ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 3, March 2023

Image Caption Generation Using Deep Learning

[1] Nikhitha Shada, [2] K Vaishnavi, [3] Medishetty Vaishnavi, [4] D. Dhana Lakshmi

[1] [2] [3] [4] Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India
Corresponding Author Email: [1] nikhitha.reddy.01@gmail.com, [2] vaisshhuureddy@gmail.com, [3] vaishnavimedishetty77@gmail.com, [4] dhanalakshmi-d@vardhaman.org

Abstract*— The process of creating descriptions for what's happening in an image is known as image captioning. Image captioning is used to provide explanations that provide context for the pictures.The analysis of huge quantity of unidentified images, the detection of unknown patterns for applications of machine learning used to drive autonomous cars, and the creation of soft- ware that aids the blind are just a few examples of the many areas where image captioning is incredibly beneficial. Deep Learning Models can be used for this image captioning.Development in Deep learning field and NLP have made it simpler than ever to create descriptions for the provided photos. Neural networks will be used in this paper's picture captioning.In order to access picture features, RESNET is utilised as an encoder, while Long Short Term Memory(LSTM)as decoder. Using built-in language and the image features, LSTM creates captions for the images.

***Index Terms*—LSTM (Long Short Term Mem- ory), RESNET, Image Captioning, Encoder, Decoder.**

I. INTRODUCTION

Captioning Images used to be a difficult process, and the captions that are produced for an image are sometimes not very useful. Many tasks that were tough and difficult to perform using Machine Learning have become simple to accomplish with the aid of Deep Learning and Neural Networks, thanks to the progress of both text processing techniques like Natural Language Processing and Neural Networks. These are partic- ularly helpful in many applications of artificial intelligence, including picture recognition, image categorization, image captioning, and many more. In essence, captioning image is the process of creating explanations of what is occurring in the input image. In essence, this model takes input in the form of pictures and generates output as a caption. The efficiency of creating image captions is developing along with technological innovation.Image captioning is very useful for various applications, including the increasingly popular self- driving cars. Image captioning is useful for many Machine Learning tasks like recommendation systems. Many methods have been put out for captioning images, such as object identification models, deep learning-based captioning images, and visual attention-based captioning images. The Inception model, the VGG model, the RESNET-LSTM model, and the traditional CNN-RNN model are some kinds of deep learning models. In this article we will describe the methodology we used for captioning the photographs using RESNET-LSTM model.

II. EXSISTING MODELS

A. Captioning Images Based on Deep Neural Networks pro- posed by Shuang Liu, Liang Bai ,Yanli Hu and Haoran Wang

CNN(Convolutional Neural Network)-RNN(Recuurent Neural Network) Based Captioning Images and CNN-CNN Based Captioning Images are two deep learning models used in the approach presented by Shuang Liu, Liang Bai ,Yanli Hu and Haoran Wang. [1]. Convolutional neural networks are used for encoding in a CNN-RNN framework, while recurrent neural networks are used for decoding. The pictures in this case are transformed into vectors using CNN, and these vectors—which are referred to as image features—are then fed as input into recurrent neural networks. The actual captions for the project are obtained using the NLTK libraries in RNN's implementation. In CNN-CNN based frame work, only CNN is utilised for encoding and decoding of the pictures.Here, using the NLTK library, a vocabulary dictionary is employed and mapped with picture attributes to obtain the precise word for provided picture. creating the caption that is free of errors. The train the continuous flowing repeatedly repetition of these methods is undoubtedly slower than the consisting of many models that are offered at the same time of convolution approaches. Less training time is required for the CNN-CNN Model than the CNN-RNN Model. As it is sequential, the CNN-RNN Model requires more training time but has lower loss than the CNN-CNN Model.

B. Image Caption Generation model proposed by A. Hani, N. Tagougui and M. Kherallah

The encoding decoding paradigm is utilised here for pic- ture captioning in the technique put out by Ansari Hani et al[2].Retrieval-based captioning and template-based caption- ing are the other two approaches for captioning images that

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 3, March 2023

are covered here. In the process of retrieval-based captioning, training pictures are stored in one area, while the captions that were created for them are stored in a different location. In the new location, correlations between the test image and the generated captions are calculated, and The caption for the supplied image is then selected from the list of available captions by the one with the maximum correlation value.They use an approach called prototype-based descripting in this article.They used the attention mechanism and GRU as the decoder in this instance, and encoder is the Inception V3 model.

