ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

A Promising Hybrid Model for Image Caption Generation Based on Deep Learning

[1]P. Venu Madhav, [2] S. Sindhusri

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract—

The usage of image captioning techniques to automatically summarize an entire picture into a natural language phrase has grown in importance in recent years due to the proliferation of social media platforms. In our digital culture, image captioning is crucial. Picture captioning involves the use of artificial intelligence algorithms to automatically generate a textual description of a picture in a natural language. The core of the picture processing system is computer vision and NLP. A subfield of computer vision, Convolutional Neural Networks (CNNs) is used for object recognition and feature extraction; concurrently, Natural Language Processing (NLP) methods contribute to the generation of the image's textual description. Because it relies on item identification, location, and the semantic linkages between them in a human-understandable language like English, producing appropriate machine-generated picture descriptions is a difficult undertaking. Our goal in writing this work is to create a hybrid image captioning method that uses VGG16, ResNet50, and YOLO as encoders and decoders. The pre-trained feature extraction models, ResNet50 and VGG16, were trained on millions of photos. YOLO is a tool for detecting objects in real time. Prior to merging the results into one file, it uses VGG16, ResNet50, and YOLO to extract picture features. Finally, the picture is described textually using LSTM and BiGRU. In order to assess the proposed model, we use the BLEU, METEOR, and RUGE scores.

Keywords—CNN; RNN; LSTM; YOLO

INTRODUCTION

Within this online environment, we see many real-life pictures on a daily basis, each of which is interpreted by an individual human being based on their own expertise. Although humans have the innate ability to translate scenes from the real world into words, machines have significant challenges in this area and are not nearly as efficient as humans. Machines still need human input and programming for optimal results, thus human-generated captions are still preferred. Computers can now perform picture captioning tasks such as object and attribute recognition, feature extraction, and the generation of syntactic and semantic captions, thanks to advancements in deep learning based approaches [1]. A lot of unanticipated changes have occurred in the world as a result of the revolutionary new ideas brought about by AI's progress in the field of image processing. Since it offers a superior platform for human-computer interaction, the picture captioning approach (Fig. 1) has broader practical applications. Picture captioning is attracting the attention of academics and researchers as a result of its growing use in image processing. Any of these descriptions might work for the image in Figure 2: two canines playing with a toy, two canines retrieving a floating toy from the ocean, or two canines racing across the water with a rope tucked between their jaws. Unlike machines, our highly-developed brains are capable of providing a nearly precise description of every given image. Therefore, the primary goal of picture captioning is to use deep learning to detect objects and their relationships in the image, and then use natural language processing to generate a textual description, and finally, use various performance matrices to assess how well the textual description was generated. In computer vision, popular Convolutional neural networks (CNNs) and recurrent neural networks (DNNs) are used for object recognition and segmentation; in natural language processing, RNNs and LSTMs are used for producing picture descriptions (Fig. 3). Convolutional neural networks (CNNs) are able to comprehend scene or picture objects and answer queries like "what," "where," and "how" about such items.

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

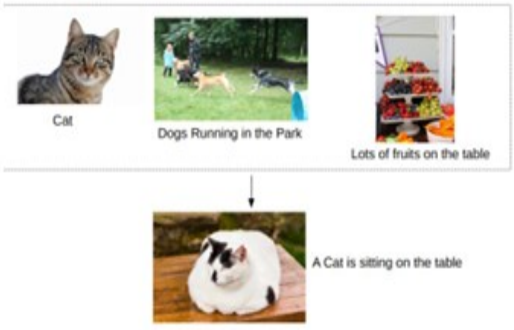

Fig. 1. Image captioning.

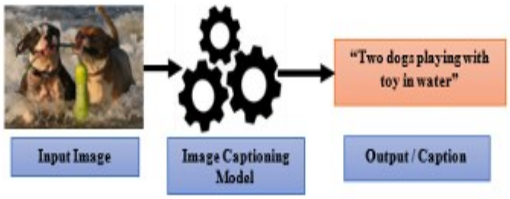

Fig. 2. Working of image captioning.

