A Review of Deep Visual-Semantic Alignments for Generating Image Descriptions

Anuja Tayal
University of Illinois
Chicago, IL
atayal4@uic.edu

Dian Jia
University of Illinois
Chicago, IL
djia7@uic.edu

Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions.IEEE Transaction of Pattern Analysis and Machine Intelligence(2017), 664–676.

ABSTRACT
Image Description is a very popular research field, which has made much progress in recent years. In this field, Deep Visual-Segmentic Alignments for Generating Image Descriptions is an important paper [12]. Karpathy et al. proposed a model that can generate descriptions of different regions in images. Their model innovatively combines Convolutional Neural Networks and Recurrent Neural Networks and the results outperform previous state-of-art methods. In this project, our goal is to review the paper in terms of motivation, related work, contribution, methods, results and along with that discuss some of the advantages and disadvantages of the paper. Finally, we will discuss the some ideas on how to extend the paper and suggest some references that paper should have cited.

INTRODUCTION
We use natural language in our daily life. When we look at an image, we can easily describe intricate details of the image with language. But such a simple task for humans is extremely difficult for computers With the rise of people's interest in identifying and describing images using computers, generating image description is an emerging field integrating computer vision, natural language processing, and machine learning. The task of Image description is to generate a summary describing the details of images. Early image description work is more like Image Classification. Pictures need to be labeled by a series of categories and described by closed vocabularies. With the development of deep learning, many deep learning approaches have been proposed to solve image description problems. The authors in [12] take a step ahead and describe regions of the image.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.XXXXXXX

PROBLEM ADDRESSED
Approaches proposed by Kulkarni et al.[16] and Farhadi et al.[7] rely on some image processing operators and sentence templates like CRF. When the details in the image become rich, these hard-coded methods can not perform well. To fix this problem, they proposed a method that can generate more dense descriptions and has better performance. The advantage or core insight of this model designed by Karpathy et al.[12] is that it treats the sentences as weak labels of different regions of images, then generates the relation between them. Thus the model will not be limited by hard-coded templates and will have better performance than other state of the art models[]. Apart from that, the authors of [12] want to design the model that learns the end to end embedding which aligns an image region with its description. Additionally, the authors learn that in the datasets available on image captioning, entities present in the image description does not point precisely to the region of the image it is present.

CONTRIBUTION
There are many claims that the authors of [12] make. The authors designed and developed a simple, not too complex, deep neural network that generates a higher level description of image regions. In particular, they combined the embeddings of image space and description space to form a multimodal embedding space by aligning the sentences and regions of an image they express. The authors in order to validate their Deep Visual Semantic Alignment model performed experiments on the Flickr8K, Flickr30K, and MSCOCO datasets. They outperformed the baselines on full images and a new dataset consisting of region level annotations.
The authors developed a multimodal RNN model to generate the descriptions by providing an image as input. With experiments, the authors validated that sentences generated by their model performed way better by producing realistic predictions with respect to their counterparts. The authors used this model to train and assess it on a dataset of annotations of regions.
Finally, the authors conducted a large scale analysis of the above RNN model on the Visual Genome dataset consisting of 4.1M descriptions and contrast the difference between image and region based captions

RELATED WORK
The literature section of the paper is carefully divided into subsections based on previous works done by the authors on basis of similar work done before, differences in annotation, generating descriptions and model complexity.

Image Captioning
Several previous works [29], [7], [20] on image captioning describe images using words to form meaning sentences, Karpathy et al [12] compare the difference in the style of the descriptions. The authors in [7], [20] fetch the most relevant annotation, while in [17], the authors combine training annotation phrases to output to form sentence. Few other works [16] [7] generate captions based on fixed template and fixed set of words. The authors feel that restricting the length of description is unnecessary. On the contrary, Karpathy et al. [12] claim to make descriptions which are not restricted in length.

Model Complexity
The authors cited [13],[14] as the approach used is similar to that of [12]. In somewhat similar work [29],[6] the authors relaxed the finite length description of image to provide detailed length description but their model is complex as opposed to the Karpathy et al [12] who developed a simpler model which generates more significant and relevant words.

