Modeling image caption based on short-time human attention

Yuze He*
yuze.he@aalto.fi
Aalto University
Finland

Abstract

The paper provides a novel image caption model which emphasises visual saliency based on short-time human attention. Under some transitory situations, people will not perceive and interpret all the content in their field of view, only noticing the stuff within their foveal vision. However, given to word by word generation methods by calculating the probability distribution of the next word based on the image features and relations between words, conventional image caption generators are always trying to explain every detail of the image they captured. In response, we present a new model which re-weights the different parts of an image based on human short-time attention before translating them into words. We combine the original image with its heatmap representing the visual saliency of human, then input the enhanced picture into a LSTM-based caption generator. This improves the quality of short-time image captioning by 1) capturing the key features that belongs to the focus of your vision, 2) reducing the impact of unnecessary content, 3) shortening the lengthy descriptions for short-time use, such as audio assistant application for blind people. While conventional image caption model may translate the picture more accurately, our model performs better when people are given a short-time viewing duration: the descriptions from real human-beings are more similar to that generated by our model than by the previous ones.

ACM Reference Format:
Yuze He. 2021. Modeling image caption based on short-time human attention. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Image caption is one of an essential research problem in computer vision and natural language processing, it is used to explain and elaborate the pictures with a few lines of text. One of the most important application of image caption is to assist the blind people. For example, Apple promoted VoiceOver and VoiceAmplifier on their iPhone12 Max to help the blinded know better about the surroundings, the users just need to hold their phone and capture the image of a scene, and the application will translated this image into a audio description. However, when there are too much information in an image, the machine may fail to locate the most key content to focus, making descriptions lengthy for containing too many details.

It has been acknowledged that, for a normal person, the visual angle of human-beings includes two parts: foveal vision and peripheral vision. While the foveal vision catches most of our attentions, the peripheral vision only provides some auxiliary information to help us understand the scene. In most situations, central vision was more efficient for scene gist recognition than the periphery on a per-pixel basis.[5] When people are walking on the street, or given a short time to look around, only foveal vision takes a lead of what they perceive and interpret, which means some content of an image can be cut down before being translated under some circumstances.

However, previous works related to image caption hardly focus on how the attention of real human-beings will impact the weight of visual features. Approaches to image caption in deep learning mainly base on word by word generation methods by maximizing the probability of the next word. The model in [6] starts with a special start symbol or any reference word, which is used to calculates the probability distribution of the next word, so that to generates the new word step by step. This cycle continues until the end symbol is generated. For each time when a word is generated, the image feature is involved in the calculation. Vinyals et al. (2014) [7] also used recurrent neural networks (RNN) based on long short-term memory (LSTM) units (Hochreiter Schmidhuber, 1997)[4] for their models. and they only showed the image to the RNN at the beginning.Their model is capable of both

generating novel captions given an image, and reconstructing visual features given an image description. In 2015, Xu et al.[8] first introduce the "attention mechanism" into image caption by learning the weight of the location variable as where the model decides to focus attention when generating the next word, but "where" the network looks next also depends on the sequence of words that has already been generated.

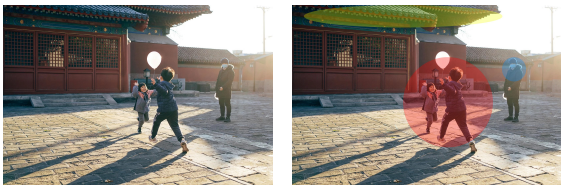
In this paper, we integrate human attention mechanism into image caption model. First, we will use the model proposed by Fosco et al. (2020) [3], which takes an image as the input and predicts saliency heatmaps for three different durations: 0.5s, 3s and 5s. Second, we will use the heatmap to re-weight different parts of the original image by conducting matrix multiplication, and input the processed image to traditional encoder-decoder framework [1] of image caption, so that to generate descriptions based on LSTM model.

The contributions of this paper can be described as below: We introduce real human attention in daily life to the pretreatment process of image caption, which extracts the key information of an image that catch the first attention of people under the scenario of short viewing time. This approach helps reduce the visual features of an image and focus on the most important things of a scene, so that to provide a more efficient way for image description and a better user experience for blind people assistant applications.

2 Related Works

In this section we provide relevant background on previous work on human attention and image caption generation. Xu et al.[8] introduce the "attention mechanism" into image caption, for which they use a multilayer perceptron conditioned on the previous hidden state, which varies as the output RNN advances in its output sequence: "where" the network looks next depends on the sequence of words that has already been generated.

