A Survey on Image Captioning datasets and Evaluation Metrics

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

Abstract. In the task of image captioning, a natural language explanation is generated for a given image. It uses the subfields of artificial intelligence: computer vision and language generation. Convolutional Neural Network (CNN) is generally applied to capture image features and language processing models such as Recurrent Neural Network for sentence generation. In this paper, various datasets and evaluation metrics which are useful for image captioning task are discussed. Also, the datasets and evaluation metrics applied by the state-of-the-art image captioning models is summarized.

Keywords: Artificial Intelligence; Convolutional Neural Network; Recurrent Neural Network; Image Captioning; Computer Vision

1. Introduction
Images are all around us. These images are coming from different sources. As a human being, we are able to understand the content of these images without the need of image descriptions. But, computer don not understand these images as we do. So, there is need of automatic generation of image explanations for a given image. This task is known as image captioning. It uses both language models and computer vision techniques. An image captioning algorithm must be able to focus n the prominent object present in the image and then generate natural language based description for the given image. For retrieving the visual features, image captioning models uses Convolutional Neural Networks (CNN). On the other hand, for sentence generation, a family of Recurrent Neural Networks is employed.

Image captioning models use different class of datasets for training, testing, and validating these models. These datasets vary in a variety of viewpoints such as the size of dataset in terms of images, reference captions given per image, caption format, and size of image. For assessing the quality of generated captions, different evaluation metrics are employed. These generated captions are then compared to the given reference captions. Every metric has its own advantages and limitations. Figure 1 shows the caption generated for a given which is more like that generated by a human.

In this paper, the popular image captioning datasets such as Flickr8K [1], Flickr30K [2], MSCOCO [3], IAPR TC-12 [4], Visual Genome [5] and FlickrStyle10 are discussed. Also, the major evaluation metrics used by image captioning models such as BLEU [6], METEOR [7], ROUGE [8], CIDEr [9] and SPICE [10] are discussed. Finally, a summary of datasets and evaluation metrics employed by state-of-the-art models is presented.
2. Encoders Image Captioning Datasets

2.1 Flickr8K: It is a popular dataset used by many image captioning models. It contains 8K images taken from Flickr. The training part includes 6K images, test part includes 1K images and development part includes 1000 images. For every image of the dataset, five captions are given by humans.

2.2 Flickr30K: It is one of the popular dataset used for automatically generating image explanations and understanding grounded language. It is collection of 30,000 images taken from Flickr. Total 158k captions are given by humans for these images. No fixed dataset splits (training, testing, and validation) are provided. Flickr30K dataset also includes common object detectors, classifier for a color and identifying large objects with bias.

Table 1: Datasets used Image captioning models (Cap-Model represents captioning model)

| Captioning Model | Flickr8K | Flickr30K | MSCOCO | IAPR TC-12 | Visual Genome | FlickrStyle10K |
|------------------|----------|-----------|--------|------------|--------------|---------------|
| Cap-Model [11]   |          |          |        |            |              | √             |
| Cap-Model [12]   |          |          |        |            |              | √             |
| Cap-Model [13]   | √        | √         |        |            |              |               |
| Cap-Model [14]   | √        | √         |        |            |              |               |
| Cap-Model [15]   | √        |           |        |            |              |               |
| Cap-Model [16]   | √        | √         | √      | √          |              |               |
| Cap-Model [17]   | √        | √         |        |            |              |               |
| Cap-Model [19]   | √        | √         |        |            |              |               |
| Cap-Model [20]   | √        | √         |        |            |              |               |
| Cap-Model [21]   | √        | √         |        |            |              |               |
| Cap-Model [22]   | √        | √         |        |            |              |               |
| Cap-Model [23]   | √        | √         |        |            |              |               |
| Cap-Model [24]   | √        |           |        |            |              |               |
| Cap-Model [25]   | √        |           |        |            |              |               |
| Cap-Model [26]   |          | √         |        |            |              | √             |
| Cap-Model [27]   | √        | √         |        |            |              | √             |
| Captioning Model | Flickr8K | Flickr30K | MSCOCO | IAPR TC-12 | Visual Genome | FlickrStyle10K |
|------------------|---------|----------|--------|------------|---------------|----------------|
| Cap-Model [28]   |         | √        |        |            |                |                |
| Cap-Model [29]   |         |          |        |            |                |                |
| Cap-Model [30]   | √       |          |        |            |                |                |
| Cap-Model [31]   |         |          | √      |            |                |                |
| Cap-Model [32]   | √       |          |        |            |                |                |
| Cap-Model [33]   | √       |          | √      |            |                |                |
| Cap-Model [34]   |         |          |        |            |                | √              |
| Cap-Model [35]   |         |          | √      |            |                |                |
| Cap-Model [36]   |         |          | √      |            |                |                |
| Cap-Model [37]   |         | √        | √      |            |                |                |
| Cap-Model [38]   | √       | √        | √      |            |                |                |
| Cap-Model [39]   | √       |          | √      |            |                |                |
| Cap-Model [40]   |         |          | √      |            |                |                |
| Cap-Model [41]   |         |          |        |            |                |                |
| Cap-Model [42]   |         |          |        |            |                |                |
| Cap-Model [43]   |         |          | √      |            |                |                |
| Cap-Model [44]   |         |          | √      |            |                |                |
| Cap-Model [45]   |         | √        | √      |            |                |                |
| Cap-Model [46]   |         |          |        |            |                | √              |
| Cap-Model [47]   |         | √        | √      |            |                |                |
| Cap-Model [48]   |         |          | √      |            |                |                |
| Cap-Model [49]   |         |          | √      |            |                |                |
| Cap-Model [50]   |         |          | √      |            |                |                |
| Cap-Model [51]   |         |          | √      |            |                |                |
| Cap-Model [52]   |         |          | √      |            |                |                |
| Cap-Model [53]   |         |          | √      |            |                |                |
| Cap-Model [54]   |         |          | √      |            |                |                |
| Cap-Model [55]   |         | √        | √      |            |                | √              |
| Cap-Model [56]   |         |          | √      |            |                |                |
| Cap-Model [57]   |         |          |        |            |                | √              |
| Cap-Model [58]   |         |          | √      |            |                |                |
| Cap-Model [59]   |         |          | √      |            |                |                |
| Cap-Model [60]   |         |          | √      |            |                |                |
| Cap-Model [61]   | √       |          |        |            |                | √              |

