A Survey on Various Deep Learning Models for Automatic Image Captioning

Gaurav1, Pratistha Mathur2
1PhD Scholar, Department of Information Technology, Manipal University Jaipur, Rajasthan 303007, India
2Professor, Department of Information Technology, Manipal University Jaipur, Rajasthan 303007, India.

Email: gauravsingla31@gmail.com

Abstract: Automatic Image captioning means the generation of a caption for an image by a machine. Image captioning is performed by recognizing objects, attributes and interconnection between them. This task involves computer vision for image understanding, natural language processing for syntax and semantics purpose and machine learning for caption generation. Preferably CNN is used to understand features of an image and RNN is used for sentence generation. Earlier, machine learning approach was used for this purpose. Input data is used to extract the features in traditional machine learning. Extracting features like handcrafted from large dataset is not so easy and feasible. Later on, various deep learning-based approaches were proposed. In deep learning, retrieval based and template-based methods were proposed but faced some issues like missing important objects and fixed length caption respectively. Then end to end learning approach based on deep learning network came into existence and image captioning task became more efficient. The objective of this paper is to study and compare various end to end learning-based framework for image captioning using standard evaluation metric and to understand how can these frameworks be used for various research applications. Along with the comparison, futuristic challenges have also been discussed.

Keywords: Image captioning, Encoder, Decoder, Attention mechanism

1. Introduction

Automatic Image captioning means the generation of a caption for an image by a machine. Image captioning is performed by recognizing objects, attributes and interconnection between them. This task involves computer vision for image understanding, natural language processing for syntax and semantics purpose and machine learning for caption generation. Preferably CNN is used to understand features of an image and RNN is used for sentence generation. There are various applications of image captioning like describing images on social media, image indexing, virtual assistants, virtual summarization and helping visually disabled people [1].

Automatic image captioning is one of the popular research areas in the field of Machine learning that deals with mainly two tasks:

- Extracting features of an image
- Describing image using these generated features
Extracting features of an image
Convolution Neural Network extracts the features from image and the feature vector is prepared with the same dimension as the dimension of Recurrent Neural Network/Long short term memory network. CNN network is used for image understanding and RNN/LSTM is used for language description.

Describing image using features
Once features are obtained from CNN, these features are converted into a dimension that will be input dimension of RNN for sentence generation.

2. Various Methods for Image Captioning
Various methods have been used for Image Captioning. Among these methods, the two broad categories are:
- Techniques based on Machine Learning
- Techniques based on Deep Learning
Figure 3. Various Techniques for Image Captioning

**Machine Learning Approach for Image Captioning:**
In traditional machine learning approach, input data is used to extract the features. Machine learning approach can be used for image captioning but it is not highly efficient. The reason is extracting handcrafted features such as LBPs (Local Binary Pattern), HOG (Histogram of oriented gradients), SIFT (Scale invariant feature transform) from large dataset is not so easy and feasible. As real-world images are more complex and dataset can be highly diversified. So, machine learning approach for image captioning is not highly efficient approach. But from last 6-7 years, various deep learning papers have been published for image captioning. Various approaches have been proposed using deep learning framework. Now, we will see various deep learning-based framework for image captioning.

**Deep Learning Approach for Image Captioning:**
- Image captioning using retrieval-based method
- Image captioning using template-based method
- Image captioning using end to end learning method

**Retrieval based approach for image captioning:**
Image captioning using retrieval approach uses Deep CNN and auto encoders to find the features from images and found various words relating to these features. This method finds similarly captioned images from the dataset and then the final caption is adjusted as per founded captions.

**Template based approach for image captioning:**
Retrieval based approach is useful approach or method for image captioning but there are few limitations with the approach as it generates caption using words found in similar images only. This approach may miss some important objects that are not available in previously found words. Therefore, researchers use the template method to overcome the issues found in retrieval approach. In this approach, few important
words representing like objects and their attributes are found. Then these generated words are then fixed with the template to finally generate a caption. It is effective approach but has limitations of generation of fixed length caption only.

