Image captioning using semantic understanding

By Jayant Parashar, Danish Khan and Maikhar Parikh

Abstract

Automatic captioning of images is a computer vision and NLP problem. The goal is to map an image to a right sequence of words. The state of the art in image captioning is an approach implementing neural network based methods. A multimodal learning based model was a pioneer in establishing neural network methods. However, in this report, we discuss performance of an encoder-decoder generative model that maximizes the likelihood of the target caption sentence given the training image. We experiment with various different ways to train the model using NLTK on training sentence captions. Our results convey that additional context helps LSTM to produce better accuracy.

1. Introduction

Image captioning is an important problem. By providing descriptions of images in real time, blind people can comprehend the environment easily. It also can drastically improve search algorithms' performance on images.

This task is harder than image classification. As a matter of fact, image captioning problem has image classification as one subset of the problem. An image caption must describe an image and what objects exist in the image as well as how are they related to one another. And a caption in English would also require the given algorithm to have a language model coupled with understanding of image.

Most approaches, namely retrieval based, template based, augmented by neural networks and multimodal learning weave together various sub problems in order to generate a caption of images. However, Google's Show and Tell model uses single joint model that takes an image I as input, and is trained to maximize the likelihood p(S|I) of producing a target sequence of words S = {S1, S2, . . .} where each word St comes from a given dictionary, that describes the image adequately.

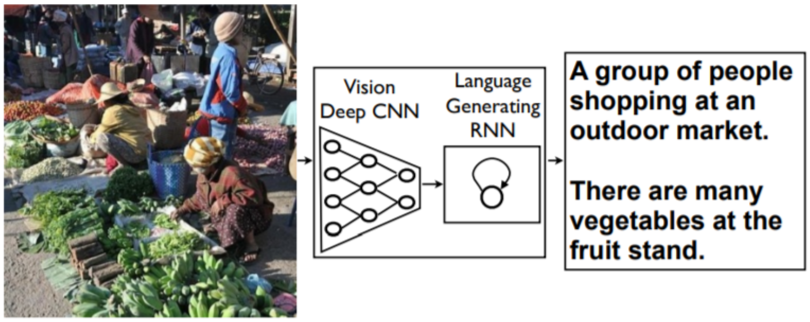

Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

Figure 1: [3]

One counter intuitive way to look at image captioning is to consider it a translation problem. This reduces the problem to merely converting a visual representation to a textual format. This opens many paths to attack the problem. Therefore recent work state of the art in machine translation has shown that translation can be done in a much simpler way using Recurrent Neural Networks (RNNs). Machine translation can be solved using two RNNs , using an “encoder” RNN to read the source sentence which is then transformed into a rich fixed-length vector representation, which in turn in used as the initial hidden state of a “decoder” RNN that generates the target sentence. The authors of “Show and Tell” refer to this as primary insight behind their new approach.

The new approach instead uses combination of CNN and RNN. The encoder RNN in machine translation is replaced by a deep convolutional neural network CNN. The effectiveness of CNNs in producing rich representation of input image by embedding it to fixed length vector is well known. Thus the choice is quite natural for what to use as encoder. Therefore the Encoder-decoder framework of " show and Tell" uses CNN as an image "encoder", by first pre-training it for an image classification task and using the last hidden layer as an input to the RNN decoder that generates sentences as shown in Figure 1[from show and tell]. We call this model the Neural Image Caption, or NIC.

With the given model, we first experiment and try to implement it on tensorflow.

Then we execute a very simple yet powerful idea. Since we wish to use external knowledge about how the world is somehow improve the performance of our model. We use NLTK library to generate some new contexts of a given image features. We take the output of CNN and input it on many trained NLP engines to generate best context words. And then we add those context words in the input to RNN. One way to look at this is "sort of hints" that help the RNN generalize better. And in some way, it provides the network a way to understand more from external NLP engines that provide context.

2. Literature Review & Related Work[3]

Method		Representative methods
Early work	Retrieval Based	Farhadi et al. [5]
	Template based	Kulkarni et al. [6]
Neural Network based	Augmenting early work by deep models	Karpathy et al. [7]
	Multimodal learning	Karpathy and Li [8]
	Encoder–decoder framework	Vinyals et al. [9]
	Attention guided	Xu et al. [10]
	Compositional architectures	Tran et al. [11]
	Describing novel objects	Mao et al. [12]

2.1 Retrieval based image captioning

Retrieval based was the first approach for image captioning problem. It analyzes a query image and retrieves a response out of a predefined sentence pool. The response can be composed of chosen best sentences or it can be a single chosen sentence from the sentence pool.

