International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 10 | Oct -2016 www.irjet.net p-ISSN: 2395-0072

Caption Generation for Images Using Deep Multimodal Neural Network

Ruturaj R. Nene[1]

[1]*Dept. Of Computer Engineering, VESIT, University of Mumbai, Maharashtra, India*

Abstract - *This paper presents a model that generates captions or descriptions for images with the help of multimodal neural networks. The model consists of two sub-networks a convolution neural network that is utilized to extract the image characteristics and a recurrent neural network for the descriptions. These sub-networks are then aligned through a multimodal embedding to form the whole model. .As compared to previous models such as temporal convolution, the recurrent models are considered to be doubly deep. These recurrent neural networks can handle variable input/output. Thus, they can directly map a variable input (e.g. video) with a variable output (e.g. a caption/description, natural language text). The results distinctly show its advantage over state-of-the-art models that are used for generating descriptions or captioning images.*

Key Words – caption generation, deep multimodal neural network, natural language, recurrent neural networks, semantics

1. INTRODUCTION

A human can describe a visual scene with huge amount of details by just having a glance at it [3, 14] .When can a machine describe an image? It would be possible only when it could generate a new caption that summarizes the salient features of the image. These features may include objects that are present, their characteristics, or their relations with other objects. Determining these features requires not just the knowledge of the contents of the image, but also it is necessary to determine the various aspects of the visual scene that may be new or rare using ones commonsense. [20, 31, 29] Apart from correctly recognizing the contents in images it is also necessary to incorporate the knowledge of the interactions and the spatial relationships between objects. Along with this information, the description of the image also needs to be relevant and grammatically correct. [19]

This work is an attempt towards achieving the goal of generating apt descriptions of images. The primary challenge towards this goal is in the design of a model that is rich enough to simultaneously reason about contents of images and their representation in the domain of natural language. The model should be free of assumptions about specific hard-coded templates, rules or categories and instead rely on learning from the training data. The second, practical challenge is that datasets of image captions are available in large quantities on the internet [21, 16, 24], but these descriptions multiplex mentions of several entities whose locations in the images are unknown. To overcome these problems the deep neural network model is modeled in such a way that it infers the latent alignment between segments of sentences and the region of the image that they describe. The model associates the two modalities through a common, multimodal embedding space and a structured objective. [14]

The proposed multimodal Recurrent Neural Networks (m-RNN) model carries out two tasks. Firstly, it generates novel descriptions for the images. Secondly it also carries out the task of image and sentence retrieval. This m-RNN architecture contains an image part, a language model part and a multimodal part. The image part uses a deep Convolution Neural Network (CNN) [8] to extract the image characteristics while the language model part learns the dense feature embedding for each word in the dictionary and maintains the semantic temporal context in recurrent layers. The structured objective or the multimodal part connects the language model and the deep CNN together by a one-layer representation. This m-RNN model is learned using a perplexity based cost function. The model parameters are updated by back-propagating the error to the three layers of the m-RNN model. In the experiments, the model is validated on two benchmark datasets: Flickr 8K [12], and Flickr 30K [9]. The model has potential extensions and it can further be improved by integrating more powerful neural networks [7].

2. RELATED WORK

The deep neural network structure develops rapidly in recent years in both the field of computer vision and natural language. Many approaches have been developed for high-level representation of images and words [14]. For computer vision, Krizhevsky et. al [8] proposed a deep convolutional neural networks having 8 layers (called as AlexNet) which was used for image classification tasks and outperformed older methods by a large margin. Girshick et. al [13] also proposed an object detection framework based on AlexNet. The Recurrent Neural Network has been extensively used for many tasks, such as speech recognition and word embedding learning [7, 10, 11].

