Image Captioning with GRU-based Attention

Mehrdad Mokhtari
Department of Computational Chemistry
Simon Fraser University
mokhtari@sfu.ca

Akbar Rafiey
School of Computing Science
Simon Fraser University
arafiey@sfu.ca

Hamid Homapour
School of Computing Science
Simon Fraser University
hamid_homapour@sfu.ca

Faezeh Bayat
School of Computing Science
Simon Fraser University
fbayat@sfu.ca

Abstract

We introduce an attention based model that automatically learns to generate a caption for images. Our model consists of a novel attention module which includes an elegant modification of GRU architecture. We validate the use of our attention model on a benchmark datasets MSCOCO, and compare its performance with other sate-of-the-art models.

1 Introduction

Automatically generating a natural language description of an image, a problem known as image captioning, is a challenging task that connects Computer Vision with Natural Language Processing which are two major fields in Artificial Intelligence. Image captioning is much harder than the well-studied image classification or object recognition tasks. Indeed, this task requires a level of image understanding that goes beyond object detection. Besides object detection, it requires understanding the relation between objects, object attributes, and the activities that objects are involved in. Moreover, it asks for an expression that relates objects to each other as well as expressing their attributes and activities.

Despite the challenging nature of this study, there has been a significant progress to address this problem. Recent works have significantly improved the quality of caption generation using a combination of Convolutional Neural Networks (CNN)s and Recurrent Neural Networks (RNN)s. They use CNNs to obtain a generic image representation, and RNNs to decode those representations into natural language sentences (see Section 3 for more details).

Visual attention is an important mechanism in the visual system of humans. Rather than compressing an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. This selective mapping allows the brain to allocate computational resources on an object at a time, guided by low-level image properties. The visual attention mechanism also plays an important role in natural language descriptions of images. In particular, one would not describe everything in an image. Instead, one tends to talk more about semantically important regions and objects in an image.

Our Contribution: InIn this work we proposed a novel attention mechanism for image captioning. Our proposed model uses a modification of Gated Recurrent Units (GRU) (Chung *et al.* [3]) that takes the regions and a probability distribution over the regions (which is a probability distribution over the importance of each region) and aims to provide a contextual representation that allows logical reasoning over the interesting regions.

2 Related Work

The problem of generating natural language descriptions from visual data has long been studying in computer vision. In this section we provide relevant background on previous works on image caption generation and attention mechanisms. As mentioned before, this task connects computer vision and natural language processing together. On the language side, models based on RNNs have been shown to produce state-of-the-art results on various tasks. RNNs are suitable to deal with sequential data of varying lengths and can learn complicated dynamics in a sequence. However, because of the vanishing gradient problem [8], they have faced difficulties in learning long-term dependencies. This problem has been addressed by introducing Long-Short Term Memories (LSTM)s [8]. On the image side, CNNs such as ResNet, GoogLeNet and VGGNet have shown great success in visual recognition tasks such as image classification and object detection. These type of models are trained on large image datasets such as ImageNet and are widely accessible.

A series of works [1, 5, 11, 13, 15] have leveraged the power of CNNs and RNNs in image captioning. Most of these works represent images as a single fc-type feature vector from the top layer of a pre-trained convolutional network. Since these methods represent the entire image with a single feature vector, the aforementioned techniques are not robust enough to background clutter and may lose spatial information relevant to the caption. Karpathy and Li [9] proposed a model based on a bidirectional RNN. Their model scores the similarity between a set of region-grounded captions and the image regions generated by Region-CNN [6]. However, the region-grounded captions cannot be used to generate a global caption, since some of them are not part of the global caption. Moreover, to generate a global caption, it is challenging to find the optimal order to feed the region-grounded captions to an RNN. We would like to emphasize their model is an attention free model. Unlike [9], there has been a significant work to improve the quality of caption generation via attention mechanisms. Xu *et al.* [17] proposed a model which can learn to fix its gaze on salient objects while generating the corresponding words. They computed an attention weight for each location in the input image using a multi-layer perceptron (MLP) conditioned on the previous generated word. In another work, You *et al.* [18] employed an attention model to combine visual features and concepts such as words and objects in an RNN which generates the caption. Instead of applying a simple soft attention, Khademi and Schulte [10] introduced a 2D spatial attention mechanism based on a Grid LSTM to dynamically attend to the important regions of an image. In this paper, we propose a new attention mechanism by introducing ATTN GRU.

3 Model

In this section we discuss our proposed model. Before diving to our model, first, we give an overview of the two simpler cases.

3.1 Warm Up

CNN+LSTM: Let us start off with an attention-free model introduced by Vinyals *et al.* [15]. Figure 1 shows the architecture of their proposed model. In their model, each image is represented by a single feature vector that is extracted from the top layer of CNN. Initially, this image representation is used as the input to the LSTM. Then, at each time step t, the last hidden state $\mathbf{h}^{t-1}$ and an embedding of the word $\mathbf{S}_t$ are used as inputs to the LSTM. Then the LSTM predicts the word for the next time-step $\mathbf{S}_{t+1}$.