Similar work
The authors of [1], [24] also proposed an intermodal embedding model on images and word to find relation between region of image and text present but they are more like region classification. On the other hand, Karpathy et al[12] text description of regions is more detailed.

Objective Function
The authors cited [8], [9] which studied the problem of understanding the image as a whole from the description as opposed to the authors of [12] whose main focus is on region captioning.

Missing References
The above paper makes use of image and word embeddings to learn inter modal or multimodal embedding. Multimodal Embedding being a recent approach being explored extensively, we feel they should have at least referenced one paper on multimodal embedding [19] to give the readers a general idea about the topic.
The papers cited by the authors are novel papers which made a huge impact on the field of image captioning, we feel that similar work is also being proposed in video datasets [28], in which the authors make use of convolution and recurrent neural networks to transcribe videos into sentences. Citing papers from different areas of the field expand the application scenario and gives the readers a broad view of the applications of the idea.

METHODS
In this section we will illustrate the methodology used to generate descriptions of different regions of images. To achieve the goal, we need to divide the method into the following steps (Figure 1). The first step is to build a model that can align sentence snippets and visual regions of images. The key insight of the first step is that the sentence descriptions corresponding to an image written by people should be related to specific but unknown regions in the image. The second step is to feed the data generated by the first step to training a model for generating image descriptions. The training input should be images and their corresponding descriptions.

Aligning Sentences and Image Regions
The first step also needs to be decomposed into several small tasks. First of all this paper represents images and sentences using same-dimension vectors. Then using the loss function, words of the sentences can be related to a particular image. After using Markov Random Field (MRF), sentence snippets can be formed out of scattered words. The details are shown as below:

Representing Images
A Classic Region-based Convolutional Neural Network(R-CNN)[3] commonly used for object detection is adapted to represent images. RCNN has been trained on ImageNet and finetuned. Top 19 classes along with the original image is further used to compute the representation based on the pixel of the bounding boxes. Finally, every image will be represented as 20 h-dimension vectors where h ranges from 1000-1600.

Representing Sentences
The paper uses Bidirectional Recurrent Neural Network(BRNN)[22] to represent the sentences. The input of BRNN is a sequence of N words that are encoded in 1-of-k representations. Then BRNN will transform the words into h-dimensional vectors, which have the same dimension as the representations of images. Unlike previous methods to generate representations, which can not leverage the context information, BRNN enriches the representations using variably-sized context around the word. Thus the h-dimensional results contain information that includes both the word and its surrounding context.

Alignment Objective
After representing both the images and the sentences, each word snippet needs to be aligned to the corresponding image. In the paper, the authors proposed an image-sentence score composed of individual region-word scores[13]. When the corresponding words support the images with confidence, the image-sentence score will have a higher matching score. The image-sentence score is showed as below:

$$S_{kl} = \sum_{t \in g_l} \sum_{i \in gk} \max\left(0, v_i^T s_t\right)$$

The authors also proposed a simplified version as below:

$$S_{kl} = \sum_{t \in g_l} max_{i \in g_k} v_i^T s_t \tag{1}$$

Using the score, the objective function was introduced:

$$\mathscr{C}(\theta) = \sum_k [\underbrace{\sum_l \max\left(0, S_{kl} - S_{kk} + 1\right)}_{\text{rank images}} + \underbrace{\sum_l \max\left(0, S_{lk} - S_{kk} + 1\right)}_{\text{rank sentences}}].$$

