Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Image Captioning Generator and Comparison Study

R. Thirumahal[1], Harshitha Prabakaran[2], G. N. Swetha[3], S. Sushmitha[4], S. Swathi[5], Chandhini Balasubramaniam[6]

[1]Assistant Professor, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, India
[2-6]B.E Computer Science and Engineering PSG College of Technology, Coimbatore, India

E-mail:[1]trk.cse@psgtech.ac.in, [2]harshitha0301@gmail.com, [3]saiswe02@gmail.com, [4]sushiii2k1@gmail.com, [5]swathissara@gmail.com, [6]chandhinibalasubramaniam@gmail.com

Abstract

Caption generation has long been of interest to researchers in the field of artificial intelligence. The ability to train a system to properly represent an image or environment, has broad applications in robotic vision, management, and many other areas. The purpose of this study is to analyze multiple transmission learning strategies and create a unique system for improving caption accuracy. To increase object relevance, image feature vectors are constructed using multiple state-of-the-art models that are input into an encoder/decoder transformation network based on attentional mechanisms. The model is evaluated for comparing datasets such as MS-COCO with criteria such as Bilingual Evaluation Understudy.

Keywords: Image retrieval, caption generator, artificial intelligence, generative adversarial network, image analysis

1. Introduction

Among the human capacities is to describe the circumstances in which they find themselves. When presented with an image, humans can easily tell all about it with a quick glimpse. One of the propelling motivations for artificial intelligence researchers is the development of systems that can perceive and interpret the actual world. In the prior literature, two basic strategies for simulating the phonetic variation have been used. A knowledge-based strategy employs phonological and linguistic information to develop phonological rules that handle speech variances, despite substantial studies in various

DOI: https://doi.org/10.36548/jiip.2022.4.009
Received: 30.11.2022, received in revised form: 03.01.2023, accepted: 18.01.2023, published: 27.01.2023

computer vision problems such as entity recognition, attribute segmentation, action classification, and image classification. Letting a computer mechanically explain a visual with sentient phrases is a relatively recent challenge, as is scene recognition.

Captioning an image requires significant comprehension of the vector data of an image as well as the ability to transfer these conceptions into a human-like language, integrating both the study areas of computer vision and natural language processing. Translating visual conceptions into natural speech complicates the task even further. Because natural languages account for most of the social interaction, regardless of whether it is written or spoken, allowing computers to define the imagery will open up a wide range of potential applications, including natural interpersonal contact, offering new learning, knowledge discovery, as well as visually impaired assistance, among others. The image captioning discipline is gaining worldwide prominence for being both difficult and vital. Image captioning has a wide range of applications, including self-driving automobiles and assistive technology for the blind.

Figure 1. A cat is swimming in the water

The procedure of image captioning is divided into two stages: visual processing and language processing. Computer vision and natural language processing techniques are used to ensure that the produced subtitles are syntactic and semantically correct. Resilient deep neural networks deliver powerful visual and language modeling algorithms. As a result, they are utilized to enhance current systems and build various new techniques. Using deep neural networks to handle image captioning produced cutting-edge results. With current innovations in domain adaptation and image captioning, a unique architecture that analyses numerous transfer learning models using various measures such as Bilingual Evaluation Understudy (BLEU) score and others has been presented in this paper.

2. Related Works

Machine learning progress has opened new avenues for using deep neural networks rather than the custom-engineered features and shallow models that were previously used. In [1], recursive neural networks used dependency-tree to convert phrases and sentences into compositional vectors in order to identify a caption for a particular image. The images were transformed into feature vectors using another deep neural network. Paper [2] proposed a multimodal Convolutional Neural Network (CNN) that evaluated connectivity between images and text at different levels of interaction. A variety of matching CNNs were employed to create the correct association between images and captions. An ensemble of the multimodal CNN decided the final matching score. The encoder decoder framework generated captions for images by considering recent advances in neural machine translation.

