Automated Neural Image Caption Generator for Visually Impaired People

Christopher Elamri, Teun de Planque
Department of Computer Science
Stanford University
{mcelamri, teun}@stanford.edu

Abstract

Being able to automatically describe the content of an image using properly formed English sentences is a challenging task, but it could have great impact by helping visually impaired people better understand their surroundings. Most modern mobile phones are able to capture photographs, making it possible for the visually impaired to make images of their environments. These images can then be used to generate captions that can be read out loud to the visually impaired, so that they can get a better sense of what is happening around them. In this paper, we present a deep recurrent architecture that automatically generates brief explanations of images. Our models use a convolutional neural network (CNN) to extract features from an image. These features are then fed into a vanilla recurrent neural network (RNN) or a Long Short-Term Memory (LSTM) network to generate a description of the image in valid English. Our models achieve comparable to state of the art performance, and generate highly descriptive captions that can potentially greatly improve the lives of visually impaired people.

1 Introduction

Visual impairment, also known as vision impairment or vision loss, is a decreased ability to see to a degree that causes problems not fixable by usual means, such as glasses. According to the World Health Organization, 285 million people are visually impaired worldwide, including over 39 million blind people [1]. Living with visual impairment can be challenging, since many daily-life situations are difficult to understand without good visual acuity.

Technology has the potential to significantly improve the lives of visually impaired people (Figure 1). Access technology such as screen readers, screen magnifiers, and refreshable Braille displays enable the blind to use mainstream computer applications and mobile phones giving them access to previously inaccessible information. Another such technology that could improve the lives of the visually impaired is image caption generation. Most modern mobile phones are able to capture photographs, making it possible for the visually impaired to make images of their surroundings. These images can be used to generate captions that can be read out loud to give visually impaired people a better understanding of their surroundings. Image caption generation can also make the web more accessible to visually impaired people. The last decade has seen the triumph of the rich graphical desktop, replete with colourful icons, controls, buttons, and images. Automated caption generation of online images can make the web a more inviting place for visually impaired surfers.

Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task. This task is significantly harder, for example, than the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community. Indeed, a description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the

Figure 1: Visually impaired people can greatly benefit from technological solutions that can help them better understand their surroundings.

activities they are involved in. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding.

In this paper, we apply deep learning techniques to the image caption generation task. We first extract image features using a CNN. Specifically, we extract a 4096-Dimensional image feature vector from the fc7 layer of the VGG-16 network pretrained on ImageNet. We then reduce the dimension of this image feature vector using Principal Component Analysis (PCA). This resulting feature vector is then fed into a vanilla RNN or a LSTM. The vanilla RNN and LSTM generate a description of the image in valid English. Both the RNN and LSTM based model achieve results comparable to those achieved by the state of-the-art models.

2 Related Work

Most work in visual recognition has originally focused on image classification, i.e. assigning labels corresponding to a fixed number of categories to images. Great progress in image classification has been made over the last couple of years, especially with the use of deep learning techniques [2, 3]. Nevertheless, a category label still provides limited information about an image, and especially visually impaired people can benefit from more detailed descriptions. Some initial attempts at generating more detailed image descriptions have been made, for instance by Farhadi et al. and Kulkarni et al. [4, 5], but these models are generally dependent on hard-coded sentences and visual concepts. In addition, the goal of most of these works is to accurately describe the content of an image in a single sentence. However, this one sentence requirement unnecessarily limits the quality of the descriptions generated by the model. Several works, for example by Li et al., Gould et al., and Fidler et al., focused on obtaining a holistic understanding of scenes and objects depicted on images [6, 7, 8, 9]. Nonetheless, the goal of these works was to correctly assign labels corresponding to a fixed number of categories to the scene type of an image, instead of generating higher-level explanations of the scenes and objects depicted on an image.

