Image Captioning using Deep Learning Approach

Tanmay Kank
1110177
MSc. in Comp. Sci.
Lakehead University, Thunder Bay
tkank@lakeheadu.ca

Ngudup Tsering
1104377
MSc. in Comp. Sci.
Lakehead University, Thunder Bay
ntsering@lakeheadu.ca

Tenzin Nyima
1119946
MSc. in Comp. Sci.
Lakehead University, Thunder Bay
tnyima@lakeheadu.ca

Khushal Paresh Thaker
1106937
MSc. in Comp. Sci.
Lakehead University, Thunder Bay
kparesh@lakeheadu.ca

***Abstract*—The ultimate goal of the study of Artificial Intelligence (AI) is to make machines understand and interact at human level. First, we make the machines understand words, and eventually the whole context behind the human actions and the intent. Natural Language Processing (NLP) is the field that make this goal possible. NLP takes advantages of the progresses made in the Machine Learning (ML) field mainly from the Deep Learning (DL) category. The application of DL is prevalent in Image Captioning. Image Captioning is a two-step-process: image extraction and the interpretation of the image or the language generation. The DL architecture used in this project is InceptionV3 and Gated Recurrent Unit (GRU). Using the knowledge transfer from ImageNet, InceptionV3 is used to extract the key features from the Microsoft Common Objects in Context (MS-COCO) dataset. GRU is trained on the captions provided by the dataset to learn to generate the captions employing word Embedding model available in Keras. The trained model produced a satisfactory result of 28.59 Bilingual Evaluation Understudy (BLEU) score on one thousand captions.**

Keywords–NLP, DL, InceptionV3, GRU, MS-COCO Convolution Neural Networks, Image Captioning, BLEU

I. Introduction

Every day, there are a lot of images that are captured and uploaded on the web from various sources. These varied sources have images that do not have any captions but humans can still understand it because of the human intelligence. Image captioning is the process of providing a suitable phrase which describes the given image. One can immediately envision that once machines can see and describe an image like humans do, then this application can be used to help visually impaired people and also make many activities easier.

Feature extraction and Caption generation are the two main components involved in generating a caption of an image using DL model. Caption generation models must not only detect the objects in an image but also frame meaningful captions with the content of the image in natural language.

For the purpose of having to resemble the human ability in compressing huge volumes of visual data into captions describing an image, this field of NLP and DL is considered to be extremely challenging. Even with the challenges, of late there has been a boom in this field of Artificial Intelligence. With advancements in neural network by Krizhevsky et al. [9] and availability of large datasets provided by Russakovsky et al. [10], there has been a major improvement in generation of captions for images using Convolution Neural Network (CNN) and RNN that present a vector representation of images and decode them into natural language sentences respectively. Corbetta et al., [10] and Rensink et al., [11] introduced the concept of attention that allowed dominant features to be visible at the foreground when needed instead of compressing an entire image and providing a static representation of it. Even though this concept was used by most previous works presented in the field of image captioning, there is a drawback of losing the important information that might generate descriptive captions.

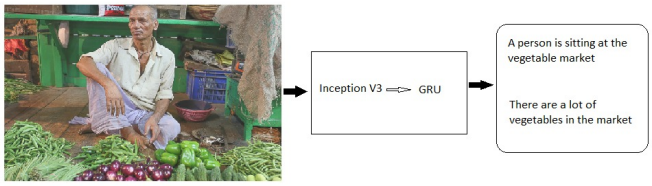


Fig. 1: Proposed model that generates complete sentences from natural language from an input image as shown in above illustration.

A typical RNN predicts one word at a time sequentially. A conditional probability distribution is predicted at every time interval. RNN govern the computations of hidden states which then generates the caption using Long Short-Term Memory (LSTM) or GRU. Models using RNN suffer from a drawback wherein there is a inherently sequential training process for every image captioning pair. Image classification is lower as well. Even with LSTM, there is a frequent problem of vanishing gradient [14]. We use NLP and DL to create an

encoder-decoder model to provide captions for the image provided as shown in Figure 1.

