

# Generating Image Captions based on Neural Embedding

Sandeep Kumar Dash<sup>1</sup>, Saurav Saha<sup>1</sup>, Partha Pakray<sup>1</sup> and Alexander Gelbukh<sup>2</sup>

<sup>1</sup> National Institute of Technology Mizoram, Aizawl, India,  
sandeep.cse@nitmz.ac.in, contact.srvsaha@gmail.com, partha.cse@nitmz.ac.in  
<sup>2</sup> Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico  
gelbukh@gelbukh.com

**Abstract.** Caption generation has long been optically discerned as a conundrum in Computer Vision and Natural Language Processing. Automatic Image Captioning could, for example, be habituated to provide descriptions of website content, or to engender frame-by-frame descriptions of video for the vision-impaired. Being able to automatically describe the content of an image utilizing verbosely composed English sentences is a challenging task, but it could have great impact by availing visually impaired people better understand their surroundings. Most modern mobile phones can help the visually impaired to capture images of their environments. These images can then be habituated to engender captions that can be read out loud to the visually impaired, so that they can get a better sense of what is transpiring around them. In this work, a model is described which is utilized to engender novel image captions for a previously unseen image by utilizing a multimodal architecture by amalgamation of a Recurrent Neural Network (RNN) and a Convolutional Neural Network (CNN). The model is trained on MSCOCO (Microsoft Common Objects in Context) image captioning dataset that projects captions and images in the same representation space, so that an image is close to its captions in that space, and far away from dissimilar captions and dissimilar images. ResNet-50 architecture is used for extracting features from images and GloVe embeddings are used along with GRUs in RNN for text representation. MSCOCO evaluation server is used for evaluation of the machine generated caption for a given image.

**Keywords:** Image Captioning, Convolutional Neural Network, Gated Recurrent Units, Multimodal Embedding

## 1 Introduction

The world relies on what gets seen. Vision modality therefore is an integral part of communication of information. But when it comes to describing it in Natural Language Text it becomes cumbersome as the limitation lies in the difference of representing the modalities, as the structure of both are different. This in turn makes emulation of human abilities to represent details of an image in simple

text a difficult task. The complexity further increases when we require efficiency of the representation. However a lot of approaches have been introduced to have a correspondence among the modalities so that they can represent each other. This often helps to generate more details about any of the involved modalities span the modality alone. The recent research are benefited by large datasets with pairing of images and their descriptions. The progress that Image Caption Generation has achieved, is mostly delved out of three main techniques:

i) Template based - The main idea here is to identify as much detail as possible about the image components such as objects, their attributes, relationship with other objects. Then further parse the sentence into phrases and learn their correspondence with the image components using models like Conditional Random Fields (CRF) [1]. Finally putting them in a fixed template as Subject-Verb-Object to generate captions. These methods perform poorly as the dependency on template fails them to generate variable length sentences.

ii) Retrieval or Transfer based - These leverage the distance in the visual space to retrieve images which are similar to the test image and then modify and combine their captions to form the semantically similar caption for the test image. These models are highly dependable on the training or seen data and need additional steps like modification and generalization to output the final caption. Also they fail to generate novel captions.

iii) Language based - These models offer greater flexibility of producing more human like captions by being independent of either templates or training data. They learn the probability distribution over a common semantic space of both image content and text. The semantic space offers measurement efficiency of both the modalities by reducing the individual structure in to a common representable platform through which the individual distance can easily be delved out. Further it offers more flexibility of representing components of both modalities to have better correspondence among them. These models are the main inspiration behind recent improvements in this direction which is undoubtedly due to the success of Neural Networks which makes the representation of the modalities in semantic space much easier. These work in a sequential encoder decoder methodology to generate output sequence from input sequence. As RNNs have proved to be efficient for text representation they are measuredly used for novel caption generation whereas CNNs have outperformed any other category for image representations, thus making it a safe bet for image encoding and decoding.