A collection of attention-based techniques [3, 4, 5] are another way to caption photos that tried to link words in the expected description with specific places in the image. Because visual attention is frequently diverted to the higher convolutional layers of CNNs, spatial localization is constrained in this situation. Furthermore, to unify the concept of global aspects, two graph convolutional networks were developed that classify the relationship between any two boxes into 11 subtypes. This idea was implemented using semantic connection graphs and spatial relationship graphs. The domains of translation [7], copywriting [8], and language understanding [9] have all witnessed significant improvements since the incorporation of the transformer architecture [7]. A transformer was employed in [10] to carry out the captioning operation.

For captions, [3] introduced an encoder/decoder framework to merge general image text embedding models with multimodal neural language models to generate word-by-word sentence output for specific images, such as language translation. For the aim of encoding, [3] employed Long Short-Term Memory (LSTM) and CNN. Post rank loss, the encoded visual data was enlarged by pairwise minimization into an embedding space encompassing the LSTM hidden states encoding the textual information. Finally, captions were generated word-by-word by decoding visual features based on contextual word feature vectors using a neural structured content language model. Inspired by the human visual attention mechanism, caption generation was guided by the attention mechanism. Subtitle generation relied on hidden states and values generated from various sections of the imagery. These were addressed by the attention mechanism, which was added to the encoder/decoder architecture.

3. Proposed Model

In captioning tasks, it is common to formulate the problem such that, given both a paragraph and a description generated so far as input, the model produces each word of the output text description. This framework can be implemented using either of the two architectures, the injection model and the merge model. The proposed model uses a soft-attention-based encoder/decoder with the added benefit of a beam search algorithm. The input image is encoded into smaller images with more training channels by an encoder with 3 color channels. This is passed to the decoder, whose input will be the encoded image. It works by generating captions for each word. Captions can be generated by keeping the encoded image hidden, with or without applying a linear transformation function. Each predicted word is used to create the next word. It mimics human tendencies by allowing machines to view images in different locations while paying attention and in accordance to the generated captions. The attention network used here computes individual pixel weights and considers the sequence generated so far to process the part of the image that needs to be described next. It uses deterministic, smooth, and differentiable soft attention, hence it is trivial to learn end-to-end using standard backpropagation. Because it uses transfer learning, the model is simple, fast to train and evaluate, and generates captions with care. This optimized model is run on the MS COCO dataset using Google Colaboratory Pro runtime engine.

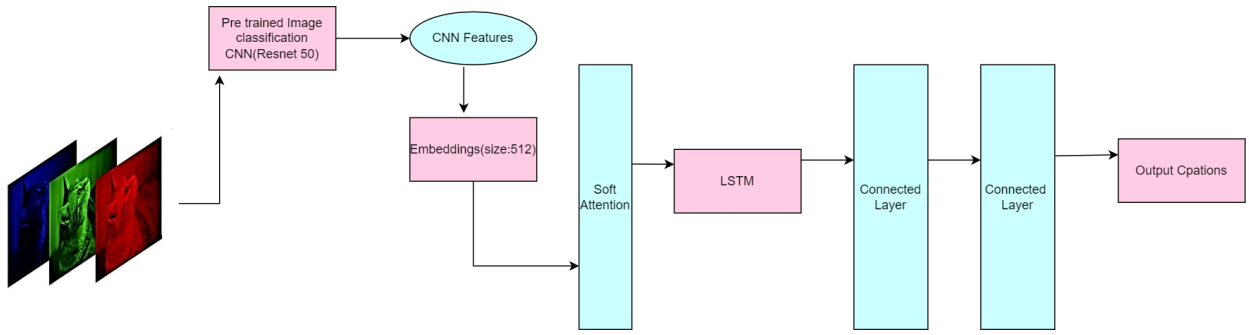

Figure 2. Architecture of the Proposed System

The first component of the system diagram consists of the pretrained Resnet-50 removing the pooling layer and linear layer to employ it specifically for encoding the image. The second component of the system diagram signifies the features in the image obtained as a result of using feature extractor and the third component resizes the encoding so that the encoder can process images of any size. The soft attention layer component consists only of linear layers and a few activation layers. The output of the encoder is obtained and resized to

fixed dimensions in which the already encoded image and the previous hidden state are taken to generate weights. The word generated before and the weighted average are fed to the decoder to generate captions word by word. The CNN module in this architecture is for feature extraction belonging to the encoder phase whereas LSTM here is the decoder used to generate captions.

