sentiments. It then takes inputs from the hidden states of both two RNNs for generating captions. This method can generate captions successfully given the appropriate sentiments.

3.6 LSTM vs. Others

Image captioning intersects computer vision and natural language processing (NLP) research. NLP tasks, in general, can be formulated as a sequence to sequence learning. Several neural language models such as neural probabilistic language model [11], log-bilinear models [105], skip-gram models [98], and recurrent neural networks (RNNs) [99] have been proposed for learning sequence to sequence tasks. RNNs have widely been used in various sequence learning tasks. However, traditional RNNs suffer from vanishing and exploding gradient problems and cannot adequately handle long-term temporal dependencies.

LSTM [54] networks are a type of RNN that has special units in addition to standard units. LSTM units use a memory cell that can maintain information in memory for long periods of time. In recent years, LSTM based models have dominantly been used in sequence to sequence learning tasks. Another network, Gated Recurrent Unit (GRU) [25] has a similar structure to LSTM but it does not use separate memory cells and uses fewer gates to control the flow of information.

However, LSTMs ignore the underlying hierarchical structure of a sentence. They also require significant storage due to long-term dependencies through a memory cell. In contrast, CNNs can learn the internal hierarchical structure of the sentences and they are faster in processing than LSTMs. Therefore, recently, convolutional architectures are used in other sequence to sequence tasks, e.g., conditional image generation [137] and machine translation [42, 43, 138].

Inspired by the above success of CNNs in sequence learning tasks, Gu et al. [51] proposed a CNN language model-based image captioning method. This method uses a language-CNN for statistical language modelling. However, the method cannot model the dynamic temporal behaviour of the language model only using a language-CNN. It combines a recurrent network with the language-CNN to model the temporal dependencies properly. Aneja et al. [5] proposed a convolutional architecture for the task of image captioning. They use a feed-forward network without any recurrent function. The architecture of the method has four components: (i) input embedding layer (ii) image embedding layer (iii) convolutional module, and (iv) output embedding layer. It also uses an attention mechanism to leverage spatial image features. They evaluate their architecture on the challenging MSCOCO dataset and shows comparable performance to an LSTM based method on standard metrics.

Wang et al. [147] proposed another CNN+CNN based image captioning method. It is similar to the method of Aneja et al. except that it uses a hierarchical attention module to connect the vision-CNN with the language-CNN. The authors of this method also investigate the use of various hyperparameters, including the number of layers and the kernel width of the language-CNN. They show that the influence of the hyperparameters can improve the performance of the method in image captioning.

4 DATASETS AND EVALUATION METRICS

A number of datasets are used for training, testing, and evaluation of the image captioning methods. The datasets differ in various perspective such as the number of images, the number of captions per image, format of the captions, and image size. Three datasets: Flickr8k [55], Flickr30k [113], and MS COCO Dataset [83] are popularly used. These datasets together with others are described in Section 4.1. In this section, we show sample images with their captions generated by image captioning methods on MS COCO, Flickr30k, and Flickr8k datasets. A number of evaluation metrics are used to measure the quality of the generated captions compared to the ground-truth. Each metric applies its own technique for computation and has distinct advantages. The commonly used evaluation

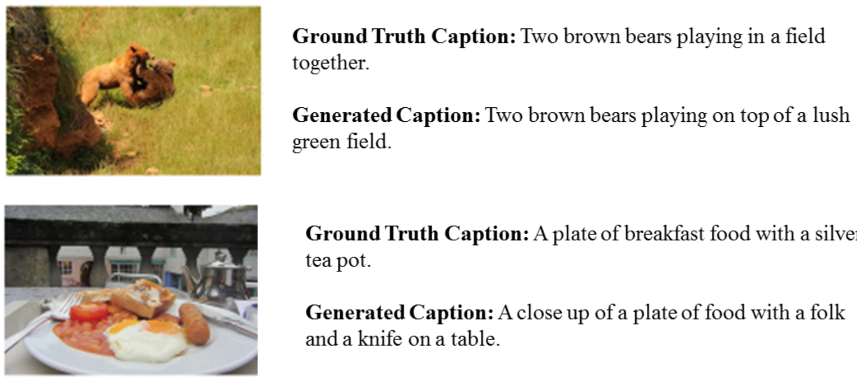


Fig. 11. Captions generated by Wu et al. [149] on some sample images from the MS COCO dataset.

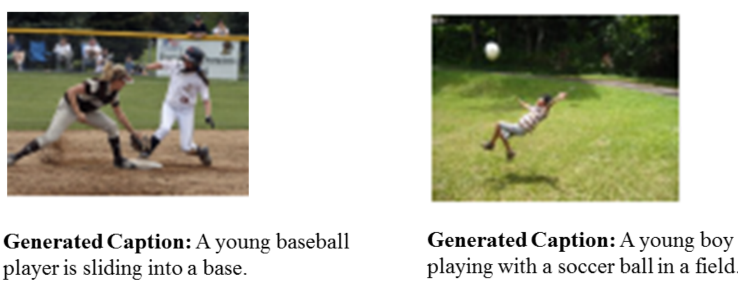


Fig. 12. Captions generated by Chen et al. [22] on some sample images from the Flickr30k dataset.

metrics are discussed in Section 4.2. A summary of deep learning-based image captioning methods with their datasets and evaluation metrics are listed in Table 2.

