Channel-wise Attention and Image Initialized LSTM for Image Captioning

Anonymous Author(s)
Affiliation
Address
email

Abstract

The introduction of attention mechanics in image caption problem successfully improves the precision of the generated sentences. Existing attentions generally cover the spatial information of the images combined with the previous word generated by the Long-short Term Memory (LSTM) cell. However, we argue that the features among channels of last layer of CNN are also important for attention mechanics. In this paper, we propose a novel channel wise attention after the features extracted by CNN, and after the multilayer perceptron we obtain the mapping natural word. We validate the proficiency of this model on the benchmarking dataset Flickr30K.

1 Introduction

Automatically image caption generation is a classical task combining computer vision and natural language processing. Not only capturing the features from the image, the model has to express the relationships of them in understandable natural language. Recently neural encoder-decoder models [11] has successfully combined the Convolutional Nerual Network (CNN) as the encoder and Recurrent Nerual Network (RNN), Long-short Term Memory (LSTM), or Gated Recurrent Unit (GRU), as the decoder for caption generation. Figure 1 demonstrates an typical structure of it.

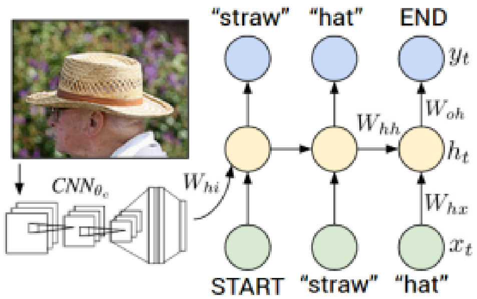


Figure 1: Typical Encoder-Decoder Model for Image Caption.

Visual attention is shown effective in such similar tasks [5]. Instead of compressing the image as the input only at the beginning of the decoder, attention mechanics is believed to dynamically select the features that needed to be expressed at the following space of the sentence. That is, the information of image participates in the generation of every word in every step. Moreover, features from an image are naturally spatial, multi-layer and channel-wise, the attention of image information is supposed

Submitted to 30th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

to consider the three characteristics comprehensively. Therefore, such attention are called spatial and channel-wise attention (SCA) [6]. However, the prediction of some words are independent of the visual signs, such as "the" or "of", and in this case the participation of image information would result in overfitting. Adaptive attention [3] would allow the decoder decide on the participation of visual signals during the word generation.

In this paper, we first realize normal LSTM encoder-decoder model without attention features, and then introduce the combination of modified channel-wise attention and image initialized LSTM in the decoder, which take advantage of the information of visual signals and the flexibilty of its participation inside the LSTM cell. We first apply the fea_fc features extracted from the distilled data. In particular, we propose a novel channel-wise attention which use a full connect network providing the adaptivity in channel weights, instead of the conventional average or outer product processing. Moreover, the initial meomory vector and hidden state in LSTM cell is also transformed by the image information. It's believed that LSTM cell would decide the attendance or absence of the image information at this word position. After the attention image is computed, the flattened vector is inputed into the multi-layer perceptron (MLP) to map the embedded word into the vocabulary dictionary. We validate the model on the classical image captioning dataset Flickr30K.

2 Related Work

In this section we investigate into the encoder-decoder framework for image caption, and the basic methods of previous related works. Early attempt [2] on this problem encoded the image with a CNN as the static input, and then used an RNN to decode the captions. However, inspired by the fact that human generally caption an image from the multiple local details inside the image instead of the whole, the concepts of attention is impressive and natural. Generally the function of attention encoder-decoder model could be described in the following categories.

1. **Spatial Attention.** As indicated in [5], spatial attention refer to selecting the region of the image of the CNN output. Specifically, stochastic "hard" attention applies an one-hot variable and select the region of the most probability. Deterministic "soft" attention uses a softmax layer in the spatial dimension, and returns the weighted average of spatial image that highlights the most likely region.

   In our model, we apply the fea_fc features that has the shape of $1 \times 1 \times 2048$, as the channel only information. The spatial attention is diminished.

2. **Channel-wise Attention.** CNN features contain not only spatial information but also channel dimension. Different semantic meaning among channels could also be interpretted as attentions. Like [6], in our model channel-wise attention are computed inter-channel. However, unlike the common processing of outer product, we find the modified dynamic weighted channel-wise attention is more reasonable and flexible.

3. **When and Where to attend.** Spatial and channel-wise attention are proved to be capable of extracting features in the visual image. However, whether to rely on the visual features are dependent on the word generated [3]. In common occasion, we use the state-of-the-art LSTM to handle the sequential generation task. In [5] the initial memory vector and hidden state of lstm cell is random initialized, and visual features are concatenated with word input as the sequential inputs. However in [6], they design the **visual sentinel** as the gate of the attendance of visual features, which is generated by the current generated word.