Each existing model has a drawback that reduces the model's effectiveness and accuracy when results are generated. The following are the limitations in all the models that have been observed:

1. We see that the CNN-CNN model, which uses CNN for both encoding and decoding purposes, has a large loss, which is unacceptable since the captions generated here won't be correct and won't apply to the provided test image.
2. The captions based on CNN-RNN model are better than the CNN-CNN model, but the duration of training is longer. The model's overall efficiency is impacted by training time, and this is where we also ran into another issue. Vanishing Gradient Problem, for example. When comparing inputs and outputs, the gradient parameter is utilised to determine the rate of loss for the specified input parameter.Artificial neural networks and recurrent neural networks are where this gradient descent problem mainly occurs. The gradient is the ratio of the weights' change to the neural network's output change in error. This gradient is also taken into account as the neural network's activation function's slope.The model's training and neural network model's rate of learning are both accelerated by high slope. As the number of hidden layers rises, the loss also does, but the gradient gradually gets less until it reaches zero. This gradient problem limits the learning of long-term sequences in RNN. The RNN's ability to learn and recall is hampered by this gradient descent difficulty. The words cannot be permanently preserved in hidden memory. It is therefore difficult for RNN to study the captions of the provided picture during training.Due to such gradient descent issue that arises during training, RNN is unable to store the words of longer captions for an extended period of time. The unknown key points in the captions are delivered to the forget gate of RNN as the hidden layer quantity rises; however, before the gradient reaches zero, it initially begins to decrease. As a result, the CNN-RNN model can provide captions for the pictures with minimal training. Finally, it is evident that the CNN-RNN model is ineffective and inaccurate for creating captions for pictures since RNN has a gradient descent issue.

III. PROPOSED METHODOLOGY

As we have shown, the vanishing gradient problem in the standard CNN-RNN paradigm prevents the recurrent neural network from learning and being trained effectively.Therefore, in order to alleviate this gradient descent difficulty, we propose this model in this study in order to boost both the efficiency and the accuracy of the caption generation for the image. The architecture of our suggested model is provided below. We will describe the Resnet-LSTM model for image captioning in this work. Here, LSTMs are utilised for decoding while

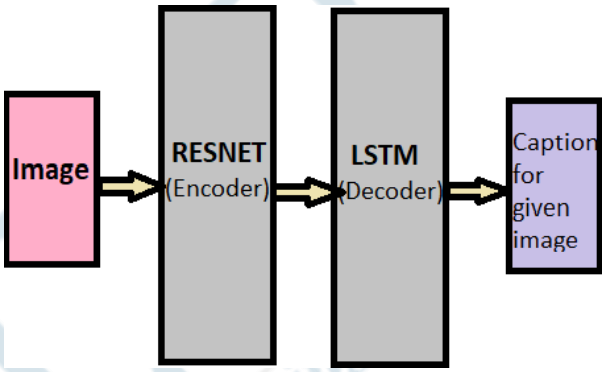

Fig. 1. Flow chart of the proposed method

Resnet Architecture is used for encoding. We will now train the model using these two parameters after sending the picture to Resnet (Residual Neural Network), which first extracts the image characteristics with the aid of vocabulary created using training captions data. We'll put the model to the test after training.The flow chart for the method we provide in this research is shown below.

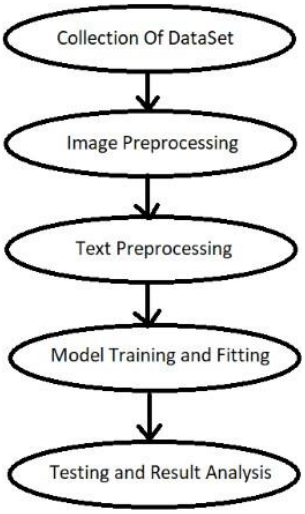

Fig. 2. Flow chart of the proposed method

A. Collection Of DataSet

There are several data sets that may be used to train a deep learning model to provide captions for photos, including ImageNet, COCO, FLICKR 8K, and FLICK 30K. For our

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 3, March 2023

model training Flickr8K data set is used. The Image Caption Generating Deep Learning Model can be trained well using the Flickr8k Dataset. There are 8019 photographs in the Flickr8K data set, around 80 percent of which are used to train a deep learning model while the remaining 20 percent are utilised to build and test the model. Five captions are included in the text data set for each image, each of which explains an action that is taking place in the image.