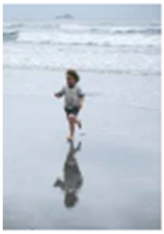
4.1 Datasets

4.1.1 MS COCO Dataset. Microsoft COCO Dataset [83] is a very large dataset for image recognition, segmentation, and captioning. There are various features of MS COCO dataset such as object segmentation, recognition in context, multiple objects per class, more than 300,000 images, more than 2 million instances, 80 object categories, and 5 captions per image. Many image captioning methods [26, 39, 61, 112, 119, 126, 135, 144, 149, 151, 156] use the dataset in their experiments. For example, Wu et al. [149] use MS COCO dataset in their method and the generated captions of two sample images are shown in Figure 11.

4.1.2 Flickr30K Dataset. Flickr30K [113] is a dataset for automatic image description and grounded language understanding. It contains 30k images collected from Flickr with 158k captions provided by human annotators. It does not provide any fixed split of images for training, testing, and validation. Researchers can choose their own choice of numbers for training, testing, and validation. The dataset also contains detectors for common objects, a color classifier, and a bias towards selecting larger objects. Image captioning methods such as [22, 65, 142, 144, 150] use this dataset for their experiments. For example, performed their experiment on Flickr30k dataset. The generated captions by Chen et al. [22] of two sample images of the dataset are shown in Figure 12.

Reference	Datasets	Evaluation Metrics
Kiros et al. 2014 [69]	IAPR TC-12,SBU	BLEU, PPLX
Kiros et al. 2014 [70]	Flickr 8K, Flickr 30K	R@K, mrank
Mao et al. 2014 [95]	IAPR TC-12, Flickr 8K/30K	BLEU, R@K, mrank
Karpathy et al. 2014 [66]	PASCAL1K, Flickr 8K/30K	R@K, mrank
Mao et al. 2015 [94]	IAPR TC-12, Flickr 8K/30K, MS COCO	BLEU, R@K, mrank
Chen et al. 2015 [23]	PASCAL, Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Fang et al. 2015 [33]	PASCAL, MS COCO	BLEU, METEOR, PPLX
Jia et al. 2015 [59]	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Karpathy et al. 2015 [65]	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Vinyals et al. 2015 [142]	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Xu et al. 2015 [152]	Flickr 8K/30K, MS COCO	BLEU, METEOR
Jin et al. 2015 [61]	Flickr 8K/30K, MS COCO	BLEU, METEOR, ROUGE, CIDEr
Wu et al. 2016 [151]	MS COCO	BLEU, METEOR, CIDEr
Sugano et at. 2016 [129]	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Mathews et al. 2016 [97]	MS COCO, SentiCap	BLEU, METEOR, ROUGE, CIDEr
Wang et al. 2016 [144]	Flickr 8K/30K, MS COCO	BLEU, R@K
Johnson et al. 2016 [62]	Visual Genome	METEOR, AP, IoU
Mao et al. 2016 [92]	ReferIt	BLEU, METEOR, CIDEr
Wang et al. 2016 [146]	Flickr 8K	BLEU, PPL, METEOR
Tran et al. 2016 [135]	MS COCO, Adobe-MIT, Instagram	Human Evaluation
Ma et al. 2016 [90]	Flickr 8k, UIUC	BLEU, R@K
You et al. 2016 [156]	Flickr 30K, MS COCO	BLEU, METEOR, ROUGE, CIDEr
Yang et al. 2016 [153]	Visual Genome	METEOR, AP, IoU
Anne et al. 2016 [6]	MS COCO, ImageNet	BLEU, METEOR
Yao et al. 2017 [155]	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Lu et al. 2017 [88]	Flickr 30K, MS COCO	BLEU, METEOR, CIDEr
Chen et al. 2017 [21]	Flickr 8K/30K, MS COCO	BLEU, METEOR, ROUGE, CIDEr
Gan et al. 2017 [41]	Flickr 30K, MS COCO	BLEU, METEOR, CIDEr
Pedersoli et al. 2017 [112]	MS COCO	BLEU, METEOR, CIDEr
Ren et al. 2017 [119]	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Park et al. 2017 [111]	Instagram	BLEU, METEOR, ROUGE, CIDEr
Wang et al. 2017 [148]	MS COCO, Stock3M	SPICE, METEOR, ROUGE, CIDEr
Tavakoli et al. 2017 [134]	MS COCO, PASCAL 50S	BLEU, METEOR, ROUGE, CIDEr
Liu et al. 2017 [84]	Flickr 30K, MS COCO	BLEU, METEOR
Gan et al. 2017 [39]	FlickrStyle10K	BLEU, METEOR, ROUGE, CIDEr
Dai et al. 2017 [26]	Flickr 30K, MS COCO	E-NGAN, E-GAN, SPICE, CIDEr
Shetty et al. 2017 [126]	MS COCO	Human Evaluation, SPICE, METEOR
Liu et al. 2017 [85]	MS COCO	SPIDEr, Human Evaluation
Gu et al. 2017 [51]	Flickr 30K, MS COCO	BLEU, METEOR, CIDEr, SPICE
Yao et al. 2017 [154]	MS COCO, ImageNet	METEOR
Rennie et al. 2017 [120]	MS COCO	BLEU, METEOR, CIDEr, ROUGE
Vsub et al. 2017 [140]	MS COCO, ImageNet	METEOR
Zhang et al. 2017 [161]	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Wu et al. 2018 [150]	Flickr 8K/30K, MS COCO	BLEU, METEOR, CIDEr
Aneja et al. 2018 [5]	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Wang et al. 2018 [147]	MS COCO	BLEU, METEOR, ROUGE, CIDEr

