A Deep Attention based Framework for Image Caption Generation in Hindi Language

Rijul Dhir*, Santosh Kumar Mishra*, Sriparna Saha, Pushpak Bhattacharyya

Department of Computer Science and Engineering, Indian Institute of Technology Patna
{rijul.cs15@iitp.ac.in,santosh_1821cs03,sriparna,pb}@iitp.ac.in

Abstract Image captioning refers to the process of generating a textual description for an image which defines the object and activity within the image. It is an intersection of computer vision and natural language processing where computer vision is used to understand the content of an image and language modelling from natural language processing is used to convert an image into words in the right order. A large number of works exist for generating image captioning in English language, but no work exists for generating image captioning in Hindi language. Hindi is the official language of India, and it is the fourth most-spoken language in the world, after Mandarin, Spanish and English. The current paper attempts to bridge this gap. Here an attention-based novel architecture for generating image captioning in Hindi language is proposed. Convolution neural network is used as an encoder to extract features from an input image and gated recurrent unit based neural network is used as a decoder to perform language modelling up to the word level. In between, we have used the attention mechanism which helps the decoder to look into the important portions of the image. In order to show the efficacy of the proposed model, we have first created a manually annotated image captioning training corpus in Hindi corresponding to popular MS COCO English dataset having around 80000 images. Experimental results show that our proposed model attains a BLEU1 score of 0.5706 on this data set.

Keywords: Image Captioning · Hindi Language · Convolutional Neural Network · Recurrent Neural Network · Gated Recurrent Unit · Attention Mechanism

1 Introduction

Modern visual classifiers can recognize many of the object categories [6][12] which have been a main focus in the computer vision community. Automatic caption generation is one of the challenging tasks in artificial intelligence because it requires recognition of not only objects but also other visual elements such as actions, attributes as well as generating sentence which describes relationships

* Both authors have equally contributed.

between objects, actions, and attributes in the image[15]. It has emerged as a challenging and important research area because of the recent advancement in statistical language modeling and image recognition [1][13]. Generating a meaningful language description of an image requires image understanding that is quite different than image classification and object recognition. This is one of the interesting problems in artificial intelligence because it connects computer vision with natural language processing.

Image captioning could have great impact by helping visually impaired people to better understand the content on the web. Image captioning came into existence with recent advancements in machine translation where the task is to transform a sentence S written in a source language into translated target language T, by maximizing $p(T|S)$.

Generating captioning of an image is very much similar to machine translation where the source language is pixels of an image, and the target language is a sentence. In this case, convolutional neural network (CNN) is used in place of RNN [15]. In recent works, it is shown that CNN can produce a rich representation of the input image by embedding it to a fixed length vector representation such that it can be used for various computer vision tasks. Hence, CNN can be used as an encoder, by first pre-training it for image classification task and then using the last hidden layer as an input to the RNN which can work as a decoder to generate the sentences (refer to Fig. 1).

Hindi is one of the oldest languages in the world which is spoken in India and South Asia. It has a direct line of evolution to Sanskrit [11]. It is the fourth most spoken language in the world. Hindi is part of one of the oldest religious and literary tradition in the world. India is a linguistically diverse country, Indian constitution has adopted 15 languages as national importance and Hindi and English as official languages. Most of the people in India use Hindi language for communication. Hence there is a need to develop systems for Hindi language so that most of the people can get the benefit of recent research works.

Currently, Government of India is promoting Hindi language in all its departments and organizations by celebrating Hindi Diwas every year. Government is motivating its employees to work in Hindi in place of English. This work can be helpful to all public organizations where captions of images in Hindi are needed. We did not find any previous work on Hindi language image captioning. Inspired by this, in the current work, we aim to develop attention based Hindi language image captioning framework utilizing Resnet 101 [6] as Encoder and GRU based deep-learning model [3] as a decoder with attention.

To the best of our knowledge, this is the first work where an image captioning framework is developed for Hindi language. Firstly we have used a pre-trained model to interpret the content of images. There are several CNN architectures available, but we have used Resnet 101 as it was shown in the literature that it attained the best result and it uses the residual connections which can solve the degradation problem [6]. Resnet 101 architecture is used to extract features of the images by saving the output of the last hidden layer as we are interested in the internal representation of images instead of classification. Here we have used

Gated Recurrent Unit (GRU) [3] as a decoder. The output of the last hidden layer of Resnet 101 is passed to attention, and the context vector generated by it is used as an input to GRU which generates a description for the image. This architecture yields better performance as compared to standard CNN and RNN architectures where Long Short Term Memory (LSTM) is used as a decoder.