Section II introduces various research carried out on image captioning literature. The related work contains various models, and datasets used in the process. The methodology is discussed in the Section III which describes the proposed model, highlighting the use of various pre-processing steps, encoder and decoder. Section IV explores the Dataset adopted along with the evaluation metrics followed by Results in Section V and Conclusion in Section VI.

II. Literature Review

The idea of captioning a still image has garnered more traction of late. The recent advancements of DL and NLP has allowed systems to generate captions with limitations in expressiveness. Farhadi et al. [1] converts triplets of image elements into text using samples with the help of detections. There have been many neural network models proposed that have sentences and images co-embedded together but do not generate important captions for it [2].

Vinyals et al. [3] make use of a Neural Image Caption (NIC) model in which CNN is used as an encoder. To perform the task of image classification, a pretrained CNN is used. The data provided by the last layer is fed as the input to the RNN which in this model is the decoder that is used to generate meaningful sentences. The NIC model also makes use of LSTM and images were shown to the RNN only at the beginning [3]. Minsi et al. [4] proposes a model in which RNN's hidden parts are divided into many same sized parts which are made to work in parallel to increase the performance and decrease the complexity. These hidden layers are provided with similar feature vectors in the forward propagation phase and then transmit the output to the RNN. Xu et al. [5] introduces a framework with an attention model which is trained by back propagation and a stochastic attention model by varying the lower bound. The model proposed does not use object detectors but learns alignment of latent from the beginning and makes use of a CNN to extract the annotation vectors. Feature extraction is carried out by using lower levels of the convolutional layer so that a correspondence between portions of a 2-D image and annotation vectors is achieved. LSTM is used in order to generate the caption sequentially. The model was then validated on Flickr8k, Flickr30k and on MS COCO dataset as well [5] [17].

The concept of using neural networks to generate captions was initially proposed by Kiros et al. [6], in which features of the images are biased with the help of multi modal log bilinear approach. Ryan et al. [7] later proposed a method that performed ranking of captions and generating the captions in a natural way where in, a LSTM was used to encode the text. They used two different paths to propose a joint embedding having the approach leading to better ranking.

Mao et al. [8] recently proposed a RNN for predicting the next word provided there is a image and a previous word. Existing approaches perfrom image captioning by either starting from an image and converting it into words or generating words for different parts of an image and then combining them. Quanzen et al. [13] proposes an alternative model in which semantic attention is used to combine both the bottom-up and top-down approach. The proposed model makes use of semantic attention to combine the input to hidden states and the output of RNN, which is then evaluated upon MS COCO and Flickr30k datasets. RNN is fed with visual features which evolve over time for caption generation. The generated word is sent back to RNN along with image features generated from CNN so that RNN has an overall idea of the image [13]. Inspired by previous works on convolutional image caption generation, Aneja et al. [14] proposed a convolutional model and compared it with the LSTM model on MS COCO dataset. In order to avoid the issues of RNN, the model proposed by Aneja et al., [14] makes use of masked convolutions which are feed forward unlike RNNs. Prediction of words rely on previously generated words. With no recurrent connections, all the ground truths are present at each given time step allowing their model to be trained parallelly. Niange et al. [15] proposes a three-level hierarchical model containing two CNNs. One CNN is used to detect the image topic and the other is used to extract the features. The CNN based Image Topic detector (ITD) detects and predicts a topic and extracts the features of an image. These features are mapped to an order-spacing space. A region embedded by the topic embedding vector and image feature embedding vector is used to extract a caption embedding vector. Image topic detection is carried out using VGG-net (Visual Geometry Group) that contains 19 layers. The predicting layer of VGG is connected with sigmoid activation layer. LSTM is then fed with input for further caption generation. While having the model tested, image is provided as an input which is driven to a sigmoid transform layer (predicted layer) which then finally helps to generate the topic prediction. A RNN encoder is used with a GRU (Gated Recurrent Units) which is like a LSTM with a forget gate but has fewer parameters than LSTM as it doesn't contain an output gate to embed the caption. LDA (Latent Dirichlet Allocation) is used to have the topics extracted.