3.1 Merge Architecture

The architecture used here to improve accuracy is the Merge architecture, which uses an encoded form of the image along with the generated text description. A decoder model then combines the encoded inputs from these two to generate the next word in sequence.

3.2 Encoder

The model takes an image and generates caption z encoded as a sequence of 1-N encoded words.

$$z=\{z_1,.....z_C\}$$

Where, z_i belongs to RN, where N is the size of the vocabulary and $_C$ is the length of the caption. The model makes use of transfer learning by using the 101 layered pre trained residual network and parameters tweaked in the process for fine-tuning to improve performance. As a result, the encoder creates a much smaller and more learned version of the original image. In this case, the encoder produces a tensor of dimensions 2048,14,14.

3.3 Decoder

Here, the encoder's output is obtained and is flattened to N, 14 * 14, 2048 dimensions in order to prevent reshaping the tensors multiple times. The decoder used in the model is LSTM whose hidden and cell states are initialized using the encoded image, which generates one word based on the previous hidden state conditioned on a context vector and previously generated words. The decoder is executed by the training process as mentioned below:

- The decoder input, encoder output, and a hidden state (initialized to 0) are given to the decoder initially.
- Calculation of loss is performed using cross entropy function with the hidden state being passed to the model. Teacher forcing is employed to determine the next decoder input.

In the framework, with the absence of attention, the encoded image could be averaged over all the pixels. Then, it is fed into the decoder as its first hidden state and a caption is generated. Finally, the generation of the succeeding words is carried out with the help of each of the predicted words.

3.4 Attention

Human intuition, which determines the significance of a particular aspect of an image, is the source of the concept of attention. By taking into account the sentence that has been generated to this point and focusing on the area of the image that has to be described next, the model imitates this quality. The weighted sum of hidden states of the encoder can be considered rather than sampling the attention region continually with stochastic attention, which is hard attention.

3.5 Beam Search

The conventional way to generate captions is the straightforward and greedy approach, which might not be the best option since the remaining part of the sentence will still be dependent on the first word. If the initial choice is not optimal, everything that proceeds the initially generated sentence is substandard.

Working of Beam Search model:

- Examine the candidates that are k in number at the first decoding step.
- Generate second words that are k in number for all the k words that arrived first.
- Update additive scores for top combinations of words that are k in number.

4. Results

To evaluate the presented models, comprehensive experiments are conducted taking batch size as 32 and by fine tuning the number of epochs to prevent overfitting of the model. All the results obtained with BLEU on the MS COCO datasets are presented.

4.1 Evaluation Metrics - BLEU

BLEU is a performance measure that is employed to determine the relationship between varying length sections of a projected or produced title and primary texts performed by individuals. Usually, this score is used to calculate the difference between predicted and actual sentences in n-grams model. BLEU-1 is computed through contrasting the forecasted

title to the authentic text in unigram, whereas BLEU-2 is determined by aligning with bigram. By statistically determining BLEU with a maximal order of four, the best correlation with subjective judgments is attained. Fluency is determined by higher n-gram scores in BLEU, while adequacy is determined by higher unigram scores.

4.2 Performance on MS-COCO Dataset

The BLEU-4 metric was used to determine the performance and effectiveness of the proposed model on the MS-COCO dataset. The software used to conduct this experiment is Google Colaboratory Pro having 180 teraflops of computational power resulting in training computational speed to be about 15 minutes for a sample count of 3,28,000 images with batch size 32. Following Andrej Karpathy's split, 5000 then 5000 images were divided for the vital approaches of validation and testing, respectively. The Resnet 50 based model trained on the MS-COCO dataset produced a validation BLEU-4 score of 33.17 and test BLEU-4 score of 33.29.

4.3 Quantitative Results

The method proposed in this paper is compared with the existing approaches and is tabulated below.

