Image Captioning using Ensemble Model

[*]Pranav Bathija[1], Harsh Chawla[1], Ashish Bhat[1], Arti Deshpande[1]
[1]Thadomal Shahani Engineering College
Mumbai, India
[*]pranavb30@gmail.com
harsh.c2699@gmail.com
munnobhat@gmail.com
arti.deshpande@thadomal.org

Abstract: Generation of Image Captions is the task of producing a description of an image using natural language. Even with the advancement in technology, the mundane task of image captioning still proves to be strenuous for a machine. The paper proposes an ensemble model of LSTM and Transformer encodings. The model works on an encoder - decoder architecture. The image is passed as input to the encoder which extracts features from the image which act as input for the decoder to output a caption from the received information. The paper evaluates the results of the implementation of the ensemble model as well as the existing model. The paper also evaluates the results on different standards of images. The ensemble model provides better results than the standalone existing models. The results were evaluated on Bleu score. The proposed model achieved a bleu score of 0.6192

Keywords: Deep Learning, Image Captioning, Long-Short Term Memory, Transformer, Natural Language Processing

1. Introduction

There is a quote “A picture is worth a thousand words”, a miniscule image contains a plethora of information. Humans are capable of processing such large amounts of information and comprehend it in an instant. Communication is an integral part of human existence, humans use written or spoken languages as a medium of communication. Humans similarly use natural languages to describe an image. This also leads to multiple possible captions for the same image. This task proves to be much more complex for machines, however, if done successfully it can prove beneficial for a variety of applications. The machine needs to have basic knowledge of natural languages, object detection and correlation to produce a caption for an image. Few of the previous models work on precise syntax and hard coded features [7] to provide a caption for an image. As a result, the comprehensive nature of the caption generated is bounded by the previous approaches. Hence, the training data should be used as guidelines to build models that are free of sentence syntax or pre-defined features [7] to overcome this limitation.

Irrespective of the limitations, the task of generating image captions has displayed a range of possible use cases such as visual aid products for improving the life of visually impaired people. Augmented reality products such as Google lens can use it for semantic image search. It has also proved useful in bringing visual

intelligence to chatbots and image tagging for social media, which with the recent surge in e-commerce makes it quite advantageous.

With the latest evolution in deep learning techniques, the feasibility to achieve a descriptive caption for an image. The proposed model consists of mainly two components - encoder and decoder [8]. The encoder extracts important features from the image that act as the input to the decoder which uses its knowledge gained during training to identify the objects individually and also find the relation shown between them in the image. The decoder is responsible to generate a caption from the word vector formed at the decoder using natural language processing.

This paper proposes a new model, which is based on an ensemble of two previous approaches, LSTM and Transformer. The paper includes a study on the existing models that generate a caption for an unknown image as well as results of execution of the suggested ensemble model. The paper also aims to compare the proposed model and the existing models on images with clear and distinct objects, versus images with clusters and vague features. The F1 score of the proposed ensemble model show a slight improvement from the results of the existing models solely based on either LSTM or transformer for the generation of the image caption.

The flow of the remaining paper is as follows. In Section 2, the paper discusses the recent research work conducted in this field. In Section 3, the dataset and the pre-processing techniques used are described. Section 4 discusses the design of the architecture of the proposed ensemble model. Section 5 evaluates the results achieved using the proposed model in comparison to the existing models. The paper concludes with direction for future work.

2. Related Work

Generation of Image Captions can be performed in mainly three different ways: template-based methods, retrieval methods, and generative methods. In template-based methods [1-4] objects, actions, scenes, attributes are prioritised to be detected. Im2Txt model [5] is one of the finest implementations of the retrieval-based method. However, the need to implement newer and better models arise out of the fact that retrieval-based methods only refer to their database images in order to rank content. This content which is ranked is then used to produce the captions. Hence these methods are incapable of generating novel captions. The generative method uses either a pipeline or an end-to-end model [6] which makes collective use of the language modelling and image recognition models. They are usually a combination of CNNs and RNNs.

In Shah et. al. [7] a Show & Tell model is used for this implementation which combines Inception-v3 model and LSTM cells [8] where the former carries out object recognition while the latter carries out language modelling [9-10]. BLEU (Bilingual Evaluation Understudy) score is used to evaluate the generated sentences which works by comparing the generated sentences to human generated sentences [11]. This implementation [7] gives an average BLEU score of 65.5. The paper [12]

uses a neural framework where Visual Group Geometry (VGG) network is used along with an LSTM network [13]. The VGG extracts the features from the images which then act as input to the LSTM network in the process of caption generation. Input to the LSTM network is an encoded fixed length vector of images and sentences. BLEU scores for the Flickr8k, Flickr30k, and MSCOCO datasets are found to be 0.53, 0.61, 0.67 respectively.