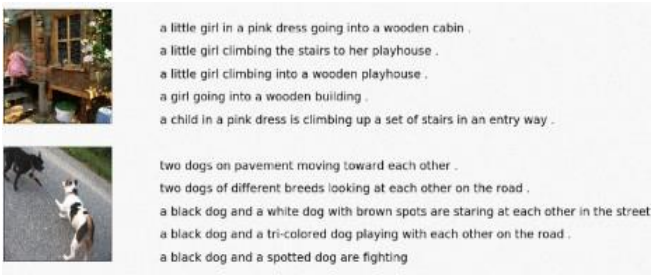

Fig. 3. Dataset

B. Image Prepocessing

In order to provide the photos as input to the ResNet, we must first preprocess them after importing the data sets. We must scale each image to the same size, which is 224X224X3, because we cannot feed various sized photos through the Con- volution layer like ResNet. Additionally, we are transforming the photos to RGB using the cv2 library's built-in capabilities.

C. Text Prepocessing

Once the captions have been added to the data set named Flickr, it is necessary to initialize them in such a way that there won't be any uncertainty or troubles while building the deep learning model's vocabulary from the captions. If there are any numbers in the captions, they must be eliminated. Next, we must remove any blank spaces and any captions that are missing from the supplied data set. In order to avoid ambiguity when developing the vocabulary and training the model, we must convert all uppercase characters in the captions to lowercase.To inform the neural network about the beginning and end of each caption during train and test phases of the model, At the start and conclusion of each caption, the phrases "startofseq" and "endofseq" are connected. This is because this model returns captions single word at a time and words that are generated earlier are utilized as inputs along with image attributes.

A neural network cannot process strings as input, we must convert captions that take the form of strings into numbers in order to feed them as input to the network. To do this, we must develop a vocabulary of numbers. This procedure is known as caption encoding.After initialization the captions included in the training set, it is mandatory to introduce a unique data structure where every word in each caption is considered. Then, the words must be numbered progressively in the dictionary order. This area is now known as the vocabulary library. We will number each caption using this vocab library to number the terms in each caption in accordance with the vocab library.Each word in a given caption is assigned a number by referencing to its value in a predefined vocabulary bank. Consider a Vocabulary Library that was created, for instance, by numbering each distinct word in the training captions.VocabDictionary=('startofseq': 1,'a': 2,.....,girl': 18,.....,'dog': 30,'and': 31,'spotted': 32, 'are': 33,....., 'playing': 36,.). Now consider the caption=a girl and a dog are spotted playing.Using the dictionary, we can now decode this caption into numbers, giving us caption=2 18 31 30 33 32 36.Now, a LSTM is supplied with this encoded caption in order to train a caption-generation model.

D. Model Training and Fitting

To train the model, we need to load the prepared photo and text data such that we may utilise it to fit the model. Then by creating the input and output sequences in batches and fitting the data to the model with model.fit(), we will train the data on all of the pictures in the Flickr8K dataset training images. The first step is to load the prepared photo and text data so that it can be used to fit the model. It is done by loading the images and their associated captions into a data structure such as a dictionary. Once the data is loaded, it is split into training and validation sets. The model training is done using a CNN(RESNET) and a Transformer(LSTM).The RESNET is used to pull out characteristics from the pictures, while the LSTM is used to produce captions based on those features. During training, the model is presented with a single image and the corresponding captions, and the model is trained to generate the captions for each image. This method is iterated till the model can generate a valid caption for every image in the training dataset.After training the model, it could be tested on the validation set and can be assured that it is able to generate valid captions for new images. Finally, the model can be deployed in production for use in real-world applications.

