Image Caption Generation using CNN and LSTM: An Adaptive Attention Based Approach

Hitesh Sapkota
Rochester Institute of Technology
Rochester, NY, USA
hxs1943@rit.edu

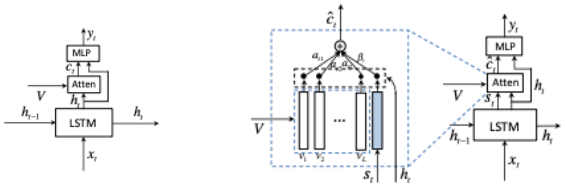
Abstract

Image captioning task has been considered as an artificial intelligence problem that requires the knowledge from both computer vision and natural language processing. Most of the state-of-the-art approaches provide equal importance on image features for every generated word. However, words such as "the", "of " require little to no image information during the decoding process. In this paper, we empirically evaluate the performance of the proposed adaptive attention approach that decides whether to look to the image or not depending on a word. Our experimentation involves Flickr8K dataset with three state-of-the-art techniques (VGG-16, ResNet, Inception-V3) for the image feature extraction and extended LSTM for the word feature extraction. The experimental result shows that the proposed approach can achieve the state-of-the-art result with a maximum BLEU-4 score of 0.225.

1. Introduction

Automatic image caption generation is considered as an interdisciplinary research problem that can serve as a huge help for visually impaired people. In order to generate high-quality caption, the model should be capable to include every fine detail present in the image. With the help of advanced training and large classification dataset, recent work has uplifted the quality of image caption generation process.

Recently, the attention-based model has shown incredible performance in image captioning task[5, 6]. The idea behind the attention model is instead of focusing on the whole image, focus only on the particular region of the image that is associated with the word. However, the underlying problem associated with the existing attention based models is that they try to attend to the image even for the words such as "the", "of " that do not have corresponding visual signals. As a result, the gradient of such non-visual words could mislead the image caption.



(a) Spatial Attention Model

(b) Adaptive Attention Model

In this paper, we propose an adaptive attention based approach that decides whether to look to the visual features and, if yes, which region of the image. Our approach involves an extended version of a Long Short Term Memory (LSTM) that produces an additional "visual sentinel" vector in addition to the hidden state vector. The proposed model uses a sentinel gate that decides amount of information from the image as opposed to relying on a sentinel vector.

2. Background

With the recent advancement in the deep learning field, powerful deep neural network based image captioning techniques have been proposed. For instance, Kiros et al. [2] proposed the encoder-decoder framework into image captioning research in order to unify the image-text embedding model and multimodal natural language model. Later Vinyals et al.[8] further improved the model by replacing RNN with more powerful LSTM as the decoder. Motivated by the fact that only image specific salient contents are responsible to create a caption, Xu et al. [9] proposed an attentive encode-decoder model that can dynamically attend salient image regions. Yao et al. [10] used the technique that involves augmentation technique that picks high-level attributes as image features in the image sentence generation. Our method is based on the Lu et al. [4] adaptive attention technique that reasons when to attend image while generating words for the sentence.

2.1. Methods

In this section, we first explain two proposed models: (1) spatial attention and (2) adaptive attention. Then, we explain the experimentation process followed by the result.

2.1.1 Spatial Attention Model

In spatial attention model, shown in Figure 1a, we compute the context vector as,

$$c_t = g(V, h_t) \tag{1}$$

where g is the attention function, $V = [v_1, ..., v_k]$ is the image feature vector and h_t is the hidden state of the LSTM. Given $V \in R^{d \times k}$ and $h_t \in R^d$, we get the attention distribution over k regions of the image as,

$$z_t = w_h^t tah(W_u V + (W_g h_t)\mathbb{1}^T) \tag{2}$$

$$\alpha_t = softmax(z_t) \tag{3}$$

where $W_v, W_g \in R^{k \times d}$ and $w_h \in R^k$ are parameters to be learned. $\alpha \in R^k$ is the attention weight over features in V. Based on the attention distribution, the context vector c_t can be obtained by:

$$c_t = \sum_{i=1}^{K} \alpha_{ti} v_{ti} \tag{4}$$

We combine c_t and h_t, using Multi Layer Perceptron (MLP), in order to predict the next word y_{t+1}.

