Deep Image Captioning: An Overview

I. Hrga*, M. Ivašić-Kos**
*Juraj Dobrila University of Pula/Faculty of Informatics, Pula, Croatia
**University of Rijeka/Department of Informatics, Rijeka, Croatia
ingrid.hrga@unipu.hr, marinai@uniri.hr

***Abstract* – Image captioning is a process of automatically describing an image with one or more natural language sentences. In recent years, image captioning has witnessed rapid progress, from initial template-based models to the current ones, based on deep neural networks. This paper gives an overview of issues and recent image captioning research, with a particular emphasis on models that use the deep encoder-decoder architecture. We discuss the advantages and disadvantages of different approaches, along with reviewing some of the most commonly used evaluation metrics and datasets.**

Keywords – image captioning, encoder-decoder, attention mechanism, deep neural networks

I. Introduction

Recent advances in deep learning methods on perceptual tasks, such as image classification and object detection [1, 2] have encouraged researchers to tackle even more difficult problems for which recognition is just a step towards to more complex reasoning about our visual world [3]. Image captioning is one of such tasks.

The aim of image captioning is to automatically describe an image with one or more natural language sentences. This is a problem that integrates computer vision and natural language processing, so its main challenges arise from the need of translating between two distinct, but usually paired, modalities [4]. First, it is necessary to detect objects on the scene and determine the relationships between them [5] and then, express the image content correctly with properly formed sentences. The generated description is still much different from the way people describe images because people rely on common sense and experience, point out important details and ignore objects and relationships that they imply [6]. Moreover, they often use imagination to make descriptions vivid and interesting.

Regardless of the existing limitations, image captioning has already been proven to have useful applications, such as helping visually impaired people in performing daily tasks. Automatically generated descriptions can also be used for content-based retrieval [7] or in social media communications.

Early image captioning approaches relied on the use of predefined templates, which were filled in based on the results of the detection of elements on the scene [8, 9]. However, the advantage of such bottom-up approaches in terms of the ability to capture details was not enough to keep them in the focus of research interest. Generated sentences were too simple, lacking the fluency of human writing. Moreover, such systems were heavily hand-designed, which constrained their flexibility. Some authors [10] have reformulated image captioning as a ranking task. Ranking-based approaches always return well-formed sentences, but they cannot generate new sentences or to describe compositionally new images [11], i.e., those containing objects that were observed during training but appear in different combinations on the test image. In contrast, today's state-of-the-art models are generative and neural networks based. They usually employ an encoder-decoder architecture by combining a Convolutional Neural Network (CNN) with a Recurrent Neural Network (RNN).

The rest of the paper is organized as follows: The next Section provides some background information on the typical architecture of image captioning systems. Section III. groups image captioning models according to the captioning task and describes relevant models for each type. Section IV. presents some of the most commonly used data sets, along with a description of how they were collected. Section V. lists the metrics and points to the problems that arise when evaluating generative approaches. The paper ends with a Conclusion.

II. Architecture and Learning Approaches

A. Encoder-Decoder Framework

Inspired by its success in Neural Machine Translation [12], many of the current state-of-the-art models for image captioning employ the encoder-decoder architecture (Fig. 1). In this architecture, the encoder is used to map the input into its real-valued fixed-dimensional vector representation. A decoder then generates output, conditioned on the representation produced by the encoder. The main advantage of such a system is that it can be trained end-to-end, meaning that the parameters of the whole network are learned together, thereby avoiding the problem of aligning several independent components.

Image captioning is often understood as a task of translating one modality, i.e. an image, into another modality, i.e. its description, so the encoder-decoder architecture has been successfully applied with a convolutional neural network (CNN) [13] on the encoder side, and a recurrent neural network (RNN) [14] on the decoder side.

A CNN acts as a feature extractor that is usually pretrained on a large dataset for a classification task [15]. A feature map from a convolutional layer or the vector representation from a fully-connected layer is then used as image representation. An RNN or one of its variants, such as the long short-term memory (LSTM) network [16], is employed for language modeling.