a) RESNET: With the advent of transfer learning, it became straightforward to use deep neural networks, such as RESNET (Residual Neural Network), which are trained in advance for numerous picture identification and categorization issues. Because ResNet was pretrained on the ImageNet data set to identify the pictures, we utilise it in place of Deep Convolutional Neural Networks. Therefore, we are minimising the cost of computation and training time by employing the transfer learning idea which would have risen if a CNN that hadn't been pre-trained is uesd. Our use of the ResNet pre- trained model further increases the accuracy of the model.50 deep convolutional neural network layers make up Resnet50. The CNN architecture that we are acquiring in our deep learn- ing model for captioning image is called ResNet50. Because there is no need of any classification output for this work, we removed the last layer of Restnet50 and obtained the image features as a single-layered vector

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 3, March 2023

output by accessing the out- put of the layer that came before it.RESNET combines residual blocks with skip connections that ultimately address CNN's vanishing gradient issue, and it also does so while reducing nput information loss in comparison to CNN. it is favoured over typical deep convolutional neural networks.ResNet out-performs classic CNN and VGG in terms of performance and accuracy while classifying images and extracting image attributes. Convolutional, ReLU (Rectified Linear Unit), and

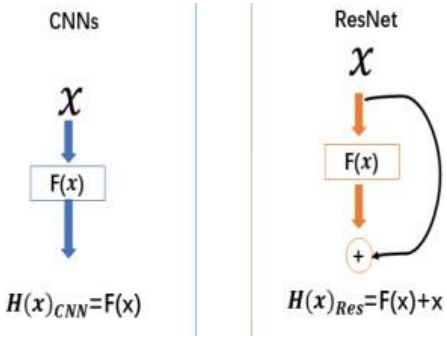

Fig. 4. Architecture of CNN and RESNET

Pooling Layers make up a conventional CNN. Following the input's passage across the conventional CNN, result acquired is as shown below: H(x) is the output value, x is considered as input, w is considered as multiplied weights, b is additional bias, and activation function is f(). H(x) = f(wx+b) or H(x) = f(x).In case of classic CNN, we can see that input and output are not equal.Therefore, if we try this to extract image characteristics or categorize pictures the accuracy acquired will be low and may contain errors in result.

The ResNet model's core element are the skip connections. In order to reach the output layer faster, the gradient uses these skip connections. When using this skip connection, the input and output becomes equal. i.e., H(x)=x+f(x), where f(x)=0, as shown in the above figure. As a conclusion, we could see that ResNet model's output from processing the images will be identical to the input, with no bias or weights applied. ResNet is therefore used to extract picture characteristics with low data loss or image properties. ResNet performs finer than the standard CNN model in extracting visual features.

Using a convolutional neural network Assigning weights to an image based on its various objects and processing it through a number of convolutional layers is how CNN operates. Convolutions are a type of mathematical operation that can extract features from an image. In a convolution, a small matrix is used to scan an image, and a mathematical operation is performed on the matrix to detect features in the image. Mathematical procedure of convolution can combine two types of information. To clarify the data and design a feature map from input data,convolution is often used.This filter's dimensions can be, for example, 3x3. This filter is sometimes referred to as a feature detector or kernel. To imple- ment convolution, element-by-element matrix multiplication is performed by kernel while iterating through input image. Every responsive area, or field where convolution occur, will be documented in the feature map with its results. We need to move the filter one more time before the feature map is complete.

The feature map is then passed through an activation func-tion, which is used to activate the neurons of the network. This helps the model to learn more features of the image. Activation functions such as ReLU, Sigmoid, and Tanh are used to ensure that the output has the desired shape.After the activation function, the feature map is passed through a pooling layer. Pooling is used to decrease the feature map dimensions and reduce the complexity of network. There are two types of pooling layers namely: max pooling and average pooling. The ultimate value of each responsive field is taken by max pooling, while average value of the responsive field is taken by average pooling.

Motive of layer which flattens the input is to compress the ResNet picture feature grid into a sole layered vector. In general, we still want to shrink the grid and transform it to single-layered attribute vector that contains picture features after the max pool operator is applied on the pattern. So in ResNet, flattening is carried out immediately following max pooling. The matrix is reduced to a single layer at this point, at which point it is considered to consist of picture features. This information is then transferred to the LSTM unit,It uses the vocabulary we have built to construct each phrase in the caption sequence.In favor of captions creation, the properties of pictures from the RESNET model are extracted and passed to the LSTM Networks in the form of a single layered vector.