2.3 *MSCOCO*: It is a very huge dataset used for image captioning, recognition and segmentation. It contains more than 30K images with 5 captions for each and every image. Also, it contains 80 object classes and more than 2 million instances.
2.4 IAPR TC-12: It includes 20K images. The sources of these images are animals, pictures of persons, games and other places around the globe. Images in this dataset include multiple objects. The captions for a given images are provided in different languages in this dataset.

2.5 Visual Genome: It contains captions for different regions of an image unlike the other datasets where captions are provided for the whole image. It contains more 108K images. For each image, 35 objects on an average, 26 features and 21 pair-wise associations between different objects are provided in the dataset.

2.6 FlickrStyle10K: It contains 10K images taken from Flickr with stylized captions. The dataset splits are performed as: 7K images to train, 2K images to validate and 1K images to test. For every image in the dataset, romantic, entertaining, and realistic captions are provided.

Table 1 summarized the datasets used by popular image captions models. Cap-Model [29] uses ReferIt dataset and Cap-Model [42] uses Instagram dataset.

3. Evaluation Metrics
3.1 Bilingual evaluation understudy (BLEU): This metric is employed to evaluate the superiority of machine produced content. A text sentence is compared with the group of a given reference captions and further scores are calculated for each one of them. These calculated scores are averaged for determining the quality of generated captions. BLEU scores are good for short generated captions. Also, in some cases high BLEU score does not imply high quality of generated captions.

| Captioning Model | BLEU | METEOR | ROUGE | CIDEr | SPICE | Human Evaluation | R@K |
|------------------|------|--------|-------|-------|-------|------------------|-----|
| Cap-Model [11]   | ✓    | ✓      |       |       |       |                  |     |
| Cap-Model [12]   | ✓    |        |       |       |       |                  |     |
| Cap-Model [13]   |      |        |       | ✓     |       |                  |     |
| Cap-Model [14]   | ✓    |        |       |       |       |                  | ✓   |
| Cap-Model [15]   |      |        |       |       |       |                  | ✓   |
| Cap-Model [16]   | ✓    |        |       |       |       |                  | ✓   |
| Cap-Model [17]   | ✓    | ✓      | ✓     | ✓     |        |                  | ✓   |
| Cap-Model [19]   | ✓    | ✓      |       |       |       |                  |     |
| Cap-Model [20]   | ✓    | ✓      |       |       |       |                  |     |
| Cap-Model [21]   | ✓    | ✓      |       |       |       |                  |     |
| Cap-Model [22]   | ✓    | ✓      |       |       |       |                  |     |
| Cap-Model [23]   | ✓    | ✓      | ✓     | ✓     |        |                  | ✓   |
| Cap-Model [24]   | ✓    | ✓      |       |       |       |                  |     |
| Cap-Model [25]   | ✓    | ✓      | ✓     |       |       |                  | ✓   |
| Cap-Model [26]   | ✓    | ✓      | ✓     | ✓     |        |                  | ✓   |
| Cap-Model [27]   | ✓    | ✓      | ✓     | ✓     |        |                  | ✓   |
| Cap-Model [28]   |      | ✓      |       |       |       |                  |     |
| Captioning Model       | BLEU | METEOR | ROUGE | CIDEr | SPICE | Human Evaluation | R@K |
|-----------------------|------|--------|-------|-------|-------|------------------|-----|
| Cap-Model [29]        | ✓    | ✓      |       |       |       |                  | ✓   |
| Cap-Model [30]        | ✓    | ✓      |       |       |       |                  |     |
| Cap-Model [31]        | ✓    | ✓      | ✓     | ✓     |       |                  | ✓   |
| Cap-Model [32]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [33]        | ✓    | ✓      | ✓     | ✓     | ✓     |                  | ✓   |
| Cap-Model [34]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [35]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [36]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [37]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [38]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [39]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [40]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [41]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [42]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [43]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [44]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [45]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [46]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [47]        | ✓    | ✓      | ✓     | ✓     |       |                  | ✓   |
| Cap-Model [48]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [49]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [50]        | ✓    | ✓      | ✓     | ✓     |       |                  | ✓   |
| Cap-Model [51]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [52]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [53]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [54]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [55]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [56]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [57]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [58]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [59]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [60]        | ✓    | ✓      | ✓     | ✓     |       |                  |     |
| Cap-Model [61]        | ✓    | ✓      | ✓     | ✓     | ✓     |                  |     |