This work simplifies the existing approaches for Image Caption Generation by means of introducing a much efficient approach through generated embedding of both the modalities. It also helps in either way retrieval as the generated embeddings correspond to each other.

## 2 Related Works

The concept of Caption Generation has mostly been an amalgamation of Computer Vision and Natural Language Processing techniques. The techniques rely on both the aspect of feature extraction from images by different Image Process-

ing methods and further analyzing the features to obtain the relation among the detected image fragments in terms of Natural Language Text. Earlier methods of Caption Generation mainly focused on generating sentences from the combinations of image annotations. The generated captions through this are the result of proper image understanding as well as Natural Language Generation. Evaluation of image sentence correspondence mapped on to an intermediate meaning space represented by (object, action, scene) template has been described in [2]. Through the intermediate space they were able to represent either modality from the other. Similar to this, work in [3][4][5] also finds objects from images along with their attributes and relation among the objects to establish a template based relation among the two modalities. However these are hard bounded to the template and are limited to the spatial or corpus based relationship among the objects identified in the image. A deviation from these approaches with densely labeled images, which incorporate object, attribute, action, and scene annotations to generate description was proposed in [6]. A three stage approach was followed in [7] without explicitly labeling the images, in the first stage they mapped the features extracted from each region of image to words likely to be present in caption. This was followed by a Maximum Entropy (ME) Language Model from a set of training image descriptions to produce high-likelihood sentences in second stage, further followed by a re-ranking stage.

Retrieval based approaches associate an image with top ranked description among the candidate descriptions of similar images. These candidate descriptions can then either be used directly (description transfer) or a novel description can be synthesized from the candidates (description generation). The retrieval of images and ranking of their descriptions can be carried out in two ways: either from a visual space or from a multimodal space that combines textual and visual information space. The first approach represents the query image as visual features and compares its similarity with the images in the candidate set which are retrieved based on their features in the feature space. Finally the candidate image description are re-ranked or their fragments are combined as per certain rules and assigned to the query image. Im2Text model [8], GIST [9], Tiny Image [10] are based on this approach. The second approach projects features of both image and text on to a common semantic space known as multimodal space. This helps to retrieve one modality given the other. Our approach mostly relates to this.

Most of the recent methods are based on Recurrent Neural Networks, inspired by the successful use of sequence-to-sequence training with deep recurrent networks in machine translation [11, 12, 13]. The first deep learning method for image captioning was proposed in [14]. The method utilizes a multimodal log-bilinear model that is biased by the features from the image. The feed-forward neural network was replaced in [15] with a Recurrent Neural Network. In [16] a LSTM (Long short-term memory) network was used, which is a refined version of a vanilla Recurrent Neural Network. Unlike models of [14] and [15], which feed in image features at every time step, in [16] the image is fed into the LSTM only at the first time step.

Top-down approaches followed in [16] and [17] used modern CNNs for encoding and replaced feed forward networks in [2] with recurrent neural networks, in particular LSTMs. The use of these models on video captioning tasks was demonstrated in [18]. One of the main contributions of [16] was that it showed that a LSTM that did not receive the image vector representation at each time step was still able to produce state-of-the-art results, unlike the earlier work in [2]. The common theme of these works is that they represented images as the top layer of a large CNN (hence the name top-down as no individual objects are detected) and produced models that were end-to-end trainable.

Bottom-up approaches implemented in [19] trains a CNN and bi-directional RNN, that maps images and fragments of captions to the same multimodal embedding, demonstrating state-of-the-art results on informational retrieval tasks. Secondly, a RNN is trained that learns to combine the inputs from various object fragments detected in the original image to form a caption. This improved on previous works by allowing the model to aggregate information on specific objects in the image rather than working from a singular image representation. However, these models were not end-to-end trainable. Recently there has been a resurgence of interest in image caption generation, as a result of the latest developments in deep learning [15, 20]. Several deep learning approaches have been developed for generating higher level word descriptions of images. Convolutional Neural 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 [21], and GloVe (Global Vectors for Word Representation) [22] and Recurrent Neural Networks can together create models that combine image features with language modeling to generate image descriptions. In [23] a novel decision-making framework for image captioning was introduced where two separate networks were used. One to provide local guidance for predicting the next word as per the current state and another to provide global and lookahead guidance by evaluating all possible extensions of the current state. Attention mechanism used in [24] enables attention to be calculated at the level of objects and other salient image regions which has produced state-of-the-art result on MSCOCO dataset [25]. Our model differs in the sense of producing captions based on nearest neighbor embedding of feature from the modalities.