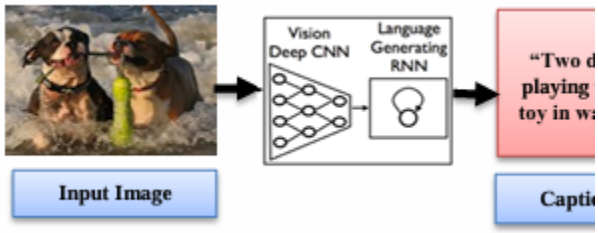

Fig. 3. Image captioning architecture.

"Dogs splashing around in the water with a toy" Caption As an example, in Figure 3, CNN is able to recognize the "dog," "toy," and "water" objects and establish a connection between them. Additionally, RNN utilizes the keywords provided by CNN, which are considered as a set of words, to provide the shape in textual form. The encoder-decoder architecture is another name for this one. Computer vision's object identification subfield makes use of a number of techniques, including YOLO, R-CNN, Mask R-CNN, Mobile Net, and Squeezed, among others, to accurately identify visual elements.

LITERATURE SURVEY

The literature review on picture captioning is provided in this part. In order to create captions that seem human, many state-of-the-art methods and models have been released in the last several years. There are a number of different techniques to image captioning, but they may be generally categorized as either template-based, retrieval-based, or encoder-decoder methods [11, 14, 17]. Using the image's geometric, conceptual, and visual characteristics, the authors of article [31] suggest a content selection method for picture description. Using CNN as its foundation, these models encode the picture, extract features, and then generate captions using RNN or LSTM. By combining picture captioning with probabilistic distributions of successor and predecessor terms, researchers in article [1] created a model for image captioning. A well-known method in picture captioning is the attention and visual oriented approach. In [2, 3], the attention mechanism is used to create the captions. Many of the publications relied on preexisting models. Some examples of these models are the well-known encoder or CNN model Unet [13], the Inception V3 [9-10] and VGG16 papers [1], [3-7], AlexNet [5], [7], ResNet [4-5], [12], and Alex Net [7]. In order to generate or decode picture captions, RNN [16], BiLSTM [7], and LSTM [8-10] and [15] are all viable options. Captions for images may also be created in a number of languages, including but not limited to German, Punjabi, Chinese, Japanese, and Hindi. The input picture is identified using a template-based technique, which makes use of preset objects, actions, and attributes. The authors estimate the image's caption using visual components such as object, action, and setting [18]. Using a Conditional Random Field (CRF) based approach; the author of [19] extracts picture characteristics. The BLUE and ROUGE scores on the PASCAL dataset were used to test the proposed model. It can't create captions for images of varying durations since it relies on pre-defined templates. Capering the image's characteristics with datasets is how retrieval-based approaches create captions. Input images are searched for captions using comparable attributes found in the dataset. In [22], the authors provide a model for query picture feature extraction by dataset searching; in [32], they suggest a technique for caption extraction using density estimation. The authors of [25] generated picture captions using semantic and visual criteria. With five descriptions per picture in the original dataset, we want to train a specific model using this data. Several pre-built image classification models use cutting-edge algorithms to effectively categorize thousands of unique objects and photos after the model has grown proficient in extracting image characteristics during the training phase. Similar to ResNet, these models provide more accurate results when it comes to picture rate categorization. These are a breeze to put into action. Machine translation and deep neural network–based picture caption synthesis both make heavy use of encoder-decoder based approaches. Both the NIC (Neural Image Caption) model and a

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

dual graph convolution network based on encoder-decoder architecture are presented in [33] and [27], respectively. Here we have a basic model that uses CNN for encoder tasks and LSTM and RNN for decoder tasks, specifically for generating picture captions.

