Contents

CHAPTER 13

Automatic Image Captioning Using Deep Learning

Toshiba Kamruzzaman
Toshiba.Zaman@gmail.com
Dept. of ECE, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh

Abdul Matin
Ammuaj.cseruet@gmail.com
Dept. of ECE, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh

Tasfia Seuti
TasfiaSeuti7427@gmail.com
Dept. of CSE, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh

Md. Rakibul Islam
Rakibul.eceruet@gmail.com
Dept. of ECE, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh

CONTENTS

Transforming an image to a text or descriptive form has recently got a lot of research attractions. Creating a sentence with correct semantics and syntactic structure is still a matter of issue. Object recognition, relations between the objects, different meanings of the same word make this task more difficult. Therefore, inspection on attention mechanism has recently achieved great progress. In this paper, we have described some existing models and then expended these models by integrating BERT, LSTM and dense models together.

After analyzing the results, we have found that while training on the same parameters, our new model has shown comparatively less amount of training time among the others and shown the better result at the all common metrics (BLEU, METEOR & CIDEr) on MS-COCO dataset.

Keywords: Image Captioning, CNN (Convolutional Neural Network), RNN (Recurrent Neural Network), NLP (Natural Language Processing), Computer Vision (CV), LSTM (Long-Short Term Memory), BERT (Bidirectional Encoder Representations from Transformers).

13.1 INTRODUCTION

Image captioning is the technique of interpreting a source image into its corresponding text version. This conversion produced by the captioning system allows the reader to quickly understand the theme of the image without going through every detail. The overall motivation of image captioning system is to convey the main meaning of the image in the form of narrative version. Image captioning is an image-to-sequence problem whose inputs are pixels (digital form of an image). These are encoded as one or multiple feature vectors in the visual encoding step, which prepares the input for a second generative step, called the language model. This produces a sequence of words or sub-words decoded according to a given vocabulary. Recent developments in image captioning have been inspired by advancements in object detection and machine translation in the past few years.

The task of image captioning involves two main aspects:

(a) resolving the object detection issue in computer vision and

(b) creating a language model that can accurately generate a sentence describing the detected objects. Image caption is one of the big fusses in the field of modern machine learning and artificial intelligence (AI). It has a large assortment of demands. It can be used for self-driving car, automatic question answering system, automatic video description system, describing medical situation to a non-medic person and so on. Moreover, more than 100 million people are visually incapacitated all over the world. So, describing the real-time scene can help them to have a safe movement.

Our contribution in this paper is following-

- We have compared between the LSTM and BERT(large) models and fine-tuned these models to enhance their performances.
- Then, we have developed a new model for image captioning system which comprise the concatenation of BERT model with LSTM and dense model.

The paper outline is organized as follows. In Section 13.2, the trending researches of image captioning systems have been presented. In Section 13.3, we have described the proposed workflow and model architecture. In Section 13.4, we have discussed our model configuration. Section 13.5 provides our experiments and result analyses Afterwards, we have talked over the constraints and remarks of our model in Section 13.6.

13.2 LITERATURE REVIEW

The early neural models for image captioning system were based on the template methods. In this method, the essential features are extracted first by using different types of classifiers such as SVM (Support Vector Machine), Naive Bayes, Decision Tree, K-Nearest Neighbor etc. According to Farhadi et al. the object, activities of the object and the whereabouts of the object are main features of an object [1]. Then the acquired attributes are converted into the textual form. These methods confide in rigid and fixed rules, for a particular situation, an individual set of rules are defined. That pattern can simply reflect the paucity of rules, which affects the sentence-versatility.

Encoder-Decoder based image captioning models solves this problem efficiently. It is now one of the most popular frameworks for image captioning system. In encoder, a CNN model is used which takes the input image and extracts the internal representations of that image by feature vector. Then this feature vector is passed to a decoder. The decoder then generates the descriptive captions from the information fetched from the CNN at one word rate at each time steps.

