International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

A Survey in Deep Learning Model for Image Annotation

Phyu Phyu Khaing[a]*, May The` Yu[b]

[a,b]*University of Computer Studies, Mandalay, Myanmar*
[a]*Email: phyuphyukhaing@ucsm.edu.mm*
[b]*Email: maytheyu@ucsm.edu.mm*

Abstract

Image annotation is generating the human-understandable natural language sentence for images. Annotating the image with sentence is one kind of the computer vision process that includes in the artificial intelligence. Annotation is working by combining computer vision and natural language processing. In image annotation, there are two types: sentence based annotation and single word annotation. Deep learning can get the more accurate sentence for the image. This paper is the survey for image annotation that applied the deep learning model. This discusses existing methods, technical difficulty, popular datasets, evaluation metrics that mostly used for image annotation.

Keywords: Image Annotation; Deep Learning Model; Datasets; Evaluation Metrics.

1. Introduction

Image annotation, converting image with natural language sentence, is still challenges in computer vision processes. Image annotation basically derived from Artificial Intelligence. In Artificial Intelligence, there are many sub fields such as computer vision, NLP, Robotics, and many others. Computer vision has two fields: image processing and image analysis. Image processing can be thought of as a transformation that takes an image into an image. The input of image processing is image and the output is also an image, such as histogram equalizing, image de-blurring. Image analysis is the extraction of meaningful information from images. So, the input is image and output is a description or a decision. Image analysis processes are object detection, image annotation, object recognition, and many others. In image annotation, there are two types: sentence based annotation, and single word annotation. Image annotation hierarchy shows in figure 1.

---

* Corresponding author.

Image annotation has two kinds of approaches: top-down and bottom-up to success the machine translation. The Top-down approaches apply the encoder-decoder network architecture (Convolutional Neural Network as encoder and LSTM as decoder). It initially takes the image into the encoder to get the feature and the features were fed into the decoder to generate the image description. The bottom-up approaches include several separated tasks, such as identifying objects or attributes, arranging words and sentences, describing sentences using a language model to generate the image caption [18].

Deep learning is also a technique that learns data from image to encourage the implementation of machine learning that is the function and structure of the brain known as artificial neural network. Deep learning is also called hierarchical learning or deep structured learning. Deep learning's neural network differ from traditional neural network because it has more hidden layers; and they can be trained in a supervised and unsupervised method for both supervised and unsupervised learning task. Neural network architecture was typically applying for deep learning. The term "deep" point the number of network layers. Although the neural networks traditionally contain only two or three layers, deep neural networks can contain hundreds. So more layers is that the deeper network.

This research paper is constructed with following sections. Section 2 presents related literature review for this research. Section 3 discusses different image annotation models that commonly used by different researchers. Section 4 describes the most famous datasets which have been applying for annotation of image. Different evaluation metrics are examined in section 5. Section 6 summarizes the annotation approaches by literature review.