3.2 **Metric for Evaluation of Translation with Explicit ORdering (METEOR):** It is mainly utilized to assess the machine translated language. Reference texts are compared to standard word parts. Apart from this, part of a sentence and word-synonyms are also used for comparing. This metric is used for establishing improved correlation at the segment or sentence stage.

3.3 **Recall-Oriented Understudy for Gisting Evaluation (ROUGE):** It's group of metrics applied for evaluating the text summary. In this metric, word strings, pair of words, and n-grams with a class of
reference summaries generated by people. Various forms of ROUGE like ROUGE-1, ROUGE-2, ROUGE-W and ROUGE-SU4 are employed for diverse applications. For example, for single document assessing, ROUGE-1 and ROUGE-W are most suitable. For short summaries, ROUGE-2 and ROUGE-SU4 give good accuracy. On the other hand, ROUGE has the limitation in assessing summary of multi-document text.

3.4 Consensus-based Image Description Evaluation (CIDEr): It is an automatic consensus metric for assessing image explanations. In most of the datasets, five captions are given for an image. The metrics discussed above deal with this smaller set of reference captions which may not be adequate enough to compute the agreement between produced captions and human generated captions. On the other hand, CIDEr attains human agreement by employing term frequency-inverse document frequency (TF-IDF).

3.5 Semantic Propositional Image Caption Evaluation (SPICE): This metric is based on semantic concept. It is a new metric for caption evaluation. It used the concept derived from graph-based semantic depiction known as scene-graph. The scene-graph captures the object’s information, their features and the relationship between these objects from image explanations.

Table 2 summarized the evaluation metrics used by popular image captions models. Figure 2 shows the captions generated by Cap-Model [58] (Left) and Cap-Model [59] (Right).

Caption: Motorcycles standing near a building.
Caption: A man is surfing on a wave.

Fig.2. Captions generated by Cap-Model [58] (Left) and Cap-Model [59] (Right)

4. Conclusion
Computer vision and language models are used by image captioning methods to generate the descriptions for a given image. The aim is to build image captioning models that can generate caption which are almost similar to the descriptions generated by humans. The applications of image captioning are in different domains such as medicine, teaching, indexing of images and helping blind people. In this paper, the different datasets used for the task of image captioning are discussed. Also, the paper explains the different evaluation metrics to assess the quality of generated captions. The paper also presents the summary of datasets and evaluation metrics used by the state-of-the-art image captioning models. In future, the researchers may work on open domain datasets for generating image captions. Also, how image captioning models can help in visual question answering is an open area of research.

References
[1] Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47,
[2] Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision (pp. 2641-2649).

[3] 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.

[4] Grubinger, M., Clough, P., Müller, H., & Deselaers, T. (2006, May). The iaprx tc-12 benchmark: A new evaluation resource for visual information systems. In International workshop onolimage (Vol. 2).

[5] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Bernstein, M. S. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123(1), 32-73.

[6] 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).

[7] 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).

[8] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

[9] 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).

[10] Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016, October). Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision (pp. 382-398). Springer, Cham.

[11] Fang, H., Gupta, S., Iandola, 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).

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

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

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

[15] 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).

[16] 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.

[17] 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).

[18] 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).

[19] 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).
[20] 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).

[21] 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 (pp. 3156-3164).

[22] 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).

[23] 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.

[24] 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).

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

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

[27] 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).

[28] 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).

[29] 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).

[30] 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.

[31] 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).

[32] 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.

[33] 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).

[34] 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).

[35] 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).

[36] 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).

[37] 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).
[38] 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).

[39] 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).

[40] 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).

[41] 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).

[42] 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).

[43] 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).

[44] 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).

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

[46] 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).

[47] 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).

[48] 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).

[49] 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).

[50] 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).

[51] 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).

[52] 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).

[53] 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).

[54] 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.

[55] 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.

[56] 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).

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

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

[59] 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.

[60] 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.

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

[62] 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.