Attention based approach was later proposed for image captioning system. Vaswani et al. recommended a stacked attention architecture [2]. Anderson et al. suggested a combination of top-down and bottom-up visual attention mechanism. They have used Faster R-CNN model to encode the input [3]. Lun Huang applied attention on both the encoder and decoder model. In encoder the attention mechanism is used to observing the alliances between the objects of the image and in decoder attention is used to filtering out the unnecessary captions [4].

Transformer architectures have played an outstanding role in different NLP tasks such as machine translation, generation of texts, speech recognition etc. X.Liu used deeper Transformers, which consists of 60 encoder layers and 12 decoder layers, to translate from English to German and English to French [5]. Gu et al. introduced a non-autoregressive translation model based on the transformer. In their models, the dependency on the previous tokens (target tokens) were eliminated [6]. Zhang et al. integrated the Transformer model alongside some rules and regulations for paraphrasing to make the sentence simple and easy [7]. Devlin et al. proposed a language

representation model called BERT, which accomplished new SOTA (state of the art) results on 11 NLP tasks [8].

13.3 MODEL ARCHITECTURE

To make our image caption generator model, we have used an encoder-decoder architecture. It is also called a CNN-RNN model.

- CNN is used for extracting features from the image.It acts as the encoder.
- LSTM is used as the decoder, which operates to gain the information from CNN for generating a description of the image.

In Figure 13.1, an overview of image caption generator has been illustrated.

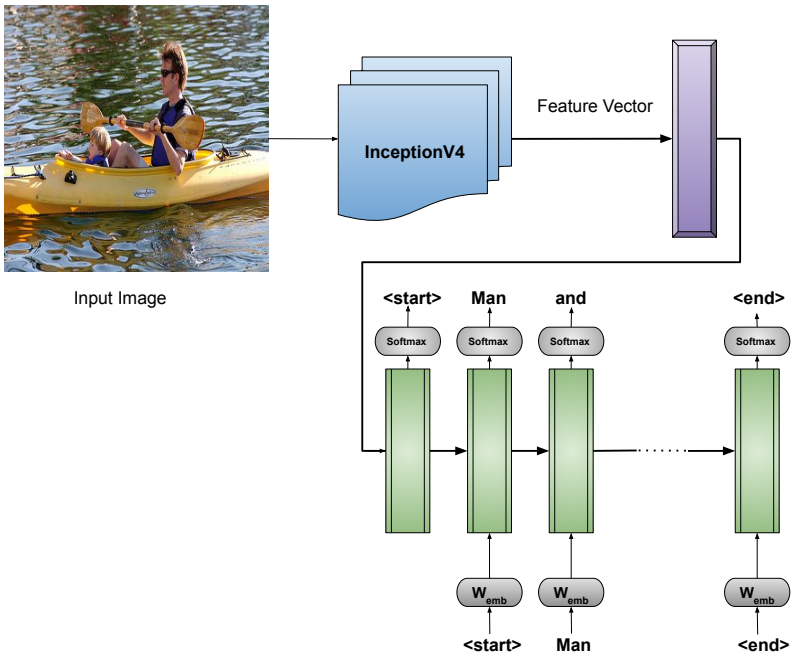


Figure 13.1 The base model image caption architecture

13.3.1 Encoder

We have used InceptionV4 as encoder. This architecture contains 22 deep layers, and including pooling layers, it is made up of 27 layers. The input is an image which consists of 299 by 299 pixels, and the output can classify 1000-categories. We have removed the last layer (fully connected layer) as we only needed to extract the features and used the SoftMax layer as our last layer. We have not performed any fine-tuning to this model.

13.3.2 Decoder

In this study, the experimentation was done with three models leveraging LSTM model, BERT model and BERT with LSTM and dense layer model.

13.3.2.1 Model-1: Base Model (LSTM: Long-Short Term Memory)

We have used an LSTM which is an upgraded version of RNN (Recurrent Neural network) to generate caption's words one step at a time by conditioning on the previous step's hidden state, the context vector, and the previously generated words.