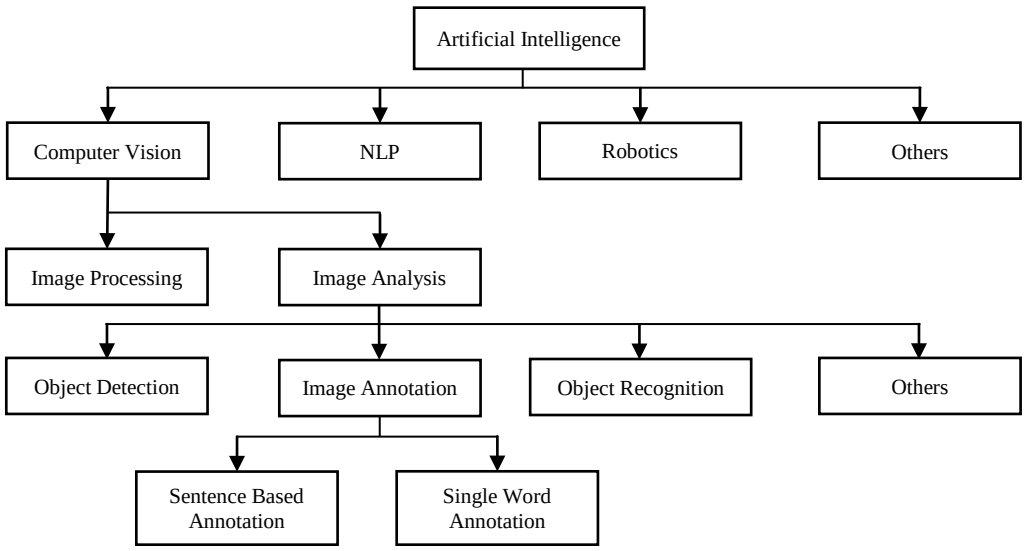

Figure 1: Image Annotation Hierarchy

2. Related Work

Cho and his colleagues (2015) described encoder-decoder networks based on attention that generate many

contents of images. This learned based on attention mechanisms that work on convolutional neural network, and gated recurrent neural networks. A bidirectional recurrent neural network (BiRNN) was used for an encoder, and a recurrent neural network learning model (RNN-LM) that based on attention was used for a decoder in neural machine translation [6]. Xu and his colleagues (2015) introduced soft deterministic attention and hard stochastic attention that are the two types of image caption generators based on attention. It attended by focusing on the visualizing of "where" and "what"; and quantitatively validated the advantages of attention for generating image caption. This operated by combining Convolutional Neural Network (CNN) for the vectors that extract features from input image and long short-term memory (LSTM) for generating word at every step on context vector [7]. Wang and his colleagues (2016) proposed deep bidirectional LSTM model that designed for image caption generation. This model is based on a deep CNN and two separate LSTM network for learning long-term interaction between image and text. This caption generation model is evaluated with Flickr8K, Flickr30K, and MSCOCO benchmark datasets. Bidirectional LSTM model achieve highly performance on both generation and retrieval tasks. As the future scope, more sophisticated language representation, multitask learning and attention mechanism can extend in the model [10]. Wang and his colleagues (2016) demonstrated parallel-fusion RNN-LSTM architecture for image captioning by combining the advantages of simple RNN and LSTM. This approach improves the performance and the efficiency by evaluating with BLEU and METEOR on Flickr8K dataset. To focus the higher performance, future work need to examine the limitation of parallel threads by using more complex image features [11].

Fu and his colleagues (2017) described automatic image captioning system by transforming images to accurate and meaningful sentences. Before generating the words, giving the other ones as the input and then it was arranged to the visual perception experience. An image was encoded with higher-level semantic information by introducing scene-specific contexts. Some benchmark datasets including Flickr8K, Flickr30K and MSCOCO were used to produce the results by applying both human evaluation and automatic evaluation metrics. The performance of the work were improved either scene-specific context or region-based attention. The combination of two modeling ingredients suggests attaining the achievement of state-of-the-art as the future scope [13].

Qu and his colleagues (2017) propounded visual attention mechanism based on long-short term memory to stare on salient object for image captioning. In this work, CNN is used for extracting features such as colors, size, and location; LSTM is used to generate a sentence; and attention mechanism is used to describe the important objects in image. CNN extract features from image with VggNet. LSTM is work with four gates (input gate, output gate, forget gate, and attend gate) and a memory cell. Attention has two aspects: color stimulus-driven and dimension stimulus-driven. The proposed model was validated on three benchmark datasets: Flick8k, Flick30k, and MS COCO and the performance show by using standard evaluation metrics: BLEU. The proposed model can generate more interpretability sentence and get more accuracy in object recognition. The future work should use unsupervised data to understand comprehensively and precisely about a whole picture [14].

Lu and his colleagues (2017) introduced an adaptive attention encoder-decoder framework for image caption generation. Adaptive attention learns when to attend and where to attend on the image. This framework tests on the Flickr30K dataset and the 2015 MSCOCO image captioning dataset to analyze the adaptive attention. The

framework is efficiently evaluated on image captioning, and it can have useful in other applications domains [15] Chen and his colleagues (2017) developed the Spatial and Channel-wise Attention in Convolutional Neural Network (SCA-CNN) for image caption generation. In multi-layer feature maps, SCA-CNN concise for the sentence generation by encoding what and where the visual attention is. This is evaluated on three benchmark datasets: Flickr8K, Flickr30K and MSCOCO. Future work intends to work temporal attention in SCA-CNN, by attending video frames features for video captioning and to increase the attentive layers without overfitting [16].