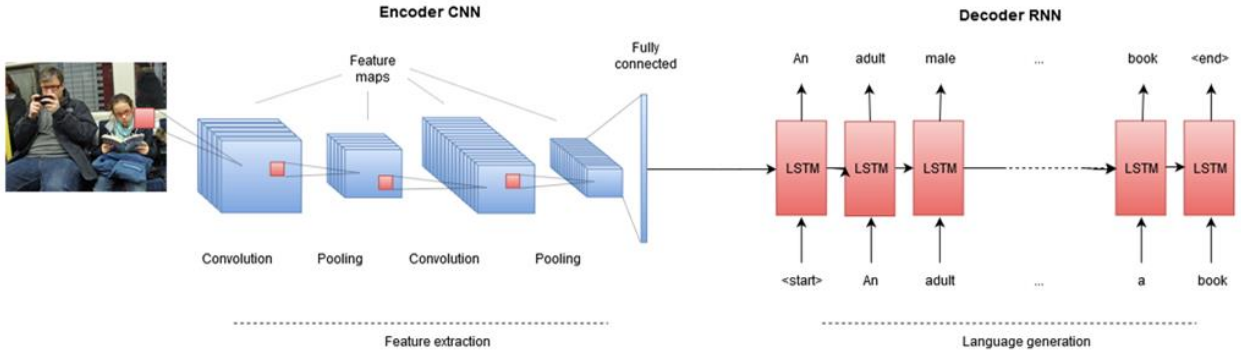

Figure 1. Encoder-decoder framework for image captioning: first a CNN encoder produces image representation (left), then an LSTM decoder generates caption conditioned on the representation produced by the encoder (right)

B. Learning

Majority of the encoder-decoder image captioning models use Maximum Likelihood Estimation (MLE) as their learning method. In the supervised learning setting, with training examples consisting of image-caption pairs, the model maximizes the probability of the correct caption given the image [17]:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{I},\boldsymbol{y}} \log p(\boldsymbol{y}|\boldsymbol{I};\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{I}$ is the image, $\boldsymbol{y} = \{y_1, \ldots, y_N\}$ is the corresponding caption of length *N*, $\boldsymbol{\theta}$ are the parameters of the model. The joint probability over words can be expressed as follows:

$$\log p(\boldsymbol{y}|\boldsymbol{I}) = \sum_{t=0}^{N} \log p(y_t|y_0, y_1, ..., y_{t-1}, \boldsymbol{I}) \tag{2}$$

where the dependency on $\boldsymbol{\theta}$is dropped for simplification.

To model $p(y_t|y_1, y_2, ..., y_{t-1}, \boldsymbol{I})$ usually an LSTM is employed, which is trained to predict the next word y_t conditioned on all the previously predicted words $(y_1, y_2, ..., y_{t-1})$ and the context vector $\boldsymbol{c}$ produced by the encoder [18, 32]:

$$p(y_t|y_1, y_2, ..., y_{t-1}, \boldsymbol{I}) = f(\boldsymbol{h}_t, \boldsymbol{c}) \tag{3}$$

where $f$is a nonlinear function that outputs the probability of y_t, $\boldsymbol{h}_t$is the hidden state of the LSTM at time step *t*.

Novel sentences can be generated by randomly sampling from the model's distribution or by using beam search [19, 17].

Although effective, some limitations of MLE learning have motivated the adoption of alternative learning methods. Reinforcement learning [20] can be used to address the exposure bias problem [21] and for the direct optimization of the standard evaluation metrics [22]. For increasing the diversity of generated captions, conditional GAN framework [23] or contrastive learning [24] were proposed.

C. Attention Mechanism

It was demonstrated in [18] that the fixed-length vector representation produced by the encoder is responsible for the degradation of the performance that occurs as the length of input increases. Regardless of the size of the input, in the basic encoder-decoder all the information is compressed into a context vector of a predefined size. Instead, the authors proposed to encode the input into a set of vectors.

The first work to employ an attention mechanism on the task of image captioning was [25]. In the proposed model, image features are extracted from a lower convolutional layer of a CNN as a set of L annotation vectors $a = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_L\}$summarizing a pre-defined spatial location of the image. To each annotation vector, a positive weight α_{ti} is assigned, indicating the amount of attention each image feature receives. The attention weight α_{ti} is computed by an attention model f_{att}:

$$e_{ti} = f_{\text{att}}(\boldsymbol{a}_i, \boldsymbol{h}_{t-1}) \tag{4}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}, \sum_{i=1}^{L} \alpha_{ti} = 1 \tag{5}$$

where $\boldsymbol{h}_{t-1}$ is the previous hidden state.