In the following decoder, we have used BERT extensions to our base model to enhance its performance.

13.3.2.2 Model-2: Transformer Model (BERT Integration)

BERT (Bidirectional Encoder Representations from Transformers) is a deep bidirectional representation from unlabeled text by jointly conditioning on both left and right context.

BERT has two distinct models, $BERT_{base}$ and $BERT_{large}$.

- $BERT_{base}$**:**Number of transformer blocks (L): 12, hidden units (H): 768, attention heads(A): 12 and parameters: 110 million
- $BERT_{large}$**:**Number of transformer blocks (L): 24, hidden units (H): 1024, attention heads(A): 16 and parameters: 340 million

Throughout our implementation, we have used $BERT_{large}$ to generate the caption's contextualized word vectors. To do so, we have done the following tasks-

1. The texts are converted to lower-case and then tokenized using WordPiece and a vocabulary size of 30,000.
2. We have used the [SEP] token to indicate the end of a sentence and [CLS] token at the beginning of our text.

13.3.2.3 Model-3: Our Model (BERT with LSTM and dense layer)

We have amalgamated the Transformer model (BERT) with the LSTM model. Our model is made up with a BERT layer, which is followed by a dropout layer. The output of this layer is fed into two distinct LSTM models, and then the outputs of these two LSTM models are concatenated. And finally, this concatenated layer is followed by several dropout and dense layers.

The structure is shown in Figure 13.2.