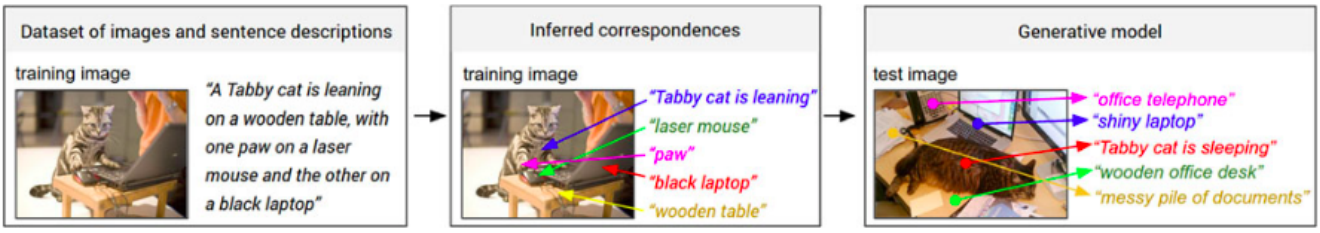

Figure 1. Model Approach

Decoding Text Snippets

The objective function proposed will form the intermediate training dataset to generate region descriptions of images. But before the training data was given as input into the RNN model, another step is required. When the intermediate data is generated, the regions in the dataset might get annotated with only a single word. To fix the problem, the authors use the latent variables of Markov Random Field(MRF) to make contiguous words match to one region in which the neighboring words through binary interaction will encourage words to align to the same regions.

Generating Descriptions

After aligning the images and text snippets, the next task is to generate descriptions. The input of the task is the set of images and their corresponding descriptions, including full images and sentences or regions and their corresponding test snippets. The key challenge of the task is to predict sequences of words that are not fixed-sized. To be more specific, the training data includes the image pixel and a set of vectors, following the process as below:

$$b_v = W_{hi}\left[CNN_{\theta_c}(I)\right]$$

$$h_t = f\left(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1)\odot b_v\right)$$

$$y_t = \text{softmax}\left(W_{oh}h_t + b_o\right)$$

The output y can be seen as the log probability of words. The loss function of the training process is to maximize the log probability assigned to the input visual regions.

EXPERIMENTAL RESULTS

In this section, we will briefly describe the experiments performed on the model proposed, and highlight on some of the important observations made on the results by the author. The authors performed intrinsic and extrinsic evaluation of the model by taking use of Recall, BLEU and METEOR performance measures. The model was not only compared with previous models, but with complex models which performed better than the model proposed.

Dataset Used

Flickr8K[11] with 8k images, Flickr30K[30] with 31k images, and MSCOCO[2] with 123k images were used to train the model. 5 sentences from each dataset were annotated using Amazon Mechanical Turk (AMT). Visual Genome Dataset[23] had 94k images which were used to generate region captions trained on region level captions. All datasets were split into training, validation, and test set and preprocessed. For preprocessing, all the sentences were converted to lowercase and filtered to occur at least 5 times.

Image Sentence Alignment

To evaluate the quality of the sentence retrieved from aligning images to sentences, the authors used the measure of Recall@K which is defined as a fraction of times an item was found within top K. A set of images and sentences were retained and sentences were sorted based on the score S_{kl}. The model had a high recall rate when compared against models [25, 13] with similar loss.

Results were not only compared with other models but also with their previous model [13]. The previous model though had different word embedding and loss; the loss was reimplemented by the authors in order to compare the model performance. They show that BRNN [22] which uses raw words gives better performance than dependency tree relations. Moreover, to show that cost function also improves performance, they replaced BRNN with dependency tree relations and showed that the simpler cost function also plays a critical role in model performance.

Qualitative

The model could align not-so-frequent objects with text such as *accordion* and is sensitive to compound words and modifiers (*red bus, yellow bus*). However, their model does not bound image regions distinctively when counting is involved. In addition, as the model learns by aligning visual appearance to text, its performance declines when the frequency of objects observed in the training data set reduces.

Discriminative Words and Regions

The better caption of the image is achieved with a higher value of visually discriminative words as compared to stop words due to the Alignment Equation (Equation 1). To illustrate this, the top 20 words with the highest and lowest magnitudes were tabulated. Apart from this, regions with discriminative entities were considered salient features and those regions were mapped to sentences, and it was demonstrated by a set of regions having high vector magnitude.

Image Caption Description

To evaluate Multimodal RNN used for generating image captions, BLEU [21], METEOR [4], CIDEr[27] scores were calculated on VGGNet [23] image features.

