

An Efficient Deep Learning based Hybrid Model for Image Caption Generation

Mehzabeen Kaur¹, Harpreet Kaur²

Ph.D Research Scholar, Department of Computer Science and Engineering, Punjabi University, Patiala¹
Faculty, Department of Computer Science & Engineering, Punjabi University, Patiala²

Abstract—In the recent years, with the increase in the use of different social media platforms, image captioning approach play a major role in automatically describe the whole image into natural language sentence. Image captioning plays a significant role in computer-based society. Image captioning is the process of automatically generating the natural language textual description of the image using artificial intelligence techniques. Computer vision and natural language processing are the key aspect of the image processing system. Convolutional Neural Network (CNN) is a part of computer vision and used object detection and feature extraction and on the other side Natural Language Processing (NLP) techniques help in generating the textual caption of the image. Generating suitable image description by machine is challenging task as it is based upon object detection, location and their semantic relationships in a human understandable language such as English. In this paper our aim to develop an encoder-decoder based hybrid image captioning approach using VGG16, ResNet50 and YOLO. VGG16 and ResNet50 are the pre-trained feature extraction model which are trained on millions of images. YOLO is used for real time object detection. It first extracts the image features using VGG16, ResNet50 and YOLO and concatenate the result into single file. At last LSTM and BiGRU are used for textual description of the image. Proposed model is evaluated by using BLEU, METEOR and RUGE score.

Keywords—CNN; RNN; LSTM; YOLO

I. INTRODUCTION

In this www world, every day in our life, all have experienced with the huge number of images in a real world which are self-interpret by the individual human being by using their wisdom. Human are naturally programmed to convert the natural scene in to text but it is the complex task for the machine as they are not much efficient like human. Still, human generated captions are considered better as machine need human intervention and programmed accordingly for the better result. Due to the recent development in deep learning-based techniques, computers are capable to handle the challenges of image captioning like detection of object, attribute and their relationship, image feature extraction and generating syntactic and semantic image caption [1].

With the advancement of AI, so many new ideas have revolutionized in the areas of image processing and it has transformed the world in a surprising way. The image captioning Approach (Fig. 1) has wider application in the real world as it provides the better platform for human computer interaction. Due to the emerging application in image

processing, image captioning becomes the topic of interest for the academicians and researchers.

By seeing the Fig. 2, picture someone guess that two dogs are playing with toy and someone might say two dogs hauling in floating toy from the ocean or two dogs run through the water with rope in their mouths, so all of these captions are appropriate to describe this picture. Our brain is so much trained and advanced that it can describe a picture almost accurate but same was not the case with machines.

Hence, the main aim of the image captioning is first identified the different objects and their relationship present in the image using deep learning-based technique, generating the textual description using the natural language processing and evaluate the performance of the natural language-based description using different performance matrices. Object detection and segmentation are the part of the computer vision and done with the help of popular CNN and DNN and generating image description (Fig. 3) are the part of natural language processing which is done by RNN and LSTM. CNN works for understanding the objects of the image or scene and provide the answers the various questions about the objects in image like what, where, how, etc.



Fig. 1. Image captioning.

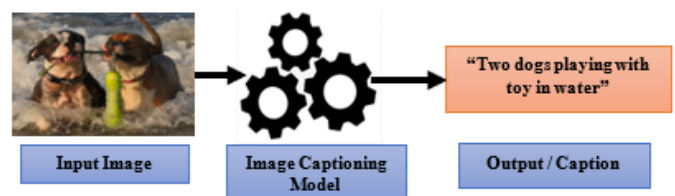


Fig. 2. Working of image captioning.

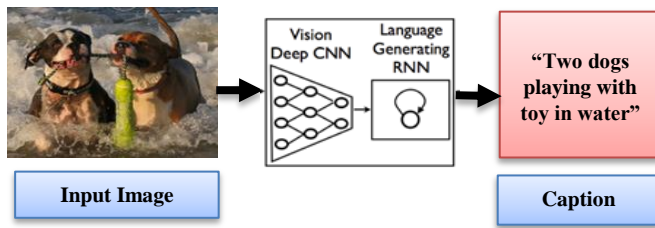


Fig. 3. Image captioning architecture.

For example, in Fig. 3, CNN identify the “dog”, “toy”, “water” and their relationship in the scene. Further RNN give the shape in textual form by using the keywords described by CNN by considering it in group of words. This one is also called the encoder-decoder architecture. Object detection is a part of computer vision which uses various algorithms, like YOLO, R-CNN, Mask R-CNN, MobileNet and SqueezeDet for detecting the different parts of the image efficiently.