As no Hindi corpus was available for solving the task of image captioning, in a part of the work we have manually created a Hindi corpus. MS COCO [9] is a publicly available English image captioning data set containing images. This is large image dataset which is used for object detection, image segmentation, caption generation etc. We have used this dataset for image captioning. In this dataset, each image has five captions. We have manually translated this corpus for Hindi language. The proposed approach is applied to this manually annotated MS COCO Hindi dataset to show model's efficiency and to prove the supremacy of the proposed approach. We have obtained the BLEU-1 score of 0.5706, BLEU-2 score of 0.3914, BLEU-3 score of 0.2935 and BLEU-4 score of 0.1726. In order to establish the efficacy of Resnet 101 as encoder and GRU as the decoder for the task of Hindi caption generation, we have also reported the results of many other combinations of encoders and decoders utilizing models of Inception v4 [14], GRU [3] and LSTM [7]. Results demonstrate the superiority of the proposed architecture.

2 Related Works

In the literature, there exist two types of approaches for generating image captioning, the first one is the top-down approach [1][13][17] and the second one is the bottom-up approach [5][8][4].

The top-down approach starts from the input image and converts it into words while the bottom up approach first comes up with words which describe the various aspects of an image and then combines them to generate the description. In the top-down approach, there is an end to end formulation from an image to sentence using the recurrent neural network, and all the parameters of the recurrent neural network are learned from training data. Limitations of the top-down approach are that sometimes fine details can not be obtained which may be important to generate the description of an image. Bottom-up approaches do not have this problem as they can be operated on any image resolution, but there is a lack of end to end formulation which is going from image to sentence.

Bottom-up approach starts with visual concepts, objects, attributes, words and phrases which are combined using the language models. Farhadi et al. [5] defined a method that can compute a score linking to an image. This score is used to attach a description to an image.

Top-down approaches are the recent ones; in these approaches, image captioning problem is formulated as a machine translation problem. In machine translation, one language is translated to another one, but in case of image captioning, the visual representation is translated to language. Sutskever et al. [13]

proposed a sequence end to end approach for sequence learning. They have used a multilayered Long Short-Term Memory (LSTM) to encode the input sequence to a vector of fixed dimensionality, and then another deep LSTM is used to decode the target sequence from the vector. Bahdanau et al. [1] developed the encoder-decoder architecture for machine translation by allowing a model to automatically search the part of the source sentence that is relevant in predicting the target word. To the best of our knowledge, there does not exist any image captioning model for Hindi language. In this work, we have proposed an attention based Encoder-Decoder model for Hindi language captioning. Also, we have used GRU instead of LSTM due to its several advantages.

3 Proposed Methodology

In this paper, we have developed a method of caption generation based on neural and probabilistic framework. Recent development in statistical machine translation has provided powerful sequence models to generate state-of-the-art results by maximizing the probability of correct translation given an input sequence in an end to end approach.

Thus, we use the probabilistic framework to directly maximize the probability of correct description given an image by using the following formulations:

$$\theta^* = arg\,\theta \max \sum_{(I,S)} log\, p(S|I;\theta) \tag{1}$$

where I is an image, θ is the parameter of the model and S is the generated description. Since S is the generated description, its length should be unbounded. Therefore we can apply the chain rule to find the joint probability over $S_0, ..., S_N$, where N is the assumed length of the description.

$$log\, p(S|I) = \sum_{i=0}^{N} log\, p(S_i|I, S_0, S_1, ...S_{i-1}) \tag{2}$$

Here θ is omitted for convenience. (S,I) is ordered pair from training example where I is an image and S is the corresponding description. Our aim is to optimize the sum of log probabilities as described in (2) over the whole training set using Adaptive Moment Estimation(Adam) optimizer.

Here $p(S_i|I, S_0, S_1, ..., S_{i-1})$ is modeled using a Recurrent Neural Network (RNN); here variable number of words is conditioned upon $i-1$ expressed by hidden state of fixed length, h_i. This hidden memory is updated after scanning a new input, x_i, by using a non-linear function, f:

$$h_{i+1} = f(h_i, x_i) \tag{3}$$

Here Gated Recurrent Unit (GRU) is used as function f which has shown state of the art performance on the sequence to sequence translation.