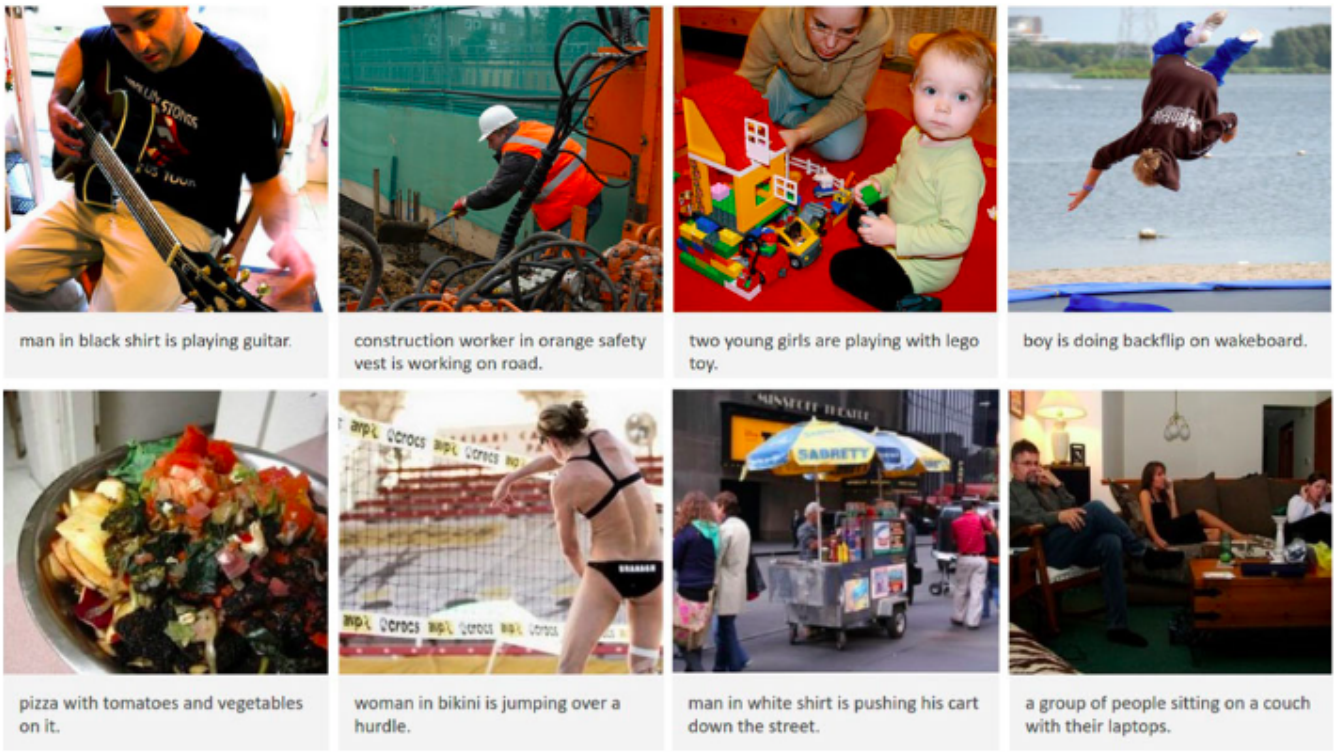

Figure 2. Image Caption Generated by Model

Aligning Sentences and Image Regions

Qualitative

The model generated reasonable image captions whose generated captions could be found in the training data either independently or in different occurrences which the model composed to form a single sentence. Among the whole test set, the authors illustrated a set of 8 images with their generated captions; out of these 8, 2 are failures cases for example (*woman in bikini is jumping over a hurdle.*). The authors reasoned that the model learns the correct essence of the depicted scene but misses a few intricate details. In addition, the model learns certain templates Figure 2 *<noun> in <noun> is <verb> in <noun>* due to which the model incorrectly outputs *jumping over hurdle* instead of *playing volleyball.*

Comparison with other models

The Multimodal RNN model which surpasses retrieval baselines was compared with many different models [5, 29]. Comparing against the nearest-neighbor model [5] on the MSCOCO dataset, the authors showed that their parametric approach is much faster with a higher BLEU score. On other hand, the complex model [29] performs better than the authors' model. However, in comparison, the model the authors proposes chooses simplicity with less speed with a little overhead on performance.

