten objects with random shapes, sizes, materials, colors and positions. The questions are also generated and each question is associated with a functional program that can be executed on an image's scene graph, with its answer also known. One example question is "What color is the cube to the right of the yellow sphere?". Answering this question requires a model to locate the "yellow sphere", then find the "right cube", and finally infer its color.

Table 4. *Image VQA Dataset statistics, including: number of question-answer pairs (#QA), number of images (#Images), question type (QType), average question length (QLen), average answer length (ALen), and evaluation type (Eval).*

Dataset	#QA	#Images	QType	QLen	ALen	Eval
DAQUAR	12,468	1,449	Human	11.5	1.2	WUPS
COCO-QA	117,684	69,172	Synthesized	8.7	1.0	Word matching
FM-IQA	316,193	158,392	Human	7.38	3.82	Turing test
Visual Madlibs	56,468	9,688	Human	4.9	2.8	Multiple-choice
VQA	614,163	204,721	Human	6.2	1.1	Open-ended
Visual7W	327,939	47,300	Human	6.9	2.0	Multiple-choice
CLEVER	100,000	999,968	Synthesized	18.0	1.0	Word matching

4.1.2 Image-QA Models

Image-QA models take as input an image, I, and question $Q = \{q_t | t = 1, ..., T\}$, made up of T words. Usually they then compute image features V using visual recognition algorithms to answer Q. For an open-ended question-answering task, the QA system could be formulated as a generation model $A = G(V, Q)$, producing a natural language sentence answer, or as a classification task $A = C(V, Q)$ to select the most likely answer from a (sometimes large) predefined set of answers. For multiple-choice QAs, candidate answers, C, are provided to the system along with I and Q as input. In these tasks, C are often fed into the model and the candidate with highest probability $C^* = \text{argmax}_{c_i} P(c_i | V, Q), c_i \in C$ is selected.

Baseline: Given the rapid development and advances in CNNs, almost all recent VQA papers use CNNs for their underlying visual feature representation; popular architectures include AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), ResNet (He et al., 2016), InceptionNet (Szegedy et al., 2017), etc.

One well-known baseline (Zhou et al., 2015) proposed a simple model for visual question answering on the VQA v1 dataset. This model, illustrated in Fig. 10, uses a visual representation produced by the last fully-connected/average-pooling output of a CNN, i.e., $V \in R^{d \times 1}$, and bag-of-words as the question representation.

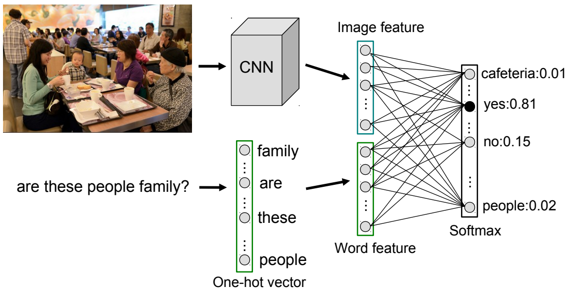


Fig. 10. Image-based VQA Baseline model used in (Zhou et al., 2015).

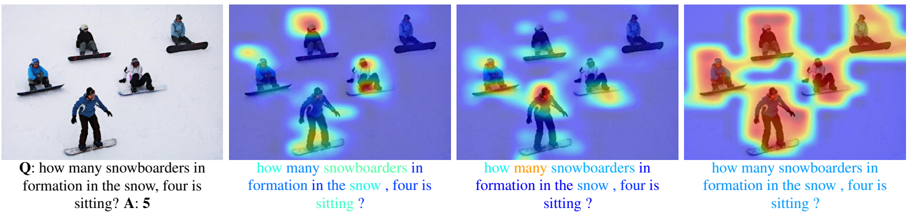

Q: how many snowboarders in formation in the snow, four is sitting? **A: 5**
how many snowboarders in formation in the snow , four is sitting ?
how many snowboarders in formation in the snow , four is sitting ?
how many snowboarders in formation in the snow , four is sitting ?

Fig. 11. Visualization of image and question co-attention maps in (Lu et al., 2016). From left to right: original image and question pairs, word-level co-attention maps, phrase-level co-attention maps, question-level co-attention maps. The attentions are scaled from red:high to blue:low.