End to End learning approach for image captioning:
End to end learning approach is the best approach so far and is used by almost every researcher for image captioning purpose. Learning of the parameter can be performed directly through the training only. Deep Convolution Neural Network (DCNN) is used as an encoder to read and convert input images into feature vectors, with the same dimension that is the input dimension of RNN that will be used to translate the vector into a sentence. [5].

3. Datasets:
There are various datasets available for image captioning and these datasets vary in terms of number of images, objects and their attributes.

- **MS COCO Dataset**: 300k images, nearly 2 million objects, 80 different categories of objects, and 05 captions for each image.
- **Flickr30K Dataset**: 30k images with 158k annotators.
- **Flickr8K Dataset**: 8K images, 6k for training and rest 2k for testing and validation.
- **MIT-Adobe 5k Dataset**: Dataset consists of 5,000 images. This dataset contains various kinds of images based on different subjects, scenes, conditions, lighting etc.
- **FlickrStyle10K Dataset**: This dataset has 10k images with stylized captions. Among these, there are 7k training images, 2k for testing and 1k for validation purpose. The images are based on romantic, humorous and some factual scenes. [4].

4. Evaluation Metrics:
The use of evaluation metric is to provide a score that can be helpful to identify the quality of generated captions for the image. There are various metric available that measure the different arena of caption like syntax and semantics and its relation with the objects available in image. The commonly used evaluation metrics are as follows:

- **BLEU**: The metric is Bilingual evaluation understudy that compares the generated caption with referenced caption and word to word matching takes place. On an average score is computed. However, syntactical correctness is not measured here.
  Generated Text: This is for test
  Reference Text: This is only test (score 0.75 or 75%)
  BLUE-1 matches word to word. Similarly, we have BLUE-2, BLUE-3, BLUE-4 evaluation metrics for pair wise matching.
- **ROUGE**: The metric is Recall-Oriented Understudy for Gisting Evaluation that measures the quality of textual summaries.
- **METEOR**: It is Metric for Evaluation of Translation with Explicit Ordering and is useful to measure stems of a sentence and synonym for words. This is a better approach to make a better correlation at the sentence or the segment level.
  e.g. Exact - Candidate-Good, Reference-Good: [Match (Yes)]
  Stemmer - Candidate-Goods, Reference-Good: [Match (Yes)]
  Synonymy - Candidate-Well, Reference-Good: [Match (Yes)]
• CIDEr is Consensus-based Image Description Evaluation metric for measuring image captions. This method achieves human consensus using TF-IDF i.e. term frequency-inverse document frequency [4].

5. Image Captioning methods based on Deep Learning
Various approaches or models have been proposed for this task and producing state of the art results. Here we will discuss the models that are deep learning based and are working on end-to-end learning approach.

5.1 Using simple CNN as encoder and LSTM as decoder
The model uses deep neural network approach that combines computer vision and machine translation to generate caption for an image. It uses CNN (Convolution Neural Network) to extract image features and then RNN that takes image features as input and generate caption. But in the model RNN is replaced by LSTM (Long short-term memory).
LSTM is highly efficient for vanishing and exploding gradients and has achieved great success in translation and sequence generation. Basic unit of the LSTM is cell C that has three different gates. The gates are input gate, output gate, and forget gate and these gates manage the LSTM and decide the sentence to be generated.
Using simple CNN as encoder and LSTM as decoder, the following results have been achieved:

| Table 1. Evaluation score (in %) using simple encoder and decoder [6] |
|--------------------------|----------------|------------|-------------|----------------|
|                         | Blue-1 | Blue-4 | METEOR | CIDEr |
| Flickr8K                 | 63     | -      |         |       |
| Flickr30K                | 66     | -      |         |       |
| MSCOCO                   | -      | 27.7   | 23.7    | 85.5  |
| Pascal                   | 59     | -      |         |       |

5.2 Visual attention model
This model uses some attention mechanism along with simple encoder and decoder. Here attention mechanism is attached with the encoder to get better and accurate image features. Using this mechanism better and efficient features are generated and passed to LSTM to get better sequence generation.
The following are the evaluation score achieved on the most popular datasets using the mentioned model:

| Table 2. Evaluation Score using visual attention model [7] |
|---------------|----------------|----------------|-------------|----------------|
|               | B-1  | B-2  | B-3  | B-4  | Meteor |
| Flickr8K      | 67   | 44.8 | 29.9 | 19.5 | 20.3   |
| Flickr30K     | 66.9 | 43.9 | 29.6 | 19.9 | 18.46  |
| MSCOCO        | 71.8 | 50.4 | 35.7 | 25.0 | 23.04  |