Table 2. An overview of methods, datasets, and evaluation metrics



Ground Truth Caption: A little boy runs away from the approaching waves of the ocean.

Generated Caption: A young boy is running on the beach.

Ground Truth Caption: A brunette girl wearing sunglasses and a yellow shirt.

Generated Caption: A woman in a black shirt and sunglasses smiles.

Fig. 13. Captions generated by Jia et al. [59] on some sample images from the Flickr8k dataset.

4.1.3 Flickr8K Dataset. Flickr8k [55] is a popular dataset and has 8000 images collected from Flickr. The training data consists of 6000 images, the test and development data, each consists of 1,000 images. Each image in the dataset has 5 reference captions annotated by humans. A number of image captioning methods [21, 59, 61, 144, 150, 152] have performed experiments using the dataset. Two sample results by Jia et al. [59] on this dataset are shown in Figure 13.

4.1.4 Visual Genome Dataset. Visual Genome dataset [72] is another dataset for image captioning. Image captioning requires not only to recognise the objects of an image but it also needs reasoning their interactions and attributes. Unlike the first three datasets where a caption is given to the whole scene, Visual Genome dataset has separate captions for multiple regions in an image. The dataset has seven main parts: region descriptions, objects, attributes, relationships, region graphs, scene graphs, and question answer pairs. The dataset has more than 108k images. Each image contains an average of 35 objects, 26 attributes, and 21 pairwise relationships between objects.

4.1.5 Instagram Dataset. Tran et al. [135] and Park et al. [111] created two datasets using images from Instagram which is a photo-sharing social networking services. The dataset of Tran et al. has about 10k images which are mostly from celebrities. However, Park et al. used their dataset for hashtag prediction and post-generation tasks in social media networks. This dataset contains 1.1m posts on a wide range of topics and a long hashtag lists from 6.3k users.

4.1.6 IAPR TC-12 Dataset. IAPR TC-12 dataset [50] has 20k images. The images are collected from various sources such as sports, photographs of people, animals, landscapes and many other locations around the world. The images of this dataset have captions in multiple languages. Images have multiple objects as well.

4.1.7 Stock3M Dataset. Stock3M dataset has 3,217,654 images uploaded by users and it is 26 times larger than MSCOCO dataset. The images of this dataset have a diversity of content.

4.1.8 MIT-Adobe FiveK dataset. MIT-Adobe FiveK [19] dataset consists of 5,000 images. These images contain a diverse set of scenes, subjects, and lighting conditions and they are mainly about people, nature, and man-made objects.

4.1.9 *FlickrStyle10k Dataset.*
FlickrStyle10k dataset has 10,000 Flickr images with stylized captions. The training data consists of 7000 images. The validation and test data consists of 2,000 and 1,000 images respectively. Each image contains romantic, humorous, and factual captions.

4.2 Evaluation Metrics

4.2.1 *BLEU.*
BLEU (Bilingual evaluation understudy) [110] is a metric that is used to measure the quality of machine generated text. Individual text segments are compared with a set of reference texts and scores are computed for each of them. In estimating the overall quality of the generated text, the computed scores are averaged. However, syntactical correctness is not considered here. The performance of the BLEU metric is varied depending on the number of reference translations and the size of the generated text. Subsequently, Papineni et al. introduced a modified precision metric. This metrics uses n-grams. BLEU is popular because it is a pioneer in automatic evaluation of machine translated text and has a reasonable correlation with human judgements of quality [20, 29]. However, it has a few limitations such as BLEU scores are good only if the generated text is short [20]. There are some cases where an increase in BLEU score does not mean that the quality of the generated text is good [82].

4.2.2 *ROUGE.*
ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [81] is a set of metrics that are used for measuring the quality of text summary. It compares word sequences, word pairs, and n-grams with a set of reference summaries created by humans. Different types of ROUGE such as ROUGE-1, 2, ROUGE-W, ROUGE-SU4 are used for different tasks. For example, ROUGE-1 and ROUGE-W are appropriate for single document evaluation whereas ROUGE-2 and ROUGE-SU4 have good performance in short summaries. However, ROUGE has problems in evaluating multi-document text summary.

