A Survey on Image Encoders and Language Models for Image Captioning

Himanshu Sharma
Department of Computer Engineering and Applications
GLA University Mathura, India
himanshu.sharma@gla.ac.in

Abstract. Generating a natural language explanation for a given image is known as image captioning. An image captioning method aims to determine the significant objects present in an image together with the relationship between these objects. Also, the model has the capability to describe an image by a syntactically and semantically correct sentence. For image encoding, convolutional neural network [CNN] is applied and for producing natural language descriptions for a given image, language models (RNN & LSTM etc.) are employed. In this paper, the image encoders and language models used by the state-of-the-art image captioning models is discussed.

Keywords: Image Captioning; Convolutional Neural Network; Recurrent Neural Network; LSTM; Artificial Intelligence

1. Introduction
On daily basis, we find lot of images around us. There are different sources of these images such as Internet, advertisements, articles and news etc. The descriptions for these images are not given but human can easily understand these images without the need of these descriptions. In order to generate image captions automatically, a machine has to understand an image as humans do. Thus, image captioning task uses both image understanding and language generation models. Image understanding focuses on the identification of objects existing in an image and the relationship between these objects. Language generation models aims to produce a syntactically and semantically accurate sentence description for a given image. Thus, these models are the broader subparts of artificial intelligence area.

Image captioning methods using deep learning techniques automatically learns image features from a huge set of training data. Convolutional Neural Networks (CNN) is widely utilized to find out visual features. For classification task, a softmax classifier can be used. However, a class of Recurrent Neural Networks (LSTM and GRU) is used to produce natural language descriptions for these images. There are many applications of image captioning. Automatic indexing of an image is one such application of image captioning. If we want to retrieve an image based on its content, image indexing plays a crucial role. This image indexing concept can be used in many fields such as education, commerce, army, web searching, digital libraries and biomedicine. Facebook and Twitter can use image captioning models to automatically generate image captions for images present on these social media platforms. Models can easily describe by what people are doing, what they are wearing and their locations like hotels, beach or café.

In this paper, image encoders and language models utilized by state-of-the-art image captioning methods is presented in detail. The image encoders used by major captioning models are AlexNet [1], [2]...
VGGNet [2], GoogLeNet [3], ResNet [4] and Inception-v3 [5]. The language models employed are RNN, LSTM [6], log-bilinear model (LBL) [7], dependency tree relations (DTR) [8], Language CNN and maximum entropy language model (MELM) [9].

2. Image Encoders

Fig. 1 depicts the general framework of convolutional neural network (CNN). It includes input layer, convolutional (Conv-layer) layer, pooling (Pool layer) layer and dense layer for flattening and finally the output layer.

![General Framework of a CNN](image)

**Fig. 1. General Framework of a CNN**

AlexNet was developed in 2012 by Alex Krizhevsky et al. [1]. It involves five Conv-layer and three fully-connected layers. All the seven layers use Rectified Linear Unit (ReLU) to achieve training faster as compare to tanh and sigmoid activation functions.

VGGNet was proposed by Simonyan et al. [2] in 2014. It consists of 19 layers. It uses 3x3 filters. It uses two fully connected layers with 4096 nodes in each which are followed by softmax layer for performing classification task.

GoogLeNet was given by Szegedy et al. [3] in 2014. It includes four conv-layers, four max-pool and three average pool-layers. Also, it contains five fully connected layer and three softmax layers. It is much deeper and wider network as compared to AlexNet with total 22 layers. All the conv-layers use Rectified Linear Unit (ReLU).

ResNet was given by He et al. [4] in 2015. It has several variations depending upon the number of layer such as 18, 34, 50, 101 and 152. ResNet is almost 8 times deeper as compared to VGGNet. Inception v-3 [5] is widely employed for the task of classification. It contains total 42 layers. It is proposed by updating the inception modules of GoogLeNet architecture.

Table 1 shows the architectures of the above mentioned convolutional neural networks.

| Parameters       | AlexNet | VGGNet | GoogLeNet | ResNet | Inception-v3 |
|------------------|---------|--------|-----------|--------|--------------|
| Year             | 2012    | 2014   | 2014      | 2015   | 2016         |
| Depth            | 8       | 19     | 22        | 152    | 48           |
| Dimension        | 227x227 | 224x224| 229x229   | 224x224| 299x299      |
| Top-1 Accuracy   | 57.1    | 70.5   | 69.8      | 75.2   | 78.8         |
| Top-5 Accuracy   | 80.2    | 91.2   | 89.3      | 93     | 94.4         |
3. Language Models

The log-bilinear language model (LBL) was given by Mnih et al. [7] in 2007. It can be assumed as a feed-forward neural network which is deterministic in nature having a single linear hidden layer. The LBL model works on word embedding vectors. It does the linear prediction by finding the next word representation vector.