It can be clearly seen that using attention model, better results are achieved as compared to previous simple encoder decoder method. Here the research issue is that better image attention model can be generated to get accurate and more semantically correct features.
5.3 DHEDN Framework:
As the model name indicates that it used multilayer CNN and multilayer LSTM. With this approach, multilevel semantics of vision and language from CNN and LSTM respectively can be generated and fused together for final caption generation and therefore model enhances the capacity of network and provides better results.
The following evaluation metrics shows the performance of the model:

|                      | B-1 | B-2 | B-3 | B-4 | Meteor | R-L | CIDEr |
|----------------------|-----|-----|-----|-----|--------|-----|-------|
| Flickr8K             | 65.1| 47.0| 32.6| 20.5| 20.5   | -   | -     |
| Flickr30K            | 65.4| 46.8| 32.9| 23.1| 19.3   | -   | -     |
| MSCOCO               | 73.1| 56.3| 42.6| 32.3| 25.8   | 53.8| 100.1 |

Using DHEDN efficient results are achieved but still the network lacks in major concerns like visual attention can be enhanced and complementary knowledge can be added to enhance image captioning.

5.4 Image captioning with Deep Bi-directional LSTMs and multitask learning:
As CNN is used for feature generation and LSTM is used for sentence generation. Bi-directional LSTM means two different LSTMs i.e. one LSTM for history context and other LSTM for future context for better semantic information. Such models have produced better captions in terms of semantic info. The table shows the evaluation metrics on different datasets using this model.

|                      | B-1 | B-2 | B-3 | B-4 | Meteor | CIDEr |
|----------------------|-----|-----|-----|-----|--------|-------|
| Flickr8K             | 66.9| 48.4| 33.3| 22.8| -      | -     |
| Flickr30K            | 63.6| 44.8| 30.4| 20.5| -      | -     |
| MSCOCO               | 68.7| 50.9| 36.4| 25.8| 22.9   | 73.9  |

Along with Bi-LSTM, visual attention can be added to make the model more efficient.

5.5 Bottom up and Top-down attention mechanism:
Bottom-up mechanism provides a set of image regions with salient features. In this model, Bottom up part is implemented by (R-CNN) Recurrent-Convolution Neural Network. Top – down approach is used to predict an attention distribution over image regions that makes the model efficient. Weighted average of all image features is calculated to generate final feature vector.
The model has been implemented only on MSCOCO Dataset. The following score shows the performance of the model:

|                      | B-1 | B-4 | Meteor | R-L | CIDEr |
|----------------------|-----|-----|--------|-----|-------|
| MSCOCO               | 79.8| 36.3| 27.7   | 56.9| 120.1 |

The model is providing better Blue score and CIDEr as compared to previous proposed models.
5.6 Image caption generation using Dual attention mechanism:
Dual attention mechanism uses visual attention as well as textual attention in a single model for image captioning purpose.

| Table 6. Evaluation score using Dual attention mechanism [11] |
|---------------------------------------------------------------|
| B-1 | B-2 | B-3 | B-4 | Meteor | R-L | CIDEr |
|---|---|---|---|---|---|---|
| AICICC | 0.741 | 0.614 | 0.509 | 0.423 | 0.350 | 0.602 | 1.236 |

However, the model has achieved state of the results but there are few research gaps that can be taken care of for futuristic point of view. Image label qualities can be enhanced for better image caption generation. In the model fusion of visual and textual attention is performed so it can be enhanced to improve image caption problem. AICICC is a Chinese benchmark dataset. The score shown in above table is obtained using AICICC dataset using the dual attention model.