Gan and his colleagues (2017) initiated a Semantic Compositional Network (SCN) for image captioning and video clip captioning. SCN detect semantic concepts from the image and use the probability of each task for parameter composition in LSTM. This is quantitatively evaluated and qualitatively analyzed on COCO, Flickr30K and Youtube2Text datasets; and the performance significantly outperforms with multiple evaluation metrics [17]. Liu and his colleagues (2017) found a quantitative evaluation metric by focusing on evaluating and improving the correctness of attention in neural image captioning. The metric evaluates between human annotations and the generated attention maps by using Flickr30K and COCO datasets. This can close the gap between human perception and machine attention and can experiment in related fields [18].

Gu and his colleagues (2017) exploited CNN language model for image caption generation. MSCOCO and Flickr30K datasets have been using to conduct the experiments for analysis. Model can generate sentence that is relevant with image but model is wrong when visual attributes are predicted. It can integrate extra attributes that learn for image captioning as future scope [19]. Vinyals and his colleagues (2017) presented a neural network system (NIC) that generates the sentence description of an image. NIC firstly uses convolutional neural network to encode the image and then it uses recurrent neural network to generate the sentence that correspond with image. This measured the performance with the standard evaluation metrics and also evaluated with human judgements on five benchmark datasets [20].

Li and his colleagues (2018) proposed the global-local attention (GLA) method for describing image caption. Features based on object-level integrated with image-level by applying attention mechanism. This used VGG16 for image feature extractor, Faster R-CNN for object detector, attention mechanism for integration of global feature and local feature, and stacked two-layer LSTM for the model of language. The proposed GLA method implemented on Microsoft COCO caption data set by checking with many favored evaluation metrics such as BLEU-1,2,3,4, CIDEr, METEOR, and ROUGE-L. This can create more appropriate and reasonable sentences that related the image context but cannot jointly the language model and train CNN part. So, the integration of image feature extractor and object detector is still as the future study to train and test of end-to-end model [21].

Ye and his colleagues (2018) initiated attentive linear transformation (ALT) for automatically generating of image caption as a novel attention framework. That model used Convolutional Neural Network (CNN) for encoding an input image to features, the high-dimensional transformation matrix for converting from the image feature to the context vector and Recurrent Neural Network (RNN) for decoding from the vector to a sentence that related with image. This experiments on the benchmark dataset such as MS COCO and Flickr30k by measuring evaluation metrics. ALT's advantage is that the linear transformation's weight can show information unless a concrete form use like feature channel or spatial region. ALT can nicely describe than existing attention

models but cannot correctly recognize words on the sign, cannot distinguish some-part-redundant object, cannot correctly count the quality of object, and mistakes the gender. This paper suggested using text detector to recognize the words and objecting detector to count the quantity of the objects as the future works [22].

Zhu and his colleagues (2018) developed a Captioning Transformer (CT) model by applying stacked attention modules without the time dependencies to address the issues of long-short-term-memory (LSTM) structure and also proposed multi-level supervision training. The encoder of this model is Convolutional Neural Network (CNN) that used ResNet and ResNext as image classification models to extract image features and the decoder is transformer model with stacked attention mechanism to decode from image features to the sentence. There are three methods for integrating image features to transformer model: 1) image spatial feature map, 2) spatial image feature map that combine the image feature with each word embedding, and 3) spatial image feature map that used image feature before the start of the text embedding. This used MSCOCO dataset and standard evaluation metrics for evaluating the performance by comparing with several start-of-the art methods. The accuracy of the study is better than original models. This pointed to study the method in the digital virtual asset security field for future scope [23].

Aneja and his colleagues (2018) explored a convolutional image captioning technique, demonstrated its efficacy on the MSCOCO dataset and the performance with baseline. The model with attention can improve the performance [24]. Wang and his colleagues (2018) discovered a framework that only employs convolutional neural networks (CNNs) to generate captions. They conduct extensive experiments on MSCOCO and investigate the influence of the model width and depth. Compared with LSTM-based models that apply similar attention mechanisms, our proposed models achieves comparable scores of BLEU-1,2,3,4 and METEOR, and higher scores of CIDEr [25].