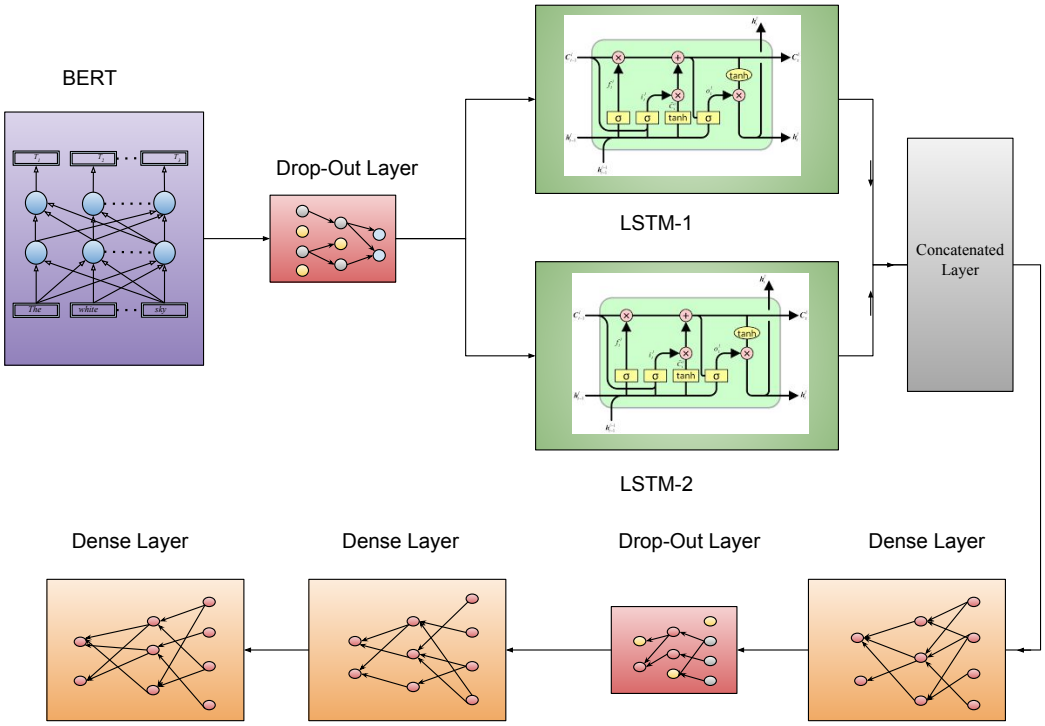


Figure 13.2 Structure of our model

13.4 EXPERIMENTAL SETUP

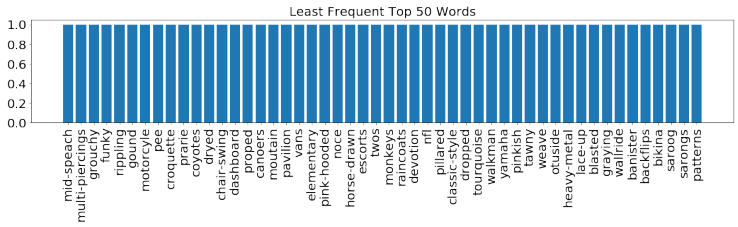
To complete the task, we have used the Kaggle environment, which provides a cloud-based service for data science. The setup is as following-

- 13 GB RAM
- 4.9 GB Hard-Disc
- 16 GB graphics memory

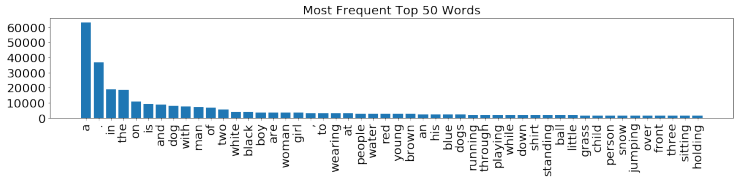
13.4.1 Dataset

Common Object in Context shortly MS-COCO is one of the largest image dataset [9]. This dataset is used for image recognition, image classification problem, image segmentation problem and also image captioning problem. Here, almost 2M images are recorded under 80 different object categories. For every image, there are 5 individual captions and more than 25000 people having key points.

The most and the least frequent 50 words in our dataset is illustrated in Figure 13.3



((a))



((b))

Figure 13.3 The most and the least frequent 50 words in our dataset.

13.4.2 Hyperparameters

During training phase, we have used the following hyperparameters as follows-

- gradient clip = 3,
- batch size = 64,
- learning rate of decoder = 0.0005,
- dropout rate = 0.4,
- epoch= 300,

13.5 RESULT ANALYSIS

Both the qualitative and quantitative analyses have been performed to interpret the results in the following segments.

13.5.1 Qualitative Analysis

Qualitative analysis gives details of the presence or non-appearance of different components in an unknown sample. It uses the subjective judgment to analyze the model's performance based on non-quantifiable information.

13.5.1.1 Model-1: Base Model (LSTM: Long-Short Term Memory)

In Figure 13.4, the base model's results have been analyzed.

We have noticed that this model is unable to generate sentences that have the same meaning as the reference sentences while using different words, it seemed that the model is attempting to copy the reference sentence word by word. This can

be explained by the fact that no pre-trained embeddings were used in this model. Therefore, it was difficult for the model to learn accurate word representations that would allow it to switch similar words.

Candidate caption:
The kids play in the wooded area near the water.

Generated caption:
Three kids are playing near the water.

Candidate caption:
Group of young kids play in the water on sunny day.

Generated caption:
Children and adults play with the water.

Candidate caption:
Several young people sitting on rail above crowded beach.

Generated caption:
Group of people sit on wall at the beach.

((a)) Examples of accurate captions generated by base model.

Candidate caption:
People stare at the orange fish.

Generated caption:
Some children watching fish in pool.

Candidate caption:
Little leaguer getting ready for pitch.

Generated caption:
Boy in white plays baseball.

Candidate caption:
Young woman in red sequined costume and feather stands on the sidewalk.

Generated caption:
Woman wearing red costume looks at two other people standing on street.

((b)) Examples of inaccurate captions generated by base model.

Figure 13.4 Examples of Image Captioning outputs generated by Encoder-Decoder Base model.