Farhadi et al. creates a meaning space that consists of object , action and scene that links image and sentences. The query image is mapped to meaning space by using a Markov random field.

[32] M. Hodosh imagine image captioning as a ranking task. The authors employ the KCCAT (Kernel Canonical Correlation Analysis technique) to project image and text items into a common space, where training images and their corresponding captions are maximally correlated. In the new common space, cosine similarities between images and sentences are calculated to select top ranked sentences to act as descriptions of query images.

Another well known retrieval method [6] captions an image by Ordonez et al. first employing global image descriptors to retrieve a set of images from a web-scale collection of captioned photographs. Then, they utilize semantic contents of the retrieved images to perform re-ranking and use the caption of the top image as the description of the query.

Many such methods work on an assumption that given a query image, a suitable description exist in their sentence pools. This assumption is seldom true in real world application. And because of this reason, retrieval based methods fail at producing novel language for novel images.

2.2 Template based image captioning

In template based methods, image captions use a syntactic and semantic process. First a specified set of visual concepts need to be detected. Then, the detected visual concepts are connected through sentence templates or specific language grammar rules or combinatorial optimization algorithms to compose a sentence.

Template based image captioning can generate sentences that are syntactically correct, and descriptions produced are also better than retrieval based methods. However, there are also disadvantages for template based methods. Because image captioning under the template based framework is limited by image contents recognized by visual models. This causes severe limitations in creativity and novelty of captions created. And if we compare the result of template based image captions to human output, they appear quite unnatural.

2.3 Deep Neural networks based methods

Progress made in deep learning have given birth to whole new set of image captioning models that outperform both retrieval based and template based methods. Many such methods can be further divided into multimodal learning ,encoder decoder framework, attention guided framework, compositional architecture.

2.3.1 Retrieval and template based methods augmented by neural networks

Hard-engineered features performed consistently bad on image captioning problem. Therefore, the inspiration was to use old retrieval based methods by using deep models to convert image captioning as multi-modality embedding and ranking problem. To search a description sentence for a query image, Socher et al. propose to use dependency-tree recursive neural networks to represent phrases and sentences as compositional vectors. They use another deep neural network [14] as visual model to extract features from images . Obtained multimodal features are mapped into a common space by using a max-margin objective function. In the end, a sentence is selected based on representations of images and text vector in common space.

Karpathy et al. used multimodal embedding to embed sentence fragments and image fragments into a common space for ranking sentences for a query image. Representing both image and sentence fragments as feature vectors, the authors create a structured max-margin objective, which includes a global ranking term and a fragment alignment term, to map visual and textual data into a common space. In the common space, similarities between images and sentences are computed based on fragment similarities, as a result sentence ranking can be conducted at a finer level.

2.3.2 Multimodal learning

Pure learning based approaches can yield original and expressive captions. Multimodal neural networks was one of the first pure learning model that was used. A general structure of the multimodal learning is given in Fig 2 [4].

Firstly, image features are extracted by a feature extractor, such as deep convolutional neural networks. Then, the generated features are forwarded to a neural language model, which maps the image feature into a common space with the word features and perform word prediction given the image feature and previously generated words.

A language model is used to generate a probability of the word wt conditioned on previously generated words w1, ... , wt−1, which is shown below: P(wt | w1, . . ., wt−1).

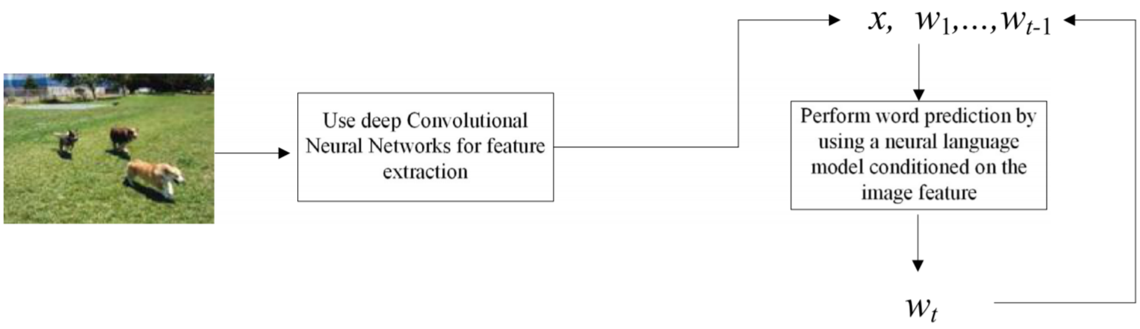

$x, \; w_1, \ldots, w_{t-1}$
Use deep Convolutional Neural Networks for feature extraction
Perform word prediction by using a neural language model conditioned on the image feature
w_t