Since this work, a series of attention mechanisms have been flourished to improve the performance. Most related and notable one is [17].

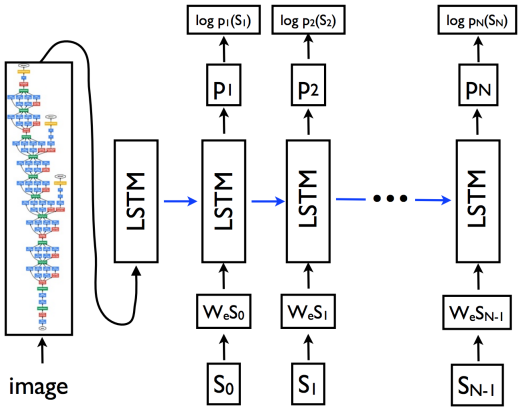


Figure 1: CNN+LSTM model with image embedder and word embedding.Figure from [15].

Attention-based Models: Generally speaking, in attention based models, one wants to assign *attention weights* to the features extracted at different image locations. Then uses these weights in a systematic way to generate a more informative inputs for the LSTM at each time step t.

Soft-Attention [17]: Let $\mathbf{R}_1, \ldots, \mathbf{R}_k$ be the features extracted from different regions of image. Moreover, let $P_1^t, \ldots, P_k^t$ denote the corresponding attention weights to each region at time step t (later we will discuss a systematic way to learn these weights). In soft-attention model, one produces a contextual vector $\mathbf{c}^t$ through a weighted summation of list of regions and corresponding probabilities P_i^t:

$$\mathbf{c}^t = \sum_{i=1}^{k} P_i^t \mathbf{R}_i$$

Then sum of context vector and word embedding $[\mathbf{c}^t + \mathbf{S}_t]$, and the previous hidden state $\mathbf{h}^{t-1}$ are used as the inputs for LSTM to predict the next word $\mathbf{S}_{t+1}$. This simple attention mechanism improves the equality of image captioning over CNN+LSTM model, see Table 1. This method, on the bright side, is easy to compute and computationally light. However the main disadvantage to soft attention is that the summation process loses both *positional* and *ordering* information about regions. Whilst multiple attention passes can retrieve some of this information, this is inefficient. Next we propose our new model and address these issues.

3.2 Our Contributions and Proposed Model

Our proposed model consists of three main components, namely a CNN, Attention Module which contains different parts, and a LSTM (Figure 2). In what follows, we elaborate on each of the mentioned components.

Deep CNN for Extracting Regions: In our model we use ResNet150 [7] to extract feature representations for different regions in the image. Opposed to the models that consider a single fc-feature vector to represent an image, we extract conv-features to represent regions of the image. This is because fc-type features do not preserve spatial information. Meaning, one cannot associate parts of the feature vector to regions in the image. Therefore, in our model, to represent regions we take the output of the last pooling layer of ResNet150 which has dimension $d = 7 \times 7 \times 2048$. The pooling layer divides the image into regions of 7×7, resulting in 49 local regional vectors of length 2048 (Figure 2). We denote these regions by $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_{49}$.

Attention Module: In order to assign attention weights to the extracted regions $\mathbf{R}_1, \ldots, \mathbf{R}_{49}$, we use a multi-layer perceptron (MLP). The shared MLP takes a single region representation as well as the previous hidden state of LSTM $\mathbf{h}^{t-1}$ as input, and learns an attention weight for each region. Parameters of the MLP are shared across all regions. The resulting attention weights are then passed