In dependency tree relations, first the model finds the objects exist in an image and then find the relationship between these objects by understanding the pair-wise interaction between these objects. Thus, a dependency tree is built that conveys the semantics of a sentence.

Language CNN models employs kernels and sequence of multiple layers to encode the context. These models do not perform pooling. Language CNNs are used widely in the vicinity of natural language understanding (NLP) as POS tagging, chunking, named entity recognition and labeling semantic roles.

Maximum entropy language model (MELM) calculates the conditional probability of the present word based on the previous words. The model is fed with a set of scores of detected words and tries to find the largest likelihood sentence that involves each word exactly one time.

Recurrent Neural Networks (RNN) has the capability to interlink the previous sequence with the current sequence. Long short term memory networks (LSTM) are the capability to remember long term dependencies which RNN are not capable of. LSTM uses three gates which are input gate (how much to add to current state), forget gate (for deciding how much information the model has to remember) and output gate (which part is for output). They also have a cell state to update the information.

Fig. 2 shows the architectures of both RNN and LSTM.

![Fig. 2. General Architecture of RNN Vs LSTM](image)

4. Image Encoders used by Image captioning models

In section 2, the different image encoders are discussed. In this section, a summary of major image captioning models that use the CNN models for image encoding is presented. Table 2 demonstrated the CNN models used by the state-of-the-art image captioning models.
### Table 2: CNN Models used in Image Captioning

| Image Captioning Model                     | AlexNet | VGGNet | GoogLeNet | ResNet | Inception-V3 |
|-------------------------------------------|---------|--------|-----------|--------|--------------|
| Captioning Model [9]                      | ✓       |        |           |        |              |
| Captioning Model [10]                     |         |        |           |        |              |
| Captioning Model [11]                     | ✓       | ✓      |           |        |              |
| Captioning Model [12]                     | ✓       |        |           |        |              |
| Captioning Model [13]                     | ✓       |        |           |        |              |
| Captioning Model [14]                     | ✓       | ✓      |           |        |              |
| Captioning Model [15]                     | ✓       |        |           |        |              |
| Captioning Model [16]                     | ✓       | ✓      | ✓         |        |              |
| Captioning Model [17]                     |         |        |           |        |              |
| Captioning Model [18]                     | ✓       |        |           |        |              |
| Captioning Model [19]                     |         | ✓      |           |        |              |
| Captioning Model [20]                     | ✓       |        |           |        |              |
| Captioning Model [21]                     |         |        |           |        |              |
| Captioning Model [22]                     | ✓       |        |           |        |              |
| Captioning Model [23]                     |         |        |           |        |              |
| Captioning Model [24]                     | ✓       |        |           |        | ✓            |
| Captioning Model [25]                     | ✓       | ✓      |           |        |              |
| Captioning Model [26]                     |         |        |           |        |              |
| Captioning Model [27]                     | ✓       |        |           |        |              |
| Captioning Model [28]                     |         |        |           |        | ✓            |
| Captioning Model [29]                     | ✓       |        |           |        |              |
| Captioning Model [30]                     |         | ✓      |           |        |              |
| Captioning Model [31]                     |         |        |           |        | ✓            |
| Captioning Model [32]                     | ✓       |        |           |        |              |
| Captioning Model [33]                     |         |        |           |        | ✓            |
| Captioning Model [34]                     |         |        |           |        | ✓            |
| Captioning Model [35]                     |         |        |           |        | ✓            |
| Captioning Model [36]                     | ✓       | ✓      |           |        |              |
| Captioning Model [37]                     |         |        |           |        |              |
| Captioning Model [38]                     | ✓       |        |           |        |              |
| Captioning Model [39]                     |         |        |           |        |              |
| Captioning Model [40]                     |         |        |           |        | ✓            |
| Captioning Model [41]                     |         |        |           |        | ✓            |
| Captioning Model [42]                     |         |        |           |        |              |
| Captioning Model [43]                     | ✓       |        |           |        |              |
| Captioning Model [44]                     |         |        |           |        | ✓            |
Table 2: Continued

| Image Captioning Model | AlexNet | VGGNet | GoogLeNet | ResNet | Inception-V3 |
|------------------------|---------|--------|-----------|--------|--------------|
| Captioning Model [45]  | √       |        |           |        |              |
| Captioning Model [46]  |         | √      |           |        |              |
| Captioning Model [47]  |         |        |           |        |              |
| Captioning Model [48]  |         |        |           |        |              |
| Captioning Model [49]  |         |        |           |        |              |
| Captioning Model [50]  |         |        |           |        |              |
| Captioning Model [51]  |         |        |           |        |              |
| Captioning Model [52]  |         |        |           |        |              |
| Captioning Model [53]  |         |        |           |        |              |
| Captioning Model [54]  |         |        |           |        |              |
| Captioning Model [55]  |         |        |           |        |              |
| Captioning Model [56]  |         |        |           |        |              |
| Captioning Model [57]  |         |        |           |        |              |
| Captioning Model [58]  |         |        |           |        |              |
| Captioning Model [59]  |         |        |           |        |              |

5. Language Models used by Image captioning models

In section 3, the different language models are discussed. In this section, a summary of major image captioning models that use the language models for image encoding is presented. Table 3 demonstrated the language models used by the state-of-the-art image captioning models.