III. Methodology

In this section, we will describe our model which is an attention-based encoder-decoder model. The attention model presented by Dzmitry et al. [18] is a mechanism that was developed to improve the performance of the Encoder-Decoder RNN on machine translation. In our project, this mechanism enables us to see what parts of the image the model focuses on as it generates a caption. Our model aims to generate a caption such as "a couple of men riding a motorcycle in a parade " for an image as shown in Figure 2. In order to achieve it, we have used MS-COCO dataset as a training corpus. The encoder-decoder model takes a subset of 50000 captions for training. The process flow of the model proposed is shown in Figure 3.

Fig. 2: Example of a Training Image.

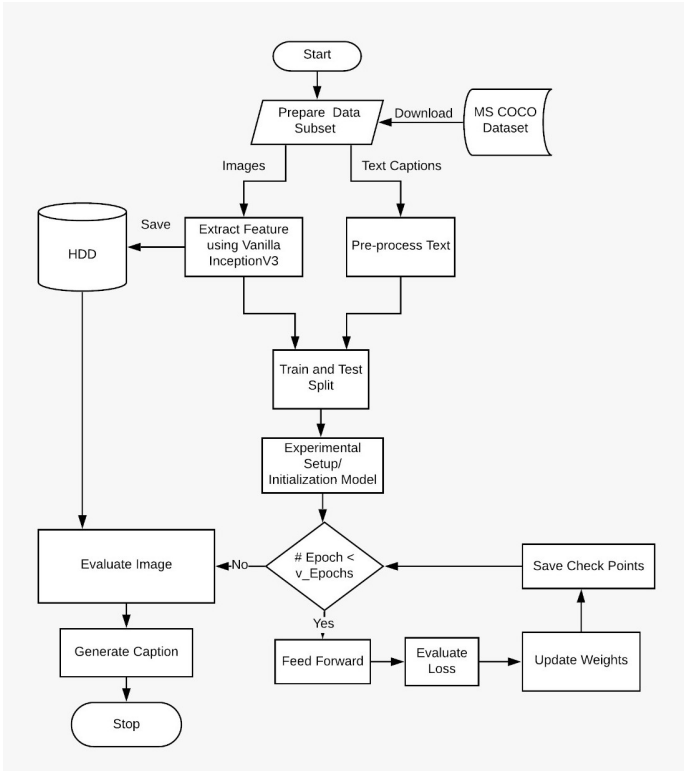


Fig. 3: Flow chart of the Process.

A. Preprocessing

The model is trained with relatively small amount of data of first 50,000 captions for about 20,000 images on account of resource constraint. MS-COCO dataset has multiple captions per image. The features of the images were extracted using the Inception V3 model pre-trained on the Imagenet [16]. Inception V3, which acts as a "multi-level feature extractor", is used to classify each image to extract features from the last convolutional layer. The images were resized to 299×299 pixels and normalized. Each image will pass through the network and the resulting vectors are stored. The duplicate images were removed and a dictionary of unique feature vectors was built. After this, the captions were pre-processed and tokenized to give a vocabulary of all of the unique words. However, on account of limited resources the vocabulary size was limited to the top 5,000 words by replacing the rest of other words with "UNK". Word-to-index and index-to-word mappings was created and finally, all the sequences are padded to be the same length as the longest one.

B. Model

The model architecture is inspired by the work proposed by the Xu et al. [5]. The features are extracted from the lower convolutional layer of InceptionV3 giving a vector of shape $8 \times 8 \times 2048$ which is then squashed to 64×2048. This vector is then passed through the CNN encoder, which consists of a two fully connected layer. Finally, GRU attends over the image to predict the next word.

1) Encoder: A single image is given to the model which generates a caption y encoded as a sequence of 1-of-K encoded words.

$$y = \{y_1, ..., y_C\}, y_i \epsilon R^K \tag{1}$$

where C is the length of the caption and K is the size of the vocabulary. CNN is used to obtain a set of feature vectors which is referred to as annotation vectors. The extractor produces L vectors, every one of which is a D-dimensional portrayal relating to a frame of the image.

$$a = \{a_1, ..., a_L\}, a_i \epsilon R^D. \tag{2}$$

The features are extracted from a lower convolutional layer to obtain a correspondence between the feature vectors and portions of the 2-D image. This permits the decoder to specifically concentrate on specific pieces of a picture by choosing a subset of all the element vectors [5]. The encoder used here was InceptionV3 for which building block of the architecture is shown in Figure 4.

