Image Caption Generation with Text-Conditional Semantic Attention

Luowei Zhou  
Robotics Institute  
University of Michigan  
luozhou@umich.edu

Chenliang Xu  
Parker Koch  
Jason J. Corso  
Electrical Engineering and Computer Science  
University of Michigan  
{cliangxu@,pakoch@,jjcorso@eecs.}umich.edu

Abstract

We propose a semantic attention mechanism for image caption generation, called \textit{text-conditional semantic attention}, which allows the caption generator to automatically learn which parts of the image feature to focus on given previously generated text. To acquire text-related image features for our attention model, we also improve the guiding Long Short-Term Memory (gLSTM) structure by back-propagating the training loss though semantic guidance to fine-tune the CNN weights. In contrast to existing gLSTM methods, such as emb-gLSTM, our fine-tuned model enables guidance information to be more text-related. This also allows jointly learning of the image embedding, text embedding, semantic attention and language model with one simple network architecture in an end-to-end manner. We implement our model based on NeuralTalk2, an open-source image caption generator, and test it on MSCOCO dataset. We evaluate the proposed method with three metrics: BLEU, METEOR and CIDEr. The proposed methods outperform state-of-the-art methods.

1 Introduction

Image caption generation is drawing increasing interest from scientists and engineers in computer vision and machine learning communities [1][2]. Basically, it requires machines to automatically describe the content of an image using an English sentence. While this task seems obvious for human-beings, it is quite complicated for machines because it requires the language model to capture implicit semantic information of an image, such as objects’ motions and actions. Another challenge for image captioning is that the generated outputs should be human-like sentences which is a challenge for some generative models.

Generally, the literature on image captioning can be divided into three categories: template-based methods [3], transfer-based methods [4] and neural network-based methods [5], respectively. Template-based methods use specific templates for generating captions, and fill in the templates with detected attributes from the target image. One critical issue of those methods is that some descriptions cannot fit into human-crafted templates. Transfer-based methods rely on image retrieval to caption the target image with descriptions of similar images in the training set. Those models have less robustness on unseen images. On the other hand, neural network-based methods do not require a fixed template structure, nor rely on adequate examples to transfer. Furthermore, they are empirically more effective than most previous methods. So, here we focus on the neural network-based methods.

Recent successes of deep neural networks in machine translation [6][7] catalyze the adoption of neural networks in solving image captioning problems. Early work of neural network-based image captioning models include the multimodal RNN [8] and LSTM [5]. In these methods, neural networks are used for both image-text embedding and sentence generating. Later, Xu et al. [9] proposed a model that integrates visual attention into the LSTM model and Jia et al. [10] proposed a guiding LSTM model (gLSTM) that uses semantic information as the learning guidance.
However, the gLSTM based models proposed in [10], such as emb-gLSTM, suffer from the following two issues. Firstly, emb-gLSTM uses Canonical Correlation Analysis (CCA) [11] to map textual features and image features into a same semantic space. Even though the embedded image features are more text-related than pre-trained CNN features, the CCA embedding can only model linear relations between image and text. Instead of using this linear semantic embedding, our method directly uses image features as the semantic guidance while fine-tuning the CNN weights during training, so that the guidance is more text-related. We name this improved framework end-to-end gLSTM (e2e-gLSTM).

Secondly, in the gLSTM model, the semantic guidance is time invariant, which means it is unrelated to the context of the generated sentence – this is inconsistent with common sense. Generally, when people are generating different parts of a sentence, they focus on different areas of an image, which means the semantic guidance should change according to the context. A very recent work [12] tackles the semantic attention problem by leveraging attribute prediction. Basically, it fuses top visual attributes extracted from images with the input and the output of LSTM, and uses them as a guidance. Even though this approach achieves state-of-the-art performance, its performance relies heavily on the quality of the pre-specified visual attributes, i.e., better attributes usually lead to better results. Also, the attribute predictor has no learning ability and is separated from the encoder-decoder language model. To this end, we propose a method to learn semantic attention automatically, called text-conditional semantic attention.

Our text-conditional semantic attention model is based on our new end-to-end gLSTM framework which has learned text-related image features. When our attention model gets a target image, it generates semantic guidance by conditioning CNN image features directly on the current text content. So, the model learns how to focus on an image given the content it has generated. If it conditions the image features on the previous word, it is basically a 1-gram word-conditional model. If it is on previous two words, we get a 2-gram word-conditional model. Similarly we have an n-gram word-conditional model. The extreme version of the text-conditional model is the sentence-conditional model, which takes advantage of all the previously generated words.