The author [14] does an analysis of the various image captioning types and the respective approaches taken in each type and provides a summary. The three types of image captioning systems mentioned are: General Image Captioning, Image Captioning with emotion, and Cross-Lingual & Multilingual Image Captioning. General image captioning involves two processes: detecting the objects, their attributes, their relationships with other objects [15-16] and modelling the appropriate language [17] which is semantically sound and best describes the image [18]. Whereas, captioning images in style by adding emotional content can be done by including the viewer's emotions and feelings towards the image [19-22] or by extracting emotions from the image itself [23]. Cross lingual captioning can be done in several ways including direct translation into the desired language, or training the dataset in the desired language. Apart from these three methods [14] also discusses various datasets like UIUC PASCAL [24], Flickr 30k [25], and Microsoft COCO Captions [26] alongside various evaluation metrics including BLEU [27], METEOR [28], ROGUE [29], CIDEr [30], and SPICE [31].

3. Data set & Pre-processing

For the task of image captioning the data required is an image to be captioned as the input and a set of captions as the output are required during the training. Several datasets exist which fulfil these requirements.

Fig 1: Example image from Flickr30k

TABLE 1. Representation of captions

i	Image feature vector	Partial caption	Target Caption
1	Image_feature_vector	[6]	5
2	Image_feature_vector	[6, 5]	4
3	Image_feature_vector	[6, 5, 4]	3
4	Image_feature_vector	[6, 5, 4, 3]	8
5	Image_feature_vector	[6, 5, 4, 3, 8]	1
6	Image_feature_vector	[6, 5, 4, 3, 8, 1]	9
7	Image_feature_vector	[6, 5, 4, 3, 8, 1, 9]	3

The most extensively used datasets are the Flickr30k and Flickr8k. Each dataset contains a lot of images with each accompanied by 5 captions on which are used to train the model. The captions for these images are obtained via crowdsourcing. Fig 1 shows an image accompanied with its 5 captions from the dataset. The captions for the picture shown in the figure are as follows:

- A man wearing a red jacket is climbing a mountain.
- A person in red jacket standing on a snowy mountain
- A person standing on a snowy mountain
- A man with a red backpack climbs a tall mountain
- A man on a mountain

Before training can begin some pre-processing steps need to be performed. The captions need to be pre-processed so that they can be passed to the model. This is done by making all the captions lower case, removing all punctuations, special characters and converting numbers to their alphabetical representation. This is followed by tokenization of the captions. Since the model predicts the caption iteratively the captions need to be converted in a format as shown in Table 1.

4. Proposed Methodology

The model presented in [8] forms the basis for the proposed model. The proposed model is an attempt to see how an ensemble model of transformer encoder model [32] and Bidirectional LSTM model [7], [33] will perform in comparison to models using just LSTMs or transformer encodings. In Fig 2 and Fig 3 the architecture of the Transformer encodings model and BiLSTM is shown. The proposed model is an ensemble model of the Transformer encodings model and BiLSTM in which the image acts an input and the output is a caption describing the image.