After obtaining the attention weights, the attention mechanism computes the context vector $\boldsymbol{z}$ as a dynamic representation of the relevant parts of the image at a given time step:

$$\boldsymbol{z}_t = \sum_{i=1}^{L} \alpha_{t,i} \boldsymbol{a}_i \tag{6}$$

The context vector is then used to update the hidden state of the decoder.

III. Image Captioning Types

We have grouped methods and models of image captioning given the task into three types: (1) standard image captioning, (2) image captioning with style and (3) cross-lingual and multilingual image captioning.

A. Standard Image Captioning

For the correct description of the image, it is necessary to:

(1) detect the content of the image in terms of objects, attributes, relationships, with the conclusion of what is new or interesting [26, 27],

(2) express the represented semantic content with properly formulated sentences [17] that are suitable for the image they describe [10].

The captions generated by most of the contemporary methods usually represent an objective and neutral description of the factual content of the scene.

An example of a model designed to generate new captions with previously unseen combinations of objects is reported in [11]. Authors proposed a multimodal Recurrent Neural Network (m-RNN) framework that is adapted to both the retrieval as well as to the sentence generation task. The model consists of a CNN and RNN, which interact with each other in a multimodal layer receiving three inputs: the word embedding layer, the recurrent layer, an image representation. A final soft-max layer generates the probability distribution of the next word.

In [17] authors introduce an end-to-end trainable Neural Image Caption (NIC) system, similar to [11] but with an LSTM variant of RNN as the decoder. The authors propose to use maximum likelihood estimation (MLE) principle for training the model. For its effectiveness, NIC became one of the most influential models, and other authors developed extensions on top of it [28, 29].

A similar end-to-end Long-Term Recurrent Convolutional Network (LRCN), combining a CNN encoder and an LSTM decoder, is introduced in [19]. Authors investigated the effects of different architectures and found that using LSTM instead of a simple RNN, combined with a more powerful CNN, contributed to better performance. Adding more LSTM layers did not bring expected improvements.

Different from the spatial attention model introduced in [25], the authors in [30] proposed a semantic attention model which combines different sources of visual information through a feedback process to attend to fine details in images while having an end-to-end trainable system. Top-down features, extracted from the last convolutional layer of a CNN, serve as a guide where to attend. A set of bottom-up attributes are detected as candidates for attention. Those with highest attention scores are then used by the attention mechanism which learns to attend to the semantically important concepts. Since irrelevant attributes may redirect attention to wrong concepts, attribute prediction plays a crucial role.

A similar approach is presented in [31] where authors combine top-down and bottom-up attention processing to calculate attention at the object-level. Instead of treating detected objects as bag-of-words that do not retain spatial information, they propose a different, feature-based approach. Bottom-up attention mechanism, based on Faster R-CNN [2], proposes a set of salient image regions. Combined with the more traditional top-down approach, this allows us to reveal the structure of the scene better and to interpret better the relationships between objects, which becomes important in dealing with compositionally new images.

Previously described attention models have a limitation in that they cannot selectively decide when to focus attention on the image. In [32] authors argue that directing attention to the image at every time step becomes unnecessary for words that do not have a corresponding visual signal such as “a “, “for” etc. They introduce an adaptive attention model that automatically decides whether to focus attention on the image or to use information stored in the decoder’s memory. An LSTM extension, called sentinel gate, produces an additional visual sentinel vector, which is used when the model decides not to attend to the image. The new context vector is modeled as a combination of the context vector of the spatial attention model and the visual sentinel vector. It was shown that the ability to decide when to attend to the image was also useful for better directing attention to the appropriate image regions, which allowed the model to achieve state-of-the-art results.

B. Image Captioning with Style

There are two lines of work focused on enriching captions with more emotional content. The first group of authors includes viewer’s attitude and emotions towards the image [29, 33, 34, 35], the second line of work includes emotional content from the image itself [36].