4.2.3 *METEOR.*
METEOR (Metric for Evaluation of Translation with Explicit ORdering) [9] is another metric used to evaluate the machine translated language. Standard word segments are compared with the reference texts. In addition to this, stems of a sentence and synonyms of words are also considered for matching. METEOR can make better correlation at the sentence or the segment level.

4.2.4 *CIDEr.*
CIDEr (Consensus-based Image Descripton Evaluation) [139] is an automatic consensus metric for evaluating image descriptions. Most existing datasets have only five captions per image. Previous evaluation metrics work with these small number of sentences and are not enough to measure the consensus between generated captions and human judgement. However, CIDEr achieves human consensus using term frequency-inverse document frequency (TF-IDF) [121].

4.2.5 *SPICE.*
SPICE (Semantic Propositional Image Caption Evaluation) [3] is a new caption evaluation metric based on semantic concept. It is based on a graph-based semantic representation called scene-graph [63, 123]. This graph can extract the information of different objects, attributes and their relationships from the image descriptions.

Existing image captioning methods compute log-likelihood scores to evaluate their generated captions. They use BLEU, METEOR, ROUGE, SPICE, and CIDEr as evaluation metrics. However, BLEU, METEOR, ROUGE are not well correlated with human assessments of quality. SPICE and CIDEr have better correlation but they are hard to optimize. Liu et al. [85] introduced a new captions evaluation metric that is a good choice by human raters. It is developed by a combination of SPICE and CIDEr, and termed as SPIDEr. It uses a policy gradient method to optimize the metrics.

The quality of image captioning depends on the assessment of two main aspects: adequacy and fluency. An evaluation metric needs to focus on a diverse set of linguistic features to achieve these aspects. However, commonly used evaluation metrics consider only some specific features (e.g., lexical or semantic) of languages. Sharif et al. [125] proposed learning-based composite metrics for evaluation of image captions. The composite metric incorporates a set of linguistic features to achieve the two main aspects of assessment and shows improved performances.

5 COMPARISON ON BENCHMARK DATASETS AND COMMON EVALUATION METRICS

While formal experimental evaluation was left out of the scope of this paper, we present a brief analysis of the experimental results and the performance of various techniques as reported. We cover three sets of results:

(1) We find a number of methods use the first three datasets listed in Section 4.1. and a number of commonly used evaluation metrics to present the results. These results are shown in Table 3.
(2) A few methods fall into the following groups: Attention-based and Other deep learning-based (Reinforcement learning and GAN-based methods) image captioning. The results of such methods are shown in Tables 4 and 5, respectively.
(3) We also list the methods that proivide top two results scored on each evaluation metric on the MSCOCO dataset. These results are shown in Table 6.

As shown in Table 3, on Flickr8k, Mao et al. achieved 0.565, 0.386, 0.256, and 0.170 on BLEU-1, BLEU-2, BLEU-3, and BLEU-4 , respectively. For Flickr30k dataset, the scores are 0.600, 0.410, 0.280, and 0.190, respectively which are higher than the Flickr8k scores. The highest scores were achieved on the MSCOCO dataset. The higher results on a larger dataset follows the fact that a large dataset has more data, comprehensive representation of various scenes, complexities, and their own natural context. The results of Jia et al. are similar for Flickr8k and Flickr30k datasets but higher on MSCOCO dataset. The method uses visual space for mapping image-features and text features. Mao et al. use multimodal space for the mapping of image-features and text features. On the other hand, Jia et al. use visual space for the mapping. Moreover, the method uses an Encoder-Decoder architecture where it can guide the decoder part dynamically. Consequently, this method performs better than Mao et al.

Xu et al. also perform better on MSCOCO dataset. This method outperformed both Mao et al. and Jia et al. The main reason behind this is that it uses an attention mechanism which focuses only on relevant objects of the image. The semantic concept-based methods can generate semantically rich captions. Wu et al. proposed a semantic concept-based image captioning method. This method first predicts the attributes of different objects from the image and then adds these attributes with the captions which are semantically meaningful. In terms of performance, the method is superior to all the methods mentioned in Table 3.

Table 4 shows the results of attention-based based methods on MSCOCO dataset. Xu et al.'s stochastic hard attention produced better results than deterministic soft attention. However, these results were outperformed by Jin et al. which can update its attention based on the scene-specific context.

Wu et al. 2016 and Pedersoli et al. 2017 only show BLEU-4 and METEOR scores which are higher than the aforementioned methods. The method of Wu et al. uses an attention mechanism with a review process. The review process checks the focused attention in every time step and updates it if necessary. This mechanism helps to achieve better results than the prior attention-based methods. Pedersoli et al. propose a different attention mechanism that maps the focused image regions directly with the caption words instead of LSTM state. This behavior of the method drives it to achieve top performances among the mentioned attention-based methods in Table 4.

