Image Caption: A Deep Learning Approach

Babandeep Singh | Rahul Gera | Robin Beura | Yash Patel | Sanghamitra Muhuri

Editor:

Abstract

Relating words to an image is a major goal in image captioning. While recent technological advances has enabled us to generate description learning from the huge corpus of words, however, more often than not we fall short of words when it comes to convey message through an image. Through this paper we are targeting to find a reasonable ground in generating customized and personalized captions after learning features of an image and generating captions. While the success of these methods is encouraging, they all share one key limitation: detail. By only describing images with a single high-level sentence, there is a fundamental upper-bound on the quantity and quality of information approaches can produce. In particular, we are interested in the generating longer, richer sentences and paragraphs that could convey a story/message rather than description of image. In this paper we explore and summaries few of the existing state-of-art techniques for image captioning and image paragraph captioning using the novel data from the Instagram. We explore standard MLE based Encoder-Decoder architecture for captioning, Hierarchical RNNs for paragraph captioning, and another less explored approach based on Conditional Generative Adversarial Networks.

1. Introduction

Recently there has been considerable interest in joint visual and linguistic problems, such as the task of automatically generating image captions. Interest has been driven in part by the development of new and larger benchmark datasets such as Flickr 8K [3], Flickr 30K [4] and MS COCO [5]. However, while new algorithms spurs considerable innovations- as text generation, image classification, or perhaps object detection [6][7], new challenges and stacking presents opportunities for advancement in AI. In this paper, we present a novel automatic image caption generation that measures the quality of generated captions by analyzing their semantic content. Our method closely resembles human annotation while offering the additional personal touch.

There have been many successful attempts to generate the description of images based on its contents based on the flickr 30K images datasets. Based on papers submitted in Conference on Computer Vision and Pattern Recognition, many similar works have been targeting the description of an image, whereas we are trying to get it more personalized by learning the syntactic and semantic structuring of captions.

Figure 1: Image of a pot with vegetables

One of the problems with description generation is that they describe the objects in the image. To illustrate the limitations, consider the following two captions for image in Figure 1.

- A shiny metal pot filled with some diced veggies.
- The pan on the stove has chopped vegetables in it.

The captions describe two very different images. However, comparing these captions, we can observe that these captions convey almost the same meaning, but exhibit low n-gram similarity as they have no words in common and generating such captions will lead to many similar meaning contextual captions. To overcome the limitations, we hypothesize that learning from personal accounts and their captions can aid in generating related and personalized captions through the capabilities of Convolution Neural Networks to learn the images and Recurrent Neural Networks to generate text.

Taking this main idea as motivation, we estimate caption quality by transforming both candidate and reference captions into a stacked Neural network. The network explicitly encodes the objects, attributes and relationships found in images and decodes the captions, abstracting away most of the lexical and syntactic idiosyncrasies of natural language in the process. This approach has been exploited and deemed to be a highly effective representation for performing complex image retrieval queries [8, 9].

2. Data and Model

2.1 Dataset

For the purpose of this study, we expected to have a data set which consists of images and human annotated captions. Even though there are many public data sets MSCOCO, Flickr30 etc available, to incorporate human touch we decided to web scrape images and captions through Instagram. Over 1 billion users across the world, made Instagram a great contender for this scraping. The proposed model uses a novel dataset consisting of 43,284 image-caption pairs created by automated web-scrapping from three Instagram feeds namely - Natgeo, Natgeotravel and Natgeoyourshot. Since the language present is diverse (high variance) to reduce the variance we focused on a

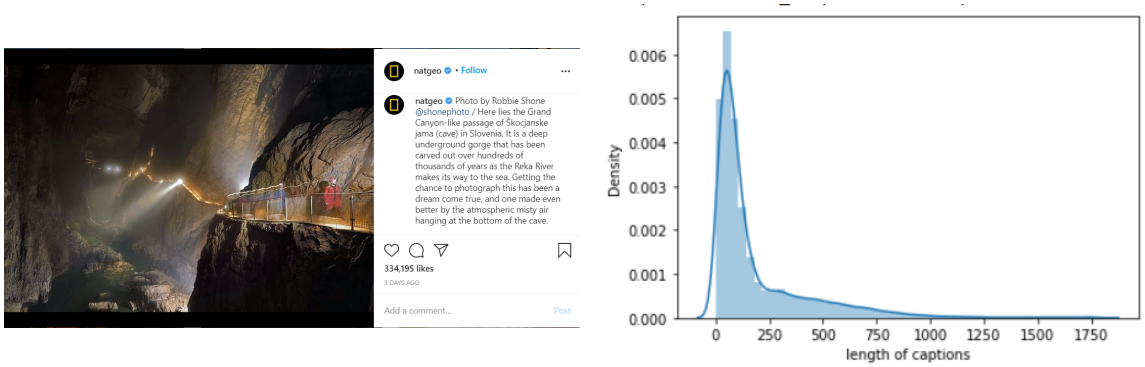

Figure 2: (a) Instagram post by Natgeo user Robbie Shone. (b) Density plot of length of captions