II. LITERATURE SURVEY

In this section, review of literature in image captioning is presented. Various state-of-the-art techniques and model have been published in previous years to generate the human like captions. Image captioning approaches [11], [14] and [17] broadly classified in to Template-based [18-21], Retrieval-based [22-26], and Encoder-decoder methods [27-30]. In paper [31] a content selection approach has been proposed for image description by using geometric, conceptual and visual features of image. All of these models work on CNN, first use encode the image and extract the feature and further use RNN or LSTM to make captions of the image. In paper [1] researchers presented an image captioning model with probabilistic distribution using successor and predecessor words and image captioning. Attention and visual based approach are very famous approach in image captioning. In [2,3] authors generate the captions using the attention mechanism. In most of the papers predefined models were used in bulk of papers like VGG16 papers [1], [3-7], YOLO [8], Inception V3 [9-10], AlexNet [5], [7], ResNet [4-5], [12] and Unet [13] are the famous encoder or CNN model used for image feature extraction. For image caption generation or decoding, LSTM [4], [8-10] and [15], BiLSTM [7], [13], RNN [16]. Image captions are also generated in various languages like Chinese, Japanese, Hindi, Punjabi and German, etc.

Template based approach uses predefined templates of objects, actions and attributes to identify the input image [18], the authors use visual elements like object, action and scene for predicting the caption of the image. In [19] author takes the advantages of Conditional Random Field (CRF) based technique extract the features of the image. The proposed model evaluated using BLUE and ROUGE score on PASCAL dataset. As it is based upon pre-defined template it is not able to generate the caption of image with variable lengths.

Retrieval based approach generate caption by capering the features of the image with the datasets. It tries to finds the caption for input image by discovering similar features in the dataset. In [22] authors proposed a model to extract feature of the query image by searching it through the dataset and in [32],

the authors propose the caption by using the density estimation method. In [25], the authors used semantic and visual features for image caption generation.

In the original dataset we have five captions for each image and our goal is to train a particular model on this dataset. After the training phase model becomes efficient for extracting the features of the particular image, various predefined image classification models are available which uses state-of-the-art algorithms for classifying the thousands of different objects/images efficiently. These models come up with better accuracy with respect to image rate classification, like ResNet. These are very easy to implement.

Encoder-decoder based approach is a most widely used for machine translation and image caption generation which is based upon deep neural networks. A dual graph convolution network based is proposed in [33] and NIC (Neural Image Caption) model based on encoder-decoder architecture is in [27]. This one is a simple model where CNN is used as a encoder, and in the decoder end LSTM and RNN are used for image caption generation.

III. RESEARCH METHODOLOGY

Here, for extracting the visual feature of the image, CNN used as an encoder which have Convolution layer, Pooling layer, and fully connected layer. Earlier AlexNet was used for compute vision problems but nowadays, the transfer learning are in trends in where several pre-trained CNN based models are available like VGGNet, Inception V3, DenseNet, ResNet etc. which are available with different convolutional neural layers and used for saving the training time of the model. Further, decoder is used to generating the final captions which gets the input from the encoder. GRU, LSTM and RNN are the most commonly used decoder. RNN are suitable for short words sequence and LSTM is best for long sequence.

This section depicts the proposed hybrid research methodology. Our main objective of the proposed model is to achieve the higher Meteor value. Our model is based on an Encoder-Decoder approach where it used the concept of transfer learning. Here in the first phase, features of the image is extracted by using VGG16, ResNet50 and YOLO (You Only Look Once) separately. YOLO is an efficient object detection algorithm in real time with is developed in 2015 Joseph Redmon et al. whereas VGG16 (Visual Geometry Group) is an object detection and classification approach which is pretrained on ImageNet dataset. This is deep Convolutional Neural Network (CNN) architecture which uses 16 convolutional layers. ResNet50 is a deep CNN with 50 convolutional layers which is able to classify more than 1000 object category.

In second phase, concatenate of the features of image extracted by the VGG16, ResNet50 and YOLO and all the duplicate words are eliminated.