We implement our model based on NeuralTalk2, an open-source implementation of Google NIC [5]. We compare our methods with state-of-the-art methods on the commonly used MSCOCO dataset [13] with publicly available splits [14] of training, validation and testing sets, and evaluate them with three metrics: BLEU, METEOR and CIDEr. The proposed methods outperform state-of-the-art methods.

The contributions are twofold. Firstly, we propose the idea of text-conditional semantic attention which allows the language model to learn context-related semantic guidance automatically. The proposed attention model learns how to focus on an image given the content it has generated. Secondly, this paper is the first to apply gLSTM to image caption generation in an end-to-end manner. Our model structure is less complicated compared to some state-of-the-art methods [10] [9] [12] owing to our reuse of the LSTM structure for semantic guidance.

2 Background

2.1 Caption Generation with LSTM

The Recurrent neural network (RNN) is a commonly used model in natural language processing [15]. Basically, in RNN model, the previous outputs of hidden layers are recursively used as one part of the inputs of hidden layers at the current time step. To avoid the well-known vanishing or exploding gradient problem of RNN, a variant of RNN called Long Short-Term Memory (LSTM) [16] is proposed by replacing the nodes in RNN with LSTM nodes. The LSTM node is the basic unit of the LSTM model, which is shown in Fig. [1]. Cell is the state of the node, which is controlled by three gates, namely, input gate, forget gate and output gate. This allows the LSTM model to learn which information to remember, and which to discard. Each gate is determined by the input and previous hidden state. The formulations of the gates are as follows [17].
\[ i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1}) \]  
\[ f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \]  
\[ o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1}) \]  
\[ c_t = f_t \odot c_{t-1} + i_t \odot h(W_{ct}x_t + W_{cm}m_{t-1}) \]  
\[ m_t = o_t \odot c_t \]

where \( W \) denote weights, \( \odot \) represents element-wise multiplication, \( \sigma(\cdot) \) is the sigmoid function, \( h(\cdot) \) is the hyperbolic tangent function, \( x_t \) stands for input, \( i_t \) for the input gate, \( f_t \) for the forget gate, \( o_t \) for the output gate, \( c_t \) for state of the memory cell and \( m_t \) for the hidden state (also output for one-layer LSTM). Their subscripts represent the time, \( t \) is current time step and \( t-1 \) is previous time step.

Vinyals et al. \[5\] propose a neural and probabilistic framework for image captioning and combine it with the LSTM model: 
\[ \Theta^* = \arg \max_\Theta \sum_{(I,S)} \log p(S|I; \Theta) \] 
where \( \Theta \) are the parameters of the model, \((I,S)\) is an image-sentence pair. Applying the Bayes chain rule, the likelihood of a sentence can be decomposed as: 
\[ \log p(S|I; \Theta) = \sum_{t=1}^N \log p(S_t|I, S_0, ..., S_{t-1}; \Theta) + \log p(S_0|I; \Theta) \]

One specific instantiation of this general framework is the Google NIC model \[5\], where the ImageNet \[18\] is used as the image embedding and LSTM is used as the language model.

### 2.2 Semantic Information and Guiding LSTM

Jia et al. \[10\] point out that sometimes the generated sentences by the LSTM model "lose track" of the original image content because LSTM accesses the image content only once, i.e., at the beginning of the learning process, and it forgets the image content after even a short period of learning. Therefore, they propose an alternative extension of the LSTM model, named guiding LSTM (gLSTM), which can extract semantic information from the target image and feed it into LSTM model in every time step as an extra guidance information. The basic gLSTM unit is shown in Fig. \[2\]. The memory cell and gates in a gLSTM node are defined as follows:

\[ i'_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ig}g) \]  
\[ f'_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fg}g) \]  
\[ o'_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{og}g) \]  
\[ c'_t = f'_t \odot c_{t-1} + i'_t \odot h(W_{ct}x_t + W_{cm}m_{t-1} + W_{ig}g) \]  
\[ m_t = o'_t \odot c'_t \]

where \( g \) represents guidance information, which is time invariant. Furthermore, Jia et al. \[10\] mention three ways to extract semantic information that can be used as guidance: 1) cross-modal retrieval-based guidance (ret-gLSTM); 2) semantic embedding guidance (emb-gLSTM); 3) image
features as guidance (img-gLSTM). In most of the cases, emb-gLSTM achieves better performances over other two approaches. In this paper, we focus on the last two methods. In emb-gLSTM, semantic representation computed by normalized CCA is used as the extra input while in img-gLSTM, the image feature is directly used as the extra input. Basically, emb-gLSTM uses a linear transformation to map images and their corresponding descriptions to similar feature vectors while img-gLSTM has no concern about the similarity of images and descriptions. However, CCA is a linear transform that cannot fully represent some nonlinear relations between the pre-trained image feature and its corresponding description. To this end, we build our model on img-gLSTM, and fine-tune the CNN weights during learning to obtain text-related image features. Another issue of gLSTM is that the semantic guidance is time invariant, as mentioned in Section 1, which we solve by text-conditional semantic attention.