verified account 'Nat-Geo' group to capture less variant caption data and structured captions in English. Figure 2 (a) is an example of the image and structured caption captured from Natgeo Instagram account. More often then not, the image caption is structured as *Photo by @username / caption*. For our image data, we tried to retain the RGB *3 * H * W* and customize images to our H*W constraints. For our text data, We initially performed data cleaning process which aimed at removing *Photo by @username /* from the caption as this does not relate to any particular image. We also removed any special mentions formatted *@username* and hashtags, for instance *#travel* present in caption. For this study we avoided removing any functional words, punctuation as they are part and parcel of the human language captions. Figure 2 (b), we observed that the caption lengths are right skewed suggesting that many captions are beyond average caption length 31. We decided to stick with captions length 36 (buffer of 5) for our study.

2.2 Model architecture

Most of the models rely on the widespread encoder–decoder framework, which is flexible and effective. Sometimes it is defined as a structure of CNN + RNN. Usually a convolutional neural network (CNN) represents the encoder, and a recurrent neural network (RNN) the decoder. The encoder is the one which "reads" an image—given an input image, it extracts a high-level feature representation. The decoder is the one which generates words—given the image representation from the encoder (encoded image), it generates words to represent the image with a full grammatically and stylistically correct sentence. [10]

2.2.1 Encoder - CNN

As there is usually only one encoder in the model, the performance is highly reliant on the CNN deployed. ResNet wins for being computationally the most efficient compared to all other convolutional networks. It is clear from the table, that ResNet

CONVOLUTIONAL NEURAL NETWORKS ARCHITECTURES					
Architecture	**#Params**	**#Multiply-Adds**	**Top-1 Accuracy**	**Top-5 Accuracy**	**Year**
Alexnet	61M	724M	57.1	80.2	2012
VGG	138M	15.5B	70.5	91.2	2013
Inception-V1	7M	1.43B	69.8	89.3	2013
Resnet-50	25.5M	3.9B	75.2	93	2015

Figure 3: Table of comparison for CNN architectures (from ref.[11])

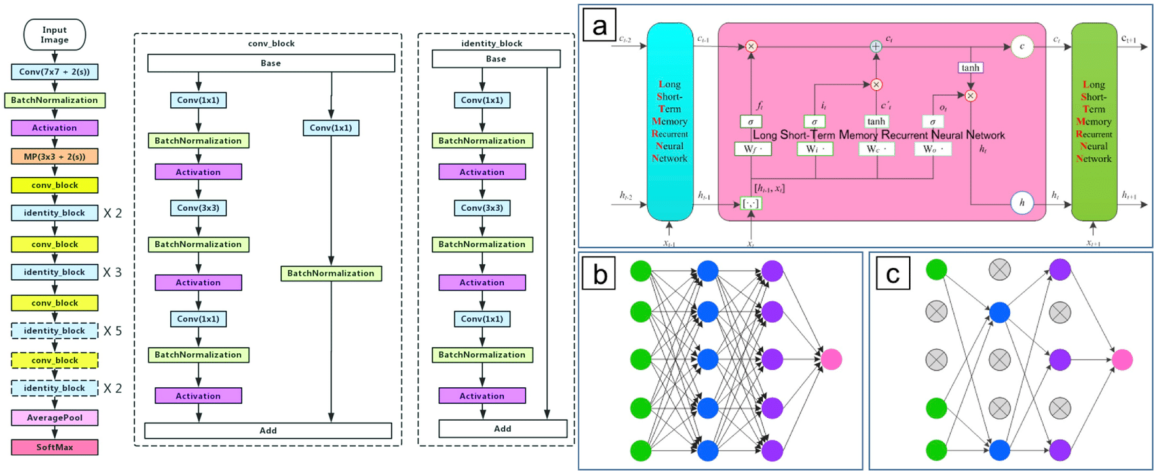

Figure 4: (Left) ResNet50 architecture. Blocks with dotted line represents modules that might be removed in our experiments. (Middle) Convolution block which changes the dimension of the input. (Right) Identity block which will not change the dimension of the input. (Right) (a) Architecture of the LSTM RNN method. (b) Standard recurrent neural network. (c) Recurrent neural network after applying Dropout algorithm.

performs best—from both Top-1 and Top-5 accuracy. It also has much fewer parameters than VGG which saves computational resources. However, being easy to implement, VGG remains popular among researchers and has the second highest result, regarding the review from ref. [11].Because of these reasons, we decided to go with ResNet model architecture to train our encoder.