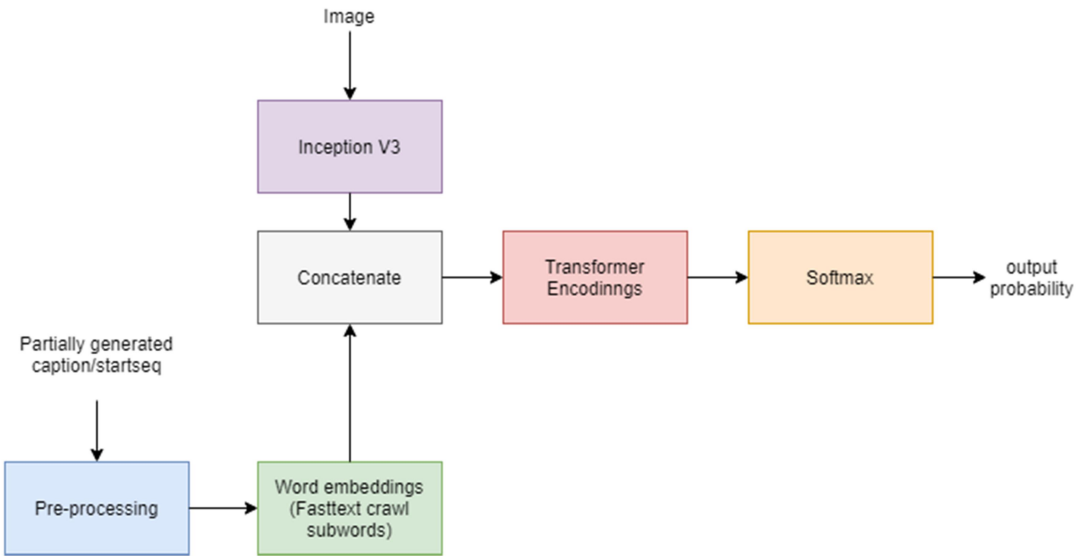


Fig 2: Transformer Encodings model Architecture

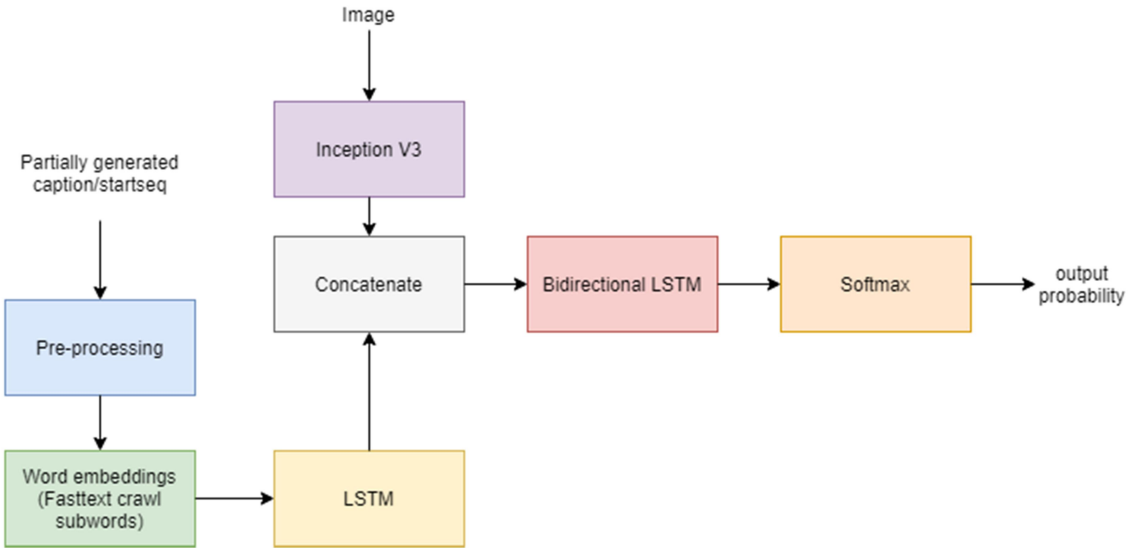


Fig 3: BiLSTM model Architecture

Before passing the partially generated caption or the startseq to the decoder it is first passed through word embeddings. For the proposed model, FastText subwords crawl pre train embeddings [34-35] have been used for this purpose. The encoder, decoder and the process of caption generation have been explained below.

4.1 Image encoder

The image encoder takes the image for which the caption is to be generated as the input and extracts the features relevant to predict the caption. For the proposed model Inception-v3 [36] has been used for extracting features from the images. CNNs have been used extensively for the process of feature extraction. Inception-v3 is a pre-trained image classification model which contains multiple CNNs and is 48 layers deep. It has been trained on the ImageNet dataset which contains more than a million images. The pre-trained network can classify images into 1000 categories but for the purpose of the proposed model this is not required which is why the last layer of the inception model is replaced with a single fully connected layer.

4.2 Image decoder

The decoder portion of the model consists of the transformer encoders and a bidirectional LSTM. The model takes as input the output from word embeddings, this is then passed to the transformer encoder. The transformer encoder model and the bidirectional LSTM model both calculate the individual probabilities of the next word of the caption. The two individuals probabilities and then combined by averaging these probabilities to output the next word of the caption.

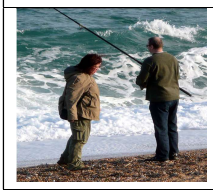
4.3 Generation of caption

For the generation of the caption the image along with the start token is passed to the model. This predicts the first word. In the next step the features from the image along with the start token and first word are passed to the encoder to generate the second word. The same process is repeated till the end token is encountered or the max length has been reached. A max length of 25 has been used as most captions in the training data have a length less than this number. The sentence produced on completion of this process is the caption generated for the given image.