### 3 Caption Generation Model

The model differentiates among relevant and irrelevant captions by projecting image and its captions to same embedding space and further measuring their distance to cast the most similar embedded caption as the corresponding caption for the image. Since the model is purely based on vector semantics, it permits both way retrieval, as relevant images for any input caption can also be identified. The model uses MSCOCO 2014 training dataset with 82,723 images in training set, each with 5 corresponding captions for training. The model's performance

is evaluated in MSCOCO 2014 validation set containing 40,504 images with 5 captions each and on MSCOCO 2014 test set containing 40,775 unseen images.

### 3.1 Model Overview

Fig. 1 describes the model which accepts three inputs an image feature (the output of the ResNet-50 [26] model), embedding of a correct caption and a noise caption respectively. As it is retrieval based, the model gets trained to have minimum distance between similar vectors and thereby maximize the distance between dissimilar vectors in space. As can be seen, both the image and caption text are embedded initially. The dot product of image with both actual and noise caption selected randomly from the dataset, is computed. The model is trained to distance noise captions from image by using the max-margin loss function. As both the modalities needs to be converted into vectors efficient neural based techniques are used to produce the embedding.

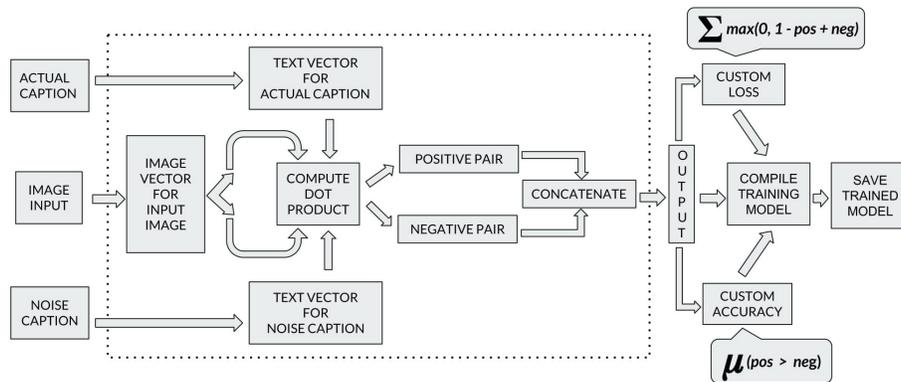


Fig. 1: Model Architecture.

### 3.2 Image Representation

CNN architectures have proved to be efficient enough in Computer Vision related tasks. So following the work by [16] the image representation model exploits CNN architecture for embedding images. It utilises the ResNet-50 [26] structure of the Imagenet Large Scale Visual Recognition Challenge [27]. The last layer of the structure is removed to retrieve the embedding matrix of shape (82783, 2048) from the images. The following figure shows the embedding generation method.

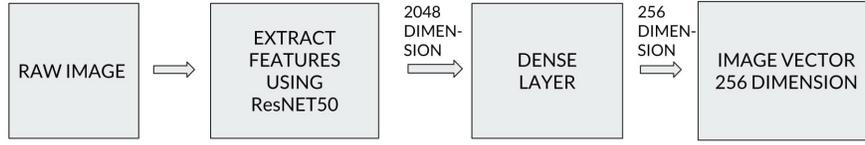


Fig. 2: Image Representation Model.