RESEARCH METHODOLOGY

In this case, Convolutional neural networks (CNNs) with pooling and fully connected layers are used as encoders to extract visual features from the picture. Prior to the rise of transfer learning, Alex Net was the go-to model for compute vision problems. However, these days, there are a plethora of pre-trained CNN based models available, such as VGGNet, Inception V3, DenseNet, ResNet, etc., each with its own unique set of Convolutional neural layers, which drastically reduces the amount of time a model needs to train. The encoder provides the data that the decoder uses to generate the final captions. The three most popular decoders are GRU, LSTM, and RNN. While LSTM is ideal for lengthy word sequences, RNNs work well with shorter ones. In this part, the suggested hybrid research approach is shown. Achieving a greater Meteor value is our primary goal with the offered approach. Utilizing the principle of transfer learning, our model is built around an Encoder-Decoder methodology. In this first step, we use VGG16, ResNet50, and YOLO (You Only Look Once) individually to extract picture characteristics. While VGG16 (Visual Geometry Group) is a method for object recognition and classification that was pretrained on the Image Net dataset, YOLO (Redmon et al., 2015) is an effective real-time object detection system. The architecture in question is a deep CNN, specifically one that makes use of sixteen Convolutional layers. One deep Convolutional neural network (CNN) that can categorize over a thousand different types of objects is ResNet50. It has fifty Convolutional layers. Phase two involves merging picture characteristics retrieved by VGG16, ResNet50, and YOLO, and removing any instances of duplicate words. Phase three involves creating captions using the BiGRU and LSTM. Natural language processing makes use of a Neural Network design known as BiGRU. For both forward and backward input, this design makes use of two GRUs. One kind of recurrent neural network design that may recognize object relationships is LSTM, or Long Short Term Memory. This network uses feedback connections to do this. Finally, the Meteor performance assessment metrics are used to compare the two captions. With the final caption, the meteor value is greater.

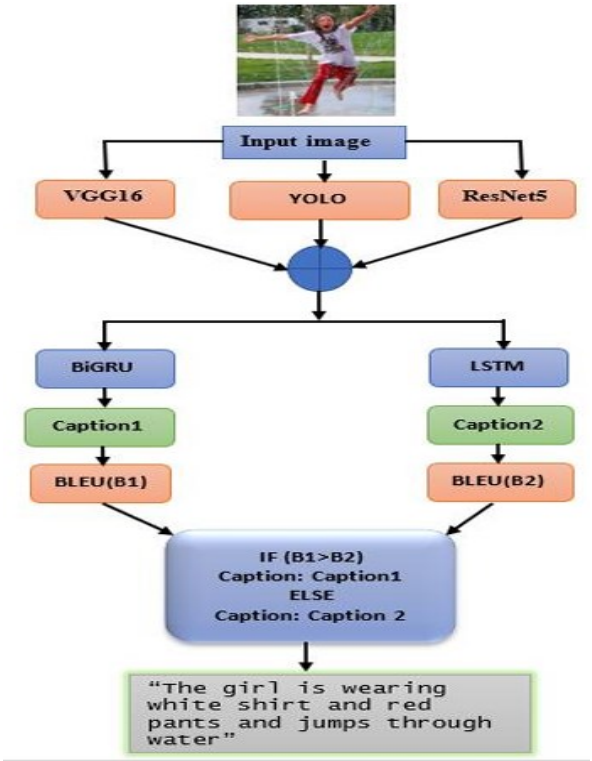

Fig. 4. Proposed image captioning architecture.

DATASETS

Any AI-based system relies on data. In recent times, image captioning has been fortunate to have access to extensive datasets such as MSCOCO, Flickr8k, Flickr30k, PASCAL, etc., where each picture is accompanied with five associated reference phrases. Various methods and grammars are used to characterize each scene. Microsoft's MSCOCO is a massive dataset with the intention of humanizing images. Prior to creating an appropriate caption, it comprehends the scenario and finishes picture identification, segmentation, and generation. There are a total of 82,783 photos in it, including 40,504 from the validation set and 40,775 from the test set. The Flickr30k dataset consists of a total of 30,000 images: 28,000 for training, 1,000 for testing, and 1,000 for validation objectives. This research uses Flickr8k as a benchmark dataset for model training. It has eight thousand pictures with five captions each, describing the quiet items in great detail. All of the