2.2.2 Decoder - RNN

An extension of RNN which is gaining grounds is a Long-Short Term Memory(LSTM) in processing of sequential data. Due to its effectiveness in memorizing long term dependencies through a memory cell, LSTM is considered the most efficient method for image captioning.

LSTM works by generating a caption by taking one word at every time step conditioned on the context vector, together with the hidden state and the earlier generated words. As a result, given the size of the vocabulary - each new word added

makes LSTM consume more time. Clearly, the recurrent processing requires a lot of storage, and deemed to be complex in maintenance. Nonetheless, the reward of using LSTM is greater than the drawbacks.

3. Experiments

Just as individuals tell stories differently, generating a paragraph caption of an image is challenging in itself. Firstly, the varsity amongst the syntactic and semantic formulations is omnipresent as captions are merely from different perceptive and perspective views. Secondly, optimizing over single annotated paragraph thus suffer from losing massive information expressed in the image. Finally, annotating images is labor-expensive process and often leading to variable length and small scale image-caption pairs which limits model generalization.

3.1 Generative Adversarial Networks - GANs

Existing image description methods are largely restricted by small sets of biased visual paragraph annotations, and fail to cover rich underlying semantics. Through a Generative Adversarial Network (GAN), an adversarial framework between a structured paragraph generator and multi-level paragraph discriminator can establish a image caption generator method. The paragraph generator generates sentences recurrently by incorporating region-based visual and language attention mechanisms at each step. The quality of generated paragraph sentences is assessed by multi-level adversarial discriminators from two aspects, namely, plausibility at sentence level and topic-transition coherence at paragraph level.

To overcome above challenges, GANs establish an adversarial training mechanism between a structured paragraph generator and multi-level paragraph discriminators, where the discriminators learn to distinguish between real and synthesized paragraphs while the generator aims to fool the discriminators by generating diverse and realistic paragraphs. Given an input image, a set of semantic regions using dense captioning method [12] are captured and represented with a visual feature vector and a short text phrase. The generator then sequentially generates meaningful sentences by incorporating the fine-grained visual and textual cues in a selective way. To ensure high-quality individual sentences and coherent whole paragraph, a sentence discriminator and topic-transition discriminator measures the plausibility and smoothness of semantic transition with preceding sentences.

We understand that attention mechanism effectively facilitates the prediction of GAN by selectively incorporating appropriate visual and language cues. Particularly, the advantages of explicitly leveraging words from local phrases suggest that transferring visual-language knowledge from more fundamental tasks (e.g. detection) is beneficial for generating high-level and holistic descriptions. The generator can sequentially output diverse and topic-coherent sentences to form a personalized paragraph for an image.

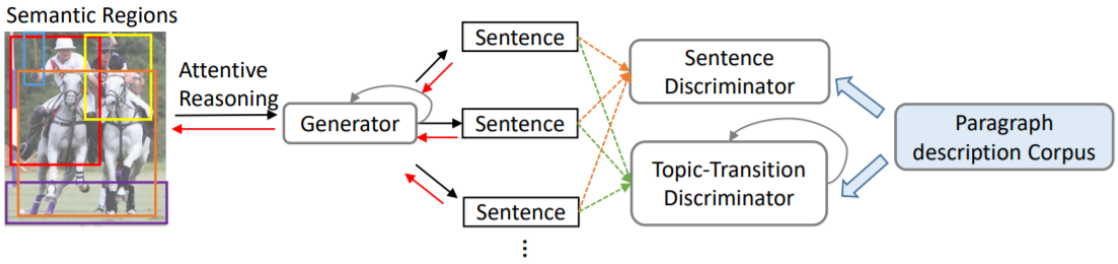

Figure 5: GAN architecture with generator and sentence and topic-transition discriminator

3.2 Hierarchical Recurrent Networks

The underlying assumptions above try to decompose both images and paragraphs into their constituent parts, detecting semantic regions in images and providing a cohesive and related captions. While the success of methods involved in caption description are encouraging, they share a key discrepancy; detail. By only describing images with a single high-level sentence, there is a fundamental upper-bound on the quantity and quality of information approaches can produce. Due to vanishing gradients, RNNs trained above with SGD often struggle to learn long-term dependencies. Although LSTM [13] alleviates this problem through a grating mechanism. Another solution is a hierarchical recurrent network, where the architecture is designed such that different parts of the model operate on different time scales.