Although this paper acquires state-of-art results by introducing the attention model, it does not consider the how the real human attention will influence the image interpretation. Heatmaps are widely used to indicate how gaze location varies on a single image, which reflects real human attention indirectly. In 2020, [3]proposed a model that takes an image as input and predicts three distinct saliency heatmaps for three different durations. They collect the CodeCharts1K dataset, which contains multiple distinct heatmaps per image corresponding to 0.5, 3, and 5 seconds of free-viewing, and they propose Temporal Excitation Module(TEM) which uses LSTM cells to generate scaling vectors that re-weight the feature maps differently for each duration. This model makes visual saliency predictable, laying a fundamental for re-weighting visual feature of an image according to real human attention under specific viewing duration.



(a) original image (b) heatmap image

Figure 1. Image caption model based on human attention

3 Approach

3.1 Model explanation

Instead of generating image caption word by word, we will re-weight the visual features of the image according to real human attention before translating them into words. We use an example to illustrate our idea. Fig11 shows two versions of a same image, (a) is the original image, (b) is the heatmap of it. People may pay their attention to the red part in the first second, then notice the blue part in the third second, and finally look at the architecture in yellow part in fifth second. When given a short time(3 seconds), human beings hardly perceived the yellow part of the image, which should be less considered accordingly for further description. In this case, our model may assign a bigger weight to the red and blue part, and smaller weight to the other two parts as a pretreatment, then send the processed image to caption model.

The concept of new model can be divided into two parts: Firstly, we use Temporal Excitation[3] Module to predict the saliency heatmap of an image, which is used to do element-wise (Hadamard) product with the original image to generate a new enhanced image. Secondly, the enhanced image can be fed into hard attention model [8] that produces the caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words. The framework of model can be seen in Fig22

3.2 Collect ground truth data

Although there is plenty of ground truth data about image caption, none of them is based on short-time glance. In order to compare the behaviour of our model with traditional one, new data based on short-time human attention should be gathered.

3.2.1 Participants. The participants include15 females and 15 males ranging in age from 20–50, with a mean of 28. All

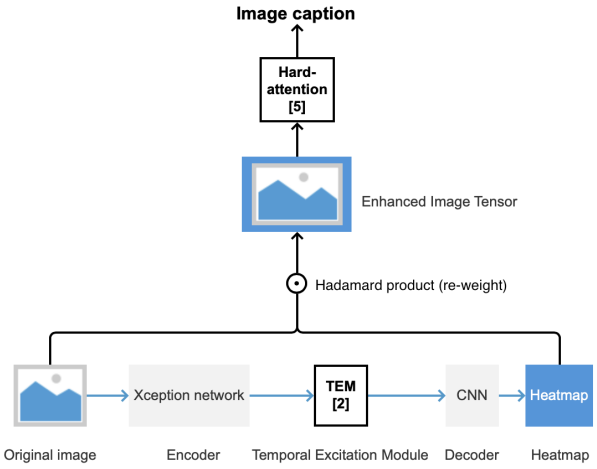


Figure 2. Image caption model based on human attention

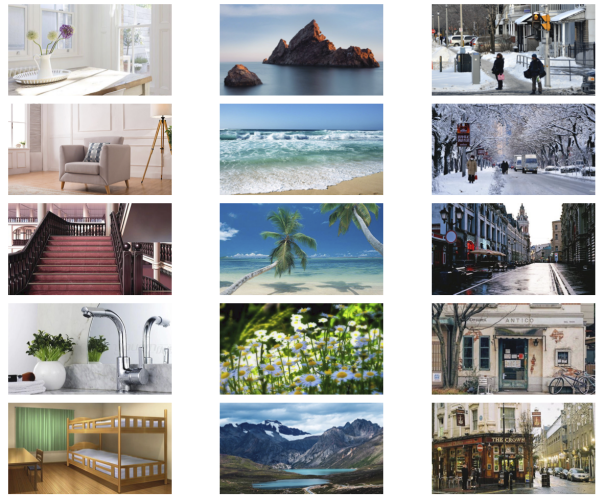


Figure 3. 15 example images

the participants are fluent in English and good at vision. The education backgrounds of the participants are undergraduate and above to ensure the basic interpretation capability. Ethics approval was obtained before we began recruiting participants. The study was advertised to college student group chats through social media application, Participants are self-selected and compensated €1.00 for completing the task.

3.2.2 Materials. We prepared 15 images of normal life Fig33, including 5 images of interior scene, 5 of city view, and 5 of nature landscape. In order to provide a similar experience with normal life, the pictures are shown on iMac 27" 5K Retina MXWV2, which is large enough and high-resolution. To capture short-time attention, we use JavaScript to make each image show for only three seconds, after which the participants are asked to describe what they have seen as soon as possible. The recording materials include Voice Memos on iPhone, Excel table and the screen recording software iShot.