   Similarly in our model, the current word, not the previous one, interacts with the image information for the attention system. Therefore, the attention contains visual features combined with sequential information from LSTM cell. However, we do not use the visual sentinel gate in attention, but instead initialized the LSTM cell with visual features and let the LSTM cell to decide the attendance of it.

3 Model and Method

The overall structure of our model is demonstrated in figure 2. We design the utilities including Image Embeder (IE), Word Embeder (WE) and MLP, each of which is an independent single hidden-layer neural network and realizes the transformation of tensors. Image embeder is viewed as the mapping

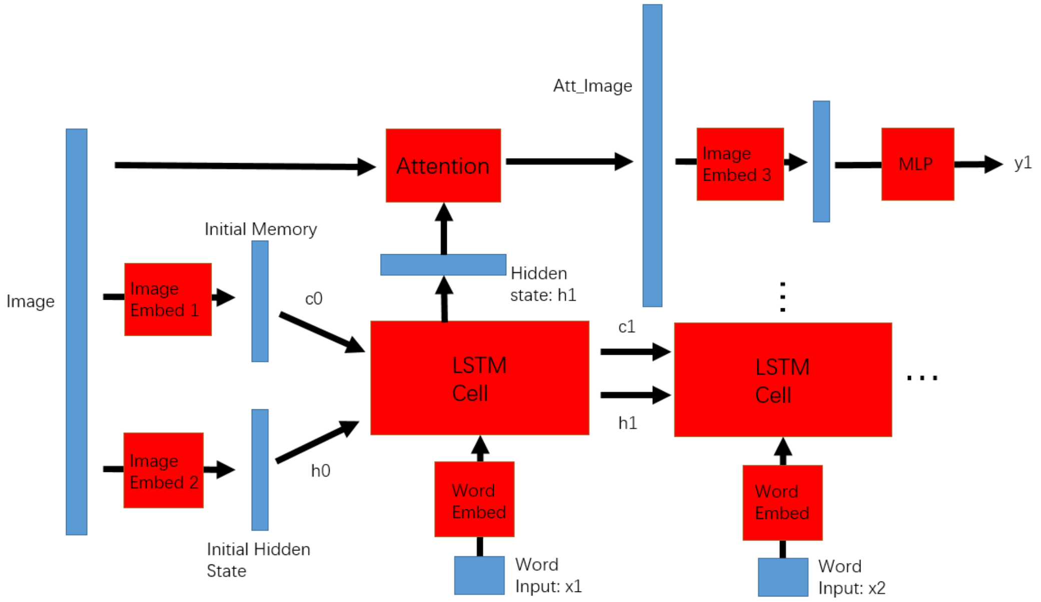


Figure 2: Structure of our Model.

function of image $V \in R^C$ to word vector $w \in R^Q$. Word embeder serves as the mapping function of raw natural word input $x_t \in R^K$ to word vector $w \in R^Q$. MLP is the mappin function of word vector $w \in R^Q$ to natural language word $y_t \in R^K$. Generally:

$$\begin{aligned} w &= IE(V) \\ w^* &= WE(x_t) \\ y_t &= MLP(w) \end{aligned} \tag{1}$$

Notice that we train three independent Image Embeder but one Word Embeder, since the three IE represent the mapping to cell memory, hidden state and transformation of attentioned image separately. However, WE serves as the same word input transformer in every step.

3.1 Objective Function

Generally for image caption task [7], we denote the size of vocabulary dictionary as K, and the length of caption as C. Then the caption could be formalized as:

$$y = \{y_1, y_2, ..., y_C\} \tag{2}$$

where $y_i \in R^K$. We denote θ as the parameters of the model and I as the image, the objective function and parameter iterative method is:

$$\theta^* = \arg\min_{\theta} \sum_{(I,y)} logp(y|I;\theta) \tag{3}$$

In this equation, we express the log-probability specifically as:

$$logp(y) = \sum_{t=1}^{T} logp(y_t|y_1, ..., y_{t-1}; I; \theta) \tag{4}$$

In our model, we firstly apply word embeder to transform the raw $x_t \in R^K$ natural word into the low-dimensional (denoted as R^Q) vector. Then it's inputed into our decoder at every step. In addition, the pre-processed features is used as extracted features from image as the output of encoder.

we make the output of encoder flattened and transformed by image embeder to serve as the initial cell memory and hidden state vector to LSTM cell in the decoder. The typical structure of LSTM cell is