Fig 2 General structure of multimodal learning based image captioning methods.

Figure 2 [4].

The authors make the language model become dependent on images through two different ways, i.e. adding an image feature as an additive bias to the representation of the next predicted word and gating the word representation matrix by using the image feature. Consequently, in the multimodal case the probability of generating a word wt is as follows: P(wt | w1, . . ., wt−1, I). (2) where I is an image feature. In their method, images are represented by a deep Convolutional Neural Network, and joint image text feature learning is lemented by back propagating gradients from the loss function through the multimodal neural network model.

2.3.3 Encoder decoder framework

The encoder–decoder framework is inspired from machine translation to generate captions for images. General structure of encoder–decoder based image captioning methods is shown in Fig. 2. This framework is described more in related work discussion.

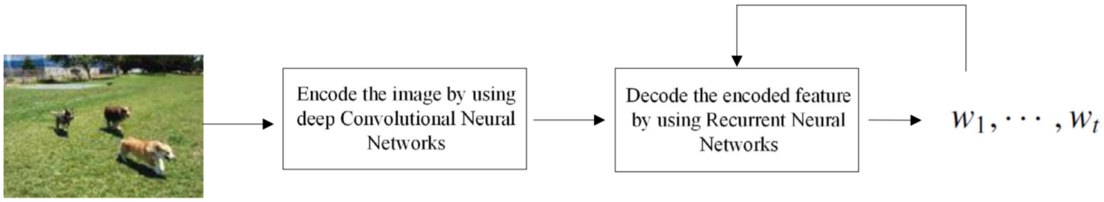

Encode the image by using deep Convolutional Neural Networks
Decode the encoded feature by using Recurrent Neural Networks
$w_1, \cdots, w_t$

Fig 3 General structure of encoder–decoder based image captioning methods.

Figure 3 [4]

2.3.4 Attention guided framework

When we look at images , we do not process whole image simultaneously. Our attention dissects the given image and then recognizes features . Similarly, when a human creates an image description, the description contains only the most useful narrative, ie we focus our attention on limited set of objects and their relations. This property of attention can be used in encoder decoder framework such that only the most salient contents are supposed to be mentioned in the description. In [15], approaches that utilize attention to guide image description generation are first proposed. By combining attention to the encoder–decoder framework, sentence generation will be conditioned on hidden states that are computed based on attention mechanism.

.

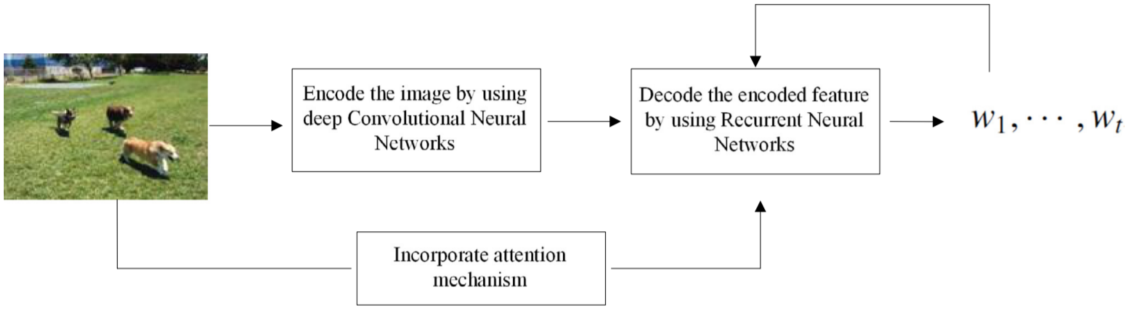

Encode the image by using deep Convolutional Neural Networks
Decode the encoded feature by using Recurrent Neural Networks
$w_1, \cdots, w_t$
Incorporate attention mechanism

Fig 4 General structure of attention guided image captioning methods.