Region Caption Description

The Multimodal RNN was trained on image regions and their corresponding text to generate region captions. The model was tested against the dataset generated via Amazon Mechanical Turk. The turkers annotated five regions on 200 images with each image annotated by 9 people generating a total of 9000 texts with an average length of 2.3. Text generated by the annotators is more extensive and has fine salient feature otherwise absent from full image captions.

The text generated of the image regions was accurate and corresponded to the training dataset. The region-level model surpassed the performance by the nearest neighbor model[5] and full-frame model used with smaller regions.

Region Caption Trained on Region Level Cations

The Multimodal RNN model consisting of LSTM[10] was trained with VGG-16[23] and fine-tuned on CNN with Visual Genome[15] dataset. Comparing with their previous model, the authors validated that METEOR performance is better if the model is trained independently on the region dataset instead of training on image caption dataset and fine tuning on regions. The METEOR metric was chosen as it correlates more with human captions. With this experiment, the authors concluded that visual and semantic statistics are different in region and image captions.

ANALYSIS OF PRESENTED APPROACH

In this section, we will critically analyse the paper to discuss the advantages and disadvantages of the presented approach and based on our analysis present some of the key extensions.

Advantage

The authors proposed a simple and novel approach to generate descriptive captions of the regions. It can be highly noted that with no independent region dataset available, the authors were able to leverage the pre-existing image caption dataset, and propose the model. Moreover, the captions generated are not restricted in length and have an open vocabulary of words. The alignment model between images and words is a novel technique taking the use of multimodal embedding. In addition, the model proposed is bisected so properly that it is easily understood by fellow researchers, reproducible, and hence used as a baseline model or extended or compared with. While inferring the correspondence between image caption and regions, the author innovatively took into account neighboring words and regions.

Apart from this, the authors performed an intrinsic and extrinsic evaluation of the model. The model was not only compared with previous models, but with complex models which performed better than the model proposed. The authors through their keen observations concluded that the visually discriminative words are given more importance than other words. They also highlighted the drawbacks which could be improved in further versions.

Disadvantage

Although the method proposed to generate descriptions of different regions is strong, there are still some limitations of their model. In the paper, the author summarizes some defects of their proposed model. First, the input pixels array must have a fixed resolution, which might lose some important information. Secondly, the image information fed into the RNN as a bias interaction, which might be too simpler and less expressive than multiplicative interactions[26]. Thirdly, every region is isolated in this model, which causes inefficiency of computation because the RNN must forward every separated region to generate descriptions. The authors also mentioned that their approach consists of several separated models. They used R-CNN to detect objects and generate regions, and used a CNN which is pre-trained on ImageNet to represent the regions along with the whole image. Then they used Bidirectional Recurrent Neural Network to represent words. Finally, there are another two modeling steps to generate the descriptions of regions and these models are trained on different datasets. A direct end-to-end method from an image-sentence dataset to region-level descriptions is still a problem and needs to be further explored. Moreover, when performing evaluation, the generated alignment consisted of descriptive words from the caption only, the enclosed region did not considered bounding, and generated captions were not 100% accurate.

In addition to these shortcomings proposed by the authors themselves, the model still has other limitations and can be improved. The first limitation is the limitation of datasets. The datasets adopted by this paper like Fliker8K[11], FLicker30K [30], and MSCOCO[2] have a problem that these datasets tend to focus on the salient parts of images. The descriptions of these datasets represent the whole image, and often only a few objects present in the image[15]. Even if they divide the images into several regions and align word snippets to each region. It is still not enough to describe every object in the image. In further research, a more complex dataset with low-level descriptions[15] needs to be collected.