These image and language representations are then concatenated and the combined feature is sent to a softmax layer to predict the answer class. Note that in this model both the open-ended and multiple-choice tasks are formulated as classification tasks. While simple, the model achieved comparable performance to several more complicated approaches at that time. Improvements over this baseline used Recurrent Neural Network (RNN) (Antol et al., 2015; Malinowski et al., 2015; Ren et al., 2015a) or a language CNN (Ma et al., 2016) to model the question (and answer).

Attention Models: Since then, most research has focused on modeling the interaction between image content and question for improving performance, as well as on model interpretability. In many cases, an answer only relates to a small portion of the image, e.g. the answer to the question *What is the color of the boy's shirt?* given an image containing a boy and a cat, only relates to *the boy*. Thus, using global image features to predict the correct answer usually leads to sub-optimal results due to noisy information introduced by the irrelevant image regions.

To address this issue, recent models (Yang et al., 2016; Xiong et al., 2016; Xu and Saenko, 2016; Shih et al., 2016; Chen et al., 2015b; Das et al., 2017; Selvaraju et al., 2017) examine different spatial regions within the image and compare their contents (and locations) to help in answering visual questions. Rather than extracting a single feature for the whole image, these models compute visual representations consisting of the last convolutional output $V \in R^{d \times G}$, where d is the feature dimension and G is the number of spatial grids. These are fed through a single layer neural network and then a softmax function generates an attention distribution over image regions:

$$\begin{aligned} q &= \mathrm{LSTM}(Q) \\ H_v &= \tanh(W_v V + W_q q) \\ a^v &= \mathrm{softmax}(w_{h,v}^T H_v) \end{aligned} \tag{5}$$

where $W_v \in R^{k \times d}$ and $w_{h,v} \in R^{1 \times k}$ are the transformation matrices. Then the weighted sum of visual representations $\widetilde{v}$ guided by the question is computed as $\widetilde{v} = \sum_{i=1}^{G} a_i^v v_i$

In addition to modelling 'where to look' through visual attention, it can also be useful to model 'what words to listen to' (Nam et al., 2017; Lu et al., 2016). A co-attention model has been proposed (Lu et al., 2016) that jointly reasons about question-guided visual attention and image-guided question attention. This model co-attends to the image and question in a hierarchical structure over word-level, phrase-level and question-level embeddings. Given the embedding $E = \{e_t | t = 1, ..., T\}$ for the input question words Q and the question-guided visual representation $\widetilde{v}$, the image-guided question representation is computed as:

$$\begin{aligned} H_q &= \tanh(W_e E + W_v \widetilde{v}) \\ a^q &= \mathrm{softmax}(w_{h,q}^T H_q) \\ \widetilde{q}^w &= \sum_{t=1}^{T} a_t^q e_t \end{aligned} \tag{6}$$

where $W_w \in R^{k \times d}$ and $w_{h,w} \in R^{1 \times k}$ are the transformation matrices. Lu et al. (2016) recursively encode the attention features for word, phrase and question. Fig. 11 shows an example, where we can see that the model jointly co-attends to interpretable regions of images and questions to predict the answer.

While most of the above work used concatenations, element-wise products or sums for interactions between the visual and textual representations, Multimodel Compact Bilinear pooling (MCB) (Fukui et al., 2016) is an alternative solution for cross-modality interaction. MCB pooling projects an outer product to a lower dimensional space and avoids computing the outer product directly. Fukui et al.'s model uses MCB twice, once to predict spatial attention and once to predict the answer, achieving state-of-art results in 2016.

Most recently, the winning model of the 2017 VQA Challenge was a bottom-up and top-down attention model (Anderson et al., 2017). The authors argued that a uniform grid of equally sized and shaped receptive fields—irrespective of the content of the image—as usually used in attention models, is sub-optimal. Instead, their bottom-up mechanism proposes a set of detected image regions for consideration, with each region represented by a pooled convolutional feature vector. These bottom-up regions are detected by Faster R-CNN (Ren et al., 2015b); a top-down mechanism then uses task-specific context to predict an attention distribution over the proposed image regions. The full VQA model is shown in Fig. 12.