The task of describing images with sentences has also been explored. A number of approaches pose the task as a retrieval problem, where the most compatible annotation in the training set is transferred to a test image or where training annotations are broken up and stitched together [21, 30, 17, 4, 26, 27, 22]. There are approaches that use

fixed templates based on the image content or generative grammars [5, 25] to generate image captions, however such approaches provide quite unoriginal outputs. Most closely related to the tasks and methods of the model, Kiros et al. [1] developed a log bilinear model that can generate full sentence descriptions for images, but their model uses a fixed window context while multimodal Recurrent Neural Network (RNN) model uses the recurrent architecture to store the temporal context, which thus allows variable context length. Multiple closely related preprints appeared on Arxiv during the submission of this work, some of which also use RNNs to generate image descriptions [7, 28, 18, 19, 29, 6]. The m-RNN is simpler compared to most of these models but also suffers in performance. This comparison is quantified in the experiments. [14]

3. PROPOSED MODEL ARCHITECTURE

The main goal of this paper is to generate novel captions for images. The proposed model combines two different neural networks, RNN and CNN, to extract the image feature and the combine with the semantic information for a successful generation of a caption for the various images. [2]

3.1 Deep Neural Networks

According to recent study, the deeper the neural network, it turns out to be more expressive which is why it is successful in computer vision. And there are some popular deep neural networks, such as deep CNNs, Deep Belief Networks (DBNs), RNNs and so on. CNN is basically a simple neural network that is extended across space, hence it is said to be deep in space. On the other hand, RNN is extended through time and hence is said to be deep in time. The proposed model uses both the CNN and the RNN due to powerful ability of CNN to extract expressive content from images.

3.1.1 Deep Convolution Neural Network:

Traditional neural networks usually consist of three layers, the input layer, the hidden layer and the output layer. Such a neural net is easy to train and understand; however, when it comes to convoluted problems the traditional neural net fails. A way to overcome this problem is to add multiple hidden layers to the neural network, which allows it to represent complex functions with fewer parameters. Also, it provides excellent ability of feature learning. Hence they proposed the deep CNNs. CNNs have multiple convolution layers and pooling layers after the input layer. These convolution layers could reduce the noise and improve the signal while the pooling layers could down sample the feature maps. This neural network then connects to full-connected layers to form a feature vector which represents the input image.

3.1.2 Deep Recurrent Neural Network:

One of the drawbacks of the traditional neural network was that it couldn't deal with sequential data. The RNN overcame this drawback. In a RNN the hidden layers of the network keeps the memory of previous inputs. This is due to the recurrent connections in the hidden layers of the neural net. This memory of the sequential inputs then influences the output of the neural network. The computations in RNN are in the following recurrent equations:

$$h_a = g(W_{xh}x_a + W_{hh}h_{a-1}) \quad (1)$$

$$y_a = g(W_{hy}h_a) \quad (2)$$

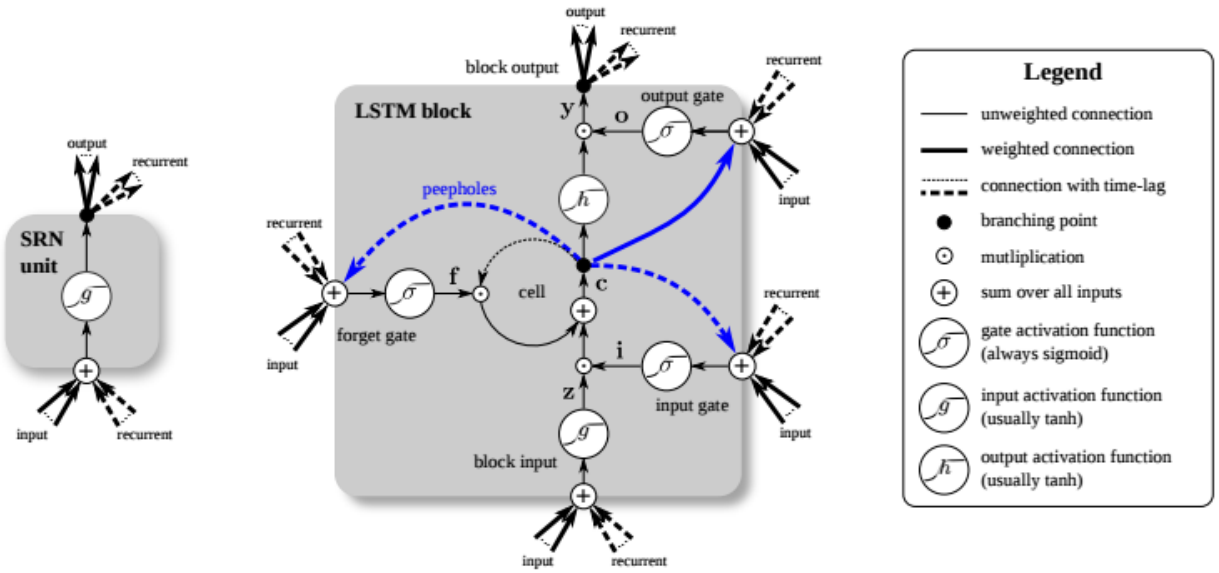

Figure 1: Long-Short Term Memory (LSTM) network

RNN's proficiency in handling sequential data and the

Where $x_a, y_a \in R^N$, y_a represent the input, hidden state and output of RNN at time a, respectively. And $g(\cdot)$ is a non-linear activation function, such as a sigmoid function or hyperbolic tangent function. The weight matrices between the layers of the RNN are represented by symbol W_{xy}, W_{hy}, W_{hh}, where *xy, hy, hh* represent the weights between the input layer-output layer, hidden layer-output layer and hidden layer-hidden layer respectively. In spite of being successfully used in the tasks of natural language processing, speech recognition and so on, RNN could not resolve the problem of "long-term dependencies" of states. This was due to the gradient vanishing and gradient exploding problem in the training stage of RNN. The solution to this problem is the use of Long-Short Term Memory networks (LSTMs). The LSTMs major components are the three "gates" and a memory cell, as we can see in Fig. 1. Each gate is a combination of a sigmoid function and a multiplication operation and it is a gateway to optionally let information through. The non-linear sigmoid function provides an output in the [0,1] range, implying how much information could go through. 0 means letting nothing go through the gate and 1 means allowing everything to go through. These three gates work in unison to control the final output of the LSTM unit. In another terms, the input gate determines whether to allow the new input to the cell, the forget gate determines whether it should forget the current cell value and the output gate controls whether to output the cell value. The definition of the gates and the update of cell information at time t are as follows:

$$i_a = \sigma(W_{xi}x_a + W_{hi}h_{a-1}) \quad (3)$$

$$f_a = \sigma(W_{xf}x_a + W_{hf}h_{a-1}) \quad (4)$$

$$o_a = \sigma(W_{xo}x_a + W_{ho}h_{a-1}) \quad (5)$$

$$g_a = \tanh(W_{xc}x_a + W_{hc}h_{a-1}) \quad (6)$$

$$c_a = f_a \circ c_{a-1} + i_a \circ g_a \quad (7)$$

$$h_a = o_a \circ c_a \quad (8)$$

Where i_a, f_a, o_a, c_a represent the outputs of input gate *i,* forget gate *f,* output gate *o* and the cell *c* at time *a,* respectively. And $\sigma(\cdot), \tanh(\cdot)$ are the sigmoid function and hyperbolic tangent function, $\circ$ represents the product with a gate value, and W represents the matrices of the parameters to train. These three control gates make it possible to train the LSTMs by dealing with the problems of exploding and vanishing gradients. [2]

3.2 Multimodal Recurrent Neural Network

In order to generate a description for an image, it is important to consider maximizing the probability of the correct descriptions for the given image info. A cost function