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

photographs include English subtitles that were added by hand. There are two groups within the dataset. The picture directory is the first, and it contains 8,000 photos with 5 captions. Six thousand photographs are used for training, while the remaining two thousand images are also utilized for training. The Flickr8k dataset contains jpg format images with resolutions ranging from 256*500 to 500*500. The average phrase length is 12 words. Section V: Discoveries and Evaluations Several assessment criteria, including BLEU, METEOR, ROUGE, CIDEr, and SPICE, are used to assess the performance of the picture captions. The suggested model is evaluated using the BLEU score, which is used to compare the predicted words with their original captions. The loss reduced significantly as the number of training epochs expanded, as seen in Fig. 4. It trained our datasets over a longer period of time, 100 epochs to be exact, allowing us to draw more accurate conclusions from our comparisons. It takes somewhere about half an epoch to a half a millisecond to compute the loss. For 10 epochs, with losses of 0.5+ and less than 0.1 epochs, respectively, the maximum and lowest values are noted. Using a visual depiction of the BLEU score, Figure 5 shows a comparison of the predicted caption with five more original captions. A significant spike from 0.50 to 0.56 BLEU score occurs between 5 and 10 epochs, after which the graph shows minor fluctuations up to 50 epochs. An additional metric known as "match words" measures the frequency with which words appear in the generated text of an image. The visual depiction shows that the match words fluctuate significantly over time. Noted as 0.49 match words for 50 epochs and 0.40 for 5 epochs. It was discovered that both Match Word and BLEU Score dipped before they reached their peaks when compared. From the fifth to the tenth epoch, the score for Match words rose from 0.500 to 0.555. Minor variations were seen in this sample during the next 50 epochs, culminating in a score of 0.575. At the 15- and 30-epoch points in time, the BLEU score peaked at 0.450 and 0.470, respectively. The graph showed a little drop at 35 (0.460), and then it reached the score at 50 (0.480).

Fig. 5. Image captions generated by proposed approach.

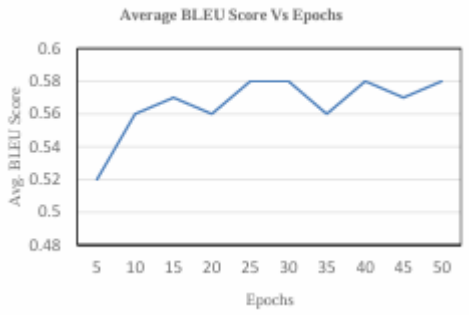

Fig. 6. Average BLEU Score vs. Epochs.

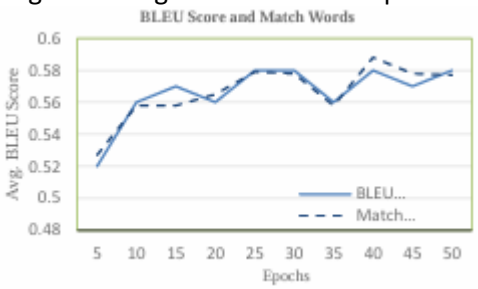

Fig. 7. BLEU Score vs. Match words.

Changes in the model's recall as a function of threshold values are shown graphically. The threshold values stayed at 1 across the range of 0.0 to

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

0.25. Following this, there was a gradual decline from 0.25 to 0.75, which got close to the 0.0 value; nevertheless, there was also a little gain of around 0.1 recall value, and the final remembered number was 64.056. Figure 6 shows the graph of the variation of accuracy with threshold values; it takes the form of a high peak at 0.500 accuracy and remains constant up to the 0.0 to 0.25v threshold value, after which it climbs straight up to 0.675 accuracy, and then it falls to 0.75 accuracy, following a similar pattern. The resulting precision is 67.052. Variations in threshold settings and model accuracy levels are shown on the graph. Despite a total precision value of 68.138, a threshold value of 0.75 is achieved by simply increasing the precision values, as opposed to the previous 0.2. In addition to 0.0–0.25, other possible beginning and ending numbers were 1.0–0.25 and 0.5–0.0–0.25. The suitable score is further shown in Fig. 7 by comparing the BLEU score with the Match score. Both had an initial average score of.52 throughout five epochs. It reaches its peak performance after 30 epochs and then declines after 35 epochs as a result of over fitting; the values are raised to 0.56 after 10 epochs. Accuracy and precise recall are shown in Figures 8, 9, and 10.