Modular Networks: The first module network for question answering was pro-

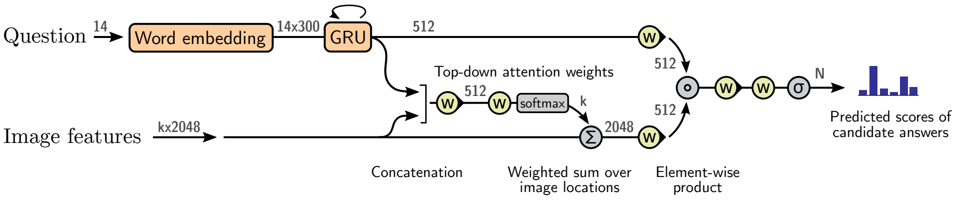


Fig. 12. Model proposed by (Anderson et al., 2017).

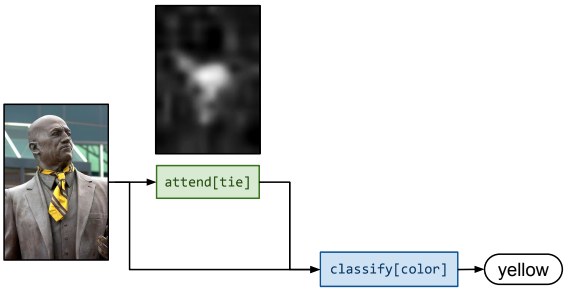


Fig. 13. NMN for answering "What color is his tie?" by (Andreas et al., 2016b)

posed in 2016 (Andreas et al., 2016a) and later extended to VQA (Andreas et al., 2016b). Neural module networks (NMN) approach the VQA task by dynamically composing networks of independent neural modules, jointly trained. Modules are selected based on a parse of the question to utilize only modules that are relevant to the particular question content. Specifically, the authors define the following modules:

1. Attention module `attend[c]` performs `Image` → `Attention` to spatially select mentioned objects `c`.
2. Re-attention module `re-attend[c]` takes an attention heatmap and maps it to another attention, i.e., `Attention` → `Attention`.
3. Combination module `combined[c]` merges two attentions into a single attention, i.e., `Attention` × `Attention` → `Attention`.
4. Classification module `classify[c]` takes an attention and image then maps them to a distribution over labels, i.e., `Image` × `Attention` → `Label`.
5. Measurement module `measure[c]` takes an attention and maps it to a distribution over labels, i.e., `Attention` → `Label`.

Given a question, modules are selected based on a language parser (De Marneffe et al., 2006), and are then mapped to a network structure assembling the relevant modules (see Fig. 13 for an example). Given the question "What color is his tie?", NMN generates its composition of `classify[color](attend[tie])`, the answer coming from the final classification module for color labels. This strategy takes advantage of the inherently compositional property of language, and inspired further work on visual reasoning (Johnson et al., 2017b; Hu et al., 2017).

4.2 VQA in Video

4.2.1 Video-based QA Datasets

TGIF-QA (Jang et al., 2017) This dataset consists of QA pairs on animated GIFs, collected using pre-defined templates. The QA pairs include three tasks: (1) counting the number of repetitions of a given action, (2) detecting a repeated action given its count, and (3) identifying state transitions, i.e. what happened before or after a specified action state. For example, "Q: What does the duck do 3 times?, A: Shake head". In addition, the authors also generate Frame-QAs for object/number/color/location questions that are answered by one of the frames.

LSMDC-QA (Maharaj et al., 2017) is built on LSMDC (Rohrbach et al., 2015), a dataset for movie description. To construct this dataset, the authors recast the video description problem as a fill-in-the-blank question-answering task. Given a video and its description with one word blanked-out, the goal is to predict the missing word, e.g. "Q: She opens the [blank]. A: door". The blanked words cover entities, actions and attributes, requiring models to understand the visual content of videos. Since each fill-in the blank answer is a single word evaluation is simple.

VideoQA (Zhu et al., 2015) The videos used in this dataset are from TACoS Multi-Level (Regneri et al., 2013) cooking dataset, MPII-MD (Rohrbach et al., 2015) movie description dataset, and TRECVID MEDTest (Over et al., 2014) web videos. The authors generate three types of QA pairs from their associated descriptions: Inferring the past, Describing the present, and Predicting the future. For each description, some phrase or words are blanked out which are used to answer the three types of questions, e.g. "Q: Predict the future. He [blank] cucumber on plate. A: places". There are 4 candidate fill-in-the-blank answers, where only one is correct, a simple multiple-choice evaluation.