Generating sentences that describe the content of images has already been explored. Several works attempt to solve this task by finding the image in the training set that is most similar to the test image and then returning the caption associated with the test image [4, 10, 11, 12, 13]. Jia et al., Kuznetsova et al., and Li et al. find multiple similar images, and combine their captions to generate the resulting caption [14, 15, 16]. Kuznetsova et al., and Gupta et al. tried using a fixed sentence template in combination with object detection and feature learning [5, 17, 18]. They tried to identify objects and features contained in the image, and based on the identified objects contained in the image they used their sentence template to create sentences describing the image. Nevertheless, this approach greatly limits the output variety of the model.

Recently there has been a resurgence of interest in image caption generation, as a result of the latest developments in deep learning [2, 19, 20, 21, 22]. Several deep learning approaches have been developed for generating higher level word descriptions of images [21, 22]. Convolutional Neu-

ral Networks have been shown to be powerful models for image classification and object detection tasks. In addition, new models to obtain low-dimensional vector representations of words such as word2vec, and GloVe (Global Vectors for Word Representation) and Recurrent Neural Networks can together create models that combine image features with language modeling to generate image descriptions [21, 22]. Karpathy et al. developed a Multimodal Recurrent Neural Network architecture that uses inferred alignments to learn to generate novel descriptions of image regions [21]. Similarly, Kiros et al. used a log-bilinear model that generates full sentence descriptions for images [22]. However, their model uses a fixed window context [22].

3 Technical Approach

Overview. We implemented a deep recurrent architecture that automatically produces short descriptions of images. Our models use a CNN, which was pretrained on ImageNet, to obtain images features. We then feed these features into either a vanilla RNN or a LSTM network (Figure 2) to generate a description of the image in valid English.

3.1 CNN-based Image Feature Extractor

For feature extraction, we use a CNN. CNNs have been widely used and studied for images tasks, and are currently state-of-the-art methods for object recognition and detection [20]. Concretely, for all input images, we extract features from the fc7 layer of the VGG-16 network pretrained on ImageNet [23], which is very well tuned for object detection. We obtained a 4096-Dimensional image feature vector that we reduce using Principal Component Analysis (PCA) to a 512-Dimensional image feature vector due to computational constraints. We feed these features into the first layer of our RNN or LSTM at the first iteration [24].

3.2 RNN-based Sentence Generator

We first experiment with vanilla RNNs as they have been shown to be powerful models for processing sequential data [25, 26]. Vanilla RNNs can learn complex temporal dynamics by mapping input sequences to a sequence of hidden states, and hidden states to outputs via the following recurrent equations.

$$h_t = f(W_{hh}h_{t-1} + W_{xh}x_t) \tag{1}$$

$$y_t = W_{hy}h_t \tag{2}$$

where f is an element-wise non-linearity, $h_t \in \mathbb{R}^N$ is the hidden state with N hidden units, and y_t is the output at time t. In our implementation, we use a hyperbolic tangent as our element-wise non-linearity. For a length T input sequence $x_1, x_2, ..., x_T$, the updates above are computed sequentially as h_1 (letting $h_0 = 0$), y_1, h_2, y_2,... h_T, y_T.

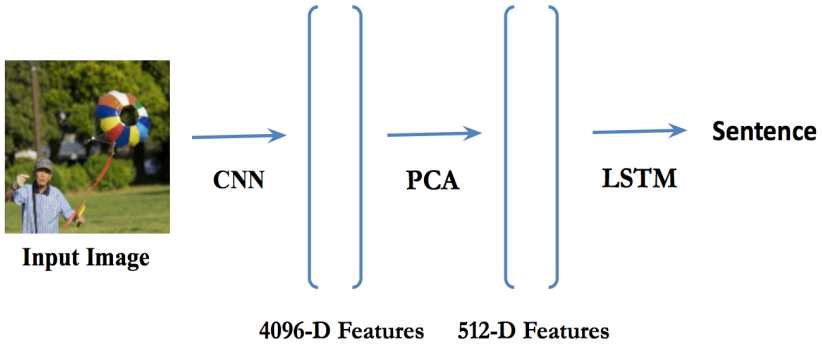

Figure 2: Image Retrieval System and Language Generating Pipeline.