3.1 Convolutional Neural Network

Convolutional Neural Network (CNN) is used for the internal representation of the image. CNN has been widely used for object recognition and object detection task in computer vision. We have used Resnet 101 convolution neural network architecture given by Kaiming He et al.[6] which won the 1st place in the ILSVRC 2015 classification competition. This is a 101 layer CNN model pre-trained on ImageNet dataset. We have removed the last layer of the model as it is used for classification but we are more interested in the internal representation of images. We have pre-processed images with the Resnet 101 model and have extracted features from block4 layer as input for further processing.

The extractor produces L vectors, each of which is a D-dimensional representation corresponding to a part of the image.

$$a = \{a_1, ..., a_L\}, a_i \epsilon R^D \tag{4}$$

Global image feature can be obtained by:

$$a_g = \frac{1}{L} \sum a_i \tag{5}$$

Image Vector and Global Image Vector can be obtained by using a single layer perceptron with rectifier activation function.

$$v_i = ReLU\big(W_a a_i\big) \tag{6}$$

$$v_g = ReLU\big(W_g a_g\big) \tag{7}$$

where W_a and W_g are the weight parameters, L is number of vectors and D is size of each vector. The transformed spatial image feature form is $V = [v1, ..., vL]$.

3.2 Attention

Attention is a process of concentrating on a discrete aspect of information while ignoring other perceivable information [2]. It is a way to instruct the model where to concentrate to generate the corresponding word rather than the whole image. At time t, based on the hidden state, the decoder would attend to the specific regions of the image and compute context vector using the spatial image features from a convolution layer of a CNN.

$$c_t = g\big(V, h_t\big) \tag{8}$$

We feed V and h_t through a single layer neural network followed by a softmax function to generate the attention distribution over the k regions of the image:

$$z_t = {W_h}^T tanh\big(W_v V + (W_g h_t) 1^T\big) \tag{9}$$

$$\alpha_t = softmax\big(Z_t\big) \tag{10}$$

where $\mathbf{1}^{\mathrm{k}}$ is a vector with all elements set to 1. $\mathbf{W}_{\mathrm{v}}$,$\mathbf{W}_{\mathrm{g}} \epsilon \mathrm{R}^{\mathrm{LXD}}$ and $\mathbf{W}_{\mathrm{h}} \epsilon \mathrm{R}^{\mathrm{L}}$ are parameters to be learnt. $\alpha \epsilon \mathrm{R}^{\mathrm{L}}$ is the attention weight over features in V. Based on the attention distribution, the context vector $\mathrm{c_t}$ can be obtained by:

$$c_t = \sum \big(\alpha_{ti} V_{ti}\big) \tag{11}$$

3.3 Gated Recurrent Unit (GRU) based Sentence Generator

x_t is the input vector and we obtain it by concatenating the word embedding vector, w_t, global image feature vector, v_g, and context vector, c_t, to get the input vector.

$$x_t = [w_t; v_g; c_t] \tag{12}$$

GRU was introduced by Kyunghyun Cho et al. [3] and applied successfully for machine translation and sequence generation. GRU is an improved version of the Recurrent Neural Network. To solve the vanishing gradient problem of standard RNN, it uses update gate and reset gate. The behaviour of the cell is controlled by these two gates.

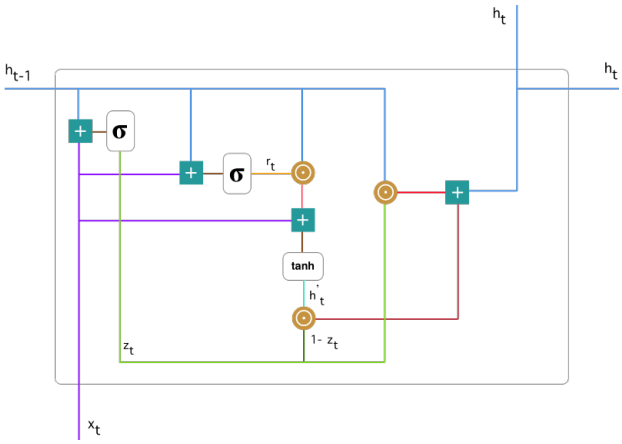

Figure 1. Gated Recurrent Unit

The core of the GRU model is memory cell which stores knowledge of every time step (what input has been observed up to present state (see Figure 1 [1])). Basically, the update gate and reset gate are the vectors responsible for deciding which information should be passed to the output. These two gates are trained to store the information from the previous time steps without wasting those as well as removing the pieces of information which are irrelevant for the prediction.