3 Methods

In this section, we first describe the text-conditional semantic attention mechanism in detail and the intuition of this design. Then, we introduce the learning framework, i.e., end-to-end gLSTM. Finally, we summarize implementation details of the proposed models.

3.1 Text-Conditional Semantic Attention

In the recent work [12], You et al. use visual attributes as the semantic attention to guide the image captioning. More specifically, the semantic guidance consists of top visual attributes of the input image, and the weight of each attribute is determined by the current word, which is the previous output of RNN. However, the attribute predictor adopted in this model has no learning ability and is separated from the encoder-decoder language model. Instead, we condition the guidance information $g_t$ on the current word $S_{t-1}$, and use the text-conditional image feature as the semantic guidance so that the model can learn which part of the image feature should focus on when seeing a specific word. For instance, when the caption generator generated a sequence as "A boy is sitting on a", its attention on the image feature should be automatically switched to objects that can be sat on, such as chairs, beds or the floor, etc. Inspired by [19], we model the text-conditional image feature as:

$$g_{t,i} = \sum_{j,k} W_{ijk} g_t s_{t-1,k} + b_i ,$$

where $g_{t,i}$ is the text-conditional guidance feature, $W$ is a 3-way tensor weight and $b$ is bias. In this model, image features are fully coupled with textual features though the 3-way tensor, which contains many redundant weights of an extra image feature embedding. Given that the model already
learns text-related image features from transferred knowledge of e2e-gLSTM, our proposed attention model only adds an embedding matrix $W_c$ for the textual feature $S_{t-1}$. Furthermore, in practice, adding one non-linear transfer function layer after the image-text feature boosts the performance. So, our proposed method ends up modeling the text-conditional feature $g'_t$ as a text-based mask on $g_t$ followed by a non-linear function:

$$g'_t = \Phi(g_t \odot W_c S_{t-1})$$  \hspace{1cm} (12)$$

where $W_c$ is the text-conditional image feature embedding, the current word $S_{t-1}$ is a one-hot vector based on the dictionary, and $\Phi(\cdot)$ is a non-linear transfer function. When $W_c$ is a all-one matrix, the conditioned feature $g_t \odot W_c S_{t-1}$ is identical to $g_t$. We transfer pre-trained model from e2e-gLSTM to initialize the CNN, language model and word embedding of our attention model. For text-conditional matrix, we initialize it with all ones. Hence, the initial state of text-conditional semantic attention model is equivalent to e2e-gLSTM when using the ReLU function as the transfer function. We show in Section 4 on transfer function selection. The gLSTM input of the model is illustrated in Fig. 4, where $W_c$ is the text-conditional embedding matrix and $\Phi(\cdot)$ is the transfer function.

We name the above model the 1-gram word-conditional semantic attention owing to the guidance feature is merely conditioned on the previous word. Similarly, we develop the 2-gram word-conditional model which utilizes previous two words, or even n-gram word-conditional model. The extreme version of the text-conditional model is the sentence-condition model, which takes advantage of all the previously generated word (see Eq. 13).

$$g'_t = \Phi(g_t \odot W_c \sum_{k=1}^{t} S_{k-1})$$  \hspace{1cm} (13)$$

One benefit of the text-conditional model is that it allows the language model to learn semantic attention automatically though the back-propagation of the training loss while attribute-based method, such as [12], represents semantic guidance by some major components of an image, but other semantic information, such as objects’ motions and locations, are discarded.

### 3.2 Model Structure

We use a fine-tuned version of img-gLSTM as the basic framework. The proposed model is shown in Fig. 5 which consists of three parts: 1) image embedding using convolutional neural network (CNN);
2) text embedding; 3) LSTM language model. Firstly, CNN features are computed for the images using VGG net [20] and bag-of-words (BoW) features are computed for the sentences. We use text embedding matrix $W_e$ to embed text features to the same feature space as image features. The text embedding matrix is initialized with Gaussian random values. Then, both of the image features and embedded text features are used as the inputs to gLSTM (see Fig. [3]). The outputs of the language model are log likelihood of each word from the target sentence and the opposite value of their sum is the loss. Finally, we back-propagate the loss to LSTM language model, the text embedding matrix and the image embedding CNN.