5.7 Hierarchical Deep Neural Network for Image Captioning:
Model consists of bottom layer and top layer. Bottom layer extracts visual and high-level semantic info from images and detected regions respectively while the top layer integrates both of them with adaptive attention mechanism for caption generation. The following table shows the performance of the model:

| Table 7. Evaluation score using Hierarchical Deep NN [12] |
|----------------------------------------------------------|
| Blue-1 | Blue-2 | Blue-3 | Blue-4 | Meteor | Rough-L | CIDEr |
|---|---|---|---|---|---|---|
| MSCOCO | 0.724 | 0.558 | 0.421 | 0.318 | 0.254 | 0.536 | 1.012 |

Regional semantic based methods can be used for image captioning.

5.8 Using Semantic attention:
This model uses the combination of top down and bottom-up methods. Convolution neural network extracts image features and image concepts like objects and their attributes. Semantic attention model along with RNN combines image features and image concepts in a RNN that generates image caption. Here attention model changes attention weights for several candidate concepts w.r.t. RNN iteration. This approach provides better results as far as semantic part of caption is concerned. The following table shows the performance of the model using Flickr30k and MSCOCO datasets.

| Table 8. Evaluation score using semantic mechanism [13] |
|-------------------------------------------------------|
| B-1 | B-2 | B-3 | B-4 | Meteor | R-L | CIDEr |
|---|---|---|---|---|---|---|
| Flickr30K | 0.824 | 0.679 | 0.534 | 0.412 | 0.269 | 0.588 | 0.949 |
| MSCOCO | 0.910 | 0.786 | 0.654 | 0.534 | 0.341 | 0.667 | 1.685 |

Results of the proposed model are on the basis of ground truth visual attributes. From futuristic point of view, experiment with phrase based visual attributes with its distributed representation can be performed. New model for semantic attention mechanism can be generated.
5.9 Attention correctness in Neural Image Captioning:
In various research papers, attention mechanism has been discussed to achieve better image features and sentence generation. But nowhere it has been discussed that the mapping of attention with objects and other properties are effective or not. Here the author proposes an attention mapping for image captioning propose [14].

5.10 Captioning for images based on news articles:
Generating a model for images feed in news articles have never been an easy task. As such types of images contain real world information and objects so highly efficient model is required for such images. Here the author proposes a model that uses multimodal attention mechanism and an efficient language model that generates a caption or a sequence for images feed in news article [15].

5.11 Framework for obtaining correct image information and describing syntactically correct caption:
Various frameworks for image captioning have been proposed but here the author has focused on syntactically correct captions. A framework has been proposed for obtaining correct image information and describing syntactically correct caption.

Table 9. Evaluation score for the framework [16]

|        | B-1 | B-4 | Meteor | R-L | CIDEr |
|--------|-----|-----|--------|-----|-------|
| MSCOCO | 78.1| 36.7| 28.5   | 58.4| 119.2 |

5.12 Using Semantic Element Embedding:
Semantic element embedding is an efficient approach that has been proved in this model. Here one such algorithm is being used that is generating local semantic features. Semantic embed method, element embedding LSTM (EE-LSTM) uses both visual mechanism as well as semantic mechanism, local and global features to fill up the gap between image and semantics of generated caption.

Table 10. Evaluation score for semantic element embedding approach for image captioning [17]

|        | B-1  | B-2  | B-3  | B-4  | Meteor | R-L  | CIDEr |
|--------|------|------|------|------|--------|------|-------|
| Flickr8K| 59.8 | 40.8 | 27.5 | 18.4 |        | -    | -     |
| Flickr30K| 59.2 | 39.1 | 25.7 | 17.0 | -      | -    | -     |
| MSCOCO | 67.5 | 49.8 | 36.4 | 26.9 | 22.6   | 80.0 | 49.9  |

This model also proposes some research issues like element denoising method to reduce loss and element pre-processing procedure can be improved.

5.13 Image captioning and comparison of different encoders:
Here various image encoders have been compared on some common parameters. Training parameters are learning rate=0.08, optimizer=RMS prop, batch size=512. Comparison is performed using Flickr8K dataset and on Blue score only [18].

\[
\begin{align*}
\text{VGG 16} &= 0.84053816 \\
\text{VGG19} &= 0.8824129834 \\
\text{Inception V3} &= 0.895487043 \\
\text{ResNet V2} &= 0.882496
\end{align*}
\]
5.14 Using Interactive Dual GAN: Dual generative adversarial networks use both retrieval-based approach and generation-based approach for image captioning methods, leading to better ensemble model for image captioning. This mechanism incorporates retrieved captions into decoding process to enhance the informative and diversity of finally generated captions.