13.5.1.2 Model-2: Transformer Model (BERT Integration)

In Figure 13.5, the BERT integration's results have been analyzed. In Figure 13.5(a)(left), the model has successfully recognized both of the dogs. The face of one dog is not clear in the image, but the model has got no difficulties to classify it perfectly. In Figure 13.5(a)(mid), the model has successfully understood the journey

interval. It has also recognized the context of the image. In Figure 13.5(a)(right), though the elements of the pictures are not clearly visible, still the model has distinctly identified the crowds and their activities.

If we go through the inaccurate captions in Figure 13.5(b), we can easily understand that though the candidate captions are not exactly the same as the reference captions, but they were grammatically and the contextually correct which is the goal of this task. BERT captions had few repetitions and were generally very well written.

13.5.1.3 Model-3: Our Model (BERT with LSTM and dense layer)

Figure 13.6 shows a direct a comparison between the three models, we have implemented. The results clearly show that our model outperforms them by large margins, though sometimes BERT model and our models yield similar results.

Our model proved to be able to generate captions that use a different style of writing than the reference captions by using different words that are similar in meaning. This change can be explained by the fact that the embeddings offer the model the ability to pick and choose the best possible word from a cluster of similar words. The first image in Figure 13.6(a) shows an example where the model has used the word "bluff" instead of a "high cliff". "Toy" was translated to "frisbee" in the generated captions in the middle image in Figure 13.6(a).

The captions, which are marked as inaccurate in Figure 13.6(b), also describe the image precisely. The generated sentences are error-free. In short, our model is very accurate in representing contextualized words.

13.5.2 Quantitative Analysis

We have used three evaluation metrices (BLEU, METEOR, CIDEr) to evaluate the performances of our model.

The BLEU score compares a sentence against one or more reference sentences and tells how well the candidate sentence matches the list of reference sentences. It gives an output score between 0 and 1 [10].

The METEOR automatic evaluation metric score's machine translation hypotheses by aligning them to one or more reference translations. METEOR was designed to explicitly address the weaknesses in BLEU identified above. It evaluates a translation by computing a score based on explicit word-to-word matches between the translation and a reference translation. If more than one reference translation is available, the given translation is scored against each reference independently, and the best score is reported [11].

CIDEr (Consensus-based Image Description Evaluation) measures the similarity of a generated sentence against a set of ground truth sentences written by humans [12].

In Table 13.1, we have compared the BLEU (BL1 & BL4), METEOR and CIDEr scores of our BERT integrated model with the existing models.

Table 13.1 Comparison of different models

Model	Evaluation Matrix			
	BL1	**BL4**	**METEOR**	**CIDEr**
Attributes & External Knowledge [13]	73	31	25	92
DR with Embedding Reward [14]	71.3	30.4	25.1	93.7
UPDOWN via personality [15]	79.3	36.4	-	124.0
GCN-LSTM+HIP [16]	-	38.0	28.6	120.3
SGAE [17]	81.0	39.0	28.4	129.1
M^2 Transformer [18]	80.8	39.1	29.2	134.5
Ours	**81.21**	**39.5**	**30.89**	**136.7**

13.6 CONCLUSION

From our experiments, we can conclude that for very small corpuses, the base LSTM model shows a better output but in large corpuses our model surpluses all the existing models. Whereas, the BERT model shows a constant output for both the small and large corpuses.

The main constraint of our model is all the captions are bound within 50 words. It cannot describe the image by a whole paragraph. Secondly, our model cannot describe any conceptual arts and cartoons. It cannot produce any metaphoric caption. The captions are not always creative, they are just logical captions. Moreover, it cannot recognize special places or famous person also.

In the future, we will implement our model in connected-images (images of same events or albums). Moreover, the effect on abstract or metaphoric images will also be inspected.

Candidate caption:
Two dogs playing in the sand at the beach.

Generated caption:
Two dogs playing together on a beach.