Our proposed image captioning architecture can be divided into the following components as shown in Figure 2:

– Resnet 101 Convolutional Neural Network which is pre-trained on the ImageNet dataset, is used to extract features from the images. Here CNN is working as feature extractor, or it can work as an encoder which converts images into fixed length vector representations. In our case, it is converted to feature vector of dimension 49×2048.

[1] https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be

- We have used a function to summarize the image and get global image vector, v_g which is passed to GRU for global information about image.
- Attention mechanism is used to obtain context vector, c_t which tells the model where to look in the image for corresponding word generation.
- There is a word embedding layer for handling the text input; it provides a dense representation of the words and their relative meanings. It is used to fit a neural network on the text data. Word-embedding vector w_t is given as input to the GRU.
- Input to GRU is Context vector, Word Embedding vector and global image vector with previous hidden state and it will predict the hidden state, h_t, based on these; here we have used GRU layer with 512 neurons that will provide a vector having 512 elements to represent a word.
- Both the context vector and the hidden state are passed to Multi-Layer Perceptron (MLP) Network to make predictions over the entire vocabulary for the next word for the caption.

Here CNN is working as feature extractor, generating features of dimension 49×2048; it is further processed by attention mechanism to produce a 2048 representation of the area of an image to look into. Global vector provides an image summary for GRU. Word Embedding provides a dense representation of the words and their relative meanings which is input to GRU. The outputs of the GRU layer and attention mechanism are then fed to MLP network which makes the prediction over the entire vocabulary for the next word in the sequence (see Fig 2).

3.4 Hyparameters of the Model

The input to the Resnet 101 convolutional neural network is 256×256 RGB images. The features are extracted from **block4** layer of the neural network with size 49×2048. The input captions are fed into an embedding layer with 512 neurons. We have used 0.5 dropout to avoid over-fitting. Here we have used softmax cross-entropy to get the loss and Adam optimizer to optimize our loss with a learning rate of $4e-4$ and batch size of 32. To train the proposed model with 80000 images on Cluster of 8 GPU Nvidia GTX 1080, approximately 14 hours were needed and to test 8000 images; it takes approximately 15 minutes to generate the description of the input images.

4 Experimental Setup

In this section, we have described the procedure of data set creation and experimental results.

4.1 Dataset Creation

We have chosen MS COCO dataset [9] for our experiment which is a commonly used data set for caption generation in English language. This data set includes

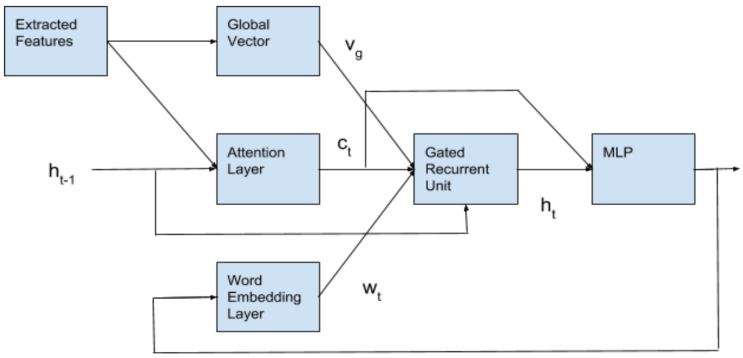

Figure 2. Process flow diagram of proposed method

complex everyday scenes with common objects in naturally occurring contexts and can be downloaded easily. Hence, it covers large possible categories of images.

As no Hindi image captioning data set is available in the literature, we have manually annotated the captions of this MS COCO data set to generate a gold standard image captioning data set for Hindi language as shown on Figure 3 and Figure 4. Firstly Google translator is applied to convert English captions to Hindi captions. We then had two annotators with the inter-annotator agreement of 84% to manually check and correct the translation on sample data of 400 captions. These annotators verify and correct the outputs obtained from Google Translator. Through this process, we convert MS COCO English dataset to Hindi language dataset.