3. Image Annotation Models

There are many image annotation models in the previous literature. Many research groups have commonly implemented little famous architecture. In image annotation processes, CNN, RNN, and LSTM are mostly used as the famous architectures. Table 1 shows an overview of the deep learning model that applied in image annotation methods. These techniques will be discussed with the subsections.

3.1. CNN

Convolutional Neural Network (CNN or ConvNet) is a feed-forward neural network for machine learning and applies in many artificial intelligences (AI) research areas. CNNs are mostly used for speech and image recognition, video analysis, pattern recognition and natural language processing. CNN is one type of deep neural network, and also used as the encoder network that extract the image features. In CNN, there are many pre-trained model such as VGGNet, AlexNet, DenseNet, MobileNet, etc.. In image annotation, CNN pre-trained model are used for the feature extraction of image.

3.2. RNN

Recurrent Neural Network (RNN) is also the neural network that based on the looping process. RNNs can apply their internal state (memory) to perform sequences of inputs, unlike feed forward neural networks. In image annotation process, RNN is used to predict the next word by learning on the current word. So, RNN is also called language model or the decoder network. RNN is also implemented as encoder-decoder network in image annotation model.

3.3. LSTM

RNN is composed with LSTM units that are often called an LSTM network. Long-and-Short-Term Memory (LSTM) is one type of recurrent neural network (RNN). LSTM is developed with the gates: input gate, output gate, forget gate, and cell. LSTM is applied for sentence representation in image annotation process. LSTM is also implemented to extract the feature for image and word.

4. Datasets

There are various kinds of datasets that applied for image detection, image classification, image recognition, and image caption generation of image. MSCOCO [4], FLICKR 8K [3], and FLICKR 30K [8] are the standard benchmark datasets that are famous and mostly used for image annotation. In table 1, we summarized the datasets that mostly used for image annotation.

4.1. MSCOCO

MSCOCO [4] dataset is mostly implemented for image annotation. There are three parts in this dataset: the training set, the testing set and the validation set. Each image is described with five sentences for training and validation but images for testing do not have annotated sentences. This dataset have many updated version by years such as 2014, 2015, and 2017. The dataset, which released in 2014, have 82,783t images for training 40,775 images for testing and 40,504 images for validation. In 2015, this cumulatively released 165,482 train images, 81,434 testing images and 81,208 validation images. The 2017 dataset release, that is the last, contain 118,287 training images, 40,670 testing images, and 5,000 validation images.

4.2. Flickr8k

The images from Flickr.com website is collected for Flickr8k [3] dataset. It consists of 8,092 images that performed the action of animals or people. There have been using 6,000 images to train, 1,000 images to test and 1,000 images to validate. Five sentences are created for each image in the dataset by characterizing with entities (animals, people and objects), situation, scenes and events. The grammar of images in the dataset is tested with the workers and spelling is checked with United State format.

4.3. Flickr30k

Flickr30k [8] is a standard benchmark dataset for sentence-based image description. In this dataset, there are 513,644 images for scene and entity and there has been working with five sentences per image. Among them,

28,000 images are to train, 1,000 images are to test and 1,000 images are to validate. This dataset emerges by combining the embedding of image and text, common object detectors, color classifier and bias that select larger objects.

5. Evaluation Metrics

For image annotation, the evaluation metrics are commonly used to evaluate the accuracy and effectiveness. The popular evaluation metrics are Bilingual Evaluation Understudy (BLEU) [1], Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [2], Metric for Evaluation based Image Description Evaluation (METEOR) [5], Consensus-based Image Description Evaluation (CIDEr) [9], and Semantic Propositional Image Caption Evaluation (SPICE) [12]. All of these methods calculate with similarity based measure between ground truth sentence and machines generated sentence. Each of these evaluation methods are introduced in the following subsections.