Dataset	Method	Category	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Mao et al. 2015 [94]	MS,SL,WS	0.565	0.386	0.256	0.170	-
	Jia et al. 2015 [59]	VS,SL,WS,EDA	0.647	0.459	0.318	0.216	0.201
	Xu et al. 2015 [152]	VS,SL,WS,EDA,AB	0.670	0.457	0.314	0.213	0.203
	Wu et al. 2018 [150]	VS,SL,WS,EDA,SCB	0.740	0.540	0.380	0.270	-
Flickr30k	Mao et al. 2015 [94]	MS,SL,WS	0.600	0.410	0.280	0.190	-
	Jia et al. 2015 [59]	VS,SL,WS,EDA	0.646	0.466	0.305	0.206	0.179
	Xu et al. 2015 [152]	VS,SL,WS,EDA,AB	0.669	0.439	0.296	0.199	0.184
	Wu et al. 2018 [150]	VS,SL,WS,EDA,SCB	0.730	0.550	0.400	0.280	-
MSCOCO	Mao et al. 2015 [94]	MS,SL,WS	0.670	0.490	0.350	0.250	-
	Jia et al. 2015 [59]	VS,SL,WS,EDA	0.670	0.491	0.358	0.264	0.227
	Xu et al. 2015 [152]	VS,SL,WS,EDA,AB	0.718	0.504	0.357	0.250	0.230
	Wu et al. 2018 [150]	VS,SL,WS,EDA,SCB	0.740	0.560	0.420	0.310	0.260

Table 3. Performance of different image captioning methods on three benchmark datasets and commonly used evaluation metrics.

Method	Category	MS COCO						
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Xu et al. 2015 [152], soft	VS,SL,WS,EDA,VC	0.707	0.492	0.344	0.243	0.239	-	-
Xu et al. 2015 [152], hard	VS,SL,WS,EDA,VC	0.718	0.504	0.357	0.250	0.230	-	-
Jin et al. 2015 [61]	VS,SL,WS,EDA,VC	0.697	0.519	0.381	0.282	0.235	0.509	0.838
Wu et al. 2016 [151]	VS,SL,WS,EDA,VC	-	-	-	0.290	0.237	-	0.886
Pedersoli et al. 2017 [112]	VS,SL,WS,EDA,VC	-	-	-	0.307	0.245	-	0.938

Table 4. Performance of attention-based image captioning methods on MSCOCO dataset and commonly used evaluation metrics.

Reinforcement learning-based (RL) and GAN-based methods are becoming increasingly popular. We name them as "Other Deep Learning-based Image Captioning". The results of the methods of this group are shown in Table 5. The methods do not have results on commonly used evaluation metrics. However, they have their own potentials to generate the descriptions for the image.

Shetty et al. employed adversarial training in their image captioning method. This method is capable to generate diverse captions. The captions are less-biased with the ground-truth captions compared to the methods use maximum likelihood estimation. To take the advantages of RL, Ren et al. proposed a method that can predict all possible next words for the current word in current time step. This mechanism helps them to generate contextually more accurate captions. Actor-critic of RL are similar to the Generator and the Discriminator of GAN. However, at the beginning of the training, both actor and critic do not have any knowledge about data. Zhang et al. proposed an actor-critic-based image captioning method. This method is capable of predicting the ultimate captions at its early stage and can generate more accurate captions than other reinforcement learning-based methods.

We found that the performance of a technique can vary across different metrics. Table 6 shows the methods based on the top two scores on every individual evaluation metric. For example, Lu et al., Gan et al., and Zhang et al. are within the top two methods based on the scores achieved on BLEU-n and METEOR metrics. BLEU-n metrics use variable length phrases of generated captions to match against ground-truth captions. METEOR [9] considers the precision, recall, and the alignments of the matched tokens. Therefore, the generated captions by these methods have good precision

**A dash (-) in the tables of this paper indicates results are unavailable*

Method	Category	MS COCO							
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Shetty et al. 2017_{GAN} [126]	VS,ODL,WS,EDA	-	-	-	-	0.239	-	-	0.167
Ren et al. 2017_{RL} [119]	VS,ODL,WS,EDA	0.713	0.539	0.403	0.304	0.251	0.525	0.937	-
Zhang et al. 2017_{RL} [161]	VS,ODL,WS,EDA	-	-	-	0.344	0.267	0.558	1.162	-

Table 5. Performance of Other Deep learning-based image captioning methods on MSCOCO dataset and commonly used evaluation metrics.