Apart from that, the network architecture can be improved further. The authors used a normal RNN to generate descriptions, which can be improved by using a more complex model. Lipton and Berkowitz[18] claim that LSTMs can outperform normal RNNs in this case. Meanwhile, they also mentioned that using random initializations to learn the word representation can outperform the word2vec embeddings, which were adopted by Karpathy and Fei-Fei. A possible reason for this is that the attributes of word2vec that can cluster similar words into embedding space do not apply in image description problems.

EXTENTION

There are many ways to extend the research because there are many stronger and more powerful methods or networks available after the paper was published. A stronger dataset is also an important factor to achieve a better result. As we mentioned before, in most datasets, the descriptions of images are too simple and not enough to describe every region in the images. Therefore, to extend this research, collecting a better dataset might be the best way to improve the performance. However, collecting a new dataset is too difficult for researchers, which may take years to write descriptions for images. Therefore collecting a new dataset may not be a good option for us.

The first step is to encode the images and the words. The authors adapt R-CNN to detect objects. Although nowadays we have some faster and more effective methods than normal R-CNN like faster R-CNN, we think there is no need to substitute R-CNN with faster R-CNN hence it does not affect the performance of image description generating. The encoding model of words uses word2vec weights to represent words. As we mentioned before, it can be substituted by random initializations which will reduce the computation and achieve better performance.

After we align the regions and word snippets, the next step is to generate the descriptions. First, the regions are feed into VGGNet to generate the bias vector at the first iteration. However, VGGNet can only extract one region's feature. The information of each region is isolated. It may be better if we add the information of other regions. Because most of the time, regions of an image often have interaction, by adding other information, the model may produce better descriptions. Therefore, we can adapt the attention model to add some information of other regions, which may lead to a better result. Then we will add the image cotext as bias into the first iteration of the RNN. We can try to add the image context into each iteration to see if it can have better results.

REFERENCES

[1] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. 2003. Matching Words and Pictures. *Journal of Machine Learning Research* (2003), 1107–1135.

[2] Xinlei Chen and C. Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption

generation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2422–2431.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[4] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 376–380.

[5] Jacob Devlin, Saurabh Gupta, Ross B. Girshick, Margaret Mitchell, and C. Lawrence Zitnick. 2015. Exploring Nearest Neighbor Approaches for Image Captioning. *ArXiv* abs/1505.04467 (2015). Authors compared the model with this paper and is not peer reviewed in any conference or journal, so cited arxiv.

[6] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2625–2634.

[7] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every Picture Tells a Story: Generating Sentences from Images. In *Computer Vision – ECCV 2010*. 15–29.

[8] Sanja Fidler, Abhishek Sharma, and Raquel Urtasun. 2013. A Sentence Is Worth a Thousand Pixels. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 1995–2002.

[9] Stephen Gould, Richard Fulton, and Daphne Koller. 2009. Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th International Conference on Computer Vision*. 1–8.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* (1997), 1735–1780.

[11] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research* (2013), 853–899.

[12] Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transaction of Pattern Analysis and Machine Intelligence* (2017), 664–676.

[13] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*. 1889–1897.

[14] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal Neural Language Models. In *Proceedings of the 31st International Conference on Machine Learning*. PMLR, 595–603.

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* (2017), 32–73.

[16] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. BabyTalk: Understanding and Generating Simple Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013), 2891–2903.

[17] Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2012. Collective Generation of Natural Image Descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 359–368.

[18] Zachary Lipton. 2015. A Critical Review of Recurrent Neural Networks for Sequence Learning. (05 2015).

[19] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *Proceedings of the 2011 Conference International Conference on Machine Learning(ICML)*. 689–696.

[20] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*, Vol. 24.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*. Association for Computational Linguistics, 311–318.

[22] M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* (1997), 2673–2681.

[23] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015*.

[24] Richard Socher and Li Fei-Fei. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 966–973.

[25] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics* (2014), 207–218.

[26] Ilya Sutskever, James Martens, and Geoffrey E. Hinton. 2011. Generating Text with Recurrent Neural Networks. In *ICML*. 1017–1024.

[27] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[28] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 1494–1504.

[29] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 3156–3164.

[30] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.