In third phase, captions are generated by using the BiGRU and LSTM. BiGRU (Bidirectional Gated Recurrent Units) is a Neural Network architecture used in NLP (Natural Language Processing). This architecture uses two GRUs for taking input in forward and backwards directions. LSTM (Long Short - Term Memory) is a type of recurrent neural network architecture which used feedback connections and capable of

identifying the relation between objects. In the last phase, both the captions are compared with the Meteor performance evaluation metrics. Final caption has the higher meteor value.

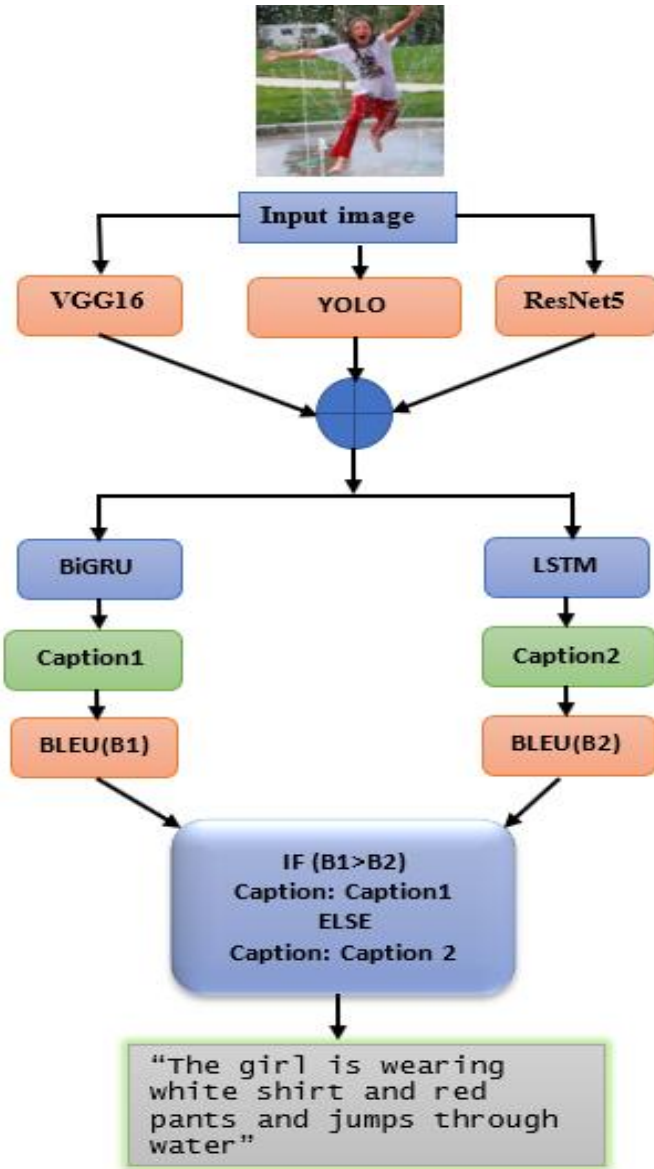


Fig. 4. Proposed image captioning architecture.

IV. DATASETS

Data are the backbone of any AI based systems. Recently image captioning is blessed with rich datasets like MSCOCO, Flickr8k, Flickr30k, PASCAL etc. in the dataset, every image is described in related five reference sentences. Every description of the scene is described by using different

algorithms and grammar. MSCOCO is a large dataset which was developed by Microsoft whose target to describe the image as a human being. It first understands the scene and complete the image recognition, segmentation and generating suitable caption of the image. It contains 82,783 images, with validation set 40,504 images, and the test set 40,775 images. Flickr30k dataset has 28000 training images, 1000 testing and 1000 validation images.

Here, in this paper a benchmark dataset Flickr8k for the training of the model. It contains 8000 images with 5 captions of each image which provides the clear descriptions of the silent objects. It has manually labelled captions for all the images in English language. The dataset is divided into two categories. First one is image directory which has 8k images with 5 captions. Out of 8000 images, 6000 are used for training and remaining 2k images are for training purpose. Images in Flickr8k dataset are in jpg format with resolution 256*500 to 500*500 and average length of sentence is 12 words.