The authors in [29] were the first to incorporate positive and negative sentiments into captions. They proposed *SentiCap*, a switching RNN model with word-level regularization which emphasizes sentiments. Two networks, consisting of a CNN and an RNN, were used to generate stylized captions. One network was trained on a large image-caption dataset to generate standard factual descriptions, and the other was trained on a small dataset with sentiment polarity. Experiments showed that *SentiCap* was able to include the appropriate sentiment in 74% of the generated sentences.

The *SentiCap* model is limited in its ability to scale because it requires words labeled with sentiment strength. To address this issue, the authors in [33] propose StyleNet, an end-to-end trainable model which generates captions in a humorous or romantic style (Figure 2). A factored LSTM is used to factorize the weight matrices to account for factual and non-factual aspects of the sentences. Multi-task learning is used to optimize the generation of factual captions, and stylized language modeling. Almost 85% of the human evaluators found the stylized captions to be more attractive than the corresponding factual descriptions.

In [34] authors propose two mechanisms to inject sentiment into captions: (1) direct injection, in which sentiment is injected as an additional dimension at each time step, (2) injection by sentiment flow, in which sentiment is provided only at the first time step and then propagated over the whole sentence by a sentiment cell. Experiments showed that both methods were able to add sentiments, but direct injection generated more captions with sentiments.

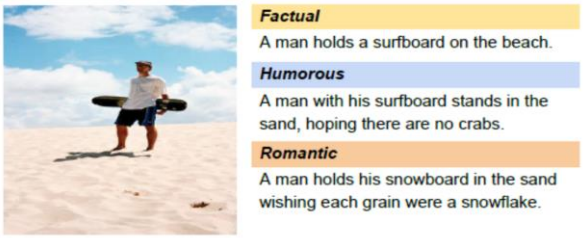

Figure 2. Comparison of factual and stylized captions [35]

FaceCap [36] model presents a different point of view and embraces the emotions detected in facial expressions of people depicted in the images. It relies on the use of a facial expression recognition model to extract facial features, which can be then used by an LSTM to generate captions. The authors observed that the model has improved in describing the actions on the scene.

C. Cross-Lingual and Multilingual Image Captioning

Cross-lingual and multilingual image captioning refers to the task of generating a caption in one language given a corpus of descriptions in one or more different languages [37].

Several approaches exist to tackle such tasks such as direct translation of generated captions, collecting a new dataset in a target language and its use for training, or learning a model from machine-translated texts. The first approach can give inferior results, among others because direct translations can worsen the errors in the generated descriptions [28]. Therefore, researchers are primarily focused on developing models that will be able to cope with different languages directly.

Authors in [38] treat the problem as a visually-grounded machine translation task in which the image is used to resolve ambiguity. In the proposed multilingual image description model, visual features are complemented with textual features of the source language (English) to generate captions in German.

The authors in [37] transfer the knowledge obtained from learning on English captions to generate captions in Japanese as the target language. The model was first pretrained on the large English dataset. Then, the trained LSTM was replaced with the new one, trained on the much smaller corpus of Japanese captions. The authors noticed that pretraining had the effect of learning on additional 10,000 images with Japanese captions. Moreover, the use of captions that are not direct translations made the model easier to scale.

Authors in [39] propose a single model capable of generating captions in multiple languages, but with a strong assumption that the images with captions in different languages are readily available. A token, provided as initial input, controls the choice of the target language.

Previous approaches rely on datasets with human-written captions in different languages. In [28] authors adopt a cheaper solution by using machine-translated text. To overcome the lack of fluency of such translations, they introduce a fluency estimation module to assign an importance score to the captions that are then chosen for training. Experiments were performed on English-Chinese datasets and showed that the model, trained on a smaller dataset from which less fluent sentence where excluded, achieved comparable results to the baseline, trained on all the machine-generated sentences (fluent or not).

IV. Datasets

The development of this research area greatly depends on the availability of large datasets that contain images with corresponding descriptions. In addition to the size of the dataset, an image captioning model benefits also significantly from the quality of captions in the spirit of natural language and their adaptation to a given task.