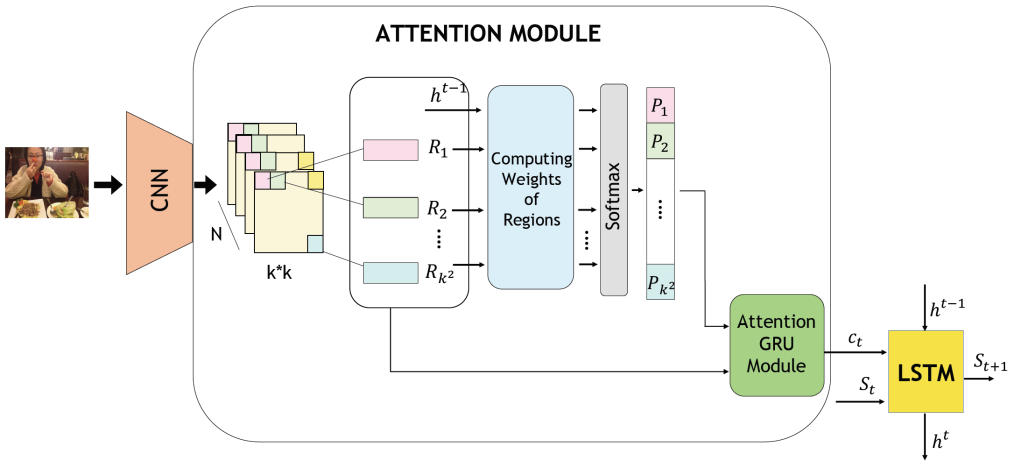


Figure 2: Our proposed attention model for image captioning, consisting of CNN, Attention Module, and LSTM. The attention module has two main components. A MLP to compute the attention weights and a attention GRU module which aims to provide a contextual representation that allows logical reasoning over interesting regions.

through a softmax layer which transfer these weights to a probability distribution over regions. Let P_i^t denote the associated probability to region $\mathbf{R}_i$ at time step t. Note that $\sum_{i=1}^{49} P_i^t = 1$.

Once we have a probability distribution over regions we use an attention mechanism to extract a contextual vector $\mathbf{c}^t$ based on the current probabilities. To do so, we introduce a new approach. It involves a novel modification of the architecture of traditional GRU.

Attention GRU Module: Inspired by [16], we want the attention mechanism to take into account both position and ordering of the input regions. An RNN would be advantageous in this situation except they cannot make use of the attention wights. In order to incorporate the attention weights P_i^t, we modify the architecture of the Gated Recurrent Unit (GRU), we call this modified version *ATTN GRU*. For the sake of completion, in the following we present the architecture of a GRU. Refer to [2, 3] for more details.

Denote the hidden state of the GRU at time step i by $\mathbf{h}'_i$. In a traditional GRU, the activation $\mathbf{h}'_i$ of the GRU at time i is a linear interpolation between the previous activation $\mathbf{h}'_{i-1}$ and the candidate activation $\tilde{\mathbf{h}}_i$. An *update* gate $\mathbf{u}_i$ governs this linear interpolation, and the candidate activation $\tilde{\mathbf{h}}_i$ depends on a *reset* gate $\mathbf{r}_i$. The exact formulation is as follows:

$$\mathbf{h}'_i = (\mathbf{1} - \mathbf{u}_i)\mathbf{h}'_{i-1} + \mathbf{u}_i\tilde{\mathbf{h}}_i \tag{1}$$

$$\mathbf{u}_i = \sigma(\mathbf{W}_u\mathbf{x}_i + \mathbf{U}_u\mathbf{h}'_{i-1}) \tag{2}$$

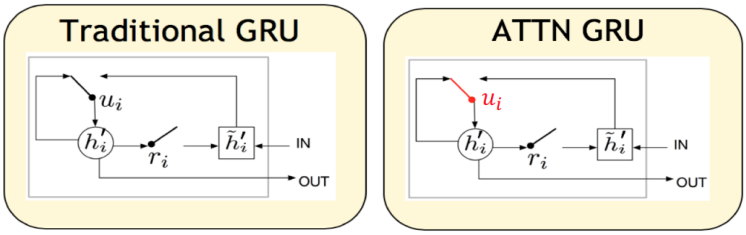
$$\tilde{\mathbf{h}}_i = tanh(\mathbf{W}\mathbf{x}_i + \mathbf{U}(\mathbf{r}_i \odot \mathbf{h}'_{i-1})) \tag{3}$$

$$\mathbf{r}_i = \sigma(\mathbf{W}_r\mathbf{x}_i + \mathbf{U}_r\mathbf{h}'_{i-1}) \tag{4}$$

In our setting, vector $\mathbf{x}_i$ is region $\mathbf{R}_i$. The update gate $\mathbf{u}_i$ in Equation 2 decides how much of each dimension of the hidden state to retain and how much should be updated with the transformed input $\mathbf{x}_i$ from the current time step.

To use the attention weights, we replace the update gate in Equation 2 with $\mathbf{u}_i^t = P_i^t \cdot \mathbf{1}$. The ATTN GRU can now use the attention probabilities for updating its hidden state. This change is depicted in Figure 3.

Now the contextual vector $\mathbf{c}^t$ is equal to the final hidden state of the ATTN GRU (i.e. $\mathbf{c}^t = \mathbf{h}'_{k^2}$, Figure 4). Note that number of state is equal to the number of regions which is 49. As before, sum



(a) $u_i = \sigma(\mathbf{W}_u \mathbf{x}_i + \mathbf{U}_u h'_{i-1})$ (b) $u_i = P_i \cdot \mathbf{1}$

Figure 3: (a) The traditional GRU, and (b) the proposed attention-based GRU model.

of context vector and word embedding $[\mathbf{c}^t + \mathbf{S}_t]$, and the previous hidden state $\mathbf{h}^{t-1}$ are used as the inputs for LSTM to predict the next word $\mathbf{S}_{t+1}$.