b) LSTM: LSTM(Long Short Term Memory) neural net-work is generally used by artificial intelligence and deep learning. Long-term data dependencies may be learned by using a this special type of RNN.This is achieved using a memory cell, an input gate, an output gate, and a forget gate.Intermediate type of storage via the memory cell is introduced by LSTM pattern, which is composed of simpler nodes in a specific configuration.Since LSTM networks can accommodate delays of unknown duration among important events in a time series, they are ideally suitable for classifying, processing, and developing predictions are depended upon time series data. Additionally, the vanishing gradient problem that arises while training the conventional RNNs is avoided by using LSTMs.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 3, March 2023

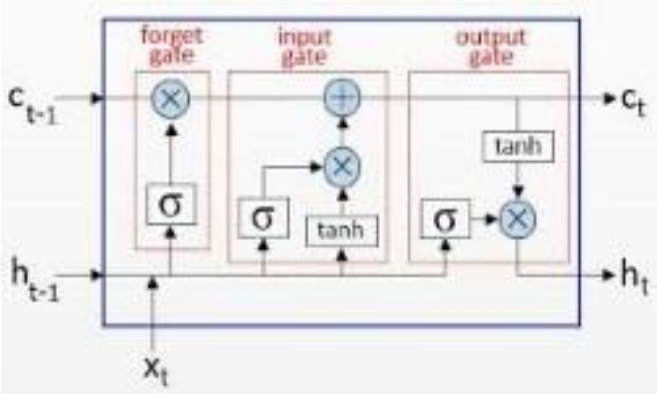

Fig. 5. Architecture of LSTM

The first step in the process is the forget gate.. Here, based on the previous concealed state as well as the fresh incoming data, we'll decide which pieces of the cell state are relevant. To do this, new input data and the prior hidden state are fed into a neural network. This network generates a vector with each component lying between [0, 1]. (ensured by using the sigmoid activation). This network has been trained to output values that are near to 0 when an input component is regarded irrelevant and closer to 1 when it is considered relevant. It is beneficial to think of each element of this vector as a type of filter that allows more information to get through as the value gets closer to 1.The input gate is responsible for deciding what new information should be added to the cell state. It takes an input at each time step, and updates the hidden state and cell state to obtain values and , using and the previous values and . The values of the input gate are calculated using the following equation: Where Xt is taken as input at the ongoing time step, Ui is the input weight matrix, Ht-1 is taken as hidden state at the preceeding time step, and Wi is the input weight matrix associated with the hidden state. The output of this equation is a number between 0 and 1, which is used to determine how much of the new information should be added to the cell state. The higher the value, the more of the new information is added. The output gate is responsible for deciding what information should be output from the cell state. It takes an input at each time step, and updates the hidden state and cell state to obtain values and , using and the previous values and . The values of the output gate are calculated using the following equation: Where Xt is input at the current time step, Uo is output weight matrix, Ht-1 is hidden condition at the earlier time step, and Wo is the output weight matrix associated with the hidden state. The result of the above calculation is a number between 0 and 1, which is used to find how much of the cell state should be produced. The higher the value, the more of the cell state is output In this way, during training, the captions are processed in the LSTM and the expressions created at each cell state are communicated to subsequent cell state before the LSTMs combine all words and create a caption for the supplied images.

E. Testing

Our model is tarined using 500 batches spread across 200 epochs. Early training epochs have shown to have less accuracy and generated captions aren't closely relevant to the images given for test. If a model is trained for at least 100 epochs, the captions generated are reasonably comparable to the test images supplied. As seen in the accompanying pictures, after the pattern is skilled for 100 epochs, an improvement in model accuracy and captions that are closely linked to the test photos can be observed.

IV. RESULTS

Automatic natural language generation evaluation metrics such as BLEU, CIDEr, ROUGE, and METEOR are used to estimate the quality of produced text, in comparision with a reference text.A precision-based metric called BLEU takes into consideration exact n-gram matching between produced and ground truth references. METEOR is an automatic metric for machine translation evaluation with improved correlation with human judgments. It is based on the harmonic mean of unigram precision and recall, with recall being calculated as a penalty for brevity. ROUGE is a recall-oriented metric which is mostly used for summarization evaluation. It is based on F-score which is harmonic mean with regards to precision includng recall. Finally, CIDEr is a metric used for evaluating image captioning, which uses the cosine similarity between generated and reference captions. These metrics all have their own strengths and weaknesses, but they provide a useful way of measuring the quality of natural language generation systems. The scores of our model for those evaluation metrics are given below.