### 3.3 Text Representation

The captions for images are initially tokenized, lower-cased and stripped of its punctuation as part of the preprocessing task using NLTK [28]. Also the texts are converted to integer sequences of the same size by padding upto 16 characters for each caption. Preprocessed texts as array of integers are now fed into the GRU based text representation model initialized with GloVe embeddings trained on 6 billion tokens from Wikipedia 2014 and Gigaword 5. The following figure shows the process followed for text embedding generation.

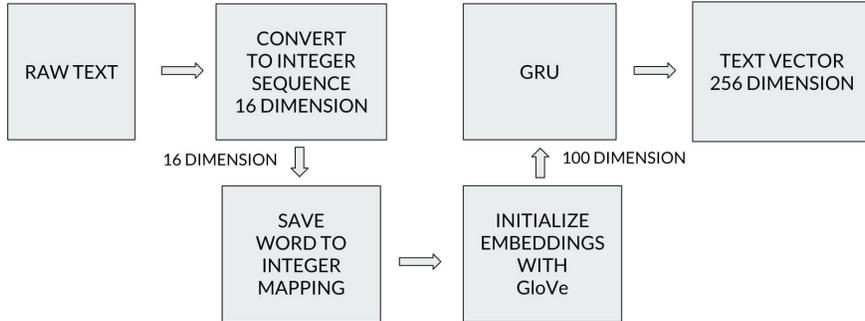


Fig. 3: Text Representation Model.

### 3.4 Training the model

The model gets trained based on maximum margin loss on positive and negative pairs of image and caption. It computes the dot product for the positive and negative pairs and aims to maximize the score for positive pairs. For this the maximum-margin loss function is used as represented in eq(1).

$$loss = \sum_i \max(0, 1 - p_i + n_i) \quad (1)$$

Here  $p_i$  refer to the score of the positive pair of the  $i$ -th image and  $n_i$  refers to the score of the negative pair of that image. The neural embedding model appears in Fig. 4. The model was trained for 500 epochs on a single Nvidia Quadro K420

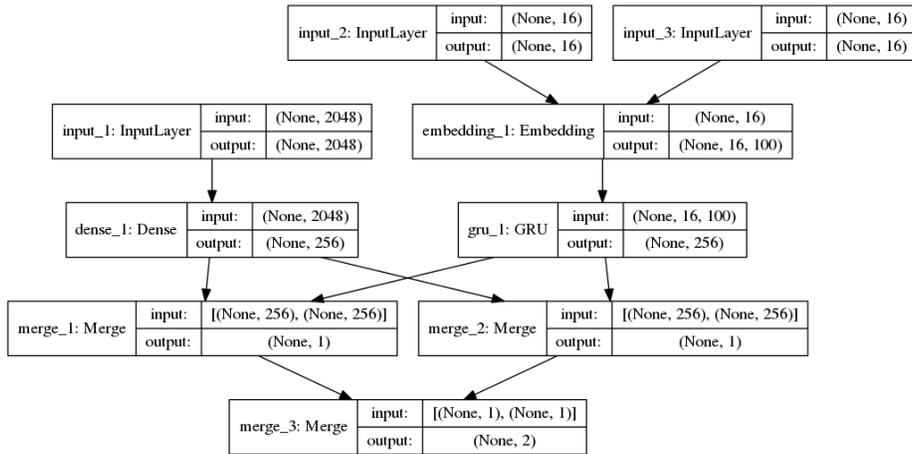


Fig. 4: Neural Embedding Model.

GPU using checkpoints on best accuracy and the best accuracy was produced in 174th epoch. The step time for each epoch was about 2 seconds and it took around a week to train the model. The accuracy and loss plots for the model are shown in Fig. 6 and Fig. 7 respectively.