5.1. BLEU

BLEU [1] is an automatic human-like evaluation and extensively used for machine translation. It is language-independence, speedy, and cheaply evaluation method. The semantic similarity between human description of image and machine generated caption can be determined by applying BLEU score. It measures n-grams' fraction that are in common between a reference and a hypothesis. The strength of BLEU [24] evaluation metrics highly correlates with the judgments of human by average of judgment errors of individual sentence. The judgment over a test corpus is divine rather than the judgment of human for every sentence.

5.2. ROUGE

ROUGE [2] is an automatic evaluation package for the comparison of the quality of a summary and human-created summaries. It is very effective for automatic evaluation of machine translation. Four different ROUGE measures are: ROUGE-L, ROUGE-N, ROUGE-S, and ROUGE-W. ROUGE-L identifies the longest common subsequence (LCS) and it has sentence-level LCS and summary-level LCS. ROUGE-N is a recall-related n-gram measure between a set of reference summaries and a candidate summary. ROUGE-S named skip-bigram co-occurrence statistics and measure the skip-bigram overlapping between a set of reference translations and a candidate translation. ROUGE-W calls the weighted longest common subsequence (WLCS) and use the polynomial function to calculate.

5.3. METEOR

METEOR [5] score has been highly applied in comparison with other metrics because of highly correlated with human subjects' annotations. It can evaluate on any target language to construct the system of statistical translation by applying the same resources. It is freely available as open source software.

5.4. CIDEr

The goal of CIDEr [9] is to automatically evaluate for image. This evaluation metrics show how many matching the consensus of image description sets with a candidate sentence. This is more suitable for the evaluation of image description generation for consensus measuring.

5.5. SPICE

SPICE [12] is an automatic evaluation metric for caption generation that captures with the human judgments. It work by comparing semantic propositional content and it is better than other automatic evaluation metrics. It tested on the MSCOCO dataset and compare with CIDEr and METEOR. It can also use for question and answering process.

Table 1: Summarization of Methods, Datasets and Evaluation Metrics

Reference	Image Encoder	Language Decoder	Datasets	Evaluation Metrics
Ref. [6]	BiRNN	RNN-LSTM	Flickr8K, Flickr30K, MSCOCO	BLEU, CIDEr
Ref. [7]	CNN	LSTM	Flickr8K, Flickr30K, MSCOCO	BLEU, METEOR
Ref. [10]	VGGNet, AlexNet	LSTM	Flickr8K, Flickr30K, MSCOCO	BLEU
Ref. [11]	VGGNet	Parallel fusion RNN-LSTM	Flickr8K	BLEU, METEOR
Ref. [13]	VGGNet, AlexNet	LSTM	Flickr8K, Flickr30K, MSCOCO	BLEU, METEOR, CIDEr, ROUGE_L
Ref. [14]	VGGNet	LSTM	Flickr8K, Flickr30K, MSCOCO	BLEU
Ref. [15]	ResNet	LSTM	Flickr30K, MSCOCO	BLEU, METEOR, CIDEr
Ref. [16]	VGGNet, ResNet	LSTM	Flickr8K, Flickr30K, MSCOCO	BLEU, METEOR, CIDEr, ROUGE_L
Ref. [17]	ResNet	LSTM	Flickr30K, MSCOCO	BLEU, METEOR, CIDEr
Ref. [18]	VGGNet	LSTM	Flickr30K, MSCOCO	BLEU, METEOR
Ref. [19]	VGGNet	Language CNN, LSTM	Flickr30K, MSCOCO	BLEU, METEOR, CIDEr, SPICE
Ref. [20]	Deep CNN	LSTM	MSCOCO	BLEU, METEOR, CIDEr, ROUGE_L
Ref. [21]	VGGNet, Faster R-CNN	LSTM	MSCOCO	BLEU, METEOR, CIDEr, ROUGE_L
Ref. [22]	VGGNet	LSTM	Flickr30K, MSCOCO	BLEU, METEOR, CIDEr, ROUGE_L, SPICE
Ref. [23]	ResNet and ResNext	LSTM	MSCOCO	BLEU, METEOR, CIDEr, ROUGE_L
Ref. [24]	VGGNet	GLU	Flickr8K, Flickr30K, MSCOCO	BLEU, METEOR, CIDEr
Ref. [25]	VGGNet	CNN	MSCOCO	BLEU, METEOR, CIDEr, ROUGE_L