Table 3: Language Models used in Image Captioning

| Image Captioning Model | LBL | LSTM | RNN | DTR | Language CNN | MELM |
|------------------------|-----|------|-----|-----|--------------|------|
| Captioning Model [9]   |     |      |     |     |              | √    |
| Captioning Model [10]  |     |      |     |     |              | √    |
| Captioning Model [11]  |     |      |     |     |              | √    |
| Captioning Model [12]  |     |      |     |     |              |      |
| Captioning Model [13]  |     |      |     |     |              |      |
| Captioning Model [14]  |     |      |     |     |              | √    |
| Captioning Model [15]  |     |      |     |     |              | √    |
| Captioning Model [16]  |     |      |     |     |              | √    |
| Captioning Model [17]  |     |      |     |     |              | √    |
| Captioning Model [18]  |     |      |     |     |              | √    |
| Captioning Model [19]  |     |      |     |     |              | √    |
| Captioning Model [20]  |     |      |     |     |              | √    |
| Captioning Model [21]  |     |      |     |     |              | √    |
| Captioning Model [22]  |     |      |     |     |              | √    |
| Captioning Model [23]  |     |      |     |     |              | √    |
| Captioning Model [24]  |     |      |     |     |              | √    |
| Captioning Model [25]  |     |      |     |     |              | √    |
| Captioning Model [26]  |     |      |     |     |              | √    |
| Captioning Model [27]  |     |      |     |     |              | √    |
Table 3: Continued

| Image Captioning Model | LBL | LSTM | RNN | DTR | Language CNN | MELM |
|------------------------|-----|------|-----|-----|--------------|------|
| Captioning Model [28]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [29]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [30]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [31]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [32]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [33]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [34]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [35]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [36]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [37]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [38]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [39]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [40]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [41]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [42]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [43]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [44]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [45]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [46]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [47]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [48]  | ✓   | ✓    | ✓   | ✓   | ✓            | ✓   |
| Captioning Model [49]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [50]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [51]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [52]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [53]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [54]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [55]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [56]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [57]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [58]  | ✓   |      |     |     | ✓            |      |
| Captioning Model [59]  | ✓   |      |     |     | ✓            |      |

6. Conclusion
Image captioning involves both image processing and language processing fields. The purpose of an image captioning model is to provide natural language explanations for a given image. In this paper, the different image encoders used by image captioning models for image feature extraction are discussed. Also, the paper discusses the language models employed by the state-of-the-art image captioning models for generating natural language explanations for a given image. In future, the aim is to build the datasets that include text in images so that this textual information can be fused with the visual component to produce more accurate captions.
References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[2] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[3] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[5] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826..

[6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[7] Mnih, A., & Hinton, G. (2007, June). Three new graphical models for statistical language modelling. In Proceedings of the 24th international conference on Machine learning (pp. 641-648).

[8] De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). Generating typed dependency parses from phrase structure parses. In Lrec (Vol. 6, pp. 449-454).

[9] Fang, H., Gupta, S., Landola, F., Srivastava, R. K., Deng, L., Dollár, P., ... & Lawrence Zitnick, C. (2015). From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1473-1482).

[10] Kiros, R., Salakhutdinov, R., & Zemel, R. (2014, January). Multimodal neural language models. In International conference on machine learning (pp. 595-603).

[11] Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539.

[12] Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2014). Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090.

[13] Karpathy, A., Joulin, A., & Fei-Fei, L. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In Advances in neural information processing systems (pp. 1889-1897).

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

[15] Chen, X., & Lawrence Zitnick, C. (2015). Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2422-2431).

[16] Zhou, L., Xu, C., Koch, P., & Corso, J. J. (2017, October). Watch what you just said: Image captioning with text-conditional attention. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017 (pp. 305-313).

[17] Jia, X., Gavves, E., Fernando, B., & Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE international conference on computer vision (pp. 2407-2415).

[18] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).

[19] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition
[20] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057).

[21] Jin, J., Fu, K., Cui, R., Sha, F., & Zhang, C. (2015). Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv preprint arXiv:1506.06272.

[22] Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., & Salakhutdinov, R. R. (2016). Review networks for caption generation. In Advances in neural information processing systems (pp. 2361-2369).