General structure of attention guided image captioning methods is given Fig. 4[4]

Attention mechanism chooses various kinds of cues from the input image and incorporates it into the encoder–decoder framework to make the decoding process focus on only some parts of the input image. [16], Xu et al. also proposes an attention based encoder–decoder model by dynamically attending salient image regions during the process of image description generation.

3.The problem[4]

The problem is to provide relevant captions of images. One counter intuitive way to look at image captioning is to consider it a translation problem. This reduces the problem to merely converting a visual representation to a textual format. Google's Show and Tell model combines CNN for image classification with recurrent networks for sequence modeling , to create a single network that generates captions of image. The RNN is trained in the "end-to-end" network. This model is built based on two RNN models of machine translation, by changing initial RNN for first language with a CNN to decipher image features. Pioneering work by Kiros et al. [17] who use a neural net, a feedforward one, to predict the next word given the image and previous words was a seminal work in the area at that time . More recent trend set by Mao et al. [18] uses a recurrent NN for the same prediction task. Google's Show and Tell is quite similar to these approaches, but however differ in following ways : RNN with more hidden layers is used, and the visual input is directly input into the RNN, which makes it RNN more powerful by mapping all objects in text and their relations generated from image. And Kiros et al. [17] constructs a joint multimodal embedding space by using a working computer vision model and an LSTM. This approach however, uses two separate pathways (one for images, one for text) to define a joint embedding, and, it is a ranking based methods.

In this report we implement encoder-decoder framework proposed by Google ‘s Show and Tell. And its sheer simplicity results in more accuracy than other models. However, we used a toned down version of the model, therefore, BLEU score accuracies also suffer due to computational restraints.

3.1 Model [4]

In statistical machine translation, state of the art results are achieved by maximizing the probability of a right translation, given a sentence. These models use RNNs which encodes the variable length input into a fixed dimensional vector, and uses this representation to “decode” it to the desired output sentence. Thus, it is natural to use the same approach where, given an image (instead of an input sentence in the source language), one applies the same principle of “translating” it into its description.

Thus, we propose to directly maximize the probability of the correct description given the image by using the following formulation:

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

where θ are the parameters of our model, I is an image, and S its correct transcription. Since S represents any sentence, its length is unbounded. Thus, it is common to apply the chain rule to model the joint probability over S0, . . . , SN , where N is the length of this particular example as :

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

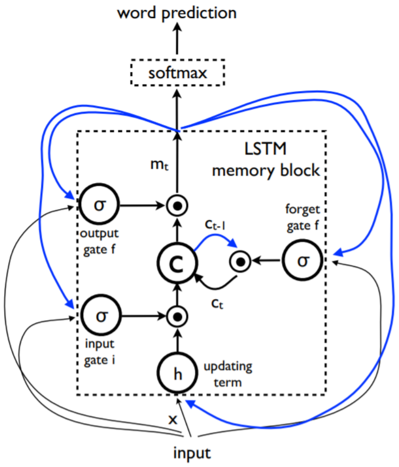


Figure 5 : LSTM: the memory block contains a cell c which is controlled by three gates. In blue we show the recurrent connections – the output m at time $t-1$ is fed back to the memory at time t via the three gates; the cell value is fed back via the forget gate; the predicted word at time $t-1$ is fed back in addition to the memory output m at time t into the Softmax for word prediction.

In Figure 5[3], the LSTM is shown.

And finally the model is implemented as shown in Figure 6[3].