The COCO dataset is an excellent object detection dataset with 80 classes, 80,000 training images and 40,000 validation images and each of the images is paired with five different captions which provide a clear description of the salient entities and events. Validation set contains 8000 images to monitor the performance of the trained model. When the performance of a trained model improves at the end of any epoch, then we save that model. At last, we can get the best-learned model on the training dataset. The dataset contains a test set which has 8000 images. We evaluated the performance of the learned model and its prediction on a test set.

4.2 Preprocessing

We truncate captions longer than 18 words present in images of COCO dataset. We then build a vocabulary of words that occur at least three times in the training set, resulting in 9034 words for COCO.

Figure 3. Example Image for Dataset Preparation wrt Fig. 4

English Captions	Google translated Captions in Hindi	Corrected Captions in Hindi
closeup of bins of food that include broccoli and bread	भोजन के डिब्बे को बंद करना जिसमें ब्रोकली और ब्रेड शामिल हैं	भोजन के वो डब्बे जिनमे ब्रोकोली और ब्रेड शामिल हैं उनकी पास की तस्वीर
a meal is presented in brightly colored plastic trays	एक भोजन चमकीले रंग की प्लास्टिक ट्रे में प्रस्तुत किया जाता है	भोजन चमकीले रंग की प्लास्टिक ट्रे में प्रस्तुत किया जाता है
there are containers filled with different kinds of foods	विभिन्न प्रकार के खाद्य पदार्थों से भरे कंटेनर हैं	विभिन्न प्रकार के खाद्य पदार्थों से भरे कंटेनर हैं
a bunch of trays that have different food	ट्रे का एक गुच्छा जिसमें अलग भोजन होता है	ट्रे का एक समूह जिसमें अलग अलग भोजन है
colorful dishes holding meat vegetables fruit and bread	मांस सब्जी फल और रोटी पकड़े रंगीन व्यंजन	रंगीन व्यंजन जिनमे मांस सब्जी फल और रोटी है

Figure 4. Example of Dataset Preparation

4.3 Chosen Technique for Comparative Analysis

BLEU (Bilingual Evaluation Understudy Score) is a popular evaluation technique to check the similarity of a generated sentence with respect to a reference sentence. BLEU score evaluation technique was proposed by Papineni et al. [10] in the year 2002. In the current work, this metric is used to evaluate the performance of our caption generation model.

5 Results and Discussions

This section will cover the quantitative and qualitative analysis of the generated descriptions of the images. During our research, we could not find any previous paper related to Hindi language image caption generation. So this is the first work in this field to the best of our knowledge. In order to compare the performance of the proposed architecture we have proposed the following baselines:

- Baseline 1: In this baseline, Resnet 101 is used as an encoder and LSTM is used as a decoder.
- Baseline 2: In this baseline, Inception v4 is used as an encoder and LSTM is used as a decoder.
- Baseline 3: In this baseline, Inception v4 is used as an encoder and GRU is used as a decoder.

Our proposed model is data-driven, and it is trained end to end. We have used manually annotated MS COCO Hindi dataset for training purpose. Hence proposed architectures can be trained on CPU and GPU. We have tested the performance of our proposed model on MS COCO test dataset.

5.1 Qualitative Analysis

Results of generated descriptions on the test images are given in Figure 5. As can be seen from the generated descriptions, these captions are pretty much close to the input images but still there is a need for improvement in the generated descriptions and we are working towards that. We can improve the quality of generated description for an image by training using larger dataset.

Figure 5. Some Examples of Generated Captions by the Proposed Model

Table 1. BLEU score of different architectures

State of Arts	BLEU-1 Score	BLEU-2 Score	BLEU-3 Score	BLEU-4 Score
Resnet 101 and GRU	**57.0**	**39.1**	**26.4**	**17.3**
Baseline 1	56.4	38.8	26.3	17.2
Baseline 2	56.3	38.4	25.8	16.9
Baseline 3	55.7	38.4	25.9	16.8

5.2 Quantitative Analysis

Although it is possible to evaluate the quality of generated descriptions of images manually, sometimes it is not very clear that generated description is successful or not given an image. Hence it is necessary to have a subjective score that decides the usefulness of each description given an image. There are many evaluation metrics in the literature which are used in machine translation. The most commonly used metric in machine translation literature is BLEU score [10].