5. Results & Evaluation

As shown in the table above the proposed methodology ensemble model shows more promising results for the first three images as compared to the existing models based either on LSTM or Transformer. All models were trained on the flickr30k dataset for 15 epochs with a learning rate of 1e-4. The models also have a dropout of 0.5 after the concatenate layer. Apart from that FastText crawl subword pre-trained word embeddings have been used to train the models. The LSTM model achieved a bleu score [11] of 0.5913 and a loss of 0.3149 on a test set consisting of 500 unseen images. The Transformer encodings model achieved a bleu score of 0.5747 and a loss

TABLE 2. Results on three clear images and three images with vague features

Image	Transformer + Lstm model Bleu score=0.6192	Transformer model Bleu score=0.5747	Lstm model Bleu score=0.5913
	Caption: a man in a blue shirt is standing on a beach Bleu score=0.909	Caption: a man in a blue shirt is standing in front of a lake Bleu score=0.593	Caption: a man in a blue shirt and black pants is standing on a beach Bleu score=0.731
	Caption: a young boy is sliding down a slide Bleu score=0.875	Caption: a young boy in a blue shirt is playing with a toy Bleu score=0.666	Caption: a little girl in a pink shirt is playing with a toy Bleu score=0.75
	Caption: two men playing soccer in a field Bleu score=1.0	Caption: a young boy in a blue shirt is playing with a soccer ball Bleu score=0.384	Caption: two soccer players are playing soccer Bleu score=0.833
	Caption: a man is sitting on a bench in front of a large amount of people Bleu score=0.266	Caption: a man in a black jacket is walking through a crowded area Bleu score=0.416	Caption: a man is standing in a crowded area with a large amount of people in the background Bleu score=0.352
	Caption: a group of people are walking down a street Bleu score=0.555	Caption: a group of people are walking down a street Bleu score=0.555	Caption: a man is standing on a city street in front of a large white building Bleu score=0.4
	Caption: a group of people are standing in front of a large building Bleu score=0.416	Caption: a group of people are walking down a street Bleu score=0.555	Caption: a group of men in orange uniforms are standing in front of a building Bleu score=0.285

of 0.321 on a test set consisting of 500 unseen images. The ensemble model of LSTM and Transformer encodings achieved a bleu score of 0.6192 shows a slight improvement over the existing models

The first three images in Table 2 are clearly well defined and have unambiguous objects or features, the flickr30k dataset used to train all the models have similar distinct images with appropriate caption. Hence, the captions generated by the proposed model for the first three images scored better than the existing models. Whereas, for the remaining three images in Table 2, the features of the images are vague, in the background as clusters and not clear in the foreground or not well lit overall. Therefore, the scores of the proposed model or existing models are sub-par.

6. Conclusion

From our findings we can conclude that the suggested ensemble model is superior at generating captions for unknown images than the existing models. The images can be categorised into the follow:

- Images with human activities
- Images of animals
- Images of scenery

To create a well-rounded model which produces good captions for each category we need sufficient data for each category. The paper also concludes that training the proposed model with a large dataset will teach it to visualise the image better, for instance if the training dataset included images having large clusters or more focus on the background instead of the foreground or dim lighting then the proposed model would accordingly give better results for such images, and communicate the information acquired by it to the user. The paper also shows that with the deep learning techniques available currently there is still scope for betterment, however, the current limitations of generating high scoring captions of distinct images and not clustered or slightly vague images needs to be overcomed.

7. References

1. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J. and Forsyth, D.: Every picture tells a story: Generating sentences from images. In European conference on computer vision (pp. 15-29). Springer, Berlin, Heidelberg.
2. Yang, Y., Teo, C., Daumé III, H. and Aloimonos, Y.: Corpus-guided sentence generation of natural images. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 444-454).
3. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C. and Berg, T.L., 2013. Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12), pp.2891-2903.
4. Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A., Berg, T. and Daumé III, H., 2012, April. Midge: Generating image descriptions from computer vision detections. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 747-756).