Method	Category	MSCOCO							
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Lu et al. 2017 [88]	VS,SL,WS,EDA,AB	***0.742***	***0.580***	**0.439**	0.332	**0.266**	-	**1.085**	-
Gan et al. 2017 [41]	VS,SL,WS,CA,SCB	**0.741**	**0.578**	***0.444***	**0.341**	0.261	-	1.041	-
Zhang et al. 2017 [161]	VS,ODL,WS,EDA	-	-	-	***0.344***	***0.267***	***0.558***	***1.162***	-
Rennie et al. 2017 [120]	VS,ODL,WS,EDA	-	-	-	.319	0.255	**0.543**	1.06	-
Yao et al. 2017 [155]	VS,SL,WS,EDA,SCB	0.734	0.567	0.430	0.326	0.254	0.540	1.00	***0.186***
Gu et al. 2017 [51]	VS,SL,WS,EDA	0.720	0.550	0.410	0.300	0.240	-	0.960	**0.176**

Table 6. Top two methods based on different evaluation metrics and MSCOCO dataset (Bold and Italic indicates the best result; Bold indicates the second best result).

and recall accuracy as well as the good similarity in word level. ROUGE-L evaluates the adequacy and fluency of generated captions, whereas CIDEr focuses on grammaticality and saliency. SPICE can analyse the semantics of the generated captions. Zhang et al., Rennie et al., and Lu et al. can generate captions, which have adequacy, fluency, saliency, and are grammaticality correct than other methods in Table 6. Gu et al. and Yao et al. perform well in generating semantically correct captions.

6 DISCUSSIONS AND FUTURE RESEARCH DIRECTIONS

Many deep learning-based methods have been proposed for generating automatic image captions in the recent years. Supervised learning, reinforcement learning, and GAN based methods are commonly used in generating image captions. Both visual space and multimodal space can be used in supervised learning-based methods. The main difference between visual space and multimodal space occurs in mapping. Visual space-based methods perform explicit mapping from images to descriptions. In contrast, multimodal space-based methods incorporate implicit vision and language models. Supervised learning-based methods are further categorized into Encoder-Decoder architecture-based, Compositional architecture-based, Attention-based, Semantic concept-based, Stylized captions, Dense image captioning, and Novel object-based image captioning.

Encoder-Decoder architecture-based methods use a simple CNN and a text generator for generating image captions. Attention-based image captioning methods focus on different salient parts of the image and achieve better performance than encoder-decoder architecture-based methods. Semantic concept-based image captioning methods selectively focus on different parts of the image and can generate semantically rich captions. Dense image captioning methods can generate region based image captions. Stylized image captions express various emotions such as romance, pride, and shame. GAN and RL based image captioning methods can generate diverse and multiple captions.

MSCOCO, Flickr30k and Flickr8k dataset are common and popular datasets used for image captioning. MSCOCO dataset is very large dataset and all the images in these datasets have multiple captions. Visual Genome dataset is mainly used for region based image captioning. Different evaluation metrics are used for measuring the performances of image captions. BLEU metric is good for small sentence evaluation. ROUGE has different types and they can be used for evaluating different types of texts. METEOR can perform an evaluation on various segments of a caption. SPICE is better in understanding semantic details of captions compared to other evaluation metrics.

Although success has been achieved in recent years, there is still a large scope for improvement. Generation based methods can generate novel captions for every image. However, these methods fail to detect prominent objects and attributes and their relationships to some extent in generating accurate and multiple captions. In addition to this, the accuracy of the generated captions largely depends on syntactically correct and diverse captions which in turn rely on powerful and sophisticated language generation model. Existing methods show their performances on the datasets where images are collected from the same domain. Therefore, working on open domain dataset will be an interesting avenue for research in this area. Image-based factual descriptions are not enough to generate high-quality captions. External knowledge can be added in order to generate attractive image captions. Supervised learning needs a large amount of labelled data for training. Therefore, unsupervised learning and reinforcement learning will be more popular in future in image captioning.

7 CONCLUSIONS

In this paper, we have reviewed deep learning-based image captioning methods. We have given a taxonomy of image captioning techniques, shown generic block diagram of the major groups and highlighted their pros and cons. We discussed different evaluation metrics and datasets with their strengths and weaknesses. A brief summary of experimental results is also given. We briefly outlined potential research directions in this area. Although deep learning-based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time.

ACKNOWLEDGEMENTS

This work was partially supported by an Australian Research Council grant DE120102960.

REFERENCES

[1] Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 115–118.

[2] Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 1250–1258.

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

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998* (2017).

[5] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5561–5570.

[6] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, 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*.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

[8] Shuang Bai and Shan An. 2018. A Survey on Automatic Image Caption Generation. *Neurocomputing*.

[9] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, Vol. 29. 65–72.

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

[11] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb, 1137–1155.

[12] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics* 22, 1 (1996), 39–71.

[13] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank, et al. 2016. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. *Journal of Artificial Intelligence Research (JAIR)* 55, 409–442.

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

[15] Cristian Bodnar. 2018. Text to Image Synthesis Using Generative Adversarial Networks. *arXiv preprint arXiv:1805.00676*.

[16] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 1247–1250.

[17] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 144–152.

[18] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. 2011. Eye movement analysis for activity recognition using electrooculography. *IEEE transactions on pattern analysis and machine intelligence* 33, 4 (2011), 741–753.