A. Collecting datasets

Images are collected primarily from photo-sharing services, mostly Flickr1 or by harvesting the web. Unlike the image gathering, obtaining appropriate descriptions turned out to be much more challenging.

As [10] pointed out, captions provided by users of photo-sharing websites are not suitable for the training of automatic image captioning systems. Such captions usually provide broader context, i.e., additional information that cannot be obtained by the image alone. Instead of using non-visual descriptions, [10] suggested focusing on general conceptual descriptions, i.e., those that refer to objects, attributes, events and other literal content of the image. Such descriptions are collected, on a large-scale, through crowdsourcing services, such as Amazon Mechanical Turk (AMT) [40, 10, 41], which involves defining a task that is performed by untrained workers [42]. Due to the low cost and high speed, crowdsourcing became the preferred way of collecting image descriptions for large-scale datasets.

B. Datasets

UIUC PASCAL Sentences [40] was one of the first image-caption datasets, consisting of 1,000 images and associated with five different descriptions collected via crowdsourcing. It was used by early image captioning systems [8], but due to its limited domain, small size, and relatively simple captions it is rarely used.

Flickr 30K [43] includes and extends previous Flickr 8K [10, 40] dataset. It consists of 31,783 images showing everyday activities, events, and scenes described by 158,915 captions obtained via crowdsourcing.

Microsoft COCO Captions [41] dataset contains more complex images of everyday objects and scenes. By adding human generated captions, two datasets were created: c5 with five captions for each of the more than 300K images and an additional, c40 dataset with 40 different captions for the randomly chosen 5K images. The c40 was created because it was observed [44] that some evaluation metrics benefit from more reference captions.

Flickr 30K and MS COCO Captions are widely accepted as benchmark datasets for image captioning by most models using deep neural networks.

V. Evaluation

Accessing the accuracy of automatically generated image captioning is a demanding task [44, 45]. The same image can be described in various ways, focusing on different parts of the image, using a different level of abstraction or different level of knowledge about objects on the scene. It is obvious that by emphasizing different aspects of the image, the resulting sentences can vary significantly while at the same time being entirely correct. In contrary, two captions can share most of the words and convey a different meaning.

Evaluation of automatically generated captions can be performed by human subjects, either by experts [10] or by untrained workers through crowdsourcing platforms [19, 22]. However, human-based evaluation creates additional costs; it is slow, subjective and difficult to reproduce [10, 46]. A better alternative is the use of automatic metrics, which, in turn, are fast, accurate and inexpensive [45]. To

[1] https://www.flickr.com

be useful, metrics should match the rating of human evaluators, but it turned out to be a goal that is difficult to achieve. Evaluation metrics should satisfy two criteria [22]: (1) captions that are considered good by humans should achieve high scores, (2) captions that achieve high scores should be considered good by humans.

Image captioning is sometimes compared [47] to language translation [8, 17] or with text summarization [48], which motivated the adaption of metrics developed initially for the evaluation of languages tasks [49, 50, 51]. All these metrics output a score indicating a similarity between the candidate sentence and the reference sentences.

BLEU [49] is a popular metric for machine translation evaluation and one of the first metrics used to evaluate image descriptions. It computes the geometric mean of n-gram precision scores multiplied by a brevity penalty in order to avoid overly short sentences.

METEOR [50] is another machine translation metric. It relies on the use of stemmers, WordNet [52] synonyms and paraphrases tables to identify matches between candidate sentence and reference sentences.

ROUGE [51] is a package of measures initially developed for the evaluation of text summaries. For image captioning, a variant ROUGEL is usually used, which computes F-measure based on the Longest Common Subsequence (LCS), i.e. a set of words shared by two sentences which occur in the same order, without requiring consecutive matches.

CIDEr [44] is a metric designed for the evaluation of automatically generated image captions. It measures the similarity between the candidate sentence and a set of human-written sentences by performing a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram.

SPICE [45] is a metric designed for image caption evaluation. It measures the quality of generated captions by computing an F-measure based on the propositional semantic content of candidate and reference sentences represented as scene graphs [53].