5. Ordonez, V., Kulkarni, G., and Berg, T. L., “Im2text: Describing images using 1 million captioned photographs,” in Advances in Neural Information Processing Systems, pp. 1143–1151, 2011
6. Karpathy, A. and Fei-Fei, L., “Deep visual-semantic alignments for generating image descriptions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137, 2015.
7. Shah, P., Bakrola, V., Pati, S., “Image Captioning using Deep Neural Architecture”, 2017 International Conference on Innovations in information Embedded and Communication Systems (ICIIECS)
8. Vinyals, O., Toshev, A., Bengio, S., and Erhan, D., “Show and tell: Lessons learned from the 2015 mscoco image captioning challenge,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PP, no. 99, pp. 1–1, 2016.
9. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., “Going deeper with convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, 2015. 28.
10. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., “Rethinking the inception architecture for computer vision,” arXiv preprint arXiv:1512.00567, 2015. 37.
11. Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311– 318, Association for Computational Linguistics, 2002.
12. Amritkar, C., Jabade, V. “Image Caption Generation using Deep Learning Technique”, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)
13. Simonyan, K., Zisserman,A.: ”Very deep convolutional networks for large-scale image recognition.” arXiv preprint arXiv:1409.1556 (2014).
14. Hrga, I., Ivašić-Kos, M.: “Deep Image Captioning: An Overview”, MIPRO 2019, May 20-24, 2019, Opatija Croatia
15. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C. and Lawrence Zitnick, C.: “From captions to visual concepts and back,” in CVPR, 2015, pp. 1473–1482.
16. Ivašić-Kos, M., Pobar, M., Ribarić,S.: “Automatic image annotation refinement using fuzzy inference algorithms”, IFSA- EUSFLAT 2015, Gijón, Asturias, (Spain) p. 242
17. Vinyals, O., Toshev, A., Bengio, S., Erhan,D.: “Show and tell: A neural image caption generator,” in CVPR, 2015, pp. 3156–3164
18. Hodosh, M., Young, P., Hockenmaier, J.: “Framing image description as a ranking task: Data, models and evaluation metrics,” JAIR, vol. 47, pp. 853–899, 2013.
19. Mathews, A., Xie, L., He, X.: “Senticap: Generating image descriptions with sentiments,” in 13th AAAI Conference on Artificial Intelligence, 2016.
20. Gan, C., Gan, Z., He, X., Gao, J., Deng, L.: “Stylenet: Generating attractive visual captions with styles,” in CVPR, 2017, pp. 3137– 3146.
21. You, Q., Jin, H., and Luo, J., “Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions,” arXiv preprint arXiv:1801.10121, 2018.

22. Chen, T., Zhang, Z., You, Q., Fang, C., Wang, Z., Jin, H. and Luo, J., "Factual or Emotional: Stylized Image Captioning with Adaptive Learning and Attention," in ECCV, 2018, pp. 519– 535.
23. Nezami, O. M., Dras, M., Anderson, P., and Hamey, L., "Face-Cap: Image Captioning Using Facial Expression Analysis," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2018, pp. 226–240.
24. Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J., "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, 2010, pp. 139–147.
25. Young, P., Lai, A., Hodosh, M. and Hockenmaier, J.: "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Transactions of the Association for Computational Linguistics, vol. 2, pp. 67–78, 2014.
26. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P. and Zitnick, C.L..: "Microsoft COCO captions: Data collection and evaluation server," arXiv preprint arXiv:1504.00325, 2015.
27. Papineni, K., Roukos, S., Ward, T. and Zhu, W. J.: "BLEU: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting on association for computational linguistics, 2002, pp. 311–318.
28. Denkowski, M. and Lavie, A.:"Meteor universal: Language-specific translation evaluation for any target language," in 9th Workshop on Statistical Machine Translation, 2014, pp. 376–380.
29. Lin, C. Y.: "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out: Proceedings of the ACL-04 workshop, 2004, vol. 8.
30. Vedantam, R., Zitnick, C., and Parikh, D.: "Cider: Consensus based image description evaluation," in CVPR, 2015, pp. 4566– 4575.
31. Anderson, P., Fernando, B., Johnson, M. and Gould, S.: "Spice: Semantic propositional image caption evaluation," in ECCV 2016, 2016, pp. 382–398.
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I: Attention Is All You Need. (2017)
33. Zhou, C., Sun, C., Liu, Z., Lau, F.: A c-lstm neural network for text classification.(2015)
34. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. In: CoRR, abs/1612.03651 (2016).
35. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. In: CoRR, abs/1607.04606 (2016)
36. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: "Rethinking the inception architecture for computer vision," arXiv preprint arXiv:1512.00567, 2015. 37.