Since all sentence captions have the goal of describing the image as a whole, they are fundamentally limited in terms of both diversity and their total information. Longer descriptions go beyond the presence of a few salient objects and include information about their properties and relationships.

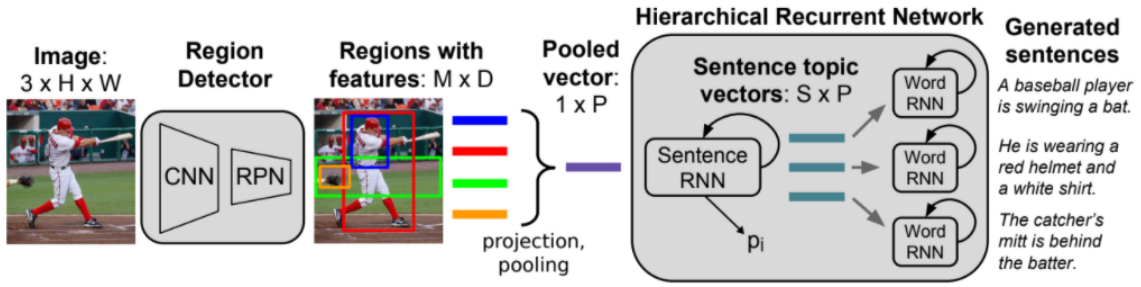

Figure 6: Hierarchical Recurrent Networks (HRNNs) architecture with generator and sentence and topic-transition discriminator

Firstly, the input image is decomposed by detecting regions, then aggregating features across these regions to produce a vector representation expressing image semantics. This vector behaves as a input to the hierarchical recurrent neural network which has 2 further levels; sentence RNN and Word RNN. The sentence RNN receives

image semantics vector, and a user defined number of output sentences generated to be included in the caption and produces a topic vector which is the input of Word RNN and specified size sentences are generated. The sentence RNN is responsible for deciding the number of sentences S that should be in the generated paragraph and for producing a P-dimensional topic vector for each of these sentences. Given a topic vector for a sentence, the word RNN generates the words of that sentence.

At each time step, the sentence RNN receives the pooled region vector v_p as input, and in turn produces a sequence of hidden states $(h_1, ..., h_S \in \mathbb{R}^H)$, one for each sentence in the paragraph. Hidden states are utilized in two distinctive ways:

- First, a linear projection from hi and a logistic classifier produce a distribution pi over the two states CONTINUE = 0, STOP = 1 which determine whether the i^{th} sentence is the last sentence in the paragraph
- Second, the hidden state h_i is fed through a two-layer fully-connected network to produce the topic vector $t_i \in \mathbb{R}^P$, which is the input to the word RNN.

The first and second inputs to the RNN are the topic vector and a special *START* token, and subsequent inputs are learned embedding vectors for the words of the sentence. At each timestep the hidden state of the last LSTM layer is used to predict a distribution over the words in the vocabulary, and a special *END* token signals the end of a sentence. After each Word RNN has generated the words of their respective sentences, these sentences are finally concatenated to form the generated paragraph.

3.3 Generative Pre-trained Transformer (GPT)

A recent study[14] summarized that training more parameters will increase accuracy of several NLP benchmarks without fine-tuning. GPT-3 is such an auto - regressive language model with 125 Billion to 175 Billion parameter variants. Baseline OpenAI suggests is that larger models captures more contextual information efficiently.

In retrospect, GPT-2 is a large transformer-based language model with 1.5 Billion parameters trained on 8 million web pages to predict next conditioned on the previous context. Since, the model is trained on huge corpora the auto-regressive models are quite attractive because we can just keep generating data forever. GPT-3 studies the model as a general solution for many downstream jobs without fine-tuning.

The GPT takes in a sequence as N words and outputs a word to be put at the end of the input sequence after we get the next word, we add it to the sequence, and get the following word. The input sequence needs to be fixed 2048 words, for which we require padding for shorter sequences. During the encoding we create a vocabulary to vector operation to input words as vectors. Since this grows exponentially with increase in vocabulary, GPT utilizes embedding to capture information within reasonable dimensions and reducing the vectors with zeroes.Through positional encoding, we can capture the sequential information of the text which acts like input for the attention mechanism. The purpose of attention is: for each output in the sequence,