The metrics above represent a standard set of metrics usually reported in papers. Their popularity can be attributed to their availability through the Microsoft COCO caption evaluation server [41], which enables a consistent comparison of different models.

However, it was shown [47, 10] that automatic metrics do not always correlate with human judgments. This was particularly evident during the Microsoft COCO 2015 Captioning Challenge in that some models outperformed human upper bound according to automatic metrics, but human judges demonstrated a preference for human-written captions [54]. It seems that "humans do not always like what is human-like" [44]. Since there is no best metric, some authors [45, 46] advise the use of an ensemble of metrics capturing various dimensions, such as grammaticality, saliency, correctness or truthfulness. In [22, 46] new evaluation metrics were proposed.

VI. Conclusion

This paper presents an overview of recent advances in image captioning research, with a particular focus on models employing deep encoder-decoder architectures. The main advantage of such architectures is in that they are trainable end-to-end, mapping directly from images to sentences.

An important extension of the basic encoder-decoder framework is the attention mechanism, which enables to focus on the most salient parts of the input image while generating the next word of the output. We group the related work into three types regarding the task: standard image captioning (with or without an attention mechanism), image captioning with style and cross-lingual or multilingual image captioning.

Large vision and language datasets have also contributed significantly to the development of the field. Additional features of the new datasets, such as emotions or descriptions in different languages, will certainly stimulate even faster advances in the periods to come.

However, there are some important tasks that are still unresolved, such as generating captions more in the spirit of the human descriptions, automatic adaptation of descriptions to the given task, and perhaps the most challenging among them, automatic assessment of the generated captions, since there are still no metrics to match human evaluation fully.

ACKNOWLEDGMENT

This research was supported by Croatian Science Foundation under the project IP-2016-06-8345 "Automatic recognition of actions and activities in multimedia content from the sports domain" (RAASS) and by the University of Rijeka under the project number 18-222-1385.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in NIPS'12, 2012, vol. 1, pp. 1097–1105.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," arXiv preprint arXiv:1506.01497, 2015.

[3] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International Journal of Computer Vision, vol. 123, no. 1, pp. 32–73, 2017.

[4] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in ICML-14, 2014, pp. 595–603.

[5] M. Ivašić-Kos, I. Ipšić, S. Ribarić, "A knowledge-based multi-layered image annotation system.", Expert systems with applications. 42 (2015), 2015; 9539-9553

[6] M. Ivašić-Kos, M. Pavlić, M. Pobar, "Analyzing the semantic level of outdoor image annotation", MIPRO 2009, IEEE Opatija, pp. 293-296

[7] M. Pobar, M. Ivašić-Kos, "Multimodal Image Retrieval Based on Keywords and Low-Level Image Features", Semantic Keyword-based Search on Structured Data, IKC 2015, Coimbra, Portugal, Springer, 2015. 133-140

[8] G. Kulkarni et al., "Babytalk: Understanding and generating simple image descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2891–2903, 2013.

[9] M. Ivašić-Kos, M. Pobar, S. Ribarić, Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme, Pattern recognition. 52 (2016); 287-305

[10] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” JAIR, vol. 47, pp. 853–899, 2013.

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

[12] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in NIPS’14, 2014, vol. 2, pp. 3104–3112.

[13] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time series,” The Handbook of Brain Theory and Neural Networks, vol. 3361, no. 10, 1995.

[14] J. L. Elman, “Finding structure in time,” Cognitive Science, vol. 14, no. 2, pp. 179–211, 1990.

[15] O. Russakovsky et al., “Imagenet large scale visual recognition challenge,” International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in CVPR, 2015, pp. 3156–3164

[18] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.

[19] J. Donahue et al., “Long-term recurrent convolutional networks for visual recognition and description,” in CVPR, 2015, pp. 2625–2634.

[20] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical sequence training for image captioning,” in CVPR, 2017, pp. 7008–7024.

[21] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, “Sequence level training with recurrent neural networks,” arXiv preprint arXiv:1511.06732, 2015.

[22] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, “Improved Image Captioning via Policy Gradient Optimization of SPIDEr,” arXiv preprint arXiv:1612.00370, 2016.

[23] B. Dai, S. Fidler, R. Urtasun, and D. Lin, “Towards diverse and natural image descriptions via a conditional gan,” in ICCV, 2017, pp. 2970–2979.