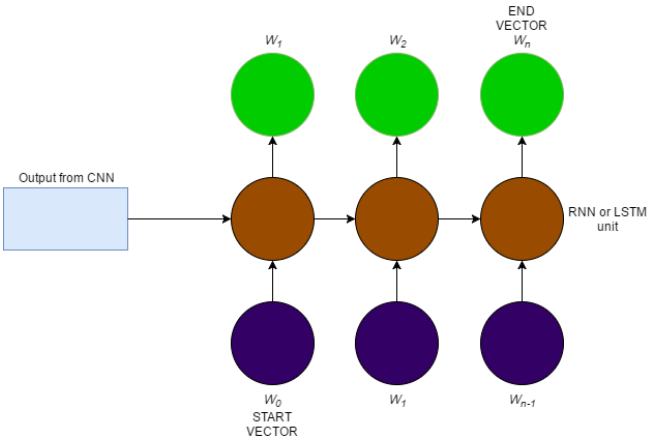

Figure 2: Training stage of Multimodal Neural Network

is adopted based on the probability of the sentences in the training set suiting their corresponding image. For each sentence corresponding to the image we can have the following formulation.

$$\gamma(w_{1:N}|I) = \max\left(\sum_{(I,S)}\left(\frac{1}{N}\sum_{n=1}^{N}\log P(w_n|w_{1:N-1}, I)\right)\right) \quad (9)$$

Where N is the length of the word sequence, I are the image pixels, $\gamma(w_{1:N}|I)$ is the perplexity of the sentence and $(I, S)$is the image-caption pair. $P(w_n|w_{1:N-1}, I)$ is the probability of generating the word w_n considering the previous words$w_{1:N-1}$ and image I. The inner summation is to sum the log probabilities of words in a single sentence while the outer is to sum all image-caption pairs in training data. Various CNNs can be utilized for the extraction of the image information while for combining the textual information a RNN or LSTMs are used. Thus this model can be said to be deep in time as well as space [2, 7].

3.2.1 Training Stage

The model takes an input set that consists of images along with their textual descriptions. These images can be full images with descriptions or part of an image with the textual info. The m-RNN takes the image pixels I and a sequence of input vectors $(x_1 \ldots\ldots x_n)$. It produces a sequence of outputs $(y_1 \ldots\ldots y_n)$ by iterating the recurrence relation for$n = 1to\ N$. The training data set consists of pair of image-captions (I, S) which is used to train the model to predict a novel description when provided with an image. Thus the RNNs or LSTMs are used to train the neural network with both images and their corresponding image descriptions. The training stage carries out the above procedure by iterating the following formulations for $n = 1to\ N$:

Table 1: Results of Flickr8K experiments. R@K stands for Recall@K.

Model	Image Annotation				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random ranking	0.1	0.5	1.2	631	0.1	0.5	1.0	500
SDT-RNN [6]	4.5	18.0	28.6	32	6.1	18.5	29.0	29
SDT-avg-RNN [30]	6.0	22.7	34.0	23	6.6	21.6	31.7	25
DeViSE [15]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
m-RNN [7]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
DeFrag-decaf [23]	5.9	19.2	27.3	34	5.2	17.6	26.5	32
DeFrag-RNN [23]	12.6	32.9	44.0	14	9.7	29.6	42.5	15
Proposed Model	13.7	34.8	40.5	12	10.6	32.0	42.6	13

$$b_0 = CNN(I), \quad (10)$$

$$h_a = g(W_1 w_a + W_2 h_{a-1} + b_0 \cdot I(a = 1)), \quad (11)$$

$$p(w_{a+1}) = softmax(W_3 h_a) \quad (12)$$

Where b_0 is the image context vector, $g(\cdot)$ is the activation function of the RNN or LSTM. The image vector b_0 is provided to the RNN only for the first iteration. Each training step receives the input word x_a which is combined with the previous hidden state h_{a-1} to provide the hidden state h_a at time=a. In simple words the input word is combined with the previous context to predict the next word. The h_a is passed through the softmax function which provides the probability distribution for the next word. The word with the highest probability is considered to be the predicted word, which further acts as the input for the next iteration. The loop continues until the neural net predicts the END vector.