Table 11. Evaluation score for the model [19]

| Dataset | B-1 (%) | B-2 (%) | B-3 (%) | B-4 (%) | Meteor (%) | R-L (%) | CIDEr |
|---------|---------|---------|---------|---------|-------------|---------|-------|
| MSCOCO  | 81.3    | 65.4    | 50.7    | 38.5    | 28.5        | 58.8    | 123.5 |

5.15 Multilayer Dense Attention Model: Here in this model, image features are extracted using faster R-CNN. Multilayer dense attention model is decoded using LSTM. Model Gradient optimization technique is used to optimize the parameters in RL.

Table 12. Evaluation score for multilayer dense attention model for image captioning [20]

| Dataset | B-1 (%) | B-2 (%) | B-3 (%) | B-4 (%) | Meteor (%) | R-L (%) | CIDEr |
|---------|---------|---------|---------|---------|-------------|---------|-------|
| Flickr8K| 47.7    | 33.4    | 23.1    | 167.5   | 46.9        |         |       |
| Flickr30K| 52.0    | 39.1    | 26.0    | 209.1   | 51.0        |         |       |
| MSCOCO  | 51      | 34      | 24.8    | 118.3   | 56.5        |         |       |

5.16 Dual CNN- for paragraph generation: Dual CNN is one of the latest approaches that uses both sentence CNN and words CNN. When paragraph generation comes into picture for an image, then even RNN can be used but time complexity of training algorithm will be higher as compared to CNN. Stanford image-paragraph dataset have been used in model.

Table 13. Evaluation score for the model [21]

| Stanford image paragraph dataset | B-1 (%) | B-2 (%) | B-3 (%) | B-4 (%) | Meteor (%) | CIDEr |
|---------------------------------|---------|---------|---------|---------|-------------|-------|
|                                 | 41.6    | 24.4    | 14.3    | 8.6     | 15.6        | 17.4  |

This model has some research issues like better visual encoding to have more semantic information.