3.3 LSTM-based Sentence Generator

Although RNNs have proven successful on tasks such as text generation and speech recognition [25, 26], it is difficult to train them to learn long-term dynamics. This problem is likely due to the vanishing and exploding gradients problem that can result from propagating the gradients down through the many layers of the recurrent networks. LSTM networks (Figure 3) provide a solution by incorporating memory units that allow the networks to learn when to forget previous hidden states and when to update hidden states when given new information [24].

At each time-step, we receive an input $x_t \in \mathbb{R}^D$ and the previous hidden state $h_{t-1} \in \mathbb{R}^H$, the LSTM also maintains an H-dimensional cell state, so we also get the previous cell state $c_{t-1} \in \mathbb{R}^H$. The learnable parameters of the LSTM are an input-to-hidden matrix $W_x \in \mathbb{R}^{4HxD}$, a hidden-to-hidden matrix $W_h \in \mathbb{R}^{4HxH}$, and a bias vector $b \in \mathbb{R}^{4H}$.

At each time step, we compute an activation vector $a \in \mathbb{R}^{4H}$ as

$$a = W_x x_t + W_h h_{t-1} + b \tag{3}$$

We then divide a into 4 vectors a_i, a_f, a_o, $a_g \in \mathbb{R}^H$ where a_i consists of the first H elements of a, a_f is the next H elements of a, etc.. We then compute four gates which control whether to forget the current cell value $f \in \mathbb{R}^H$, if it should read its input $i \in \mathbb{R}^H$, and whether to output the new cell value $o \in \mathbb{R}^H$, and the block input $g \in \mathbb{R}^H$.

$$i = \sigma(a_i) \tag{4}$$
$$f = \sigma(a_f) \tag{5}$$
$$o = \sigma(a_o) \tag{6}$$
$$g = tanh(a_g) \tag{7}$$

where σ is the sigmoid function and $tanh$ is the hyperbolic tangent; both are applied element-wise.

Finally, we compute the next cell state c_t which encodes knowledge at every time step of what inputs have been observed up to this step, and the next hidden state h_t as

$$c_t = f \circ c_{t-1} + i \circ g \tag{8}$$
$$h_t = o \circ tanh(c_t) \tag{9}$$

where $\circ$ represents the Hadamard product. The inclusion of these multiplicative gates permits the regulation of information flow through the computational unit, allowing for more stable gradients and long-term sequence dependencies [24]. Such multiplicative gates make it possible to train the LSTM robustly as these gates deal well with exploding and vanishing gradients. The non-linearities are sigmoid $\sigma()$ and hyperbolic tangent $tanh()$.

Procedure. Our LSTM model takes the image I and a sequence of inputs vectors $(x_1, ..., x_T)$. It then computes a sequence of hidden states $(h_1, ..., h_t)$ and a sequence of outputs $(y_1, ..., y_t)$ by following the recurrence relation for $t = 1$ to T:

$$b_v = W_{hi}[CNN(I)] \tag{10}$$
$$h_t = f(W_{hx} x_t + W_{hh} h_{t-1} + b_h + 1(t = 1) \circ b_v) \tag{11}$$
$$y_t = Softmax(W_{oh} h_t + b_o) \tag{12}$$

where W_{hi}, W_{hx}, W_{hh}, W_{oh}, x_i, b_h, and b_o are learnable parameters and $CNN(I)$ represents the image features extracted by the CNN.

Training. We train our LSTM model to correctly predict the next word (y_t) based on the current word (x_t), and the previous context (h_{t-1}). We do this as follows: we set $h_0 = 0$, x_1 to the $START$ vector, and the desired label y_1 as the first word in the sequence. We then set x_2 to the word vector corresponding to the first word generated by the network. Based on this first word vector and the previous context the network then predicts the second word, etc. The word vectors are generated using the word2vec embedding model as described by Mikolov et. al [27]. During the last step, x_T represent the last word, and y_T is set to an END token.