3.2.3 Procedure. To avoid bias caused by individual difference, we choose the within-group experiment. All participants are told that the experiment would take a maximum of thirty minutes. After arriving at the laboratory individually, they were assured confidentiality, and they provided informed consent.To alleviate aesthetic fatigue, each participant is allowed to rest for 5 minutes after captioning every 5 pictures. During the test, each image are randomly presented and shown for three seconds, the participants just need to orally describe the image in English, and the observer is responsible for recording but not revising the original sentences from participants. In the end of the experiment, we received 15 * 30 terms of descriptions of images.

3.3 Model implementation

After collecting the ground truth data, we implemented our model by inputting the same 15 image into it. We use a state-of-the-art backbone as our encoder: the Xception network [2]. For the decoder, our experiments showed that a simple module composed of 3 sets of convolution, up-sampling and dropout layers are sufficient for this task. Besides, we use architecture of the Temporal Excitation Module(TEM)[3] between encoder and decoder.

After we complete the saliency prediction by getting the output (heatmap of the original image), we conduct Hadamard product between heatmap and original image to get the new matrix, which is fed to the traditional hard attention model [8] of image caption. In this way, we get the description of 15 images produced by our model.

3.4 Model evaluation

To compare the behaviour of our model with traditional image caption model, we set the cosine-similarity between the model's description and ground truth description as the metric of evaluation. The more the cosine-similarity value is close to 1, the better the model behavior. We transfer descriptions into vectors using word2vec https://code.google.com/archive/p/word2vec/and get 15(images) * 30(participants) * 3(ground truth data, descriptions from traditional hard attention model, descriptions from our model)data, and compared them by calculating cosine-similarity value.

3.5 Results

Although we fail to validate the model, the result table can be organized by comparison of two metrics: cosine-similarity between ground truth data and descriptions from traditional

hard attention model (Cos(G and H)) and cosine-similarity between ground truth data and descriptions from our model (Cos(G and O)). We choose the average value of Cos(G and H) and Cos(G and O) of 30 participants and fill them into the form. Beside, we also separate the images according to their types, in order to see whether the difference will vary among different kinds of pictures.

Category	No.	Cos(G and H)	Cos(G and O)
Interior scene	image1		
	image2		
	image3		
	image4		
	image5		
Nature landscape	image6		
	image7		
	image8		
	image9		
	image10		
City view	image11		
	image12		
	image13		
	image14		
	image15		

3.6 Data analysis

For explore whether there's significant difference between the performance of two models, we compare the average of Cos(G and H) and Cos(G and O). As the sample size is 30 and an same image is processed by two different model, we choose paired sample t-test(dependent sample t-test). Like many statistical procedures, the upper-tailed alternative hypothesis (H0) assumes that d[(Cos(G and O))-(Cos(G and H))] is smaller than zero, and significance level is set as 0.05. If we had validate the model successfully, there might be three kinds of results:

1) If p-value<0.05, we reject H0, which means (Cos(G and O)) is grater than (Cos(G and H), which means our model performs better than hard attention model under short-time viewing;
2) If p-value0.05, we cannot reject H0, which means we do not have enough evidence to prove our model performs better than hard attention model; Under this situation, we can 1) separate the data according to image categories and see whether there is significant difference under some image categories; 2) collect more data.

4 Discussions

Guided by the insight that what you interpret depends on what you focus, not what you "see", we design a new model that translates the image contents based on their proportion weighted by human short-time attention. We combine the Temporal Excitation Module[3] with hard attention model [8], and conduct empirical experiment to collect ground truth data of three types of image: Interior scenes, Nature landscape and City view. Finally, We find the metric to evaluate our model: the cosine-similarity between the description from ground truth data and it from image caption models.

There are a lot of space for us to improve. First, we fail to validate the model due to the lack of original code from hard attention model, but other image caption models with open code resources should be tried if possible. Secondly, the ground truth data we collect is still small (n=30), more participants are supposed to be invited if we invest more time and money.

Despite of the flaws, we have provided a new concept for image caption based on many literature review, and have thought of how would we conduct the research and model evaluation of the whole process. The possible application of our model may include timing billboard design, audio assistant application for blind people and other short-time image caption situations.

References

[1] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[2] François Chollet. "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[3] Camilo Fosco et al. "How much time do you have? Modeling multi-duration saliency". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 4473–4482.

[4] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[5] Adam M Larson and Lester C Loschky. "The contributions of central versus peripheral vision to scene gist recognition". In: *Journal of Vision* 9.10 (2009), pp. 6–6.

[6] Junhua Mao et al. "Explain images with multimodal recurrent neural networks". In: *arXiv preprint arXiv:1410.1090* (2014).

[7] Oriol Vinyals et al. "Show and tell: A neural image caption generator". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3156–3164.

[8] Kelvin Xu et al. "Show, attend and tell: Neural image caption generation with visual attention". In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057.