6. Comparative Analysis of Various Models

Table 14. Comparative analysis of various models for image captioning

| S. No. | Methods | Dataset | Evaluation Metrics | Scope of Improvements |
|--------|---------|---------|--------------------|-----------------------|
|        |         |         | B-1 (%) | B-2 (%) | B-3 (%) | B-4 (%) | METEOR (%) | R-L (%) | CIDEr |
| 1      | Using Simple CNN as Encoder and LSTM as Decoder | Flickr8K | 63 |       |       |       |         |        |    |
|        |         | Flickr30K | 66 |       |       |       |         |        |    |
|        |         | MSCOCO   | 27.7 | 23.7 |       |       |         |        |    |
| 2      | Visual Attention | Flickr8K | 67 | 44.8 | 29.9 | 19.5 | 20.3 |        |    |
|   | Model                          | Flickr30K | 66.9 | 43.9 | 29.6 | 19.9 | 18.46 | can be generated for better accuracy |
|---|-------------------------------|-----------|------|------|------|------|-------|-------------------------------------|
|   |                               | MSCOCO    | 71.8 | 50.4 | 35.7 | 25   | 23.04 |                                     |
| 3 | DHEDN Framework               | Flickr8K  | 65.1 | 47   | 32.6 | 20.5 | 20.5  | A Better framework can be used for high accuracy |
|   |                               | Flickr30K | 65.4 | 46.8 | 32.9 | 23.1 | 19.3  |                                     |
|   |                               | MSCOCO    | 73.1 | 56.3 | 42.6 | 32.3 | 25.8  | 53.8 | 100.1                               |
| 4 | Deep Bi-Directional LSTM      | Flickr8K  | 66.9 | 48.4 | 33.3 | 22.8 | -     | Visual attention can be added to Bi-LSTM to enhance the model |
|   |                               | Flickr30K | 63.6 | 44.8 | 30.4 | 20.5 | -     |                                     |
|   |                               | MSCOCO    | 68.7 | 50.9 | 36.4 | 25.8 | 22.9  | 73.9                               |
| 5 | Bottom-Up and Top-Down attention | MSCOCO   | 77.2 |      | 36.2 | 27.0 | 56.4  | A Better framework can be used for high accuracy |
| 6 | Dual Attention Mechanism      | AICICC    | 74.1 | 61.4 | 50.9 | 42.3 | 35    | Image Label qualities can be improved |
| 7 | Using Hierarchical Deep Neural Network | MSCOCO  | 72.4 | 55.8 | 42.1 | 31.8 | 25.4  | Regional semantic based methods can be used |
| 8 | Using Deep RL Model           | MSCOCO    | 71.3 | 53.9 | 40.3 | 30.4 | 25.1  | A Better framework can be used for high accuracy |
| 9 | By Exploring visual relationshio p | MSCOCO  | 77.4 |      | 37.1 | 28.1 | 57.2  | Relationship modeling can be generalized for better performance |
| 10| Using Semantic Mechanism     | Flickr30K | 70.8 | 53.4 | 38.8 | 27.6 | 22.2  | Phrase based visual attributes can be used. New model for semantic attention mechanism can be improved |
|   |                               | MSCOCO    | 76.6 | 61.7 | 48.4 | 37.7 | 27.9  | 58.2 | 123.7                               |
| 11| Attention Balance Mechanism  | MSCOCO    | 78.1 |      | 36.7 | 28.5 | 58.4  | Fusion of Visual and Textual Attention |
| 12| Pointing Novel Objects        | MSCOCO    | 78.1 |      | 36.7 | 28.5 | 58.4  | Expansion of vocabulary and placement of copying novel |
|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
|   | Using Semantic Element Embedding | Flickr8K | 59.8 | 40.8 | 27.5 | 18.4 |
|   | Flickr30K | 59.2 | 39.1 | 25.7 | 17 | - | - |
|   | MSCOCO | 67.5 | 49.8 | 36.4 | 26.9 | 22.6 | 80 | 49.9 |
|   |   |   |   |   |   |   |
|   | Using Interactive Dual GAN | MSCOCO | 81.3 | 65.4 | 50.7 | 38.5 | 28.5 | 58.8 | 123.5 |
|   |   |   |   |   |   |   |
|   | Multilayer Dense Attention Model | Flickr8K | 47.7 | 33.4 | 23.1 | 46.9 | 167.5 |
|   | Flickr30K | 52 | 39.1 | 26 | 51 | 209.1 |
|   | MSCOCO | 51 | 34 | 24.8 | 56.5 | 118.3 |
|   | MSCOCO (Inception V3) | 69.5 | 51.8 | 38 | 27.7 | 23.5 | 89.4 |
|   |   |   |   |   |   |   |
|   | Using Convolution Language Decoder | Stanford Image paragraph dataset | 41.6 | 24.4 | 14.3 | 8.6 | 15.6 | 17.4 |

7. Futuristic Research Directions:
- Generating a model for an open domain dataset is still a research problem.
- Generating a model using unsupervised technique and reinforcement technique can be useful as it is not always feasible to obtain supervised data.
- The problem of automatic understanding of information in multimode is yet to be solved.
- To replace pre trained CNN model, combination of top down and bottom-up approach may be used for efficient results.
- Better image attention mechanism-based model can be generated for high level image features.
- Cascading based models can be generated for better evaluation score.

8. Conclusion
Here we have seen various image captioning models with their evaluation metrics on various datasets. Initially we discussed simple encoder and decoder approach. Then we discussed attention mechanism and found that better results are achieved. In attention mechanism, we applied visual attention as well as textual attention known as dual attention and found more better results. Along with these methods, semantic embedding methods have been used. GAN methods have been used for image caption generation. From the comparative table, it can be seen few approaches like dual generative adversarial networks and dual attention mechanism-based methods are providing state of the art results. From research point of view, even better model can be generated. A model can be generated for unsupervised data as it is not always possible to have dataset with images and their captions. An attention mechanism-based model can be generated that provides better evaluation score on almost every standard dataset.
References