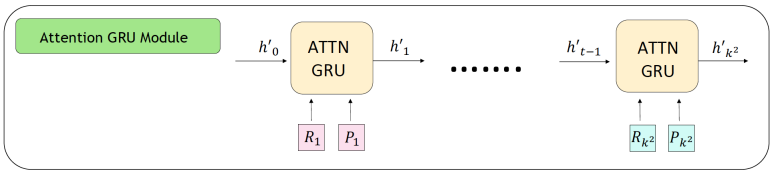


Figure 4: Attention GRU Module which has k^2 states. The output of the last state i.e. $\mathbf{h}'_{k^2}$ is used as context vector.

Bringing in the Attributes: In order to initialize the LSTM, we use the softmax layer of ResNet. The output of this layer provides a probability distribution over 1000 object classes. This improves the result by informing the LSTM about the presence of objects in the image. Vinyals *et al.* [15] observed that feeding the image at each time step as an extra input yields inferior results, as the network can explicitly exploit noise in the image and overfits more easily.

4 Experiments

In this section, we firstly introduce the dataset and evaluation metrics that we use in our experiments. Then, we explain our experimental set up and methodology. Finally, the experimental results are presented and discussed.

Dataset: We performed experiments on MS COCO dataset [12], contains complex day-to-day scenes of common objects in their natural context. The dataset contains 82,783 training images, 40,504 validation images, and 40,775 test images. Each image is annotated with 5 sentences using Amazon Mechanical Turk. Since, there is an ongoing competition on this dataset, annotation for test dataset is not available. To train our model, we have used both training and validation sets. To test the proposed model, we have hold out 5000 samples of the validation set. The same split is used for all the experiments. We did not use the test set for evaluation since there is a limited number of submissions available per day.

Metric: We use METEOR [4] and BLEU [14] as evaluation metrics, which are popular in the machine translation literature and used in recent image caption generation papers. The BLEU score is based on n-gram precision of the generated caption with respect to the references. The METEOR is based on the harmonic mean of uni-gram precision and recall, and produces a good correlation with human judgment.

Implementation Details: The models are implemented with Tensorflow and are trained using the RMSprop optimizer for 100 epochs with batch size 40 and learning rate 0.0001.

For fairness, we re-implemented two baselines models in [15, 17] and trained them and our proposed model with the same setting. The CNN used in all the models is ResNet150 [7] and we used the pre-trained model on ImageNet for initializing the weights.

4.1 Experimental Results

All the experiments are trained by teacher forcing methodology. In teacher forcing methodology, at each time-step the ground-truth word is fed to the model. But in the test phase, at each time-step we use the previously generated word as the input to our model for generating the next word. Table 1 shows the experiment results that compares our proposed attention model with the baselines. Our model outperforms both of the baselines. Comparing with the CNN+LSTM baseline, we can see that attention mechanism is helping the model to capture more focused visual that helps the model to generate more reasonable captions. In comparison to soft-attention (Soft-Attention), we believe that our model is better in capturing the spatial information relevant to the caption. Observe that ATTN-GRU+Attribution outperforms ATTN-GRU. This shows incorporating image attributes improves the results. Figure 5 shows some captioning examples that are generated by our model and the baselines, demonstrating the effectiveness of our proposed model.

Table 1: BLEU-1,2,3,4/METEOR metrics compared to other methods on MS COCO dataset. Models with * are trained on both train set and validation set.

Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**
BiRNN [9]	62.5	45.0	32.1	23.0	19.50
LRCN [5]	62.8	44.2	30.4	21.0	-
Google NIC [15]	66.6	46.1	32.9	24.6	-
CNN+LSTM*	66.7	51.1	39.0	29.92	22.1
Soft-Attention*	72.5	55.6	41.4	30.6	24.6
ATTN-GRU*	73.7	57.1	43.1	32.4	25.6
ATTN-GRU+Attributes*	**74.0**	**57.5**	**43.6**	**33.1**	**27.7**

5 Conclusion

We have presented a new attention mechanism for image caption generation by introducing ATTN GRU (a modified version of traditional GRU). Unlike soft-attention mechanism, our attention model preserves the spatial information as well as the order of the regions in the image. Experimental results on MS COCO dataset shows the effectiveness of our model in image captioning task.

References

[1] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6298–6306. IEEE, 2017.

[2] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[4] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

a young boy is playing a video game

a bedroom with a bed and a window

a group of people playing a game of soccer

a woman sitting at a table with a plate of food

Figure 5: Example captions generated by our model.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[9] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[10] M. Khademi and O. Schulte. Image caption generation with hierarchical contextual visual spatial attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1943–1951, 2018.

[11] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *International Conference on Machine Learning*, pages 595–603, 2014.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[16] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406, 2016.

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[18] 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.