MovieQA (Tapaswi et al., 2016) This dataset contains more diverse sources of information compared with the other video-based VQA datasets, including plot synopses, videos, subtitles, DVS and scripts. Here, the authors use plot synopses to collect questions about movies. During data collection, each annotator is shown a paragraph from a plot synopsis then asked to provide questions and answers related to the provided plot. This often results in complex high-level questions that require a great deal of understanding to answer. For, example, "Why does Cypher betray Morpheus?" in the Matrix movie. Multiple-choice is used for evaluation. The whole dataset contains 408 movies and 14,944 QAs, but only 140 videos have video to plot alignment, resulting in 16,066 video clips (Table. 5)).

PororoQA (Kim et al., 2017) Different from above, the media domain of PororoQA is cartoons, sampled from the popular children's series '*Pororo*'. This show has a simple, clear, and coherent story structure and a small environment compared to dramas or movies. Each of the 16,066 video clips contain dialogues and each clip is short (34 frames). All questions and answers were written by people and evaluation is multiple-choice question answering, where each question is coupled with 5 possible answers (1 correct and 4 incorrect). This dataset allows for reasoning about characters that carry over the whole dataset, e.g. "Q: What does Pororo think when he hides behind the tree? A: Pororo thinks Loopy can't find him".

Table 5. *Video Question Answering Datasets information including number of question-answer pairs, number of videos, average video length and source domain.*

Dataset	#QA	#videos	avg. video length	domain
TGIF-QA	165,165	71,741	3.1s	Social media GIFs
LSMDC-QA	348,998	111,744	4.8s	Movie
VideoQA	390,744	109,895	-	Cooking/Movie/Web
MovieQA	6,462	6,771	200s	Movie
PororoQA	8,913	16,066	~1s	Cartoon
MarioQA	187,757	187,757	<6s	Game

MarioQA (Mun et al., 2017) MarioQA is a synthetic video QA dataset, constructed on Super Mario Bros. gameplay videos. Questions are synthesized using templates, asking about event-centric questions, counting questions, and state questions, e.g. "Q: What enemy did Mario kill by stomping?, A: Para Goomba". These questions are split into different levels of reasoning complexity: questions without temporal relationships (NT), questions with easy temporal relationships (ET) and questions with hard temporal relationships (HT). These event-centric questions are especially suited to evaluate the temporal reasoning capability of algorithms.

4.2.2 Video-QA Models

Frame representation: Each video is composed of a set of frames $F = \{F_1, F_2, ..., F_N\}$. Similar to image-based VQA models, CNNs are typically used for extracting visual representations. We denote each frame feature as $F_n = \{f_{n,i} | i = 1, ..., G\}$, where n denotes the n-th frame and G is the number of regions. Note, the frame feature here is not restricted to CNN features on RGB images. Some recent works also consider using optical flow or spatial-temporal features via C3D (Jang et al., 2017; Jang et al., 2017; Hendricks et al., 2017).

The simplest way to abstract the representation of each F_n is via mean pooling. Additionally, spatial attention models can be used to learn which regions of F_n to attend to for a given question Q. The spatial attention score for each region can be computed as (Zhao et al., 2017; Jang et al., 2017; Yu et al., 2016b):

$$s_{ni} = w\tanh(W_{qs}q + W_{fs}f_{ni} + b_s) \tag{7}$$

where W_{qs} and W_{fs} are transformation matrices and b_s is a bias term. For each region f_{ni}, the normalized attention is computed as $\alpha_{ni} = \frac{\exp(s_{ni})}{\sum_i \exp(s_{ni})}$, where $q = \text{LSTM}(Q)$ is the question feature as in Eqn. 5. Then, the spatially attended visual representation for each frame is computed as: $v_n = \sum_i \alpha_{ni} f_{ni}$.

Video representation: Given frame features (with or without attention), the next step is to encode the whole video. One method uses mean-pooling, f_n (Venugopalan et al., 2015) $\widetilde{v} = \frac{1}{N}\sum_n f_n$, as the final video representation, but this weights the importance of each frame equally ignoring information about what portion of the video the question focuses on. Some authors (Yu et al., 2016b; Zhao et al., 2017; Jang et al., 2017) model video as a temporal sequence and use an RNN to encode its information. For example, if we use an LSTM to encode the video, then a corresponding sequence of hidden states $\widetilde{v} = h_N$ can be computed as (Yu et al., 2016b): $h_n = \mathrm{LSTM}(v_n, h_{n-1})$. The final output is then the final video representation $\widetilde{v}$.