[1] Saloni Kalra, Alka Leekha, “Survey of convolutional neural networks for image Captioning”, Journal of Information and Optimization Sciences, Taylor and Francis, ISSN: 0252-2667, P.No. 239-260, Feb-2020.
[2] https://wiki.pathmind.com/convolutional-network.
[3] https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/.
[4] MD. ZAKIR HOSSAIN, FERDOUS SOHEL, MOHD FAIRUZ SHIRATUDDIN, and HAMID LAGA, “A Comprehensive Survey of Deep Learning for Image Captioning”, ACM Computing Surveys, Vol. 51, No. 6, Article 118. Feb-2019.
[5] Xiaoxiao Liu, Qingyang Xu, Ning Wang, “A survey on deep neural network-based image captioning”, The Visual Computer, Springer Nature, https://doi.org/10.1007/s00371-018-1566-y, June-2018.
[6] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, “Show and Tell: A Neural Image Caption Generator”, arXiv:1411.4555v2 [cs.CV], Apr-2015.
[7] Kelvin xu, Yoshua Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, arXiv:1502.03044v3 [cs.LG], Apr-2016.
[8] Xinyu Xiao, Lingfeng Wang, Kun Ding, Shining Xiang, and Chunhong Pan, “Deep Hierarchical Encoder-Decoder Network for Image Captioning,” IEEE Transaction on Multimedia, Apr-2018.
[9] CHENG WANG, HAOJIN YANG, and CHRISTOPH MEINEL, “Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning”, ACM Trans. Multimedia Comput. Commun. Appl., Vol. 14, No. 2s, Article 40. April-2018.
[10] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”, cvpr, computer vision foundation, May-2019.
[11] Maofu Liu, Lingjun Li, Huijun Hu, Weili Guan, Jing Tian, “Image caption generation with dual attention mechanism”, Information Process and Management - 57, Elsevier, Nov-2019.
[12] Yuting Su, Yuqian Li, Ning Xu, An-An Liu, “Hierarchical Deep Neural Network for Image Captioning”, Neural processing letters, Springer Nature, 2019.
[13] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, “Image Captioning with Semantic Attention”, CVPR, Computer Vision Foundation, IEEE Explore, 2017.
[14] Chenxi Liu, Junhua Mao, Fei Sha, Alan Yuille, “Attention Correctness in Neural Image Captioning”, AAAI Conference on Artificial Intelligence (AAAI-17), 2017.
[15] Alasdair Tran, Alexander Mathews, Lexing Xie, “Transform and Tell: Entity-Aware News Image Captioning”, CVPR, Computer Vision Foundation, IEEE Explore, 2020.
[16] ZHENYU YANG, QIAO LIU, “ATT-BM-SOM: A Framework of Effectively Choosing Image Information and Optimizing Syntax for Image Captioning”, IEEE Access, March-2020.
[17] Xiaodan Zhang, Shengfeng He, Xinhang Song, Rynson W.H. Lau, Jianbin Jiao, Qixiang Ye, “Image captioning via semantic element embedding”, Neurocomputing, Elsevier, June-2019.
[18] Ankit Pal, Subasish Kar, Anuveksh Tanega, Vinay Kumar Jadoun, “Image Captioning and Comparison of Different Encoders”, Journal of Physics: Conference Series, doi:10.1088/1742-6596/1478/1/012004, 2020.
[19] Junhao Liu, Kai Wang, Chunpu Xu, Zhou Zhao, Rui Feng Xu, Ying Shen, Min Yang, “Interactive Dual Generative Adversarial Networks for Image Captioning”, Association for the Advancement of Artificial Intelligence, 2020.
[20] Eric ke wang, Xun zhang, Fan wang, Tsu-yang wu, and Chien-ming chen, Multilayer Dense Attention Model for Image Caption, IEEE Access, June-2019.
[21] Ruifan Li, Haoyu Liang, Yihui Shi, Fangxiang Feng, Xiaojie Wang, “Dual-CNN: A Convolutional language decoder for paragraph image captioning”, Neurocomputing, Elsevier, Feb-2020.