International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Image Caption Generating Deep Learning Model

Aishwarya Maroju[#1], Sneha Sri Doma[#2], Lahari Chandarlapati[#3]
Department Of Electronics and Computer Engineering
J.N.T.U, Hyderabad ,
Sreenidhi Institute of Science And Technology,
Ghatkesar, Yamnampet, Hyderabad,India

***Abstract*—Image captioning is the process of generating descriptions about what is going on in the image. By the help of Image Captioning descriptions are built which explain about the images. Image Captioning is basically very much useful in many applications like analyzing large amounts of unlabeled images and finding hidden patterns for Machine Learning Applications for guiding Self driving cars and for building software that guides blind people. This Image Captioning can be done by using Deep Learning Models. With the advancement of deep learning and Natural Language Processing now it has become easy to generate captions for the given images. In this paper we will be using Neural Networks for the image captioning. Convolution Neural Network (ResNet) is used as encoder which access the image features and Recurrent Neural Network (Long Short Term Memory) is used as decoder which generates the captions for the images with the help of image features and vocabulary that is built.**

Keywords—Deep Learning; Image Captioning;Convolutional Neural Networks; Recurrent Neural Networks;ResNet;Long Short Term Memory; insert (key words)

1. INTRODUCTION

In earlier days Image Captioning was a tough task and the captions that are generated for the given image are not much relevant. With the advancement of Neural Networks of Deep Learning and also text processing techniques like Natural Language Processing, Many tasks that were challenging and difficult using Machine Learning became easy to implement with the help of Deep Learning and Neural Networks. These are very much useful in image recognition, Image classification, Image Captioning and many other Artificial Intelligence applications. Image Captioning is basically generating descriptions about what is happening in the given input image. Basically ,this model takes image as input and gives caption for it. With the advancement of the technology the efficiency of image caption generation is also increasing. This Image Captioning is very much useful for many applications like Self driving cars which are now talk of the town. Image captioning can be used in many Machine Learning tasks for Recommendation Systems. There are many models proposed for image captioning like object detection model, visual attention- based image captioning and Image Captioning using Deep Learning. In Deep Learning also there are different deep learning models like Inception model, VGG Model , ResNet-LSTM model, traditional CNN-RNN Model. In this paper we are going to explain about the model we have followed for captioning the images .i.e; ResNet-LSTM model

2.LITERATURE SURVEY

In method proposed by Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran et al. [1], two models of deep learning namely, Convolutional Neural Network-Recurrent Neural Network(CNN-RNN) Based Image Captioning, Convolutional Neural Network-Convolutional Neural (CNN-CNN)Based Image Captioning. In CNN-RNN Based frame work, Convolutional Neural Networks for encoding and Recurrent Neural Networks for the decoding process. Using CNN the images here are converted to vectors and these vectors are called image features these are passed into Recurrent neural networks as input. In RNN's se NLTK libraries are used to get the actual captions for the project. In the CNN-CNN based frame work only CNN is used for both encoding and decoding of the images. Here vocab dictionary is used and it is mapped with Image features to get the exact word for the given image using NLTK library. Thus generating the error free caption. Consisting of many models that are given at the same time of convolution techniques simultaneously is certainly quicker compared to the train the continuous flowing recurrently repetition of this techniques. CNN-CNN Model has less training time as compared to the CNN-RNN Model. The CNN-RNN Model has more training time as it is sequential but it has less loss compared to the CNN-CNN Model.

In the method proposed by Ansari Hani et al[2] Here they have used encoding decoding model for image captioning. Here they have mentioned two more models for image captioning they are: Retrieval based captioning and template based captioning. Retrieval based captioning is the process where training images are placed in one space and their corresponding captions which are generated are placed in another scope now in the new scope the correlations are calculated for the test image and captions the highest valued correlation caption is retrieved as caption for the given image from the given set of captions dictionary. Prototype based descripting is the technique is done by them in this paper .Here they have used Inception V3 model as their encoder and they have used attention mechanism and GRU as their decoder to generate the captions.