TABLE 1. SCORES OF OUR MODEL

Sentences	*BLEU*	*CIDER*	*ROUGE*	*METRICS*
s1	0.579	0.600	0.396	0.195
s2	0.404	0.658	0.274	0.256
s3	0.279	0.599	0.400	0.172
s4	0.191	0.677	0.450	0.137

The results of our model using the BELU, CIDER, ROUGE, and METEOR meteors were compared to those of the prior models that were already in use. These are the results that are displayed as a graph.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 3, March 2023

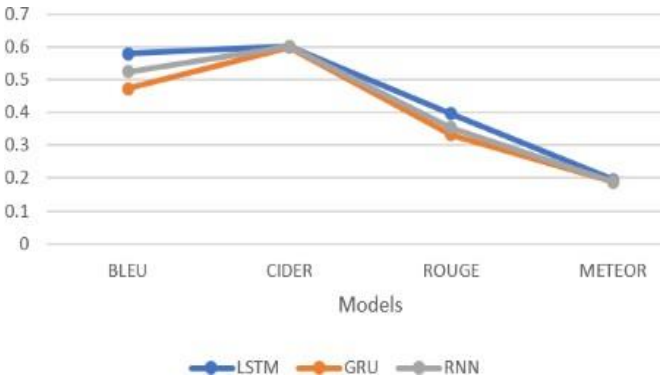

Fig. 6. Comparison graph of different models

Model is trained over 200 epochs with a batch size of 500.The trained model has got an accuracy of 0.8885 and had a loss of 0.3469.Then the model is fed with the testing images.The results we received for the test data are shown below.Thus this generated model can be used in many different fields, such as virtual assistants, image indexing, and recommendation systems. It can also be used to help the visually impaired, as it can generate captions that accurately describe the contents of an image. Additionally, it can be used in automated editing applications, providing accurate descriptions of the images without any manual input.

Fig. 7. Output 1

Fig. 8. Output 2

V. CONCLUSION

In this article, a deep learning model for image captioning is suggested. For each of the provided images, we have created a caption using the RESNET-LSTM model. The model has been trained using data from the Flickr 8k dataset. The convo- lution layer's architecture is called RESNET. This RESNET architecture is utilised to extract picture characteristics, and these image features are fed into LSM units, which construct captions with the use of language created during training. Comparing this ResNet-LSTM model to CNN-RNN and GRU Model, we can say that it is more accurate. In order to guide self-driving cars and create software to assist the blind, it is essential to analyse vast volumes of unstructured and unlabeled data. This is where this picture captioning deep learning model comes in extremely handy.

REFERENCES

[1] Shuang Liu, Liang Bai, Yanli Hu, Haoran Wang Image Captioning Based on Deep Neural Networks MATEC Web of Conferences. 232. 01052. 10.1051/matecconf/201823201052.

[2] G. Geetha1, T. Kirthigadevi1, G.Godwin Ponsam1, T. Karthik1 and M.

[3] Safa1 Image Captioning Using Deep Convolutional Neural Networks.

[4] Karen Simonyan Andrew Zisserman VERY DEEP CONVOLUTION- ALNETWORKS FOR LARGE-SCALE IMAGE RECOGNITION Vi- sual Geometry Group, Department of Engineering Science, University of Oxford karen,az@robots.ox.ac.u.

[5] Fang, H., et al. "From captions to visual concepts and back." Computer Vision and Pattern Recognition IEEE, 1473-1482. (2015).

[6] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 770-778. (2016).

[7] "Region-based Convolutional Networks for Accurate Object Detection and Segmentation," by Ross Girshick et al. 38.1:142–158 IEEE Trans- actions on Pattern Analysis Machine Intelligence (2015).

[8] Learning to Evaluate Image Captioning Yin Cui1,2 Guandao Yang Andreas Veit 1,2 Xun Huang 1,2 Serge Belongie 1,2 1 Department of Computer Science, Cornell University Cornell Tech.

[9] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, Yejin Choi CLIPScore: A Reference-free Evaluation Metric for Image Captioning.