2.1.2 Adaptive Attention Model

Figure 1b shows the proposed adaptive attention model. We obtain visual sentinel vector s_t as follow:

$$g_t = \sigma(w_x x_t + w_h h_{t-1}) \tag{5}$$

$$s_t = g_t \odot tah(m_t) \tag{6}$$

where, w_x and w_h are weight parameters to be learned, x_t is the input to LSTM at time t, g_t is gate applied on a memory cell m_t, and σ is the logistic sigmoid activation.

We compute new adaptive content vector as,

$$\hat{c}_t = \beta_t s_t + (1 - \beta_t) c_t \tag{7}$$

where β_t is the new sentinel gate at time t which is obtained from the last element (k+1) of the $\hat{\alpha}_t$ which is defined as,

$$\hat{\alpha}_t = softmax([z_t; w_h^T tanh(W_s s_t + (W_g h_t))]) \tag{8}$$

where ; indicates concatenation, W_s and W_s are weight parameters.

We combine $\hat{c}_t$ and h_t using MLP to get the new word y_{t+1}.

2.2. Results

2.2.1 Datasets

We experiment with the Flickr8k dataset which consists of 6000 training images, 1000 validation images, and 1000 test images. Each of the images has five ground truth captions.

2.2.2 Training Details

We experiment on Flickr8k dataset with three different image feature extraction techniques: Resnet [1], VGG16 [3], and Inception-V3 [7]. For embedding, we use LSTM with a hidden size of 512. Our training involves Adam optimizer with the learning rate of $4e^{-4}$ for the language model. The momentum and weight decay in our experimentation are 0.8 and 0.999 respectively. Further, we use a batch size of 64 and trains up to 20 epoch with early stopping if the validation BLEU score had not improved over the last 6 epochs. Experimentation involves beam size of 3 for the image caption generation.

2.2.3 Evaluation Results

The table below demonstrates the BLEU score obtained using three different spatial feature extraction techniques. Using VGG16 technique we get the best performance with BLEU-4 score of 0.225. Resnet also provides comparable performance whereas, Inception-V3 provides the least performance with BLEU-4 score of 0.116.

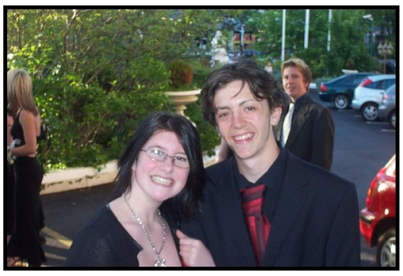
Table 1: Performance on Flickr8k dataset.

Method/Accuracy	B-1	B-2	B-3	B-4
Resnet	0.650	0.451	0.350	0.214
VGG16	0.662	0.470	0.364	0.225
Inception-V3	0.540	0.334	0.224	0.116

The result reveals the fact that the performance of the proposed spatial attention technique outperforms the state-of-the-art Soft and Hard attention techniques. The demonstration of the caption generated using the proposed technique is presented in the appendix section.

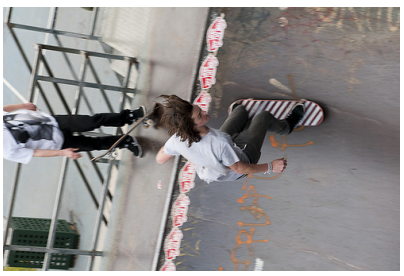
3. Conclusions

In this paper, we propose a sentinel-based adaptive attention technique for the image caption. The proposed technique decides whether to attend to the image features or not for a particular word using sentinel gate. The evaluation result on the Flickr8k shows that the proposed approach can achieve a state-of-the-art result.



(a) a man in a black shirt and a woman in a crowd

(b) a man in a white shirt and tie is sitting on a bench



(c) a boy in a blue shirt is jumping on a skateboard

Figure 2: Caption generated using the best performed model (with VGG16 as CNN) trained on Flickr8k.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[2] R. Kiros, R. R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[3] S. Liu and W. Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, Nov 2015.

[4] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250, 2017.

[5] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 289–297, USA, 2016. Curran Associates Inc.

[6] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[9] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2048–2057. JMLR.org, 2015.

[10] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4904–4912, 2017.