Candidate caption:
Several people are taking a break while on a snowmobile ride.

Generated caption:
Three people are taking a break while on a snowmobile ride.

Candidate caption:
People watching hot air balloons

Generated caption:
Crowd watching air balloons at night.

((a)) Examples of accurate captions generated by BERT model.

Candidate caption:
Several hikers rest with their gear in front of mountain.

Generated caption:
Group of people lay on the dry ground.

Candidate caption:
Korean man sells soda.

Generated caption:
A man is sitting in front of soda cans.

Candidate caption:
The young man with the white tshirt and jeans is rappeling up the rock wall.

Generated caption:
Person with blue helmet and purple pants is rock climbing.

((b)) Examples of inaccurate captions generated by BERT model.

Figure 13.5 Examples of Image Captioning outputs generated by BERT model.

Candidate caption:
Skier is overlooking snow-covered mountain.

Generated caption:
Base: Skier is overlooking snow-covered landscape.
BERT: Skier is overlooking the beautiful white snow covered landscape.
Ours: Hiker standing high on bluff overlooking the mountains below.

Candidate caption:
Black and white dog jumps after yellow toy.

Generated caption:
Base: Dog jumps to catch toy.
BERT: Black dog is jumping up to catch purple and green toy.
Ours: Black and white dog jumps to get the frisbee.

Candidate caption:
Boy jumping in fountain.

Generated caption:
Base: Little boy playing in the water.
BERT: Boy plays in the fountains.
Ours: A young boy is jumping in fountain.

((a)) Examples of accurate captions generated by our model.

Candidate caption:
Young child is swung by his or her hands while another child sits on grass watching.

Generated caption:
Base: The little girl is being swung around by her arms.
BERT: Little girl in sweater is swung around by an unseen hand.
Ours: Child is sprawled underneath blanket in midair.

Candidate caption:
Girl sitting in dark bar.

Generated caption:
Base: Dark room with chairs.
BERT: Girl sits on bar stool.
Ours: There several people in dark bartype room.

Candidate caption:
Woman with crazy hairdo is shopping outside

Generated caption:
Base: Man in feather hat looking down.
BERT: The person wearing earrings is wearing feathered hat.
Ours: Woman in floral print dress and shaved head at store.

((b)) Examples of inaccurate captions generated by our model.

Figure 13.6 Examples of Image Captioning outputs generated by our model.

Bibliography

[1] Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D. (2010, September). Every picture tells a story: Generating sentences from images. In European conference on computer vision (pp. 15-29). Springer, Berlin, Heidelberg.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[3] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).

[4] Huang, L., Wang, W., Chen, J., Wei, X. Y. (2019). Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4634-4643).

[5] Liu, X., Duh, K., Liu, L., Gao, J. (2020). Very deep transformers for neural machine translation. arXiv preprint arXiv:2008.07772.

[6] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... Chen, T. (2018). Recent advances in convolutional neural networks. Pattern Recognition, 77, 354-377.

[7] Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., Liu, Y. (2018). Improving the transformer translation model with document-level context. arXiv preprint arXiv:1810.03581.

[8] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[9] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[10] Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

[11] Banerjee, S., Lavie, A. (2005, June). 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 (pp. 65-72).

[12] Vedantam, R., Lawrence Zitnick, C., Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).

[13] Wu, Q., Shen, C., Wang, P., Dick, A., Van Den Hengel, A. (2017). Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine intelligence, 40(6), 1367-1381.

[14] Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L. J. (2017). Deep reinforcement learning-based image captioning with embedding reward. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 290-298).

[15] Shuster, K., Humeau, S., Hu, H., Bordes, A., Weston, J. (2019). Engaging image captioning via personality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12516-12526).

[16] Yao, T., Pan, Y., Li, Y., Mei, T. (2019). Hierarchy parsing for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2621-2629).

[17] Yang, X., Tang, K., Zhang, H., Cai, J. (2019). Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10685-10694).

[18] Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10578-10587).