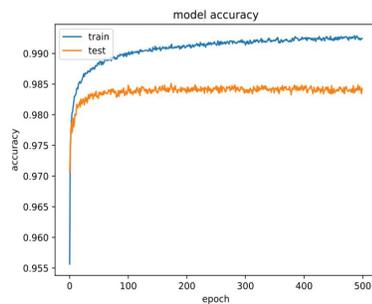


Fig. 5: Accuracy plot over 500 epochs

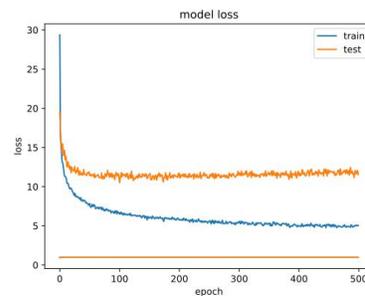


Fig. 6: Loss curve plot over 500 epochs

### 3.5 Generating Captions

The diagram in Fig. 7 describes the typical steps involved in generating a caption for any novel input image. The neural embeddings generated through the image and text representation models create an embedding space where each vector is reduced to same size of 256 dimension. For any raw image its embedding vector is projected onto the embedding space and the nearest vector from the common embedding space is retrieved as its relevant caption.

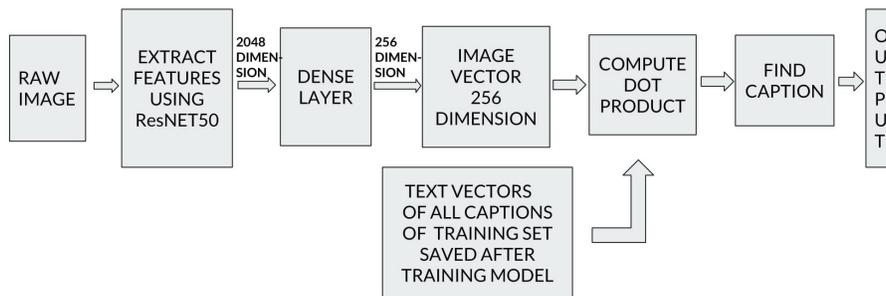


Fig. 7: Caption Generation procedure for any input image.

## 4 Result Analysis

### 4.1 Evaluation Metric

The effectiveness of the model is tested on 40,775 images contained in the MSCOCO 2014 test dataset. Also, to avoid overfitting the model, MSCOCO validation dataset is used consisting of 40,504 images. As in machine translation, the BLEU (Bilingual Evaluation Understudy) [29] Score is the main metric used to evaluate image-to-text translations. The BLEU-n score computes a modified n-gram precision for words in the candidate translation compared to the reference translations. In particular, the n-gram precision is computed by taking the count of each n-gram in the reference translation, and clipping this count by the maximum number of times that the n-gram appears in a reference translation. This clipped count is divided by the unclipped count in the reference translation to produce a score. CIDEr [30], Meteor [31], Rouge-L [32] are also used for evaluation purpose in this work.

### 4.2 Qualitative Analysis

The model produces better result when objects in the training images correspond to similar objects in the test image whereas it slightly faults when totally

untrained objects come up. Performance of the model can be seen in Fig. 8-10 which are produced on MSCOCO validation dataset.