In the method proposed by Subrata Das, Lalit Jain et al[3] This model is mainly based on how the deep learning models are used for Military Image captioning. It mainly uses CNN-RNN based frame work.They have used Inception model for encoding the images and to decrease the gradient descent problem they have used Long Short Term Memory (LSTM'S) Networks.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

In the method proposed by G Geetha et al[4] they have used CNN-LSTM model for image captioning. The entire flow of the model was explained from data set collection to caption generation. Here Convolutional Neural Networks was used as encoder and LSTM's was used as decoder for generating the captions

3.DISADVANTAGES OF EXISTING MODEL

As we have seen in the literature survey there are many drawbacks of the existing model. Each existing model has its own disadvantage making the model less efficient and less accurate when the results are generated. The observed drawbacks in all the existing models are as follows:

1)In CNN-CNN based model where CNN is used for both encoding and decoding purpose we observe that CNN-CNN model has high loss which is not acceptable as the generated captions won't be accurate and the captions generated here will be irrelevant to the given test image.

2)While in the case of CNN-RNN based captions there might be less loss compared to the CNN-CNN based model but the training time is more. Training time effects the whole efficiency of the model and here we also encountered another problem. i.e.; Vanishing Gradient Problem. Gradient is the parameter which is used to calculate the rate of loss per the given input parameter comparing both inputs and outputs.
This Gradient Descent Problem occurs mainly in Artificial Neural Networks and Recurrent Neural Networks. Gradient is the ratio of change in the weights with respect to change in the error in the output of the neural network. This gradient is also considered as slope of the activation function of the neural network. If the slope is high then the training for the model is faster and the neural network model learns faster. As the hidden layers increases the loss increase whereas gradient decreases and finally gradient becomes zero. This gradient problem hinders the learning of long term sequences in Recurrent Neural Networks. This Gradient Descent problem hinders the RNN in learning and remembering process. The words cannot be stored in hidden memory for long term use .Hence it becomes hard for the RNN to analyze the captions for the given image during the training purpose. RNN cannot store the words of larger captions for longer period due gradient descent problem during the training time, As the number of hidden layers increases the gradient starts to decrease and finally it reaches to zero where the hidden key words in the captions are sent to forget gate of RNN. Hence CNN-RNN model be trained efficiently for generating captions for the images. Finally we can conclude that as RNN have gradient descent problem generation of captions for the images using CNN-RNN model is not efficient and accurate.

4.PROPOSED MODEL

As we have observed that using traditional CNN-RNN model there is vanishing gradient problem which hinders the Recurrent Neural Network to learn and get efficiently trained. So in order to reduce this gradient descent problem ,In this paper we are proposing this model so as to increase the efficiency of generating captions for the image and also to increase the accuracy of the captions. Given below is the architecture for our proposed model.[Figure 1]

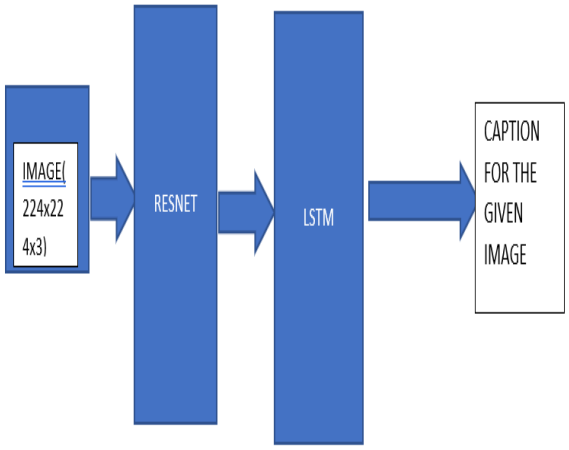

Fig 1:Architecture of ResNet-LSTM Model

In this paper, We are going to explain Resnet-LSTM model for the image captioning process. Here Resnet Architecture is used for encoding and LSTM's are used for decoding .Once when the image is sent to Resnet (Residual Neural Network) it extracts the image features then with the help of vocabulary that is built using training captions data ,We will now train the model with these two parameters as input .After training ,We will test the model. Given below is the flow diagram of our proposed model in this paper.[Figure 2]

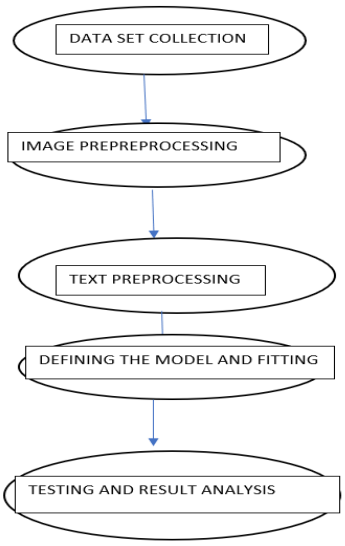

Fig 2:Model Implementation Flow Chart

4.1.DATA SET COLLECTION

There are many data sets which can be used for training the deep learning model for generating captions for the images like ImageNet, COCO,FLICKR 8K,FLICK 30K .In this paper, We are using FLICKR 8K data set for training the model. FLICKR 8K data set works efficiently for training the Image Caption Generating Deep Learning Model. The FLICKR 8K data set consists of 8000 images in which 6000 images can be used for training the deep learning model and 1000 images for development and 1000 images for testing the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

model. Flickr Text data set consists of five captions for each given image which describes about the actions performed in the given images