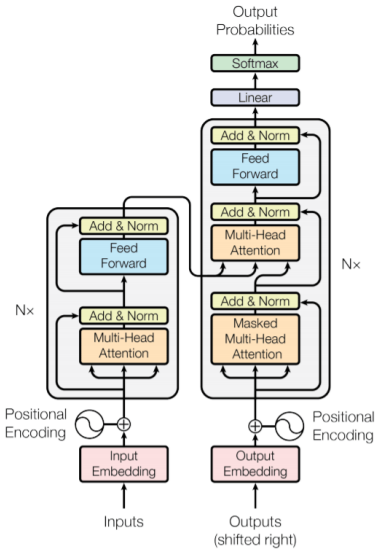

Figure 7: Generative Pre-trained Transformer (GPT) Architecture [16]

predict which input tokens to focus on and how much. In GPT, multi-head attention is implemented suggesting that the above focusing process is repeated many times each with a different learned query, key, value projection weights. During the Feed-Forward block, the input is multiplied with learned weights and added with learned bias, repetitively to get results and normalized. Once the sequences are learned, they are reverse mapped to initial word encoding.

4. Results

Figure 8 (a) shows an example of automatic caption generated through Encoder - Decoder model as *a woman stands in front of a buddhist temple in seoul south korea discriminator*. Although it captured nuances of the woman and standing, it failed to capture the regions appropriately to distinguish a temple and a woman in dusky leafs and a car. With a BLEU score of 9.418382295637229e-232. Closer to 1 BLEU score meaning a more apt overlaping of the words which is difficult in this situation.

Another example is Figure 8 (b) where the the generated text is able to understand that there is a crane present in the picture and still misses out the granular details. With a BLEU score of 9.206597977384398e-232 we see there are many overlapping instances suggesting that the model is able to capture the intricacies of the image and language

Figure 8:
(a) **Ground Truth** :a group of dusky leaf is resting after a morning spent feeding.
Prediction:a woman stands in front of a buddhist temple in seoul south korea discriminator.
(b) **Ground Truth** : a sandhill cranes flies over the platte river in nebraska joelsartore beautiful photooftheday nebraska
Prediction : a sandhill cranes flies across the sandhill atlas in switzerland this is the first aerial photo of this site i have seen although it is a large and very large cranes that are usually very small they fly on the serengeti pillars and can dive into

5. Conclusion

Through this study we were able to generate human like captions, and how they can be improved through HRNNs, GANs or GPT3. All of them rely on dense captioning which initially segments the image and captures the regions of image and generate few sentences to relate to those regions. Following which the partially generated sentences are feed into the RNN architectures which generates the human like captions.

Although the captions lack coherence in basic Encoder-Decoder architecture, in future we expect to explore more Pre-trained encoders enabling to train a model quickly or without much computational resources. Alternatively, building a corpora vocabulary of captions for individual users can help in more personalized captions.Further implications suggests that utilizing regions of an image gives a more control over variance and correlation with the caption the model generates therefore, building on top of ideology of object detection, huge vocabulary, one can achieve quality and qualified captions.

Acknowledgments

This paper and the research behind it would not have been possible without the exceptional support of our supervisor, Theja Tulabandhula and Teaching Assistant Seyed-Danial Mohseni-Taheri. Their knowledge and exacting attention to detail have been an inspiration and kept our work on track from my first encounter. Also, Instaloader API for aiding in smooth webscraping of Instagram posts.

References

1. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015)

2. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention (2015).

3. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. JAIR 47, 853–899 (2013)

4. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. TACL 2, 67–78 (2014)

5. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D.,

Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Heidelberg (2014)

6. Dicke, Robert H. "Object detection system." U.S. Patent No. 2,624,876. 6 Jan. 1953.

7. Anderson, Peter, et al. "Spice: Semantic propositional image caption evaluation." European Conference on Computer Vision. Springer, Cham, 2016.

8. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Image retrieval using scene graphs. In: CVPR (2015)

9. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: EMNLP 4th Workshop on Vision and Language (2015)

10. Staniūtė, Raimonda, and Dmitrij Šešok. "A systematic literature review on image captioning." Applied Sciences 9.10 (2019): 2024.

11. KOUSTUBH. ResNet, AlexNet, VGGNet, Inception: Understanding Various Architectures of Convolutional Networks. Available online: Here

12. J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In CVPR, pages 4565–4574, 2016. 2, 4, 6

13. S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 1997.

14. Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

15. https://medium.com/analytics-vidhya/openai-gpt-3-language-models-are-few-shot-learners-82531b3d3122 [accessed on 2^{nd} December 2020]

16. https://dugas.ch/artificial_curiosity/GPT_architecture.html [accessed on 2^{nd} December 2020]