[19] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 97–104.

[20] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the Role of Bleu in Machine Translation Research.. In *EACL*, Vol. 6. 249–256.

[21] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 6298–6306.

[22] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, Adapt and Tell: Adversarial Training of Cross-domain Image Captioner. In *The IEEE International Conference on Computer Vision (ICCV)*, Vol. 2.

[23] Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2422–2431.

[24] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Association for Computational Linguistics*. 103–111.

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

[26] Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. 2017. Towards Diverse and Natural Image Descriptions via a Conditional GAN. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2989–2998.

[27] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. IEEE, 886–893.

[28] Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, Vol. 6. Genoa Italy, 449–454.

[29] Etienne Denoual and Yves Lepage. 2005. BLEU in characters: towards automatic MT evaluation in languages without word delimiters. In *Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing*. 81–86.

[30] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809* (2015).

[31] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[32] Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1292–1302.

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

[34] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*. Springer, 15–29.

[35] Alireza Fathi, Yin Li, and James M Rehg. 2012. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*. Springer, 314–327.

[36] William Fedus, Ian Goodfellow, and Andrew M Dai. 2018. Maskgan: Better text generation via filling in the _. *arXiv preprint arXiv:1801.07736*.

[37] Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. 2013. Learning distributions over logical forms for referring expression generation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1914–1925.

[38] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.

[39] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3137–3146.

[40] Chuang Gan, Tianbao Yang, and Boqing Gong. 2016. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 87–97.

[41] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 1141–1150.

[42] Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.

[43] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

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

[45] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[46] Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 410–419.

[47] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*. Springer, 529–545.

[48] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[49] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. DRAW: A recurrent neural network for image generation. In *Proceedings of Machine Learning Research*. 1462–1471.

[50] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, Vol. 5. 10.

[51] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2017. An empirical study of language cnn for image captioning. In *Proceedings of the International Conference on Computer Vision (ICCV)*. 1231–1240.

[52] Yahong Han and Guang Li. 2015. Describing images with hierarchical concepts and object class localization. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 251–258.

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

[54] Sepp Hochreiter and JÃijrgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[55] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[56] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*. 5967–5976.

[57] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. In *Advances in Neural Information Processing Systems*. 2017–2025.

[58] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*.

[59] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*. 2407–2415.

[60] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1072–1080.

[61] Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. 2015. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv preprint arXiv:1506.06272*.

[62] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4565–4574.

[63] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3668–3678.

[64] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410* (2016).

[65] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[66] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*. 1889–1897.

[67] S Karthikeyan, Vignesh Jagadeesh, Renuka Shenoy, Miguel Ecksteinz, and BS Manjunath. 2013. From where and how to what we see. In *Proceedings of the IEEE International Conference on Computer Vision*. 625–632.

[68] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes.. In *EMNLP*. 787–798.

[69] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 595–603.

[70] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. In *Workshop on Neural Information Processing Systems (NIPS))*.

[71] Vijay R Konda and John N Tsitsiklis. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*. 1008–1014.

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

[73] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[74] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.

[75] Akshi Kumar and Shivali Goel. 2017. A survey of evolution of image captioning techniques. *International Journal of Hybrid Intelligent Systems* Preprint, 1–19.

[76] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 359–368.

[77] Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, and Yejin Choi. 2014. TREETALK: Composition and Compression of Trees for Image Descriptions. *TACL* 2, 10 (2014), 351–362.

[78] RÃĬmi Lebret, Pedro O Pinheiro, and Ronan Collobert. 2015. Simple image description generator via a linear phrase-based approach. *Workshop on International Conference on Learning Representations (ICLR)*.

[79] Yann LeCun, LÃĬon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[80] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 220–228.

[81] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, Vol. 8. Barcelona, Spain.

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

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

[84] Chenxi Liu, Junhua Mao, Fei Sha, and Alan L Yuille. 2017. Attention Correctness in Neural Image Captioning. In *AAAI*. 4176–4182.

[85] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE Internationa Conference on Computer Vision (ICCV)*, Vol. 3. 873–881.

[86] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[87] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[88] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 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)*. 3242–3250.

[89] Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *International Conference on Learning Representations (ICLR)*.

[90] Shubo Ma and Yahong Han. 2016. Describing images by feeding LSTM with structural words. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*. IEEE, 1–6.

[91] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR)*.

[92] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 11–20.

[93] Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L Yuille. 2015. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *Proceedings of the IEEE International Conference on Computer Vision*. 2533–2541.

[94] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *International Conference on Learning Representations (ICLR)*.

[95] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090* (2014).

[96] Oded Maron and TomÃąs Lozano-PÃľrez. 1998. A framework for multiple-instance learning. In *Advances in neural information processing systems*. 570–576.

[97] Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. SentiCap: Generating Image Descriptions with Sentiments.. In *AAAI*. 3574–3580.

[98] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[99] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

[100] Ajay K Mishra, Yiannis Aloimonos, Loong Fah Cheong, and Ashraf Kassim. 2012. Active visual segmentation. *IEEE transactions on pattern analysis and machine intelligence* 34, 4 (2012), 639–653.