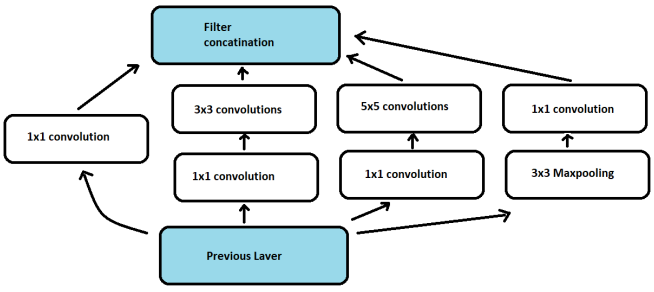


Fig. 4: Inception Block.

2) Decoder: The decoder is prepared to anticipate the following word y'_t given the setting vector c and all the recently anticipated word $\{y_1, ..., y'_{t-1}\}$

$$p(y) = \Pi_{t=1}^{T} p(y_t | \{y_1, .., y_{t-1}\}, c), \tag{3}$$

In this way, the decoder characterizes a likelihood over the interpretation y by decaying the joint likelihood into the arranged conditional probability:

$$p(y_t|\{y_1, ..., y_{t-1}\}, c) = g(y_{t-1}, s_t, c), \tag{4}$$

where $y = \{y_1, ..., yT_y\}$. The decoder here is identical to one in the attention model proposed by Bahdanau et al. [18]. The decoder used is GRU.

C. Training

We extract the features which are present in saved file. These features are then sent through the encoder, the output of which, along with the decoder input and the hidden state is passed to the decoder. The hidden states of the decoder and its predictions are returned as output by the decoder. These hidden states are then sent to the model and loss is calculated using the predicted values. We make use of teacher forcing technique to have the model decide the next input to the decoder wherein a teacher forcing technique allows the next input of the decoder to be the target word. Finally, gradients are calculated and applied to the optimizer and have a backpropagation process carried out. We utilize sparse categorical cross-entropy as the loss for the model as the anticipated likelihood wanders from the genuine label. Adam is selected as the optimizer for the model because it combines the positives of stochastic gradient descent with momentum and RMSprop. As our model makes use of an adaptive learning rate method, Adam is more useful. Figure 5 demonstrates the gradual descent of the loss during the training of the proposed model for 20 epochs.

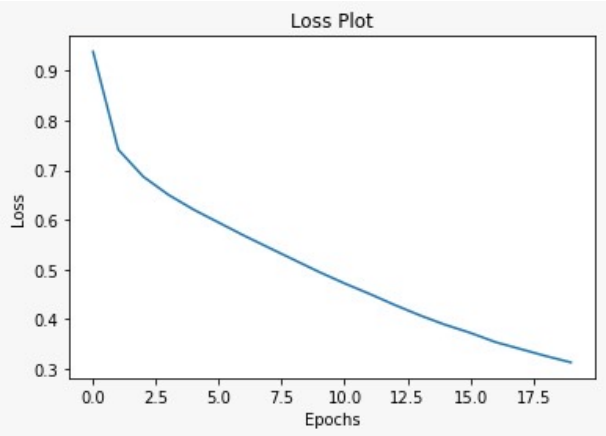


Fig. 5: Loss graph.

IV. Dataset And Evaluation Metrics

A. Dataset

We report results on the MS-COCO dataset which has more than 80,000 images. COCO is a huge corpus of object identification, division, and captioning dataset. This form of the dataset contains pictures, bounding boxes, and marks for the 2014 adaptation. The train and validation sets have few images that don't have valid explanations. Even though train, validation and test sets are different, COCO 2014 and 2017 utilizes similar images. The test split contains only the pictures and doesn't have any explanations for the image. COCO characterizes 91 classes however the information just uses 80 classes. Panoptic annotations define 200 classes but only uses 133. This dataset has images with more than one captions and also some of the images have over five references. So we decided to choose a subset of 50000 captions and its related images.