In addition to spatial attention, temporal attention can also play an important role for localizing what portion of the video content is useful for answering a given question. (Zhao et al., 2017; Jang et al., 2017) consider applying a temporal attention model, computing the relevance scores over each hidden state h_n:

$$s_n^{(t)} = w^{(t)}\tanh(W_{qt}q + W_{ht}h_t + b_t) \tag{8}$$

The attention score for each frame (hidden state) is thus:

$$\beta_n = \frac{\exp(s_n^{(t)})}{\sum_n \exp(s_n^{(t)})} \tag{9}$$

The attentional pooled feature $\widetilde{v} = \sum_n \beta_n h_n$ is regarded as a question-driven video representation.

Question answering: Similar to image-based QA, the inference model depends on the type of question-answer pairs. As in image-based VQA, for the open-ended question-answering task the model is formulated as a generation/classification model producing a sentence answer, while for multiple-choice QAs a classification model is typically used. Taking the classification model as an example, given the video representation $\widetilde{v}$ and question representation q, one approach is to first fuse the video and question modalities (Jang et al., 2017): $\widetilde{v}_q = \tanh(W_v\widetilde{v}) \oplus q$, where $\oplus$ is an element-wise sum and W_v is a transformation matrix to make the dimensions of the two modalities equal. A linear classifier can be defined that takes as input the video-question vector $\widetilde{v}_q$, computing the confidence score for the c-th answer as $\mathrm{s}_c = \mathrm{softmax}(W_c\widetilde{v}_q + b_c)$, where W_c and b_c are model parameters. At inference time, the solution is simply selected as $c^* = \mathrm{argmax}_{c\in C}\mathrm{s}_c$.

5 Conclusion

We have reviewed recent work in language and vision tasks, including datasets and methods for producing general natural language descriptions of images (Section 2), referring expression generation and comprehension (Section 3), and visual question answering (Section 4). There has been a great deal of progress on each of these tasks, largely due to the growing availability of large labeled datasets and neural learning based methods. Moving forward, we foresee vision and language tasks moving into the real world where intelligent agents collaborate and communicate with people. This implies a need for algorithms that can produce not just static language about fixed physical objects and scenes, but also adaptively interact with people through dialogue and exploration. As a result there will be new data and evaluation challenges that will be exciting to investigate.

References

Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *ECCV'16*, pages 382–398.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2017). Bottom-up and top-down attention for image captioning and VQA. *arXiv preprint arXiv:1707.07998.*

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. (2016a). Learning to compose neural networks for question answering. *NAACL'16.*

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. (2016b). Neural module networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual Question Answering. In *ICCV'15.*

Belz, A. (2009). That's nice... what can you do with it? *Computational Linguistics*, 35(1):118–119.

Belz, A. and Hastie, H. (2014). Comparative evaluation and shared tasks for NLG in interactive systems. In *Natural Language Generation in Interactive Systems.* CUP, Cambridge.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. of Artificial Intelligence Research*, 55:409–442.

Chen, J., Kuznetsova, P., Warren, D., and Choi, Y. (2015a). Déjà image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 504–514.

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

Chen, X. and Zitnick, C. L. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431.

Dai, B., Lin, D., Urtasun, R., and Fidler, S. (2017). Towards diverse and natural image descriptions via a conditional gan. *arXiv preprint arXiv:1703.06029.*

Dale, R. and Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science (CogSci)*, 19:233264.

Dale, R. and Reiter, E. (2000). *Building natural language generation systems.* CUP.

Das, A., Agrawal, H., Zitnick, L., Parikh, D., and Batra, D. (2017). Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Genoa Italy.

De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., and Courville, A. (2017). Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR.*

Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., and Mitchell, M. (2015). Language models for image captioning: The quirks and what works. *CoRR*, abs/1505.01809.