One significant benefit of back-propagating error through guidance information for fine-tuning the CNN is that the model can allow guidance information being more similar to its corresponding text description, which helps text-conditional semantic understanding as mentioned in Section 3.1. Even though the CCA based semantic embedding proposed in [10] has the same function, it can only have a linear transformation of pre-trained image features and text features, which cannot represent nonlinear similarity. Apparently, our proposed model can learn a nonlinear relation by leveraging the existing CNN-LSTM structure. To compute the gates, cell state and hidden state for our model, we use following formulations:

$$i_t' = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{iq}g_t) \quad (14)$$

$$f_t' = \sigma(W_{if}x_t + W_{im}m_{t-1} + W_{iq}g_t) \quad (15)$$

$$o_t' = \sigma(W_{io}x_t + W_{im}m_{t-1} + W_{iq}g_t) \quad (16)$$

$$c_t' = f_t' \odot c_{t-1} + i_t' \odot h(W_{ci}x_t + W_{cm}m_{t-1} + W_{cq}g_t) \quad (17)$$

$$m_t = o_t' \odot c_t' \quad (18)$$

where $g_t$ represents guidance information at time step $t$. Given that the semantic guidance (output of CNN) keeps changing, we model the guidance information as a time variant variable.

### 3.3 Implementation Details

We implement our model based on Karpathy’s NeuralTalk2 [14], which is an open source implementation of Google NIC method. For e2e-gLSTM, we concatenate the image features and text features as a long vector, and forward it to the gLSTM. Then, we back-propagate the loss from gLSTM to both word embedding and CNN (see Fig. [5]). The implementation details of text-conditional semantic attention method are shown in Fig. [4].

Several transfer learning mechanisms are adopted to accelerate the learning process. NeuralTalk2 initializes the VGG net with the pre-trained model from [21]. Also, inspired by curriculum learning [22], we train our model in two stages though transfer learning. Firstly, we train the model without CNN fine-tuning to obtain approximate values of $W_{iq}$. After a short period of learning, around 16 hours, we activate the CNN fine-tuning to get better image embedding, meanwhile, fine-tune the weights in the language model.

For the text-conditional methods, we train them based on the pre-trained model of e2e-gLSTM but no longer fine-tune the CNN weights to allow text-conditional matrix to learn a fixed mask on the image feature for each word.

### 4 Experiments

#### 4.1 Datasets, Metrics and Baselines

We use MSCOCO [13] as our dataset, which includes 123287 color images. We use the commonly adopted [14] splits of the dataset which assign 113287 images for training, 5000 for validation and 5000 for testing. Three metrics are adopted: BLEU, METEOR and CIDER. We use img-gLSTM and Google NIC (NeuralTalk2) as the baselines of e2e-gLSTM and also compare it with state-of-the-art non-attention-based model emb-gLSTM [10]. Then we compare our text-conditional semantic attention model with state-of-the-art attention-based methods such as Soft-Attention, Hard-Attention [9], emb-gLSTM [10] and ATT-FCN [12].
Table 1: Experimental results. For some competing methods, we extract their performance from the corresponding papers. We highlight the top score for each metrics. See text for details.

| Methods                | Bleu_1 | Bleu_2 | Bleu_3 | Bleu_4 | METEOR | CIDEr |
|------------------------|--------|--------|--------|--------|--------|-------|
| img-gLSTM [10]         | 64.7   | 45.9   | 31.1   | 21.4   | 20.4   | 67.7  |
| emb-gLSTM, Gaussian [10]| 67.0   | 49.1   | 35.8   | 26.4   | 22.7   | 81.3  |
| NeuralTalk2 [14]      | 70.5   | 53.2   | 39.2   | 28.9   | 24.3   | 92.3  |
| Hard-Attention [9]    | 71.8   | 50.4   | 35.7   | 25.0   | 23.0   | -     |
| Soft-Attention [9]    | 70.7   | 49.2   | 34.4   | 24.3   | 23.9   | -     |
| ATT-FCN [12]          | 70.9   | 53.7   | 40.2   | 24.3   | 24.3   | -     |
| e2e-gLSTM             | 71.2   | 54.0   | 40.1   | 29.2   | 24.4   | 95.1  |
| 1-gram word-conditional | 71.6   | 54.1   | 39.7   | 29.1   | 24.4   | 93.3  |
| 2-gram word-conditional | 71.3   | 53.9   | 39.7   | 29.4   | 24.5   | 94.9  |
| sentence-condition    | 72.0   | 54.6   | 40.4   | 29.8   | 24.5   | 95.9  |

4.2 Model Selection

We mentioned in Section 3 that we use a non-linear transfer function $\Phi(\cdot)$ in our attention model (see Eq.12). We test several different functions for $\Phi(\cdot)$, namely, Softmax, ReLU and Sigmoid. For all of them, we initialize the text-conditional embedding matrix with all ones. So, for ReLU, the initial state of the attention model is equivalent to the pre-trained e2e-gLSTM model. We base our experiments on the 1-gram word-conditional model and conclude that the model achieves best performance when $\Phi(\cdot)$ is a Softmax function (see Fig. 6). It is worth noting that other transfer functions might lead to better results than those mentioned in our experiments.