Here we have used the BLEU score for evaluation of generated description of an image. In the dataset, each image has been annotated by five sentences that are relatively visual and unbiased. There are 8000 images in the test set of this dataset to test the performance of the proposed model. BLEU has been recorded for the test dataset. We have evaluated BLEU-1, BLEU-2 BLEU-3 and BLEU 4 scores of our model and all the three baselines which are given in Table-1. Results show that our proposed model outperforms all the baselines. The proposed method has obtained BLEU-1 score of 57.0, BLEU-2 score of 39.4, BLEU-3 score of 26.4 and BLEU-4 score of 17.3, which are significantly better than other three baselines as shown in Table-1.

5.3 Statistical Test

Welch's t-test, [16] statistical hypothesis test, has been conducted at 5%(0.05) significance level to show that performance improvements attained by our proposed method over other baselines are statistically significant, not happened by chance. BLEU scores are obtained by 10 consecutive runs of each architecture. Smaller p-values indicate better performance of the proposed architecture compared to other architectures. The p-values obtained by the statistical test are reported in Table 2. Results obtained established the statistical significance of the improvements obtained.

Table 2. p-values produced by Welch's t-test comparing our proposed method with other baselines

State of Arts	BLEU-1 Score	BLEU-2 Score	BLEU-3 Score	BLEU-4 Score
Baseline 2	6.86249e-34	1.58517e-23	2.68755e-36	1.67851e-10
Baseline 2	2.68755e-36	2.68755e-36	6.86249e-34	1.0313e-27
Baseline 3	3.67333e-46	2.68755e-36	1.82938e-29	4.42441e-31

5.4 Error Analysis

In this section, we have tried to analyze the errors made by our proposed system. In Fig-5 (a) generated description reports the action of 'sitting on the tree' in place of 'sitting on a bear', This may happen because there are many images in training data where the bear is associated with the tree. More frequent words

in the vocabulary have more probabilities during softmax prediction. This is the main reason for an error in the generated description. In Fig-5 (b), generated descriptions predicted 'cat' in place of 'dog' because the head of the dog is not visible. In Fig-5 (c), the generated caption is very close to activity and object inside the image, but the terms 'glass' and 'bowl' are missed in the generated description. In Fig-5 (d), the model is not able to differentiate between the terms 'sofa' and 'chair'. In Fig-5 (e), the model has correctly predicted objects of the input image. For the description of Fig-5 (f), the model has predicted people sat on the 'table' in place of 'chair'; this is because 'chair' is not visible in the input image. Below we have enumerated the possible reasons for getting errors:

- Most of the errors occurred due to non-presence of certain words in the training captions. Increasing the size and diversity of the training data may help in reducing these types of errors.
- Most of the errors occurred because certain phrases are frequently occurring in the training data. This can be solved by the use of a suitable unbiased training set.

6 Conclusions and Future Work

We have presented an attention-based image captioning model trained end to end that can view an image and generate a description in the Hindi language. The proposed model is based on a convolution neural network to encode an image into a fixed length vector representation which is further used by attention layer to produce context vector which is further used by the recurrent neural network. As there was no previous work done for Hindi language Image Captioning, we have tried various combinations for encoder and decoder and treated them as baseline models. The best result we attained is from a combination of Resnet 101 as an encoder and GRU as a decoder. Although GRU and LSTM provide similar results, GRU is more efficient due to its less complex structure and fewer number of gates. In future, we will explore the possibility of an ensemble of various possible encoder and decoder architectures to improve the result further. We will also try different sampling techniques to remove the bias towards certain phrases.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Cho, K., Courville, A., Bengio, Y.: Describing multimedia content using attention-based encoder-decoder networks. IEEE Transactions on Multimedia **17**(11), 1875–1886 (2015)
3. Cho, K., Van Merriënboer, B., 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)

4. Elliott, D., Keller, F.: Image description using visual dependency representations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1292–1302 (2013)
5. 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. pp. 15–29. Springer (2010)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
8. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating image descriptions. In: Proceedings of the 24th CVPR. Citeseer (2011)
9. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
11. Rubino, C.R.G., Garry, J., Rubino, C.: Facts about the world's languages: An encyclopedia of the world's major languages, past and present. Hw Wilson (2001)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (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. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. vol. 4, p. 12 (2017)
15. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
16. Welch, B.L.: The generalization ofstudent's' problem when several different population variances are involved. Biometrika **34**(1/2), 28–35 (1947)
17. Wu, Q., Shen, C., Liu, L., Dick, A., van den Hengel, A.: What value do explicit high level concepts have in vision to language problems? In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 203–212 (2016)