3.2.2 Testing Stage:

This stage is similar to the training stage. The image representation is provided as input to the trained model and the model computes the word over each time step. The iteration continues until the end token is generated and all the predicted words form the description of the image [2, 14].

4. EXPERIMENTS

This section summarizes the datasets used in the experiments. These experiments are evaluated with the help of the evaluation metrics. Finally the results of the proposed model are compared with the results of other experiments.

4.1 Datasets

The model is tested on two datasets Flickr8K [12], and Flickr30K [9]. The Flickr8K dataset consists of 8,000 images that are extracted from Flickr. Each image is independently annotated up to 5 sentence annotations. The standard separation provided by the dataset is adopted. For this dataset 6000 images are used for training while 1000 images are used for testing and validation each. The Flickr8K dataset is extended to form the Flickr30K dataset. Similar to the Flickr8K dataset each image is also provided with five annotated sentences. It consists of around 150,000 crowd-sourced captions describing 30,000 images. The grammar and style for the annotations of this dataset is similar to Flickr8K. This dataset uses the same separation as the previous dataset. Flick8K as well as Flickr30K dataset are often used for the image-sentence retrieval tasks. [7, 19, 14]

4.2 Evaluation Metrics

The same evaluation metrics is adopted for Flickr8K and Flickr30K datasets, as adopted by previous models [30, 23, 15] for image as well as sentence retrieval. [7] The evaluation was performed using Recall@K, which provides the mean number of images that had correct caption ranked within the top-K retrieved results (and vice-versa for sentences). In other words it provides the fraction of times the top K results had a correct item. [19] They used R@K (K = 1, 5, 10) as the measurements, which are the recall rates of the first retrieved groundtruth images (image retrieval task) or sentences (sentence retrieval task). Higher R@K corresponds to a better retrieval performance. The R@K with small values of K are quite important as top-ranked retrieved results are the most sought out for. The Med r is another score that is used, which shows the median rank of the first retrieved images or sentences. A better performance can also be provided by lower Med r. [7]

Table 2: Results of Flickr30K experiments. R@K stands for Recall@K.

	Image Annotation				Image Retrieval			
Model	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random ranking	0.1	0.5	1.2	631	0.1	0.5	1.0	500
SDT-RNN [30]	9.6	29.8	41.1	16	8.9	29.8	41.1	16
DeViSE [15]	4.5	18.1	29.2	26	6.7	21.9	32.7	25
m-RNN [7]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
DeFrag [23]	14.2	37.7	51.3	10	10.2	30.8	44.2	14
DeFrag-Finetune CNN [23]	16.4	40.2	54.7	8	10.3	31.4	44.5	13
Proposed Model	13.9	37.5	49.9	9	11.0	32.0	43.2	13

The results of the proposed model were compared with the following models from previous work. The deep visual semantic embedding model (DeViSE) [15] was proposed as a way of performing zero-shot object recognition and was used as a baseline by [23]. In this model, sentences are represented as the mean of their word embeddings and the objective function optimized matches ours. The semantic dependency tree recursive neural network (SDT-RNN) [30] is used to learn sentence representations for embedding into a joint image-sentence space. The same objective is used. Deep fragment embeddings (DeFrag) [23] were proposed as an alternative to embedding full-frame image features and take advantage of object detections from the R-CNN [13] detector. Descriptions are represented as a bag of dependency parses. Their objective incorporates both a global and fragment objectives, for which their global objective matches ours. The multimodal recurrent neural network (m-RNN) [7] is a recently proposed method that uses perplexity as a bridge between modalities, as first introduced by [1]. Unlike all other methods, the m-RNN does not use a ranking loss and instead optimizes the log-likelihood of predicting the next word in a sequence conditioned on an image.