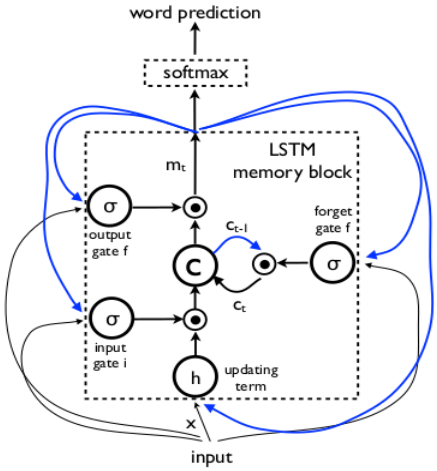

Figure 3: LSTM unit and its gates

Testing. To predict a sentence, we obtain the image features b_v, set $h_0 = 0$, set x_1 to the $START$ vector, and compute the distribution over the first word y_1. Accordingly, we pick the argmax from the distribution, set its embedding vector as x_2, and repeat the procedure until the END token is generated.

Softmax Loss. At every time-step, we generate a score for each word in the vocabulary. We then use the ground truth words in combination with the softmax function to compute the losses and gradients. We sum the losses over time and average them over the minibatch. Since we operate over minibatches and because different generated sentences may have different lengths, we append $NULL$ tokens to the end of each caption so that they all have the same lengths. In addition, our loss function accepts a mask array that informs it on which elements of the scores counts towards the loss in order to prevent the $NULL$ tokens to count towards the loss or gradient.

Optimization. We use Stochastic Gradient Descent (SGD) with mini-batches of 25 image-sentence pairs and momentum of 0.95. We cross-validate the learning rate and the weight decay. We achieved our best results using Adam, which is a method for efficient stochastic optimization that only requires first-order gradients and computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients [28]. Adam's main advantages are that the magnitudes of parameter updates are invariant to rescaling of the gradients, its step-size is approximately bounded by the step-size hyperparameter, and it automatically performs a form of step-size annealing [28].

4 Experiments

4.1 Dataset

For this exercise we will use the 2014 release of the Microsoft COCO dataset which has become the standard testbed for image captioning [29]. The dataset consists of 80,000 training images and 40,000 validation images, each annotated with 5 captions written by workers on Amazon Mechanical Turk. Four example images with captions can be seen in Figure 4. We convert all sentences to lower-case, and discard non-alphanumeric characters.

Figure 4: Example images and captions from the Microsoft COCO Caption dataset.

4.2 Evaluation Metric

For each image we expect a caption that provides a correct but brief explanation in valid English of the images. The closer the generated caption is to the captions written by workers on Amazon mechanical Turk the better.

The effectiveness of our model is tested on 40,000 images contained in the Microsoft COCO dataset. We evaluate the generated captions using the following metrics: BLEU (Bilingual Evaluation Understudy) [30], METEOR (Metric for Evaluation of Translation with Explicit Ordering) [31], and CIDEr (Consensus-based Image Description Evaluation) [32]. Each method evaluates a candidate sentence by measuring how well it matches a set of five reference sentences written by humans. The BLEU score is computed by counting the number of matches between the n-grams of the candidate caption and the n-grams of the reference caption. METEOR was designed to fix some of the problems found in the more popular BLEU metric, and also produce good correlation with human judgement at the sentence or segment level [30]. METEOR differs from the BLEU metric in that BLEU seeks correlation at the corpus level [31]. The CIDEr metric was specifically developed for evaluating image captions [32]. It is a measure of consensus based on how often n-grams in candidate captions are present in references captions. It measures the consensus in image captions by performing a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram, because frequent n-grams in references are less informative [32]. For all three metrics (i.e. BLEU, METEOR, and CIDEr) the higher the score, the better the candidate caption is [30][31][32].

4.3 Quantitative Results

We report the BLEU, METEOR and CIDEr scores in Figure 5 and compare it to the results obtained in the literature. Both our RNN and LSTM model achieve close to state-of-the-art performance. Our LSTM model performs slightly better than our RNN model; it achieves a higher BLEU, METEOR, and CIDEr score than the RNN model.

4.4 Qualitative Results

Our models generates sensible descriptions of images in valid English (Figure 6 and 7). As can be seen from example groundings in Figure 5, the model discovers interpretable visual-semantic correspondences, even for relatively small objects such as the phones in Figure 7. The generated descriptions are accurate enough to be helpful for visually impaired people. In general, we find that a relatively large portion of generated sentences (60%) can be found in the training data.