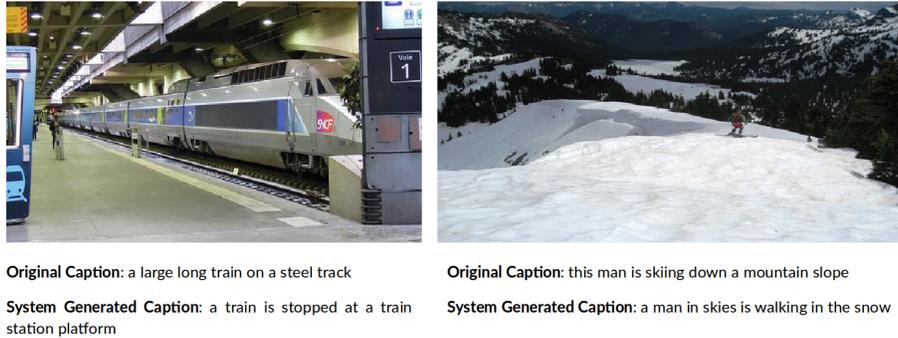


Fig. 8: Successfully generated relevant captions by the model

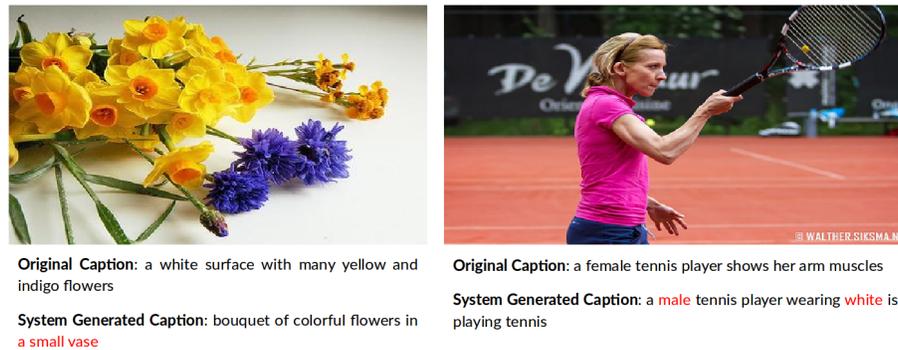


Fig. 9: Partial successful result from the model putting forth the limitation of the model as it misses out the individual local features in image resulting in partial success.

Results shown in Fig. 11 shows the performance of the system on some of the images from MSCOCO Testset whose context were learned, thereby producing relevant captions. However the result slightly defaults when it gets totally unseen images as shown in Fig. 12.

The model generalizes very well for unseen images which can be observed from the results of unseen images in test set. This can be observed from the Table 1 in section 4.3 where the performance on unseen test image was nearly same as that of validation dataset which shows that our model is generalized for unseen images. Also, since the batch size was kept high (256), the gradient



**Original Caption:** a bunch of soccer players are playing a game

**System Generated Caption:** a baseball field with players and a crowd of spectators



**Original Caption:** a skateboarder riding their board in a skate park

**System Generated Caption:** diners at a cafe overlooking a sandy beach

Fig. 10: Un-successful result generated by the model as the image context as 'skate park' etc appeared very less in the training set resulting in un-successful result.



a man in a blue jacket is traveling on snowshoes through snowy woods



a boy jumping his skateboard on a city plaza



a daredevil motorcyclist performs a wheelie with his legs over the handle bars

Fig. 11: Successfully generated captions from MSCOCO Testset



a small compact car parked in an alley with a bull dog sitting on top of it



a close up of a cat in a sink in a bath room

Fig. 12: Partial-successfully generated captions from MSCOCO Testset.

updates during training were done after every batch of 256 images was seen, thus resulting in a more generalized model. Hence the model is able to gen-

erate captions for images which are semantically similar in context. Since the automatic metrics look for exact matching of tokens for evaluating captions, so even if the model generated a caption which is semantically correct, the performance was not very high due to the use of different words in the caption than expected in ground truth captions. This can be observed from the results of BLEU/CIDEr/METEOR/ROGUE-L metric of c40 in test set, where the caption generated by system is tested against 40 ground truth captions instead of just 5 ground truth captions, c5 as in validation dataset which concludes that our model gathers much semantic information during captioning and hence score increases when number of ground truth captions are high, which can be seen in Table 2.

### 4.3 Quantitative Analysis

The system results are reported using COCO captioning evaluation tool which reports metrics such as BLEU, Meteor, Rouge-L and CIDEr. Table 1 shows the comparison of individual scores for metrics obtained from the MSCOCO evaluation server on validation set and test set respectively where 5 images were held for ground truth caption(c5). Table 2 shows the comparative analysis of performance on the evaluation metrics on MSCOCO Testset, where system generated caption is evaluated against 5 ground truth captions in c5 and 40 ground truth captions in c40. The higher value of c40 against c5 for all metrics prove that our model carries more semantic information and hence performance is high when varieties of captions are used to describe the same scenario. Hence, our model is highly robust for novel image captioning.