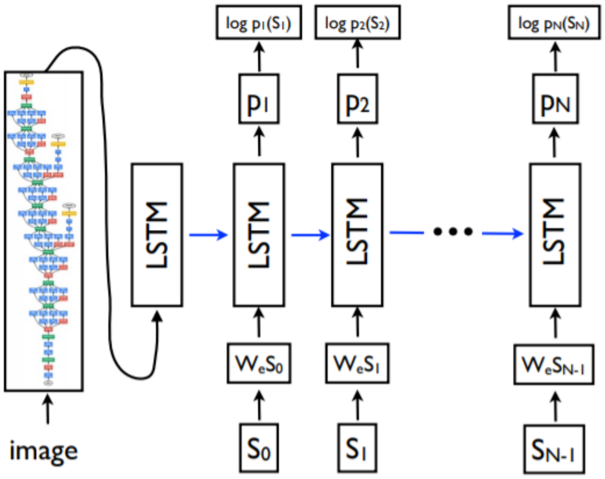
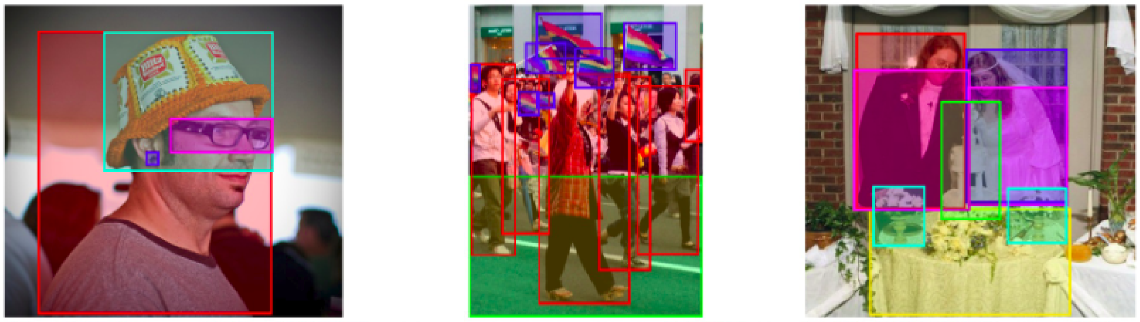


Figure 6 LSTM model combined with a CNN image embedder and word embeddings. The unrolled connections between the LSTM memories are in blue and they correspond to the recurrent connections in Figure 5. All LSTMs share the same parameters.

4. Dataset

4.1 Training Dataset

We have used Flickr30K dataset as training dataset because it has become the standard dataset for description based image dataset. We downloaded it from [2] which had results.token file which had the descriptions for the image. The Flickr30K dataset has 31,783 images in it. Every image in the dataset has 5 associated sentences or descriptions of the image itself. On an average, every image has about 8.7 objects. There are in total 44,518 categories of the objects in all the images in the dataset. Every category has about 6 objects roughly. Each description of the image can be separated and associated with a certain region in the image. This helps in determining the significance of the object in the image. This significance and all the 5 descriptions can be used to create a common suitable caption for the image.



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .

A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

Figure 7: Examples from the Flickr30K dataset [1][2]

These are some of the examples from the dataset. As we can see from the descriptions, the objects are linked to particular regions of the images. As we can in the first image, “man” is bounded in the red box while the glasses are the pink box. This creates association between the descriptions and the images and help in generating a new suitable caption.

For the words in the caption, we first created a vocab. For this words greater than a certain threshold are only used. Once the vocab is created, the words are converted to vectors. There is an array of word vectors of the size of word count.

4.1 Testing Dataset

We download Flickr 8k dataset which also has 5 captions for every image. It contains 8,000 images of which we only use 1000 for testing purposes due to computational restraints.

5. Experiments

In Image captioning, there are many ways to test the accuracy of a model. But there are some standout methods such BLEU score, cosine similarity and Jaccard Similarity.

5.1 Testing Mechanisms

In image captioning, we get a caption as an output. And to test the accuracy of that output, we can compare it with given 5 captions in testing dataset. However, there are many ways to write a sentence in natural language. This poses a problem for testers.

So, many automatic testers are formulated that tests the captions for human like descriptions based on given 5 captions. The best score from 5 captions is chosen. The reason for not taking an average is that, every one of the 5 captions represent a way to describe the image that is relevant to humans. And if a model produces a caption which is semantically similar in vector space to any of the captions, it should be given high score. The best types of tests for semantic similarity are shown below.

BLEU score

BLEU, or the Bilingual Evaluation Understudy, is a score for comparing a candidate translation of text to one or more reference translations. Although developed for translation, it can be used to evaluate text generated for a suite of natural language processing tasks. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0. It offers compelling benefits like quick and inexpensive to calculate, easy to implement and is widely used to test accuracy of an Image captioning model.

NLTK provides the sentence_bleu() function for evaluating a candidate sentence against one or more reference sentences. We provide a list of true captions i.e 5 for every image, and compare/calculate BLEU score against the caption generated by the system.

We get a mean BLEU score of 0.643 on testing dataset on base model.