[23] Sugano, Y., & Bulling, A. (2016). Seeing with humans: Gaze-assisted neural image captioning. arXiv preprint arXiv:1608.05203.

[24] Mathews, A., Xie, L., & He, X. (2015). Senticap: Generating image descriptions with sentiments. arXiv preprint arXiv:1510.01431.

[25] Wang, C., Yang, H., Bartz, C., & Meinel, C. (2016, October). Image captioning with deep bidirectional LSTMs. In Proceedings of the 24th ACM international conference on Multimedia (pp. 988-997).

[26] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4565-4574).

[27] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 11-20).

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

[29] Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., ... & Sienkiewicz, C. (2016). Rich image captioning in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 49-56).

[30] Ma, S., & Han, Y. (2016, July). Describing images by feeding LSTM with structural words. In 2016 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[31] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4651-4659).

[32] Yang, L., Tang, K., Yang, J., & Li, L. J. (2017). Dense captioning with joint inference and visual context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2193-2202).

[33] Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., & Darrell, T. (2016). Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-10).

[34] Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2017). Boosting image captioning with attributes. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4894-4902).

[35] Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 375-383).

[36] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (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 (pp. 5659-5667).

[37] Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., ... & Deng, L. (2017). Semantic compositional networks for visual captioning. In Proceedings of the IEEE conference on...
computer vision and pattern recognition (pp. 5630-5639).

[38] Pedersoli, M., Lucas, T., Schmid, C., & Verbeek, J. (2017). Areas of attention for image captioning. In Proceedings of the IEEE international conference on computer vision (pp. 1242-1250).

[39] 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).

[40] Chunseong Park, C., Kim, B., & Kim, G. (2017). Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 895-903).

[41] Wang, Y., Lin, Z., Shen, X., Cohen, S., & Cottrell, G. W. (2017). Skeleton key: Image captioning by skeleton-attribute decomposition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7272-7281).

[42] Tavakoli, H. R., Shetty, R., Borji, A., & Laaksonen, J. (2017). Paying attention to descriptions generated by image captioning models. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2487-2496).

[43] Liu, C., Mao, J., Sha, F., & Yuille, A. (2016). Attention correctness in neural image captioning. arXiv preprint arXiv:1605.09553.

[44] Gan, C., Gan, Z., He, X., Gao, J., & Deng, L. (2017). Stylenet: Generating attractive visual captions with styles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3137-3146).

[45] Dai, B., Fidler, S., Urtasun, R., & Lin, D. (2017). Towards diverse and natural image descriptions via a conditional gan. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2970-2979).

[46] Shetty, R., Rohrbach, M., Anne Hendricks, L., Fritz, M., & Schiele, B. (2017). Speaking the same language: Matching machine to human captions by adversarial training. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4135-4144).

[47] Liu, S., Zhu, Z., Ye, N., Guadarrama, S., & Murphy, K. (2017). Improved image captioning via policy gradient optimization of spider. In Proceedings of the IEEE international conference on computer vision (pp. 873-881).

[48] Gu, J., Wang, G., Cai, J., & Chen, T. (2017). An empirical study of language cnn for image captioning. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1222-1231).

[49] Yao, T., Pan, Y., Li, Y., & Mei, T. (2017). Incorporating copying mechanism in image captioning for learning novel objects. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6580-6588).

[50] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7008-7024).

[51] Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T., & Saenko, K. (2017). Captioning images with diverse objects. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5753-5761).

[52] Zhang, L., Sung, F., Liu, F., Xiang, T., Gong, S., Yang, Y., & Hospedales, T. M. (2017). Actor-critic sequence training for image captioning. arXiv preprint arXiv:1706.09601.

[53] 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.

[54] Aneja, J., Deshpande, A., & Schwing, A. G. (2018). Convolutional image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5561-5570).

[55] Wang, Q., & Chan, A. B. (2018). Cnn+ cnn: Convolutional decoders for image captioning.
arXiv preprint arXiv:1805.09019.

[56] Yang, X., & Xu, C. (2019). Image Captioning by Asking Questions. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(2s), 1-19.

[57] Wang, J., Wang, W., Wang, L., Wang, Z., Feng, D. D., & Tan, T. (2020). Learning visual relationship and context-aware attention for image captioning. Pattern Recognition, 98, 107075.

[58] Yu, N., Hu, X., Song, B., Yang, J., & Zhang, J. (2018). Topic-oriented image captioning based on order-embedding. IEEE Transactions on Image Processing, 28(6), 2743-2754.

[59] Sharma, H., & Jalal, A. S. (2020). Incorporating external knowledge for image captioning using CNN and LSTM. Modern Physics Letters B, 34(28), 2050315.

[60] Sharma, H., Agrahari, M., Singh, S. K., Firoj, M., & Mishra, R. K. (2020, February). Image Captioning: A Comprehensive Survey. In 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC) (pp. 325-328). IEEE.