V. RESULT AND ANALYSIS

Performance of the image captions are evaluated by using different evaluation BLEU, METEOR, ROUGE, CIDEr and SPICE metrics. When analyzing the proposed model and matching the predicted words to their original captions, the BLEU score is applied. Fig. 4 illustrates how the loss gradually decreased as the number of training epochs grew. it could train our datasets across more epochs to get better descriptions, and here it did so for 100 epochs to enable comparison study. The loss value is between 0.5 and 0.1 epochs. Maximum and minimum values are observed for 10 epochs with losses of 0.5+ and less than 0.1 epoch, respectively. In Fig. 5, the comparison of the predicted caption with five additional original captions using a graphic representation of the BLEU score is illustrated. From 5 to 10 epochs, a sharp increase is observed from 0.50 to 0.56 BLEU score, then the graph experiences slight ups and downs till 50 epochs. Another score called "match words" counts the words that match up with the produced text of a picture. As shown in the graphical representation, the match words undergo significant upswell with changes as time passes. Witnessed as 0.49 match words in the case of 50 epochs and 0.40 in the case of 5 epochs. When Match Word and BLEU Score were compared, it was found that both inclined before reaching the heights. In the instance of Match words, the score increased from 0.500 to 0.555 from 5 to 10 epochs. After that, this sample saw minor changes through 50 epoch, reaching a score of 0.575. When discussing the BLEU score, it had two distinct peaks at 0.450 and 0.470 score at the 15 and 30 epochs. At 35, the graph had a slight decline (0.460), and at 50, it finally hit the score (0.480).









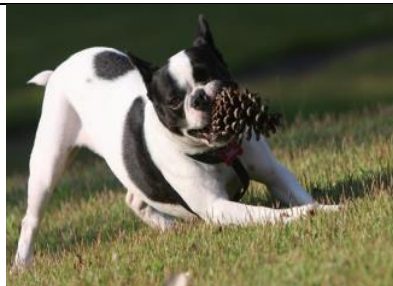






		
“a brown puppy is walking in snow” BLEU Score: 75	“A man flying with skateboard” BLEU Score: 72	“A girl is running on beach” BLEU Score: 73
		
“a player in white uniform is running with ball”, BLEU Score: 73	“a white dog runs around in grass”, BLEU Score: 75	“a man in black dress rides bike on hill”, BLEU Score: 69
		
“a puppy is hopping in a grassy area”, BLEU Score: 70	“three person standing under umbrella”, BLEU Score: 72	“a spotted dog is running with a ball”, BLEU Score: 73
		
“a black dog playing with a ball”, BLEU Score: 75	“a person is climbing a snowy mountain”, BLEU Score: 74	“two old woman in red dress smile”, BLEU Score: 74
		
“a woman is smiling and swinging”, BLEU Score: 72	“a small girl in pink is sitting with a dog”, BLEU Score: 74	“a black dog jumping over a log”, BLEU Score: 76

Fig. 5. Image captions generated by proposed approach.

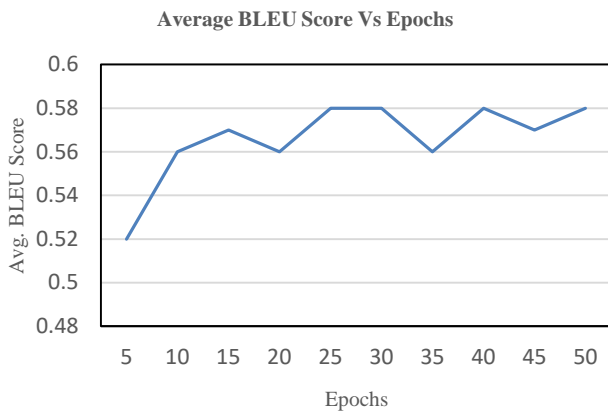


Fig. 6. Average BLEU Score vs Epochs.

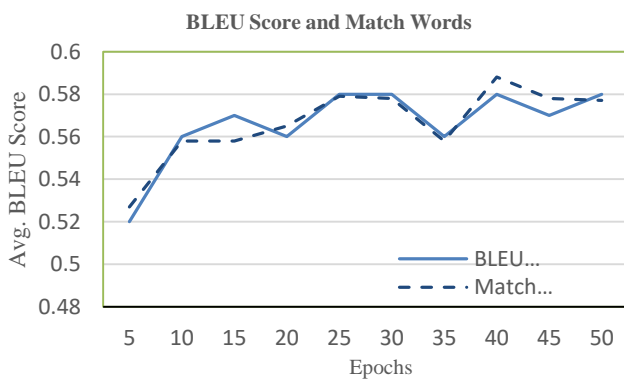


Fig. 7. BLEU Score vs Match words.