The proposed LSTM model consists of a single layer with 250 units and weights set between the range from [-0.06, 0.06]. The margin α was set to α = 0.3, which showed a decent performance on both the datasets. The neural network adopted an initial learning rate of 1 which decreased exponentially. The training was carried out using the stochastic gradient descent algorithm. Apart from the models mentioned above many other models were used for the comparison of the results. [19]

4.3 Results for Flickr8K Dataset

The Flickr8K dataset was considered as a benchmark dataset of image and sentence retrieval. The proposed model performs quite well compared to that of the m-RNN, while for some metrics the proposed model is out performed by a few. The R@K and Med r of different methods are shown in Table 1. [7]

4.4 Results for Flickr30K Dataset

The extension of the Flickr8K dataset i.e. the Flickr30K dataset naturally has only a few methods report their retrieval results as this dataset is relatively new. The R@K evaluation metric for the different methods are shown in Table 2. [7]

5. CONCLUSION

This paper reviews the various methods for caption generation and also proposes a deep multimodal neural network to generate novel captions for images. Combination of different CNNs and RNN or LSTM is modeled to provide state of the art performance on image-caption generation. The proposed m-RNN is capable of being extended to use more complex image context and language model.

REFERENCES

[1] Ryan Kiros, Richard S Zemel, and Ruslan Salakhutdinov. Multimodal neural language models. ICML, 2014.
[2] Bo Qu, Xuelong Li, Dacheng Tao, Xiaoqiang Lu, "Deep Semantic Understanding of High Resolution Remote Sensing Image" *IEEE Conf,* 2016. (P2)
[3] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we perceive in a glance of a real-world scene? Journal of vision, 7(1):10, 2007.
[4] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In ICCV, 2011.
[5] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daume, III. Midge: Generating image descriptions from computer vision detections. In EACL, 2012.
[6] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654, 2014.

[7] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Alan L. Yuille, "Explain Images with Multimodal Recurrent Neural Networks" *arXiv:1410.1090v1 [cs.CV]* 4 Oct 2014 (P3)

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks" *NIPS*, pages 1097–1105, 2012.

[9] P. Y. A. L. M. Hodosh and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.

[10] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur. Extensions of recurrent neural network language model. In ICASSP, pages 5528–5531, 2011.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, pages 3111–3119, 2013.

[12] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's mechanical Turk. In NAACL-HLT workshop 2010, pages 139–147, 2010.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[14] Andrej Karpathy, Li Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions" *IEEE,* 05 August 2016. (P6)

[15] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In NIPS, pages 2121–2129, 2013.

[16] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2014.

[17] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In NIPS, 2011.

[18] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach,S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. arXiv preprint arXiv:1411.4389, 2014.

[19] Ryan Kiros, Ruslan Salakhutdinov, Richard S. Zemel, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models" *arXiv:1411.2539v1 [cs.LG],* 10 Nov 2014. (P4)

[20] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In CVPR, 2013.

[21] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013.

[22] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In CoNLL, 2011.

[23] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *arXiv*:1406.5679, 2014.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. arXiv preprint arXiv:1405.0312, 2014.

[25] M. Yatskar, L. Vanderwende, and L. Zettlemoyer. See no evil, say no evil: Description generation from densely labeled images. Lexical and Computational Semantics, 2014.

[26] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In ACL, 2012.

[27] P. Kuznetsova, V. Ordonez, T. L. Berg, U. C. Hill, and Y. Choi. Treetalk: Composition and compression of trees for image descriptions. Transactions of the Association for Computational Linguistics, 2(10):351–362, 2014.

[28] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555, 2014.

[29] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, et al ."From Captions to Visual Concepts and Back" *IEEE Conf. on Computer Vision and Pattern Recognition,* 2015. (P1)

[30] R. Socher, Q. Le, C. Manning, and A. Ng. Grounded compositional semantics for finding and describing images with sentences. In TACL, 2014.

[31] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In ICCV, 2013.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In CVPR, 2009.

[33] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. In AAAI, 2010.

[34] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. CoRR, abs/1411.5726, 2014.

[35] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, et al. Video in sentences out. arXiv preprintarXiv:1204.2742, 2012.

[36] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329, 2014.