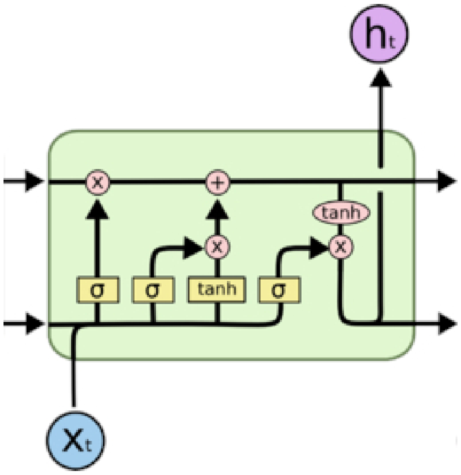


Figure 3: Structure of typical LSTM cell.

demonstrated in figure 3. For convenience, we assign that the memory and hidden state of LSTM cell has the same dimension as word embedding vector. We denote the h_t as the hidden state of LSTM, x_t as the word input and m_t as the cell memory vector at the t-th cell. The following equation explains the generation requirement:

$$h_t = LSTM(h_{t-1}, x_t, m_{t-1}) \tag{5}$$

3.2 Attention Model

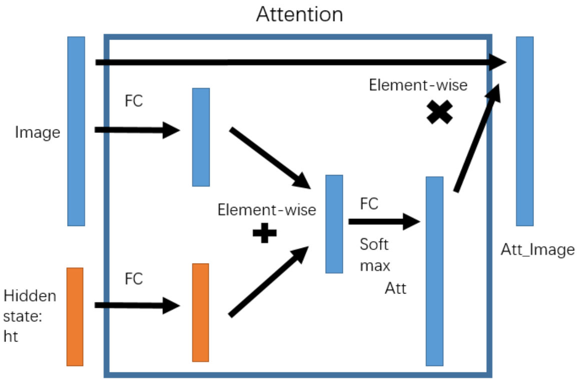


Figure 4: Structure of Attention.

Figure 4 demonstrates that our attention model includes the channel-wise attention, and provides the current hidden state of LSTM cell to decide the attendance of visual features. Channel-wise attention refers to the emphasized channel of an image when a certain word is under decision. We denote the channel-wise attention function as $g(.)$, $V \in R^C$ as the image input, and $h_t \in R^Q$ as the hidden state of LSTM cell. Then:

$$ch_t = g(V, h_t) \tag{6}$$

The function $g(.)$ is also a two-layer nerual network:

$$\begin{aligned} z_c &= tanh(V * W_{vc} + h_t * W_{hc} + B_{vc}) \\ ch_t &= softmax(tanh(W_c * z_c + B_c)) \\ Vat_t &= ch_t \otimes V \end{aligned} \tag{7}$$

Where $W_{vc} \in R^{C \times Q}$, $W_{hc} \in R^{Q \times Q}$, $B_{vc} \in R^Q$, and $W_c \in R^{Q \times C}$, $B_c \in R^C$. These are the parameter matrices to be optimized. $\otimes$ refer to the element-wise product of two vectors of the same shape. After the softmax layer, the channel-wise attention $ch_t \in R^C$ vector represents the weights of the matching channel of the image in each region.

Afterwards, the image after the attention are transformed by the image embeder and multi-layer perceptron (MLP) to generate the natural language word. The equation is expressed as:

$$y_t = MLP(IE(Vat_t)) \tag{8}$$

4 Experiments

4.1 Settings

We validate our model with the dataset: Flickr30k. It contains 31783 images, and with split information of 2783 test samples, and 29000 for train. Each image has 5 relative captions. We randomly choose 2894 images as validation dataset for hyper-parameter selection. As preprocessing, we truncate the captions who have the length of more than 22 words for convenience in train dataset, and limited our generated caption in this way. The vocabulary dictionary has 8638 words.

In our model, we denote the size of output of LSTM cell $Q = 32$, and batch size as 512. Moreover, after hyper-parameter selction, we assign learning rate as 10^{-3} where we find the best performance in validation dataset. We use Adam optimizer for parameter update. For pure LSTM model, we denote size of LSTM cell $Q = 126$, with the same optimizer.

We used the COCO captioning evaluation tool for validation and test performance, which includes BLEU(B@1, B@2, B@3, B@4), METEOR, CIDEr(CD) [9] and ROUGE-L(RG). They are 4 mainstream evaluation metrics for Natural Language Processing. Though BLEU, ROUGE-L and METEOR was first designed for translation evaluation, which mainly considers precision and recall rate, the CIDEr (Consensus-based Image Description Evaluation) mainly considers similarity to human-written languages. So, in most Nerual Network models, performance on this metric is always worst.