Elliott, D. and de Vries, A. (2015). Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 42–52.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017). Findings of the second shared task on multimodal machine translation and multilingual image description. *CoRR*, abs/1710.07177.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual English-German image descriptions. *arXiv preprint arXiv:1605.00459.*

Elliott, D. and Keller, F. (2013). Image description using visual dependency representations. In *Proc. 18th Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1292–1302, Seattle.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollazr, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Zitnick, C. L., and Zweig., G. (2015). From captions to visual concepts and back. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1473–1482, Boston.

Fang, R., Liu, C., She, L., and Chai, J. (2013). Towards situated dialogue: Revisiting referring expression generation. In *EMNLP'13*.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *ECCV'10*, pages 15–29.

Feng, Y. and Lapata, M. (2008). Automatic image annotation using auxiliary text information. *Proceedings of ACL-08: HLT*, pages 272–280.

FitzGerald, N., Artzi, Y., and Zettlemoyer, L. (2013). Learning distributions over logical forms for referring expression generation. In *Empirical Methods on Natural Language Processing (EMNLP)*.

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

Gan, C., Gan, Z., He, X., Gao, J., and Deng, L. (2017a). Stylenet: Generating attractive visual captions with styles. In *CVPR*.

Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., and Deng, L. (2017b). Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., and Xu, W. (2015). Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304.

Gatt, A. and Belz, A. (2010). *Introducing Shared Tasks to NLG: The TUNA Shared Task Evaluation Challenges*, pages 264–293. Springer Berlin Heidelberg, Berlin, Heidelberg.

Gella, S., Lapata, M., and Keller, F. (2016). Unsupervised visual sense disambiguation for verbs using multimodal embeddings. *arXiv preprint arXiv:1603.09188.*

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Grice, H. P. (1975). Logic and conversation. page 4158.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006a). The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 5, page 10.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006b). The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 5, page 10.

Gupta, A., Verma, Y., and Jawahar, C. V. (2012). Choosing linguistics over vision to describe images. In *Proc. 26th AAAI Conf. on Artificial Intelligence*, pages 606–612.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proc. IEEE conf. on computer vision and pattern recognition*, pages 770–778.

Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T., Mao, J., Huang, J., Toshev, A., Camburu, O., et al. (2016). Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hendricks, L. A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., and Russell, B. (2017). Localizing moments in video with natural language. In *ICCV*.

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

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Hu, R., Andreas, J., Rohrbach, M., Darrell, T., and Saenko, K. (2017). Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 804–813.

Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., and Darrell, T. (2016). Natural language object retrieval. In *CVPR*. IEEE.

Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., et al. (2016). Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Jang, Y., Song, Y., Yu, Y., Kim, Y., and Kim, G. (2017). Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *CVPR*.

Jia, X., Gavves, E., Fernando, B., and Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2407–2415. IEEE.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. (2017a). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE.

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C. L., and Girshick, R. (2017b). Inferring and executing programs for visual reasoning. *ICCV*.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3128–3137, Boston.

Karpathy, A., Joulin, A., and Fei-Fei, L. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897.

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

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., and Erdem, E. (2017). Re-evaluating automatic metrics for image captioning. 1:199–209.

Kim, K.-M., Heo, M.-O., Choi, S.-H., and Zhang, B.-T. (2017). Deepstory: video story qa by deep embedded memory networks. *IJCAI*.

Kinghorn, P., Zhang, L., and Shao, L. (2018). A region-based image caption generator with refined descriptions. *Neurocomputing*, 272:416–424.

Kong, C., Lin, D., Bansal, M., Urtasun, R., and Fidler, S. (2014). What are you talking about? text-to-image coreference. In *CVPR*.

Krahmer, E. and van Deemter, K. (2012). Computational generation of referring expressions: A survey. In *Computational Linguistics*, volume 38, page 173218.

Krause, J., Johnson, J., Krishna, R., and Fei-Fei, L. (2016). A hierarchical approach for generating descriptive image paragraphs. *arXiv preprint arXiv:1611.06607*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017a). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, pages 1–42.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017b). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1601–1608, Colorado Springs.

Kuznetsova, P., Ordonez, V., Berg, T. L., and Choi, Y. (2014). Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2(10):351–362.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., and Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *Proc. 15th Conf. on Computational Natural Language Learning*, pages 220–228, Portland, Oregon.

Li, X., Lan, W., Dong, J., and Liu, H. (2016). Adding Chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 271–275.