[24] B. Dai and D. Lin, “Contrastive learning for image captioning,” in Advances in Neural Information Processing Systems, 2017, pp. 898–907.

[25] K. Xu et al., “Show, attend and tell: Neural image caption generation with visual attention,” in ICML, 2015, pp. 2048–2057.

[26] H. Fang et al., “From captions to visual concepts and back,” in CVPR, 2015, pp. 1473–1482.

[27] M. Ivašić-Kos, M. Pobar, S. Ribarić, “Automatic image annotation refinement using fuzzy inference algorithms”, IFSA- EUSFLAT 2015, Gijón, Asturias, (Spain) p. 242

[28] W. Lan, X. Li, and J. Dong, “Fluency-guided cross-lingual image captioning,” in Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1549–1557.

[29] A. P. Mathews, L. Xie, and X. He, “Senticap: Generating image descriptions with sentiments,” in 13th AAAI Conference on Artificial Intelligence, 2016.

[30] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in CVPR, 2016, pp. 4651–4659.

[31] P. Anderson et al., “Bottom-up and top-down attention for image captioning and vqa,” arXiv preprint arXiv:1707.07998, 2017.

[32] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via A visual sentinel for image captioning,” arXiv preprint arXiv:1612.01887, 2016.

[33] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, “Stylenet: Generating attractive visual captions with styles,” in CVPR, 2017, pp. 3137–3146.

[34] Q. You, H. Jin, and J. Luo, “Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions,” arXiv preprint arXiv:1801.10121, 2018.

[35] T. Chen et al., “Factual or Emotional: Stylized Image Captioning with Adaptive Learning and Attention,” in ECCV, 2018, pp. 519–535.

[36] O. M. Nezami, M. Dras, P. Anderson, and L. Hamey, “Face-Cap: Image Captioning Using Facial Expression Analysis,” in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2018, pp. 226–240.

[37] T. Miyazaki and N. Shimizu, “Cross-lingual image caption generation,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, vol. 1, pp. 1780–1790.

[38] D. Elliott, S. Frank, and E. Hasler, “Multilingual image description with neural sequence models,” arXiv preprint arXiv:1510.04709, 2015.

[39] S. Tsutsui and D. Crandall, “Using artificial tokens to control languages for multilingual image caption generation,” arXiv preprint arXiv:1706.06275, 2017.

[40] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting image annotations using Amazon’s Mechanical Turk,” in Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, 2010, pp. 139–147.

[41] X. Chen et al., “Microsoft COCO captions: Data collection and evaluation server,” arXiv preprint arXiv:1504.00325, 2015.

[42] R. Bernardi et al., “Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures,” JAIR, vol. 55, pp. 409–442, 2016.

[43] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” Transactions of the Association for Computational Linguistics, vol. 2, pp. 67–78, 2014.

[44] R. Vedantam, C. L. Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in CVPR, 2015, pp. 4566–4575.

[45] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Spice: Semantic propositional image caption evaluation,” in ECCV 2016, 2016, pp. 382–398.

[46] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, “Re-evaluating automatic metrics for image captioning,” arXiv preprint arXiv:1612.07600, 2016.

[47] D. Elliot and F. Keller, “Comparing automatic evaluation measures for image description,” in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers, 2014, pp. 452–457.

[48] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, “Corpus-guided sentence generation of natural images,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 444–454.

[49] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meeting on association for computational linguistics, 2002, pp. 311–318.

[50] M. Denkowski and A. Lavie, “Meteor universal: Language-specific translation evaluation for any target language,” in 9th Workshop on Statistical Machine Translation, 2014, pp. 376–380.

[51] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out: Proceedings of the ACL-04 workshop, 2004, vol. 8.

[52] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, “Introduction to WordNet: An on-line lexical database,” International Journal of Lexicography, vol. 3, no. 4, pp. 235–244, 1990.

[53] J. Johnson et al., “Image retrieval using scene graphs,” in CVPR, 2015, pp. 3668–3678.

[54] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: Lessons learned from the 2015 MS COCO image captioning challenge,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 652–663, 2017.