Table 1. Results on the MS-COCO Dataset

Method	BLEU-4
m-RNN [11]	25
Soft attention [3]	24.3
Hard attention [3]	25
g-LSTM [12]	26.4
Proposed model	33.29

Based on the comparative analysis of all the models listed above in Table 1, it is found that the proposed model has proved to attain better results with a BLEU-4 score of 33.9, as it incorporates beam search always to find the most optimal sequence of words with the best score at each decode step.

5. Conclusion

This overview summarizes all aspects of creating image captions, discusses framework of the model proposed in recent years for solving the main task that is description task, examines the various algorithmic essences of the attention mechanisms, and finally concludes how the attention mechanism is used here. The present paper summarizes the datasets that are being used commonly these days and illustrates the criteria used in evaluation. So, it's to be noted that the results of the experiment must be improved in the future due to the following three concerns. Firstly, it must be able to generate complete sentences in all possible natural languages like a human does. Secondly, it must be able to generate a grammatically correct sentence without generating incomplete sentences. Lastly, it should be able to write captions that are more reliable and consistent and should match with the context of the image.

Thus, the minute improvements to be done for the model to work more accurately are:

- For the development of corpus description language texts from several languages, an image description system that can handle numerous language families must be created.
- The model needs to be properly trained, tested, and the caption generation should be adjusted in order to increase performance.
- Since the photos are known to be content-rich, the model should be able to produce descriptions that are compatible with numerous principles from a variety of target images rather than sticking to just one.

References

[1] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, και A. Y. Ng, ‘Grounded Compositional Semantics for Finding and Describing Images with Sentences’, Transactions of the Association for Computational Linguistics, τ. 2, σσ. 207–218, 04 2014.

[2] M. Barraco, M. Stefanini, M. Cornia, and S. Cascianell, “CaMEL: Mean Teacher Learning for Image Captioning,” paperswithcode.com, Feb. 21, 2022.

[3] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C] ,International Conference on Machine Learning. 2015: 2048-2057.

[4] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4651–4659, 2016.

[5] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov. Review networks for caption generation. In Advances in Neural Information Processing Systems, pages 2361–2369, 2016.

[6] T. Yao, Y. Pan, Y. Li, and T. Mei. Exploring visual relationship for image captioning. In European Conference on Computer Vision, pages 684–699, 2018.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 1:8, 2019.

[10] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[11] Mao J, Xu W, Yang Y, et al. Deep captioning with multimodal recurrent neural networks (m-rnn)[J]. arXiv preprint arXiv:1412.6632, 2014.

[12] Jia X, Gavves E, Fernando B, et al. Guiding the long-short term memory model for image caption generation[C],Proceedings of the IEEE International Conference on Computer Vi-sion. 2015: 2407-2415.

[13] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, “Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention,” Hindawi.com, Oct. 21, 2020.

[14] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, “Image Captioning: Transforming Objects into Words,” Jul. 21, 2019.

[15] G. Sharma, P. Kalena, N. Malde, and A. Nair, “Visual Image Caption Generator Using Deep Learning,” SSRN Electronic Journal, Jan. 2019, doi: 10.2139/ssrn.3368837.

[16] M.Cornia, M. Stefanini, L. Baraldi, και R. Cucchiara, ‘Meshed-Memory Transformer for Image Captioning’. arXiv, 2019.

[17] Z. Zohourianshahzadi and J. K. Kalita, "Neural attention for image captioning: review of outstanding methods," Artificial Intelligence Review, Nov. 2021, doi: 10.1007/s10462-021-10092-2.

[18] Muhammad Abdelhadie Al-Malla, Nada Ghneim, and Assef Jafar, "Image captioning model using attention and object features to mimic human image understanding," Feb. 14, 2022.

[19] MindSpore, "Image Classification Using ResNet-50 Network,"Aug.2019. https://www.mindspore.cn/tutorial/training/en/r1.1/advanced_use/cv_resnet50.html

[20] Chenliang Li,Haiyang Xu, and Junfeng Tian, "mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections," Papers with code, May 24, 2022.

[21] S.Vinodababu,"a-PyTorch-Tutorial-to-Object-Detection,"github.com,Aug.08,2020.