Model	BLEU	METEOR	CIDEr
Nearest Neighbor	48.0	15.7	38.3
Google NIC [33]	66.6	-	-
Karpathy et al. [21]	62.5	19.5	66.0
Chen and Zitnick [34]	-	20.4	-
MS Research [35]	-	20.7	-
LRCN [36]	62.8	-	-
Our RNN Model	62.2	19.3	65.6
Our LSTM Model	62.5	19.4	65.8

Figure 5: Evaluation of full image predictions on 1,000 test images of the Microsoft COCO 2014 dataset.

Figure 6: Example image descriptions generated using the RNN structure.

Figure 7: Example image descriptions generated using the LSTM structure.

5 Conclusion

We have presented a deep learning model that automatically generates image captions with the goal of helping visually impaired people better understand their environments. Our described model is based on a CNN that encodes an image into a compact representation, followed by a RNN that generates corresponding sentences based on the learned image features. We showed that this model achieves comparable to state-of-the-art performance, and that the generated captions are highly descriptive of the objects and scenes depicted on the images. Because of the high quality of the generated image descriptions, visually impaired people can greatly benefit and get a better sense of their surroundings using text-to-speech technology. Future work can include this text-to-speech technology, so that the generated descriptions are automatically read out loud to visually impaired people. In addition, future work could focus on translating videos directly to sentences instead of generating captions of images. Static images can only provide blind people with information about one specific instant of time, while video caption generation could potentially provide blind people with continuous real time information. LSTMs could be used in combination with CNNs to translate videos to English descriptions.

Acknowledgments

We would like to thank the CS224D course staff for their ongoing support.

References

[1] "Visual Impairment and Blindness." *World Health Organization.* (2014). Web. 10 Apr. 2016

[2] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision Int J Comput Vis 115.3* (2015): 211-52. Web. 19 Apr. 2016

[3] Everingham, Mark, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. "The Pascal Visual Object Classes (VOC) Challenge." *International Journal of Computer Vision Int J Comput Vis 88.2* (2009): 303-38. Web. 22 May 2016

[4] Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. "Every Picture Tells a Story: Generating Sentences from Images." *Computer Vision ECCV 2010 Lecture Notes in Computer Science* (2010): 15-29. Web. 5 Apr. 2016

[5] Kulkarni, Girish, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. "Baby Talk: Understanding and Generating Simple Image Descriptions." *Cvpr 2011* (2011). Web. 27 May 2016

[6] Li, Li-Jia, R. Socher, and Li Fei-Fei. "Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework." *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009). Web. 21 Apr. 2016

[7] Gould, Stephen, Richard Fulton, and Daphne Koller. "Decomposing a Scene into Geometric and Semantically Consistent Regions." *2009 IEEE 12th International Conference on Computer Vision* (2009). Web. 6 May 2016

[8] Fidler, Sanja, Abhishek Sharma, and Raquel Urtasun. "A Sentence Is Worth a Thousand Pixels." *2013 IEEE Conference on Computer Vision and Pattern Recognition* (2013). Web. 18 May 2016

[9] Li, Li-Jia, and Li Fei-Fei. "What, Where and Who? Classifying Events by Scene and Object Recognition." *2007 IEEE 11th International Conference on Computer Vision* (2007). Web. 10 Apr. 2016

[10] Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni. "Combining Language and Vision with a Multimodal Skip-gram Model." *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2015). Web. 23 May 2016

[11] Hodosh, Young, and Hockenmaier. "Framing image description as a ranking task: data, models and evaluation metrics." *Journal of Artificial Intelligence Research* (2013). Web. 3 Apr. 2016

[12] Socher, Richard, Andrej Karpathy, Quoc V. Le, Christopher Manning, and Andrew Y. Ng. "Grounded compositional semantics for finding and describing images with sentences." *Transactions of the Association for Computational Linguistics (TACL)* (2014). Web. 24 May 2016