Jaccard Similarity:

Jaccard similarity or intersection over union is defined as size of intersection divided by size of union of two sets. In order to calculate similarity using Jaccard similarity, we will first perform lemmatization to reduce words to the same root word.

We get mean Jaccard similarity score 0.500 on testing dataset on base model.

Cosine Similarity:

Cosine similarity calculates similarity by measuring the cosine of angle between two vectors. This is calculated as:

Similarity: Cos(Θ) = $A.B\ /\ ||A||.||B||$

With cosine similarity, we need to convert sentences into vectors. One way to do that is to use bag of words with either TF (term frequency) or TF-IDF (term frequency- inverse document frequency).

We get Mean Cosine similarity score 0.171 on testing dataset on base model.

We design our experiments to provide extra context to LSTM with NLTK library. The output of CNN is concatenated with extra context. This is a simple yet powerful idea, because it improves the chances of input correctly mapping to a correct caption. We make following changes.

1. We provide base form words of every object using Lemma of synset of NLTK.

2. We also do sentiment analysis of CNN output. And we concatenate the output of sentiment analysis with the output of 1. The sentiment usually comes out to be negative or positive. This can help divide the feature space based on perceived sentiment of our given algorithm. We accomplish this using Sentiment Analyzer method from NLTK.

3. Next we input CNN output into a semantic reasoner, which provides us a theory of what the objects' semantic meaning is. This theory is an ungrammatical mashup of words which can be a bad caption for our image also. However, we concatenate it to output of CNN to provide additional context to

6. Results

Let's denote changes described in above section as 1,2 and 3 for convenience. When we make all 1,2,3 changes , we get following results :

We get a mean BLEU score of 0.679 on testing dataset.

We get mean Jaccard similarity score 0.550.

We get Mean Cosine similarity score 0.130.

When we make only 1) change , we get almost similar results:

We get a mean BLEU score of 0.640 on testing dataset.

We get mean Jaccard similarity score 0.110.

We get Mean Cosine similarity score 0.800.

When we make only 2) change , we get worse than base model results :

We get a mean BLEU score of 0.563 on testing dataset.

We get mean Jaccard similarity score 0.440.

We get Mean Cosine similarity score 0.071.

When we make only 3) changes, we get better than base model results :

We get a mean BLEU score of 0.682 on testing dataset.

We get mean Jaccard similarity score 0.550.

We get Mean Cosine similarity score 0.210.

The results are marginally improved from base model.

References

[1] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik, Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models, IJCV, 123(1):74-93, 2017

[2] https://drive.google.com/file/d/0B_PL6p-5reUAZEM4MmRQQ2VVSlk/view

[3] Dumitru Erhan, Samy Bengio, Alexander Toshev, Oriol Vinyals, Show and Tell: A Neural Image Caption Generator

[4] Shan An, Shuang Bai, A survey on automatic image caption generation, 2017

[5]] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: Generating sentences from images, in: Proceedings of the European Conference on Computer Vision„ 2010, pp. 15–29.

[6]] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, BabyTalk: understanding and generating simple image descriptions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2891–2903

[7] A. Karpathy, A. Joulin, F. Li, Deep fragment embeddings for bidirectional image sentence mapping, in: Proceedings of the Twenty Seventh Advances in Neural Information Processing Systems (NIPS), 3, 2014, pp. 1889–1897.

[8]] A. Karpathy, F. Li, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.

[9]] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.

[10] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, arXiv:1502.03044v3 (2016).

[11]] K. Tran, X. He, L. Zhang, J. Sun, Rich image captioning in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 434–441

[12] J. Mao, X. Wei, Y. Yang, J. Wang, Learning like a child: fast novel visual concept learning from sentence descriptions of images, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2533–2541.

[13] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: data, models and evaluation metrics, J. Artif. Intell. Res. 47 (2013) 853–899

[14] Q.V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A.Y. Ng, Building high-level features using large scale unsupervised learning, in: Proceedings of the International Conference on Machine Learning, 2012.

[15] R. A.Rensink, The dynamic representation of scenes, Vis. Cognit. 7 (1) (2000) 17–42.

[16] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv:1409.0473v7 (2017).

[17]R. Kiros and R. Z. R. Salakhutdinov. Multimodal neural language models. In NIPS Deep Learning Workshop, 2013.

[18] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. Explain images with multimodal recurrent neural networks. In arXiv:1410.1090, 2014.