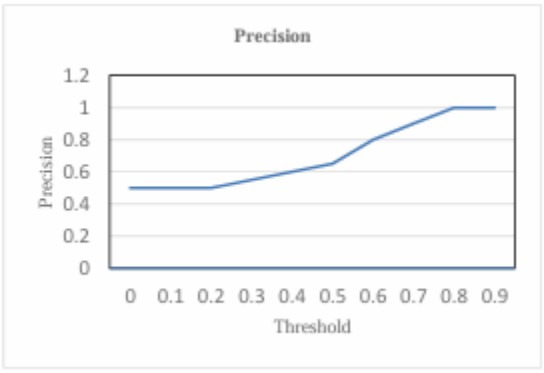

Fig. 8. Precision of proposed systems.

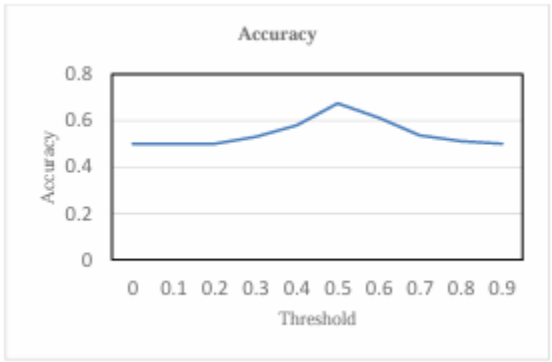

Fig. 10. Accuracy of proposed systems.

You can see the loss and the epochs on the graph. The given scale indicates that on 0.0 epochs, maximum values are achieved by 1.0. At 1.0 epoch, the loss had reached 0.75. Continuing with a graph that shows the loss as a function of time, we see that the loss came to a halt at 17.5 epochs, when the loss value was 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

The provided the output of a long short-term memory (LSTM) based decoder model trained on the flickr8k dataset using a signal encoder is shown in Tables I and II. Inception V3, Res Net50, VGG19, and the proposed hybrid approach are the five encoders shown in the table chart. Each encoder represents a different value of BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, and METEOR. The suggested Hybrid Approach Encoder achieves a maximum value of 0.67 while considering BLEU-1 data. In BLEU-2, however, Res Net50 holds the minimal value. The suggested Hybrid Approach Encoder exhibits the highest value, while ResNet50 sends the lowest values, 0.18 and 0.12, when looking at the data in BLEU-3 and BLEU-4. Information for Inception V3, VGG16, Res Net50, VGG19, and the Proposed Hybrid method is numbered 0.21, 0.23, 0.27, 0.21, and 0.31 in ROUGE-L, in that order. Conversely, when it came to METEOR, the value that was comparable to VGG16 and VGG19 was 0.22.

CONCLUSION

This work presents a technique that uses the Flickr8k dataset to create excellent picture captions using hybrid encoder-decoder architecture. The suggested approach extracted picture features during the encoding phase using a transfer learning-based model, such as VGG16, ResNet5o, and YOLO. To merge the features and get rid of the duplication, a concatenate function is used. The whole picture caption is obtained during decoding using BiGRu and LSTM. Both the BiGRU and LSTM captions are further tested for BLEU value. When the METEOR value is high, the final caption is evaluated. In addition, METEOR and ROUGE assess the suggested model. On the Flickr8k dataset, the suggested model obtained BLUE-1: 0.67, METEOR: 0.54, and ROUGE: 0.31. Comparing BLUE, METEOR, and ROUGE against other state-of-the-art models, the experimental findings reveal that they provide superior outcomes. Additionally, the model is useful for real-time caption generation.

REFERENCES

[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/matecconf/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.