Li, Z., Tao, R., Gavves, E., Snoek, C. G., Smeulders, A., et al. (2017). Tracking by natural language specification. In *CVPR*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014a). Microsoft coco: Common objects in context. In *ECCV'14*, pages 740–755.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014b). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Liu, J., Wang, L., Yang, M.-H., et al. (2017). Referring expression generation and comprehension via attributes. In *CVPR*.

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6.

Lu, J., Yang, J., Batra, D., and Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In *NIPS'16*, pages 289–297.

Ma, L., Lu, Z., and Li, H. (2016). Learning to answer questions from image using convolutional neural network. In *AAAI*, volume 3, page 16.

Maharaj, T., Ballas, N., Rohrbach, A., Courville, A., and Pal, C. (2017). A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. *CVPR'17*.

Malinowski, M. and Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690.

Malinowski, M., Rohrbach, M., and Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1–9. IEEE Computer Society.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Mason, R. and Charniak, E. (2014). Nonparametric method for data-driven image captioning. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 592–598, Baltimore, Maryland.

Mathews, A. P., Xie, L., and He, X. (2016). Senticap: Generating image descriptions with sentiments. In *AAAI*, pages 3574–3580.

Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., and Daumé, III, H. (2012). Midge: Generating image descriptions from computer vision detections. In *Proc. of the 13th Conf. of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France.

Mitchell, M., Reiter, E., and van Deemter, K. (2013a). Typicality and object reference. In *Cognitive Science (CogSci)*.

Mitchell, M., van Deemter, K., and Reiter, E. (2010). Natural reference to objects in a visual domain. In *International Natural Language Generation Conference (INLG)*.

Mitchell, M., van Deemter, K., and Reiter, E. (2011). Two approaches for generating size modifiers. In *European Workshop on Natural Language Generation*.

Mitchell, M., van Deemter, K., and Reiter, E. (2013b). Generating expressions that refer to visible objects. In *NAACL'13*.

Miyazaki, T. and Shimizu, N. (2016). Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1780–1790.

Mun, J., Seo, P. H., Jung, I., and Han, B. (2017). Marioqa: Answering questions by watching gameplay videos. In *ICCV*.

Muscat, A. and Belz, A. (2017). Learning to generate descriptions of visual data anchored in spatial relations. *IEEE Computational Intelligence Magazine*, 12(3):29–42.

Nagaraja, V. K., Morariu, V. I., and Davis, L. S. (2016). Modeling context between objects for referring expression understanding. In *ECCV*. Springer.

Nam, H., Ha, J.-W., and Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor et al., editor, *Advances in Neural Information Processing Systems 24*, pages 1143–1151. Curran Associates, Inc.

Ortiz, L. G. M., Wolff, C., and Lapata, M. (2015). Learning to interpret and describe abstract scenes. In *NAACL'15*, pages 1505–1515.

Over, P., Fiscus, J., Sanders, G., Joy, D., Michel, M., Awad, G., Smeaton, A., Kraaij, W., and Quénot, G. (2014). Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID*, page 52.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *Proc. NAACL'10 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., and Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Ren, M., Kiros, R., and Zemel, R. (2015a). Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961.

Ren, S., He, K., Girshick, R., and Sun, J. (2015b). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Ren, Y., Van Deemter, K., and Pan, J. Z. (2010). Charting the potential of description logic for the generation of referring expressions. In *International Natural Language Generation Conference (INLG)*.

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., and Schiele, B. (2016). Grounding of textual phrases in images by reconstruction. In *ECCV*. Springer.

Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B. (2015). A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Rosenfeld, A. (1978). Iterative methods in image analysis. *Pattern Recognition*, 10(3):181–187.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *ICCV*.

Shih, K. J., Singh, S., and Hoiem, D. (2016). Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4613–4621.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. *Computer Vision–ECCV 2012*, pages 746–760.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:27–218.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., and Fidler, S. (2016). Movieqa: Understanding stories in movies through question-answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

Unal, M. E., Citamak, B., Yagcioglu, S., Erdem, A., Erdem, E., Cinbis, N. I., and Cakici, R. (2016). Tasviret: A benchmark dataset for automatic turkish description generation from images. In *Signal Processing and Communication Application Conference (SIU), 2016 24th*, pages 1977–1980. IEEE.