The graphical representation illustrates the model's recall changes with threshold values. Threshold values from 0.0 to 0.25 remained constant at 1. After then, a steady fall was observed from 0.25 to 0.75 and approached 0.0 value until a very little increase with around 0.1 recall value was noted too and final recalled value is accounted as 64.056. The graph that depicts the variation in accuracy with threshold values changes the shape of a sharp peak that is constant at 0.500 accuracy up until 0.0 to 0.25v threshold value, then a straight climb up to 0.675 accuracy, followed by a similar value fall up until 0.75 threshold value (Fig. 6). And resultant accuracy is 67.052. The graph shows model precision levels as well as variations in threshold settings. Although the precision value overall is 68.138, changes are seen from a 0.2 threshold value to a 0.75 with a simple increase in the precision values. Other starting and ending values were 1.0 from .075 to 0.25 and 0.5 from 0.0 to 0.25. Further in Fig. 7, BLEU score and Match score are compared which shows the compatible score. First average score of both are .52 on 5 Epochs. At 10 Epochs the values are increased to 0.56. it shows its best performance in 30 Epochs and decreases in 35 Epochs due the overfitting. In Fig. 8, 9 and 10 precision recall and accuracy are shown.

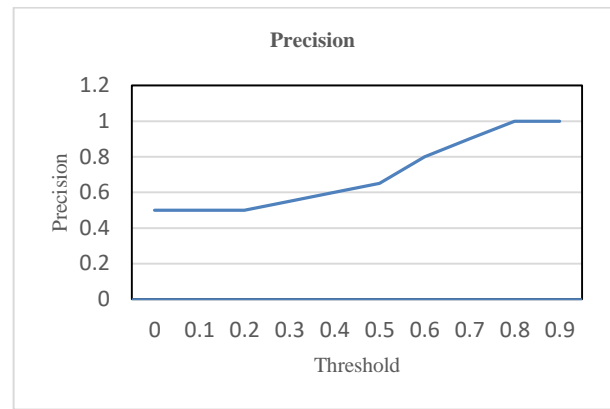


Fig. 8. Precision of proposed systems.

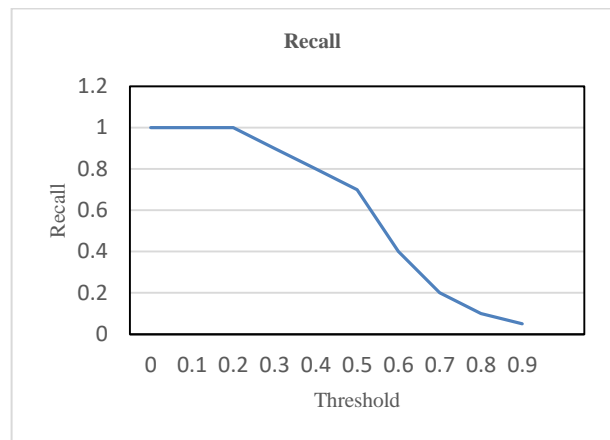


Fig. 9. Recall of proposed systems.

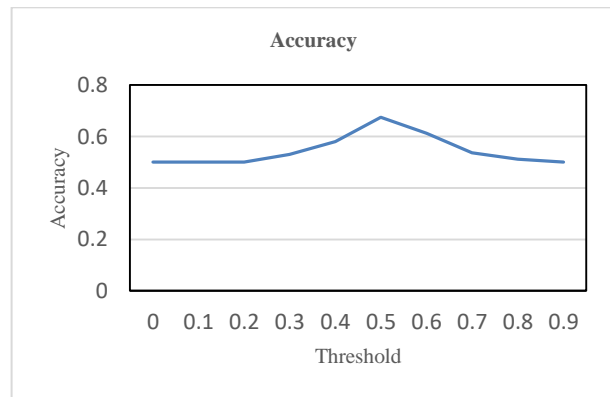


Fig. 10. Accuracy of proposed systems.

The represented graph illustrates the loss and the epochs. According to the provided scale, maximum values are attained by 1.0 on 0.0 epochs. The loss reached a value of 0.75 at 1.0 epochs. Moving further with a curved change value of loss and epochs graph, the loss stopped at 17.5 epochs when the value of loss was witnessed as 0.3.