Figure 6: Model selection of the transfer function. Results are normalized by the highest score of each metric.

4.3 Quantitative Results

We summarize the quantitative results in Tab. 1. Unsurprisingly, CNN fine-tuning in the e2e-gLSTM boosts the performance compared to our baseline method img-gLSTM. e2e-gLSTM achieves better performance than emb-gLSTM, which empirically demonstrates that the non-linear transformation of pre-trained CNN features can capture more semantic information than CCA-based linear transformation. e2e-gLSTM also leads NeuralTalk2 by 1.0%, 0.4% and 3.0% in Bleu_4, METEOR and CIDEr respectively, showing the effectiveness of adding semantic guidance information in the traditional encoder-decoder image captioning model.
We further compare the text-conditional methods with state-of-the-art attention-based methods. For 1-gram word-conditional method, the attention on the image feature guidance is merely determined by the previously generated word. Apparently, this results in semantic information loss. This is consistent with our result that 1-gram word-conditional method gets lower scores in some of the metrics, such as Bleu_4 and CIDEr, than e2e-gLSTM. We found a similar problem for 2-gram word-conditional method, though their performances are still on par with or better than state-of-the-art attention-based methods, such as Hard-Attention and ATT-FCN. We then upgrade the word-conditional model to the sentence-conditional one which leads to around 1% improvement in Bleu and CIDEr compared to 2-gram word-conditional method.

4.4 Qualitative Results

We sample some testing images from MSCOCO and use different generators to generate the image caption (see Fig. 7). For the three images in the first column, all the caption generator show well performance. We also show three advantaged results of the sentence-conditional method in the second column. Our attention model can better capture details in the target image, such as the surfboards in the first image and the river in the third image. Also, the captions generated by sentence-conditional method are more human-like than those by NeuralTalk2 and e2e-gLSTM, as in the third image, the caption generated by sentence-conditional method directly uses “riding elephants” instead of “riding on the backs of elephants”, as generated by NeuralTalk2.

5 Conclusion

In this paper, we propose a semantic attention mechanism for image caption generation, called text-conditional semantic attention. We also improve the existing gLSTM framework by adding CNN fine-tuning, which provides better textual-related image features for our proposed attention model. We show in our experiments that the proposed methods significantly improve the baseline method and outperform state-of-the-art methods. However, there are several critical issues in our method. The first one is overfitting. As the Google NIC paper [5] pointed out, we cannot access enough training samples, even for the relatively huge dataset MSCOCO. One possible solution is combining weakly annotated images [23] with current dataset. The second issue is time complexity. We fine-tune the VGG net, which requires extra training time. Even though we use transfer learning techniques to make the time gap shrink, efforts should still be put in building fast learning systems.
References

[1] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. In Proceedings of the 24th CVPR. Citeseer, 2011.

[2] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632, 2014.

[3] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Computer Vision–ECCV 2010, pages 15–29. Springer, 2010.

[4] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. Generalizing image captions for image-text parallel corpus. In ACL (2), pages 790–796. Citeseer, 2013.

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru 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.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112, 2014.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[8] Andrej Karpathy and Li 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.

[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044, 2015.

[10] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding long-short term memory for image caption generation. arXiv preprint arXiv:1509.04942, 2015.

[11] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. International journal of computer vision, 106(2):210–233, 2014.

[12] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiiebo Luo. Image captioning with semantic attention. arXiv preprint arXiv:1603.03925, 2016.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014, pages 740–755. Springer, 2014.

[14] Andrej Karpathy. [https://github.com/karpathy/neuraltalk2] 2015.

[15] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 129–136, 2011.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[17] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. arXiv preprint arXiv:1503.04069, 2015.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[19] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In Advances in Neural Information Processing Systems, pages 2845–2853, 2015.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[21] Karen Simonyan and Andrew Zisserman. http://www.robots.ox.ac.uk/~vgg/research/very_deep, 2014.

[22] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48. ACM, 2009.

[23] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In Computer Vision–ECCV 2014, pages 529–545. Springer, 2014.