Van Deemter, K., Gatt, A., van Gompel, R. P., and Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. In *Topics in Cognitive Science*, volume 4(2), page 166183.

van Deemter, K., van der Sluis, I., and Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In *International Conference on Natural Language Generation (INLG)*.

van Miltenburg, E., Elliott, D., and Vossen, P. (2017). Cross-linguistic differences and similarities in image descriptions. *CoRR*, abs/1707.01736.

Vedantam, R., Zitnick, C. L., and Parikh, D. (2014). Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726.

Venugopalan, S., Hendricks, L. A., Mooney, R., and Saenko, K. (2016). Improving LSTM-based video description with linguistic knowledge mined from text. In *EMNLP'16*, pages 1961–1966.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., and Saenko, K. (2015). Translating videos to natural language using deep recurrent neural networks. *NAACL'15*.

Viethen, J. and Dale, R. (2008). The use of spatial relations in referring expression generation. In *International Natural Language Generation Conference (INLG)*.

Viethen, J. and Dale, R. (2010). Speaker-dependent variation in content selection for referring expression generation. In *Australasian Language Technology Workshop*.

Viethen, J., Mitchell, M., and Krahmer, E. (2013). Graphs and spatial relations in the generation of referring expressions. In *ENLG'13*.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Wang, L., Li, Y., Huang, J., and Lazebnik, S. (2018). Learning two-branch neural networks for image-text matching tasks. *IEEE Trans. Pattern Analysis and Machine Intelligence.*

Wang, L., Li, Y., and Lazebnik, S. (2016a). Learning deep structure-preserving image-text embeddings. In *CVPR.*

Wang, L., Schwing, A., and Lazebnik, S. (2017). Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5756–5766. Curran Associates, Inc.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2016b). Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*. Springer.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, 3(1):1191.

Wu, Q., Shen, C., Wang, P., Dick, A., and van den Hengel, A. (2017). Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence.*

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

Xiong, C., Merity, S., and Socher, R. (2016). Dynamic memory networks for visual and textual question answering. In *Int. Conf. on Machine Learning*, pages 2397–2406.

Xu, H. and Saenko, K. (2016). Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Yagcioglu, S., Erdem, E., Erdem, A., and Cakici, R. (2015). A distributed representation based query expansion approach for image captioning. In *Proc. ACL-IJCNLP'15*, volume 2, pages 106–111.

Yang, Y., Teo, C. L., Daumé III, H., and Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Proc. 16th Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 444–454, Edinburg, Scotland.

Yang, Z., He, X., Gao, J., Deng, L., and Smola, A. (2016). Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29.

Yatskar, M., Vanderwende, L., and Zettlemoyer, L. (2014). See no evil, say no evil: Description generation from densely labeled images. In *Proc. 3rd Joint Conference on Lexical and Computational Semantics*, pages 110–120, Dublin, Ireland.

Yoshikawa, Y., Shigeto, Y., and Takeuchi, A. (2017). Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823.*

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., and Berg, T. L. (2018). Mattnet: Modular attention network for referring expression comprehension. *arXiv preprint arXiv:1801.08186.*

Yu, L., Park, E., Berg, A. C., and Berg, T. L. (2015). Visual madlibs: Fill in the blank description generation and question answering. In *The IEEE International Conference on Computer Vision (ICCV).*

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. (2016a). Modeling context in referring expressions. In *ECCV'16*, pages 69–85.

Yu, L., Tan, H., Bansal, M., and Berg, T. L. (2017). A joint speakerlistener-reinforcer model for referring expressions. In *CVPR*.

Yu, Y., Ko, H., Choi, J., and Kim, G. (2016b). End-to-end concept word detection for video captioning, retrieval, and question answering. *CVPR*.

Zhao, Z., Yang, Q., Cai, D., He, X., and Zhuang, Y. (2017). Video question answering via hierarchical spatio-temporal attention networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2.

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

Zhu, L., Xu, Z., Yang, Y., and Hauptmann, A. G. (2015). Uncovering temporal context for video question and answering. *arXiv preprint arXiv:1511.04670*.

Zhu, Y., Groth, O., Bernstein, M., and Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004.

Zitnick, C. L. and Parikh, D. (2013). Bringing semantics into focus using visual abstraction. In *CVPR '13*, pages 3009–3016.