We compare our performance with the other state-of-the-art methods in offline evaluation, such as **DeepVS** [1], **GoogleNIC** [7], **m-RNN** [4], **Soft-Att** and **Hard-Att** [5], **emb-gLSTM** [10], **ATT** [8], **SCA-CNN-ResNet** [6] and **Adapt-Att** [3].

4.2 Results

4.2.1 Convergence Diagnose

We denote an epoch as we explore the whole train dataset for train, which is about 450 batches, and validate the performance of the epoch every 50 epoches using the evaluation tools on validation dataset. The curve of training loss and validation evaluation performance is shown in figure 5.

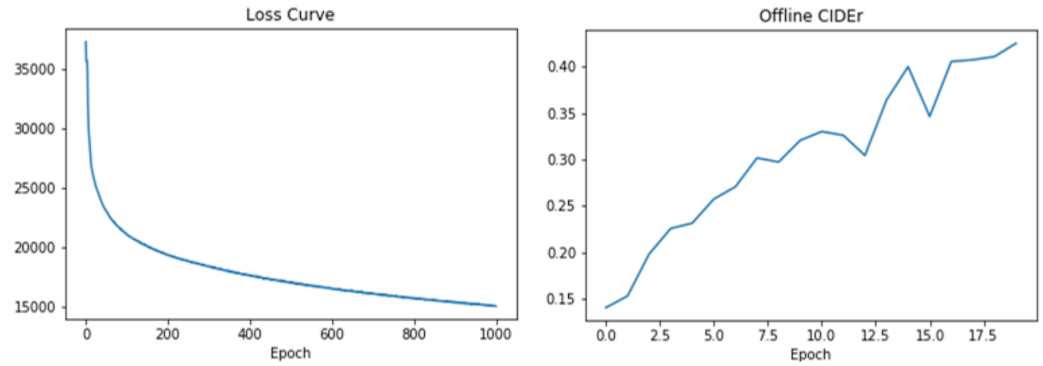


Figure 5: Loss function CIDEr performance in validation dataset.

As is indicated, the loss function reveals a typical convergence curve when the order of epoch increases, which decrease fast at first while slower later. CIDEr and other performance keeps increasing gradually though it seems that better performance is available if more training time is permitted. However, as order of epoch increases, the increase of metric grades slows down, even goes nonmonotonic at some points. This is because the model is approaching its best and the uncertainty of random sampling from the validation dataset.

4.2.2 Comparison

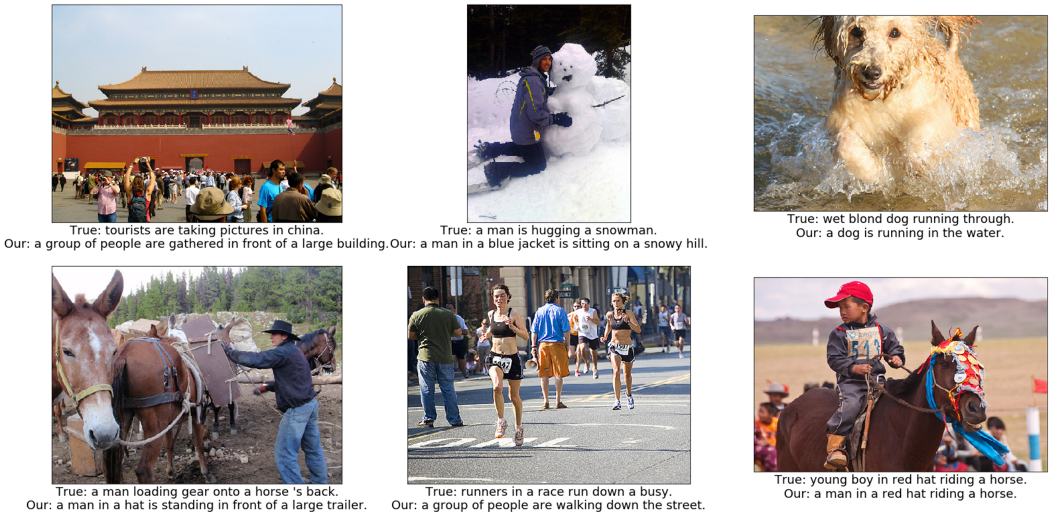
model	B@1	B@2	B@3	B@4	METEOR	CIDE
Deep VS	57.3	36.9	24.0	15.7	15.3	24.7
Google NIC	66.3	42.3	27.7	18.3	-	-
m-RNN	60.0	41.0	28.0	19.0	-	-
Soft-Att	66.7	43.4	28.8	19.1	18.5	-
Hard-Att	66.9	43.9	29.6	19.9	18.5	-
emb-gLSTM	64.6	44.6	30.5	20.6	17.9	-
ATT	64.7	46.0	32.4	23.0	18.9	-
SCA-ResNet†	66.2	46.8	32.5	22.3	19.5	-
Adapt-Att†	67.7	49.4	35.4	25.1	20.4	53.1
our LSTM	56.9	36.6	23.2	15.1	15.5	23.2
our LSTM-Att	61.3	42.9	29.7	20.4	18.5	42.5