B. Sparse Cross Entropy Loss

Cross entropy is a loss function to measure the dissimilarity between the distribution of observed class labels and the predicted probabilities of classes. Categorical cross entropy refers to the possibility of more than two classes occurring. Sparse signifies having to use a single integer from zero to the number of classes minus one for a class label, instead of a dense one-hot encoding of the class label. The Loss function is given by

$$L(y, \widehat{y}) = \sum_{0}^{N} \sum_{0}^{M} (y_{i,j} * \log(\widehat{y_{i,j}})) \tag{5}$$

where $\widehat{y}$ is the predicted expected value and y is the observed value.

C. BLEU Score

BLEU is a measure that compares a unique translation of text to one or more reference translations. Even though, it was meant for translation, BLEU can be used to evaluate the text generated for various NLP tasks as well.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C} Count(n-gram')} \tag{6}$$

V. Results and Discussions

Once the model is trained, four images are randomly fed to check the model performance.

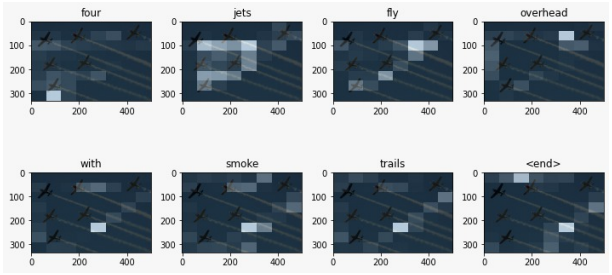
The first prediction as shown in Figure 6(a) is very human like interpretation except it counted four instead of five. It also very intelligibly avoided using the word "sky", as "fly" would infer the meaning of being in the sky. This is a high scoring interpretation.

The second caption as shown in Figure 7(a) is not as good as the first one. "Fly" and "sky" both in same sentence which is redundant. A human's interpretation would be "Two planes in the sky" or "Two planes flying". This is an average scoring performance.

It is interesting to note that the patches of attention on the parts of images as shown in Figure 6(b) and Figure 7(b) have focused on the objects as well as the change of patterns (edges and corners) which implies the attention model is well-trained to identify special features.

Figure 8 should show a caption "A small bird sitting on a rock". The model used "wall" instead of "rock". This could be a piece of concrete for a wall, but for the human lingual, "rock" is the better choice of word.

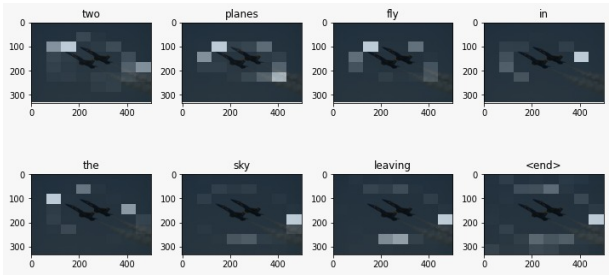
(a) Predicted Caption: ”four jets fly overhead with smoke trails”.



(b) Patches of Attention on Parts of an Image.

Fig. 6: Output of Test Image 1.

(a) Predicted Caption: ”two planes fly in the sky leaving”.



(b) Patches of Attention on Parts of an Image.

Fig. 7: Output of Test Image 2.

Fig. 8: Prediction Caption: the small bird sitting by a wall on an area

Fig. 9: Prediction Caption: a split picture of puppies are laying on well decorated

The last image caption as shown in Figure 9 is very far from the reality. The dog’s age is way passed the puppy’s age, and ”puppies” is a plural form. There is only one dog, and it got the count wrong. This time, there is also the grammar issue as well, which the first three had no issue with.

The results of our attention-based encoder-decoder model is demonstrated in the Table I. We have used the BLEU metric, a standard widely for imagine captioning, to evaluate our model. BLEU score is an algorithm for evaluating the quality of text in the DL model. Our BLEU score was recorded to be 28.59 on one thousand captions.