[13] Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. "Im2text: Describing images using 1 million captioned photographs." *NIPS: 1143-1151* (2011). Web. 29 Apr. 2016

[14] Jia, Yangqing, Mathieu Salzmann, and Trevor Darrell. "Learning Cross-modality Similarity for Multinomial Data." *2011 International Conference on Computer Vision* (2011). Web. 28 May 2016

[15] Kuznetsova, Polina, Vicente Ordonez, Alexander C. Berg, Tamara Berg, and Yejin Choi. "Collective generation of natural image descriptions." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics 1* (2012): 359:368. Web. 30 Apr. 2016

[16] Li, Siming and Kulkarni, Girish and Berg, Tamara L. and Berg, Alexander C. and Choi, Yejin. "Composing simple image descriptions using web-scale n-grams." *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: 220-228* (2011). Web. 27 Apr. 2016

[17] Kuznetsova, Polina, Vicente Ordonez, Tamara Berg, Yejin Choi. "TREETALK: Composition and Compression of Trees for Image Descriptions." *Transactions of the Association for Computational Linguistics 2* (2014): 351-362. Web. 1 Apr. 2016

[18] Gupta and Mannem. "From image annotation to image description. In Neural information processing." *Springer* (2012). Web. 7 Apr. 2015

[19] LeCun, Bottou, Bengio, and Haffner. "Gradient- based learning applied to document recognition." *Proceedings of the IEEE* (1998): 86(11):22782324. Web. 27 May 2016

[20] Krizhevsky, Sutskever, and Hinton. "Imagenet classification with deep convolutional neural networks." *NIPS* (2012). Web. 28 Apr. 2016

[21] Karpathy, Andrej, and Li Fei-Fei. "Deep Visual-semantic Alignments for Generating Image Descriptions." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015). Web. 29 May 2016

[22] Kiros Ryan, Rich Zemel, and Ruslan Salakhutdinov. "Multimodal neural language models." *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*: 595-603 (2014). Web. 21 May 2016

[23] Simonyan, Karen and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *CoRR* (2014). Web. 28 May 2016

[24] Hochreiter, Sepp, and Jrgen Schmidhuber. "Long Short-Term Memory." *Neural Computation 9.8* (1997): 1735-780. Web. 23 Apr. 2016

[25] Graves, Alex. "Generating sequences with recurrent neural networks." *CoRR* (2013). Web. 30 May 2016

[26] Graves, Alex and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks." *Proceedings of the 31st International Converence on Machine Learning (ICML-14): 1764-1772* (2014). Web. 28 May 2016

[27] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." *Advances in Neural Information Processing Systems (NIPS) 26: 3111-3119* (2013). Web. 29 Apr. 2016

[28] Kingma, Diederik and Jimmy Ba. "Adam: A method for stochastic optimization." *CoRR* (2015). Web. 19 May 2016

[29] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. "Microsoft COCO: Common Objects in Context." Computer Vision ECCV 2014 Lecture Notes in Computer Science (2014): 740-55. Web. 27 May 2016

[30] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu, Bleu: a method for automatic evaluation of machine translation." *Proceedings of the 40th Annual Meeting on Association for Computation Linguistics (ACL): 311-318* (2002). Web. 24 May 2016

[31] Denkowski, Michael, and Alon Lavie. "Meteor Universal: Language Specific Translation Evaluation for Any Target Language." *Proceedings of the Ninth Workshop on Statistical Machine Translation* (2014). Web. 22 Apr. 2016

[32] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "CIDEr: Consensus-based Image Description Evaluation." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015). Web. 24 May 2016

[33] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and Tell: A Neural Image Caption Generator." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015). Web. 25 May 2016

[34] Chen, Xinlei and C. Lawrence Zitnick. Learning a Recurrent Visual Representation for Image Caption Generation. *CoRR abs/1411.5654* (2014). Web. 19 May 2016

[35] Fang, Hao, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. "From Captions to Visual Concepts and Back." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015). Web. 27 Apr. 2016

[36] Donahue, Jeff, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. "Long-term Recurrent Convolutional Networks for Visual Recognition and Description." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015). Web. 20 Apr. 2016