Table 1: Offline Performance on Flickr30k test splits. † indicates an ensemble model results. - means no public record available.

Chanllenges include that the fact that the encoders in some of these benchmarking models are Inception or ResNet, while our encoder is unknown. The precision of encoder would differ the performance even with the same decoder structure.

As is shown in table 1, our model significantly outforms Deep VS, m-RNN and common soft or hard attention models. It is proved that the introduction of our attention significanly increases the evaluation compared with simple encoder-decoder model, and the channel-wise attention together with our initialization method enhance the quality of captioning.

4.2.3 Examples of generated captions



True: tourists are taking pictures in china.
Our: a group of people are gathered in front of a large building.

True: a man is hugging a snowman.
Our: a man in a blue jacket is sitting on a snowy hill.

True: wet blond dog running through.
Our: a dog is running in the water.

True: a man loading gear onto a horse 's back.
Our: a man in a hat is standing in front of a large trailer.

True: runners in a race run down a busy.
Our: a group of people are walking down the street.

True: young boy in red hat riding a horse.
Our: a man in a red hat riding a horse.

Figure 6: Some typical examples in test dataset and our captioning.

Some typical examples of test images and the caption generated from our model is indicated in figure 6. Generally, little grammar mistakes or baffling expressions are generated, which indicate the effect of LSTM. Moreover, campared to single encoder-decoder model, we find that in our model with attention, the image features are mostly captured. For example, the only LSTM caption for the fourth picture is "a man is standing in front of a horse". It could be explained that the LSTM-only model miss the attention of "hat", while the attention model successfully contains more details.

The successful description of different objects in an image means that our attention model is not only able to decide the different focus in the image when generating a certain word, but also finds more detailed focus.

5 Conclusion

In this paper, we propose an channel-wise attention in encoder-decoder LSTM model. Moreover as the innovation, we initalize the hidden state of LSTM cell using image features as the adaptive judgement of the attendence of visual features, and using the novel structure inside the channel-wise attention. Our model validate its significance with the standard COCO test tool for evaluation. For further improvement, the structures of our designed utilities require further studies in more training time and other image sets.

References

[1] L. Fei-Fei A. Karpathy. Deep visual-semantic alignments for generating image descriptions. *CVPR*, 2015.

[2] S. Guadarrama M. Rohrabach S. Venugopalan K. Saenko T. Darrell J. Donahue, L.A. Hendricks. Long-term recurrent convolutional networks for visual recogonition and description. *CVPR*, 2015.

[3] D. Parikh R. Socher J. Lu, C. Xiong. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *CVPR*, 2017.

[4] Y. Yang J. Wang Z. Huang A. Yuille J. Mao, W. Xu. Deep captioning with multimodal recurrent networks(m-rnn). *ICLR*, 2015.

[5] R. Kiros K. Cho A. Courville R. Salakhutdinov R. S. Zemel Y. Bengio K. Xu, J. Ba. Show, attend and tell: Nerual image caption generation with visual attention. *ICML*, 2015.

[6] J. Xiao L. Nie J. Shao W. Liu T. Chua L. Chen, H. Zhang. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *CVPR*, 2017.

[7] S. Bengio D. Erhan O. Vinyals, A. Toshev. Show and tell: A neural image caption generator. *CVPR*, 2015.

[8] Z. Wang C. Fang J. Luo Q. You, H. Jin. Image captioning with semantic attention. *CVPR*, 2016.

[9] D. Parikh R. Vedantam, C.L. Zitnick. Cider: Consensus-based image description evaluation. *CVPR*, 2015.

[10] B. Fernando T. Tuytelaars X. Jia, E. Gavves. Guiding the long-short term memory model for image caption generation. *ICCV*, 2015.

[11] Y. Wu R. Salakhutdinov W. W. Cohen Z. Yang, Y. Yuan. Encode, review, and decode: Reviewer module for caption generation. *NIPS*, 2016.