Table 1: Performance comparison of our model on MSCOCO val2014 and test2014

DataSet	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROGUE-L
Validation Set	0.384	0.208	0.113	0.064	0.200	0.150	0.310
Test Set	0.382	0.204	0.107	0.058	0.191	0.148	0.306

### 4.4 Error Analysis

The system generates relevant captions for images in those cases where the model sees objects in the similar context. As the model gets trained on the whole image as feature input, it cannot detect the individual objects present in the image resulting in the generation of partial and unsuccessful result. As in cases it can identify the 'tennis court' context but fails to identify whether the player is a male or female. Similarly it clearly fails to identify a skate park whereas it has identified the beach which is in the context but not in attention.

Table 2: Performance Comparison on MSCOCO test set

Metric	c5	c40
BLEU-1	0.381	0.569
BLEU-2	0.203	0.371
BLEU-3	0.107	0.229
BLEU-4	0.057	0.141
CIDEr	0.191	0.212
METEOR	0.147	0.198
ROGUE-L	0.306	0.398

## 5 Conclusion

The model describes neural embedding based method to project image and its corresponding captions to the same vector space. The max-margin loss function utilized, reduces the distance between image and its relevant caption. The model generates better captions for images in seen context. As future prospective attention based mechanism for generating object based features from image along with Recurrent Neural Network based method for text generation will be used to improve the performance of the model. Also, a better image model will be used for identifying local and global features from image which will be aid in generating more accurate captions.

## References

1. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: 18th International Conference on Machine Learning, pp. 282-289 (2001)
2. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: European conference on computer vision, Springer, Berlin, Heidelberg, pp. 15-29 (2010)
3. Yang, Y., Teo, C.L., Daum III, H., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 444-454 (2011)
4. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating image descriptions. In: 24th CVPR, pp. 1601-1608 (2011)
5. Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., Daum III, H.: Midge: Generating image descriptions from computer vision detections. In: 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 747-756 (2012)
6. Yatskar, M., Galley, M., Vanderwende, L., Zettlemoyer, L.: See No Evil, Say No Evil: Description Generation from Densely Labeled Images. In: Third Joint Conference on Lexical and Computational Semantics, pp. 110-120 (2014)

7. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollr, P., Gao, J.: From captions to visual concepts and back. In: IEEE conference on computer vision and pattern recognition, pp. 1473-1482 (2015)
8. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: Advances in Neural Information Processing Systems, pp. 1143-1151 (2011)
9. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. In: International Journal of Computer Vision, vol.42, no. 3, pp. 145-175 (2001)
10. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.30, no. 11, pp. 1958-1970 (2008)
11. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
12. Cho, K., Merriënboer, B.V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
13. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104-3112 (2014)
14. Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal Neural Language Models. In: 31st International Conference on Machine Learning (ICML-14), pp. 595-603 (2014)
15. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632 (2014)
16. Oriol V., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156-3164 (2015)
17. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: IEEE conference on computer vision and pattern recognition, pp. 2625-2634 (2015)
18. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014)
19. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128-3137 (2015)
20. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. arXiv preprint arXiv:1612.01887 (2016)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
22. Pennington, J., Socher, R., Manning, C.: "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543 (2014)
23. Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L.: Deep Reinforcement Learning-based Image Captioning with Embedding Reward. arXiv preprint arXiv:1704.03899 (2017)
24. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and VQA. arXiv preprint arXiv:1707.07998 (2017)

25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, pp. 740-755 (2014)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp. 770-778 (2016)
27. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z.: Imagenet large scale visual recognition challenge. In: International Journal of Computer Vision, vol. 115, no. 3, pp. 211-252 (2015)
28. Loper, E., Bird, S.: NLTK: The natural language toolkit. In: ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1, pp. 63-70 (2002)
29. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: 40th annual meeting on Association for Computational linguistics, pp. 311-318. (2002)
30. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: IEEE conference on computer vision and pattern recognition, pp. 4566-4575 (2015)
31. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Ninth workshop on statistical machine translation, pp. 376-380 (2014)
32. Lin, C. Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out: Proceedings of the ACL-04 workshop, vol. 8 (2004)