TABLE I. COMPARATIVE ANALYSIS OF PROPOSED APPROACH WITH SINGLE MODEL

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Inception V3	0.65	0.43	0.29	0.17	0.21	0.41
VGG16	0.66	0.38	0.30	0.16	0.23	0.22
Res Net50	0.56	0.31	0.18	0.12	0.27	0.51
VGG19	0.61	0.35	0.28	0.18	0.21	0.22
Proposed Hybrid Approach	0.67	0.46	0.35	0.26	0.31	0.54

TABLE II. COMPARATIVE ANALYSIS OF PROPOSED APPROACH WITH HYBRID MODEL

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Densenet169 + LSTM [34]	63.73	45.00	30.87	21.13	46.41	19.95
Resnet101 + LSTM [35]	62.77	44.11	30.62	21.10	43.54	18.79
VGG-16 + LSTM [36]	60.56	41.98	28.66	19.51	44.82	19.04
Densenet121 + Attention + LSTM[34]	65.00	46.99	32.83	22.56	47.57	20.44
ResNet152 + Attention + LSTM [37]	65.26	47.55	33.72	23.67	47.54	20.94
VGG-16 + Attention + LSTM [36]	63.81	45.77	32.35	22.55	46.72	20.19
Proposed Hybrid Approach	0.67	0.46	0.35	0.26	0.31	0.54

REFERENCES

The given Tables I and II are the results from an LSTM based decoder model using a signal encoder on the flickr8k dataset. There are five encoders (Inception V3, VGG16, Res Net50, VGG19, and Proposed Hybrid Approach) given each represents their own values of BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, and METEOR in the table chart. The maximum value in terms of BLEU-1 data is 0.67 for the proposed Hybrid Approach Encoder. However, in BLEU-2, the minimum value is held by Res Net50. Considering the data in BLEU-3 and BLEU-4, the minimum is seen in the case of ResNet50 as 0.18 and 0.12, whereas the maximum is witnessed in the case of the proposed Hybrid Approach Encoder. In ROUGE-L, data is numbered as 0.21, 0.23, 0.27, 0.21, and 0.31 for Inception V3, VGG16, Res Net50, VGG19, and Proposed Hybrid approach, respectively. On the other hand, 0.22 was the value which was similar to VGG16 and VGG19 in the case of METEOR.

VI. CONCLUSION

In this paper, a hybrid encoder-decoder based model to generate the effective caption of the image by using the Flickr8k dataset. During the encoding phase, the proposed model used transfer learning-based model like VGG16 and ResNet50 and YOLO for extracting the image features. A concatenate function is used to combine the feature and removes the duplicate one. For the decoding, BiGRU and LSTM are used to get the complete caption of the image. Further BLEU value is evaluated of both the captions generated by BiGRU and LSTM. Final caption is considered whose METEOR value is high. The proposed model is also evaluated by METEOR and ROUGE. The proposed model achieved score BLUE-1: 0.67, METEOR: 0.54 and ROUGE: 0.31 on Flickr8k dataset. The experimental results show the better results through BLUE, METEOR and ROUGE when compared to another state-of-art models. The model is also helpful in generating the captions at real time.

- [1] J. Gu, G. Wang, J. Cai, and T. Chen, "An Empirical Study of Language CNN for Image Captioning," Proc. IEEE Int. Conf. Comput. Vis., vol. 2017-October, pp. 1231–1240, 2017, doi: 10.1109/ICCV.2017.138.
- [2] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 5561–5570, 2018, doi: 10.1109/CVPR.2018.00583.
- [3] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." Available: <http://proceedings.mlr.press/v37/xuc15>.
- [4] K. Xu, H. Wang, and P. Tang, "Image Captioning With Deep Lstm Based On Sequential Residual Department of Computer Science and Technology , Tongji University , Shanghai , P . R . China Key Laboratory of Embedded System and Service Computing , Ministry of Education ,," no. July, pp. 361–366, 2017.
- [5] S. Liu, L. Bai, Y. Hu, and H. Wang, "Image Captioning Based on Deep Neural Networks," MATEC Web Conf., vol. 232, pp. 1–7, 2018, doi: 10.1051/mateconf/201823201052.
- [6] R. Subash, R. Jebakumar, Y. Kamdar, and N. Bhatt, "Automatic image captioning using convolution neural networks and LSTM," J. Phys. Conf. Ser., vol. 1362, no. 1, 2019, doi: 10.1088/1742-6596/1362/1/012096.
- [7] C. Wang, H. Yang, and C. Meinel, "Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning," ACM Trans. Multimed. Comput. Commun. Appl., vol. 14, no. 2s, 2018, doi: 10.1145/3115432.
- [8] M. Han, W. Chen, and A. D. Moges, "Fast image captioning using LSTM," Cluster Comput., vol. 22, pp. 6143–6155, May 2019, doi: 10.1007/s10586-018-1885-9.
- [9] H. Dong, J. Zhang, D. McIlwraith, and Y. Guo, "I2T2I: Learning Text To Image Synthesis With Textual Data Augmentation."
- [10] Y. Xian and Y. Tian, "Self-Guiding Multimodal LSTM - When We Do Not Have a Perfect Training Dataset for Image Captioning," IEEE Trans. Image Process., vol. 28, no. 11, pp. 5241–5252, 2019, doi: 10.1109/TIP.2019.2917229.
- [11] K. Xu, H. Wang, and P. Tang, "Image Captioning With Deep Lstm Based On Sequential Residual" Department of Computer Science and Technology , Tongji University , Shanghai , P . R . China Key Laboratory of Embedded System and Service Computing , Ministry of Education ,," no. July, pp. 361–366, 2017.