6. Conclusion and Recommendations

This paper presents the comprehensive study of deep learning model for image annotation process. For image annotation, deep learning is very useful and powerful to annotate the image with sentence description. In recent

years, deep learning based image annotation achieves remarkable progress. In deep learning models, CNN is mostly used to extract image feature and RNN/LSTM is commonly applied to generate the sentence description. This paper also presented a brief review of the benchmark datasets and standard evaluation metrics that mostly applied for image annotation. According with the comparison table, CNN and LSTM is mostly used for image encoder and language decoder respectively. MSCOCO dataset is commonly applied for image annotation; and BLEU, METEOR, and CIDEr are usually worked as evaluation metrics. This study only learns from the previous literatures and also attentively explains the deep learning models, datasets, and evaluation metrics that most applied for image annotation. This paper, however, did not emphasize about the attention mechanism which used on deep learning model to more accurate the performance.

References

[1] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation", Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[2] X.Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", Text Summarization Branches Out, vol.8, 2004, pp. 1-8.

[3] M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, 47:853-899, 2013.

[4] T.Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollar, "Microsoft COCO: Common Objects in Context", European Conference on Computer Vision, 2014, pp. 740-755.

[5] M. Denkowshi, and A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language", Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376-380. 2014.

[6] K. Cho, A. Courville, and Y. Bengio, "Describing Multimedia Content Using Attention-Based Encoder-Decoder Networks", IEEE Trans. on Multimedia, 17(11):1875-1886, 2015.

[7] K. Xu, J.L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. S. Zemel and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", International conference on machine learning, 2015, pp. 2048-2057.

[8] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models", Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2641-2649.

[9] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation", Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566-4575.

[10] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional LSTMs", In Proceedings of the 2016 ACM on Multimedia Conference, 2016, pp. 988–997.

[11] M. Wang, L. Song, X. Yang, and C. Luo, "A parallel-fusion RNN-LSTM architecture for image caption generation", In 2016 IEEE International Conference on Image Processing (ICIP'16), 2016, pp. 4448–4452.

[12] P. Anderson, B. Fernando, M. Johnson, and S. Gould. "SPICE: Semantic Propositional Image Caption

Evaluation", In European Conference on Computer Vision, Springer, Cham, 2016, pp. 382-398.

[13] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning Where to see and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts", IEEE Trans. on Pattern Analysis and Machine Intelligence, 39(12):2321-2334, 2017.

[14] S. Qu, Y. Xi, and S. Ding, "Visual Attention Based on Long-Short Term Memory Model for Image Caption Generation", Control and Decision Conference (CCDC), 2017 29th Chinese, IEEE, May 2017, pp. 4789-4794.

[15] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), 2017, pp. 3242–3250.

[16] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, and T.S. Chua, "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'17), 2017, pp. 6298–6306.

[17] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), 2017, pp. 1141–1150.

[18] C. Liu, J. Mao, F. Sha, and A. L. Yuille, "Attention Correctness In Neural Image Captioning", In AAAI, 2017, pp. 4176–4182.

[19] J. Gu, G. Wang, J. Cai, and T. Chen, "An Empirical Study Of Language CNN For Image Captioning", In Proceedings of the International Conference on Computer Vision (ICCV'17), 2017, pp. 1231–1240.

[20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challeng", IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, Apr. 2017, pp. 652-663.

[21] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, "GLA: Global-Local Attention for Image Description", IEEE Trans. on Multimedia, 20(3):726-737, 2018.

[22] S. Ye, J. Han, and N. Liu, "Attentive Linear Transformation for Image Captioning", IEEE Trans. on Image Processing, 27(11):5514-5524, 2018.

[23] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu, "Captioning Transformer with Stacked Attention Model", Applied Sciences, 8(5):739, 2018.

[24] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5561–5570.

[25] Q. Wang and A. B. Chan, "CNN+ CNN: Convolutional Decoders For Image Captioning", arXiv preprint arXiv:1805.09019, 2018.