[101] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 747–756.

[102] Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *Proceedings of the 6th international natural language generation conference*. Association for Computational Linguistics, 95–104.

[103] Margaret Mitchell, Kees Van Deemter, and Ehud Reiter. 2013. Generating Expressions that Refer to Visible Objects.. In *HLT-NAACL*. 1174–1184.

[104] Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*. ACM, 641–648.

[105] Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*. ACM, 641–648.

[106] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 160–167.

[107] Timo Ojala, Matti PietikÃďinen, and Topi MÃďenpÃďÃď. 2000. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision*. Springer, 404–420.

[108] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*. 1143–1151.

[109] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. 2014. Training object class detectors from eye tracking data. In *European Conference on Computer Vision*. Springer, 361–376.

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

[111] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 6432–6440.

[112] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of Attention for Image Captioning. In *Proceedings of the IEEE international conference on computer vision*. 1251–1259.

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

[114] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on learning Representations (ICLR)*.

[115] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *Proceedings of Machine Learning Research*, Vol. 48. 1060–1069.

[116] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[117] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. 2015. Multi-instance visual-semantic embedding. *arXiv preprint arXiv:1512.06963* (2015).

[118] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. 2016. Joint image-text representation by gaussian visual-semantic embedding. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 207–211.

[119] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep Reinforcement Learning-based Image Captioning with Embedding Reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 1151–1159.

[120] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 1179–1195.

[121] Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation* 60, 5 (2004), 503–520.

[122] Hosnieh Sattar, Sabine Muller, Mario Fritz, and Andreas Bulling. 2015. Prediction of search targets from fixations in open-world settings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 981–990.

[123] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, Vol. 2.

[124] Karthikeyan Shanmuga Vadivel, Thuyen Ngo, Miguel Eckstein, and BS Manjunath. 2015. Eye tracking assisted extraction of attentionally important objects from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3241–3250.

[125] Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. 2018. Learning-based Composite Metrics for Improved Caption Evaluation. In *Proceedings of ACL 2018, Student Research Workshop*. 14–20.

[126] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In *IEEE International Conference on Computer Vision (ICCV)*. 4155–4164.

[127] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

[128] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2 (2014), 207–218.

[129] Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203* (2016).

[130] Chen Sun, Chuang Gan, and Ram Nevatia. 2015. Automatic concept discovery from parallel text and visual corpora. In *Proceedings of the IEEE International Conference on Computer Vision*. 2596–2604.

[131] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[132] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*. 1057–1063.

[133] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[134] Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying Attention to Descriptions Generated by Image Captioning Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2487–2496.

[135] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich image captioning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 49–56.

[136] Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth International Natural Language Generation Conference*. Association for Computational Linguistics, 130–132.

[137] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*. 4790–4798.

[138] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[139] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[140] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1170–1178.

[141] Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*. Association for Computational Linguistics, 59–67.

[142] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[143] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence* 39, 4 (2017), 652–663.

[144] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 988–997.

[145] Heng Wang, Zengchang Qin, and Tao Wan. 2018. Text Generation Based on Generative Adversarial Nets with Latent Variables. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 92–103.

[146] Minsi Wang, Li Song, Xiaokang Yang, and Chuanfei Luo. 2016. A parallel-fusion RNN-LSTM architecture for image caption generation. In *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 4448–4452.

[147] Qingzhong Wang and Antoni B Chan. 2018. CNN+ CNN: Convolutional Decoders for Image Captioning. *arXiv preprint arXiv:1805.09019*.

[148] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W Cottrell. 2017. Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 7378–7387.

[149] Qi Wu, Chunhua Shen, Anton van den Hengel, Lingqiao Liu, and Anthony Dick. 2015. Image captioning with an intermediate attributes layer. *arXiv preprint arXiv:1506.01144* (2015).

[150] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence* 40, 6, 1367–1381.

[151] Zhilin Yang Ye Yuan Yuexin Wu and Ruslan Salakhutdinov William W Cohen. 2016. Encode, Review, and Decode: Reviewer Module for Caption Generation. In *30th Conference on Neural Image Processing System(NIPS)*.

[152] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. 2048–2057.

[153] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. 2016. Dense Captioning with Joint Inference and Visual Context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1978–1987.

[154] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating copying mechanism in image captioning for learning novel objects. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5263–5271.

[155] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision (ICCV)*. 4904–4912.

[156] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4651–4659.

[157] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu Seqgan. 2017. Sequence generative adversarial nets with policy gradient.. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.
[158] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J Zelinsky, and Tamara L Berg. 2013. Exploring the role of gaze behavior and object detection in scene understanding. *Frontiers in psychology* 4 (2013).
[159] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.
[160] Gregory J Zelinsky. 2013. Understanding scene understanding. *Frontiers in psychology* 4 (2013).
[161] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. 2017. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*.

Received April 2018; Revised October 2018; Accepted October 2018