- [12] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain Images with Multimodal Recurrent Neural Networks," pp. 1–9, 2014. [Online]. Available: <http://arxiv.org/abs/1410.1090>.
- [13] W. Cui et al., "Landslide image captioning method based on semantic gate and bi-temporal LSTM", *ISPRS Int. J. Geo-Information*, vol. 9, no. 4, 2020, doi: 10.3390/ijgi9040194.
- [14] H. Dong, J. Zhang, D. McIlwraith, and Y. Guo, "I2T2I: Learning Text To Image Synthesis With Textual Data Augmentation."
- [15] C. Liu, F. Sun, and C. Wang, "MMT: A multimodal translator for image captioning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10614 LNCS, p. 784, 2017.
- [16] Q. You, H. Jin, Z. Wang, C. F.-P. of the I., and undefined 2016, "Image captioning with semantic attention," *openaccess.thecvf.com* Available: <http://openaccess.thecvf.com/>.
- [17] X. Liu, Q. Xu, and N. Wang, "A survey on deep neural network-based image captioning", *The Visual Computer*, 35(3):445–470, 2019.
- [18] A. Farhadi, M. Hejrati, M. Amin Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images", In *European conference on computer vision*, pages 15–29. Springer, 2010.
- [19] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg, "Baby talk: Understanding and generating simple image descriptions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.
- [20] Y. Yang, C. Lik Teo, H. Daum'e, and Y. Aloimonos, "Corpus-guided sentence generation of natural images", *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, (May 2014):444–454*, 2011.
- [21] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daum'e, "Midge: Generating image descriptions from computer vision detections", *EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, pages 747–756, 2012.
- [22] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg, "Im2Text: Describing images using 1 million captioned photographs", *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, pages 1–9, 2011.
- [23] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk", In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, 2010.
- [24] N. Gupta and A. Singh Jalal, "Integration of textual cues for fine-grained image captioning using deep cnn and lstm", *Neural Computing and Applications*, 32(24):17899–17908, 2020.
- [25] C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora", *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter:2596–2604, 2015.
- [26] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics", *IJCAI International Joint Conference on Artificial Intelligence, 2015-Janua(Ijcai):4188–4192*, 2015.
- [27] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan., "Show and tell: A neural image caption generator", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June:3156–3164*, 2015.
- [28] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique", *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, pages 1–4, 2018.
- [29] A. Ghosh, D. Dutta, and T. Moitra, "A Neural Network Framework to Generate Caption from Images", *Springer Nature Singapore Pte Ltd.*, pages 171–180, 2020.
- [30] J. Donahue, L. Anne Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691, 2017.
- [31] G. Barlas, C. Veinidis, and A. Arampatzis, "What we see in a photograph: content selection for image captioning", *The Visual Computer*, 37(6):1309–1326, 2021.
- [32] R. Mason and E. Charniak, "Nonparametric Method for Data-driven Image Captioning", pages 592–598, 2014.
- [33] X. Dong, C. Long, W. Xu, and C. Xiao, "Dual graph convolutional networks with transformer and curriculum learning for image captioning", *arXiv preprint arXiv:2108.02366*, 2021.
- [34] H., Gao, Z. Liu, L. Van Der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks", In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708. 2017.
- [35] He, Kaiming, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. 2016.
- [36] Simonyan, Karen and Zisserman, Andrew, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.
- [37] He, Kaiming, X. Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition", In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. 2016.