TABLE I: BLEU score Results.

Captions	**BLEU**
10	28.34
100	28.53
1000	28.59

VI. Conclusion

We have presented an image captioning model, an encoder-decoder system, that can automatically view an image and generate captions for it. InceptionV3 is used to extract the key features and GRU is used to train the model on the

caption based on the MS-COCO dataset and generate the captions using embedding model. Grading the correctness of the image captions generated by the models could be subjective in general. BLEU score uses the captions given by people who assembled the dataset. Based on our evaluation, the first prediction is "very good", the second prediction is acceptable, the third one is quite "average" and the last one is "very poor". From our assessment, the captions generated are average. This goes to show Image Captioning falls under the difficulty level high. The model needs to get the syntax, the semantic, and the sentiment correct. The future work includes continuous optimization of the performance. Model fusions where each model bringing transfer-knowledge from other datasets like ImageNet, SUN397, Places365 and flickr30K would enhance the performance. The BLEU score we have is 28.59 which is fair.

VII. Contribution

Tanmay and Tenzin worked on the Related works and chipped in with important code modules for pre-processing and evaluation. Ngudup and Khushal worked on the model initialization and execution for train and test sets. Both the work was combined and checked by each other so as to have an equal contribution in the project. Every task was collaborated through intensive zoom meetings and telephonic conversations. The project could be found in the given link: https://github.com/NTsering/ImageCaption.git

References

[1] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV, 2010.

[2] R. Socher, A. Karpathy, Q. V. Le, C. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. In ACL, 2014.

[3] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, 2015.

[4] Wang, Minsi, Li Song, Xiaokang Yang, and Chuanfei Luo. "A parallel-fusion RNN-LSTM architecture for image caption generation." In 2016 IEEE International Conference on Image Processing (ICIP), pp. 4448-4452. IEEE, 2016.

[5] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In International conference on machine learning, pp. 2048-2057. 2015.

[6] Kiros, Ryan, Salahutdinov, Ruslan, and Zemel, Richard. Multimodal neural language models. In International Conference on Machine Learning, pp. 595–603, 2014.

[7] Kiros, Ryan, Salakhutdinov, Ruslan, and Zemel, Richard. Unifying visual-semantic embeddings with multimodal neural language models. arXiv:1411.2539 [cs.LG], November 2014.

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

[9] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey. ImageNet classification with deep convolutional neural networks. In NIPS. 2012.

[10] Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, Berg, Alexander C., and Fei-Fei, Li. ImageNet Large Scale Visual Recognition Challenge, 2014.

[11] Corbetta, Maurizio and Shulman, Gordon L. Control of goaldirected and stimulus-driven attention in the brain. Nature reviews neuroscience, 3(3):201–215, 2002.

[12] Rensink, Ronald A. The dynamic representation of scenes. Visual cognition, 7(1-3):17–42, 2000.

[13] You, Quanzeng, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. "Image captioning with semantic attention." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4651-4659. 2016.

[14] Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. "Convolutional image captioning." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5561-5570. 2018.

[15] Yu, Niange, Xiaolin Hu, Binheng Song, Jian Yang, and Jianwei Zhang. "Topic-oriented image captioning based on order-embedding." IEEE Transactions on Image Processing 28, no. 6 (2018): 2743-2754.

[16] Christian Szegedy and Wei Liu and Yangqing Jia and Pierre Sermanet and Scott Reed and Dragomir Anguelov and Dumitru Erhan and Vincent Vanhoucke and Andrew Rabinovich, 'Going Deeper with Convolutions',2014,arXiv.

[17] X. Chen and H. Fang and TY Lin and R. Vedantam and S. Gupta and P. Dollár and C. L. Zitnick,"Microsoft COCO Captions: Data Collection and Evaluation Server",2015

[18] Dzmitry Bahdanau and Kyunghyun Cho and Yoshua Bengio,'Neural Machine Translation by Jointly Learning to Align and Translate',2014,arXiv.