4.2.IMAGE PREPROCESSING

After loading the data sets we need to preprocess the images in order to give this images as input to the ResNet. As we cannot pass different sized images through the Convolution layer like ResNet we need to resize every image so that they are in same size i.e;224X224X3 .We are also converting the images to RGB by using inbuilt functions of cv2 library.

4.3.TEXT PREPROCESSING

After loading the captions for the images using FLICKR text data set we need to preprocess those captions so that there is no ambiguity or difficulty while generating vocabulary from the captions and also while training the deep learning model.We need to check whether the captions contain any numbers if found they must be removed and after that we need to remove white spaces and also missing captions in the given data set.We need to change all the upper case letters in the captions to the lower case in order to eliminate ambiguity during vocabulary building and training of the model. As this model will generate captions one word at a time and previously generated words are used as inputs along with the image features as input ,<start seq> and <end seq> are attached at the starting and end of each of the caption to signal the neural network about the starting of the caption and ending of the captions during the training and testing of the model.

4.4.VOCABULARY BUILDING

We cannot pass the string captions directly as input to the neural network because neural network cannot process string as input so the captions which are in the form of strings to numbers for that process we need to build a vocabulary of numbers. This process is called encoding of captions. Firstly, After preprocessing of the captions given in the training data set we need to create new space where all words in every caption are taken. Now we have to give numbers to the words sequentially in the dictionary order. Now this space is called vocabulary library. With the help of this vocab library we will number each captions by numbering their words accordingly with vocab library. For a given caption each word is numbered by referring their values in already defined vocab library. For example: Let us consider a Vocab Library that we have built by numbering every unique word of the given training captions .Vocab Dictionary={a:1,aa:11,aam:2,……..,cat:450,……
……..is:890……
……on:1120,……
……..table:3770,……
……..,the:5000,……..}
Now consider the caption={the cat is on the table}.Now this caption can be encoded into numbers using the dictionary and we can encode this caption as caption={5000 450 890 1120 5000 3770}. Now this encoded caption is passed into neural network(LSTM) for training the model to generate captions.

4.5.DEFINING AND FITTING THE MODEL

After collecting the data set and preprocessing the images and captions and building vocabulary. Now we have to define the model for generation of captions. Our proposed model is ResNet(Residual Neural Network)-LSTM(Long Short Term Memory) model. In this model Resnet is used as encoder which extract the image features from the images and converts them into single layered vector and pass them as input to LSTM's . Long Short Term Memory is used as decoder which takes image features as input and also vocabulary dictionary to generate each word of the caption sequentially.

4.5.1.RESNET 50

With the introduction of transfer learning (using knowledge gained in training network on one type of problem and applying the knowledge in another problem of same pattern) using deep neural networks like RESNET(Residual Neural Network) which is a pretrained model for many image recognition and classification became easy. We use this ResNet model in place of Deep Convolutional Neural Network because ResNet is a pretrained model on ImageNet data set to classify the images. So by using the concept of transfer learning we are reducing the computation cost and training time. If we have used CNN which is not pretrained then the computation cost would have increased and the model takes more time to learn. By using ResNet pretrained model we are also increasing the accuracy of the model.
Resnet50 consists of 50 deep convolutional neural network layers. ResNet50 is the architecture of Convolutional Neural Network that we are using in Image Caption Generation Deep Learning Model. The last layer of Restnet50 is removed as it gives classification output and we are accessing the output of the o layer before the last one in order to get the image features as output single layered vector because we don't need classification output in this paper. The ResNet is preferred compared to traditional deep convolutional neural networks because the ResNet contains residual blocks which have skip connections that ultimately reduce the vanishing gradient problem in CNN and ResNet also decreases the loss of input features compared to CNN. ResNet is having better performance and accuracy in classification of images and extracting image features compared to traditional CNN ,VGG. Below is the figure representing the working of ResNet block and its importance compared to traditional CNN.

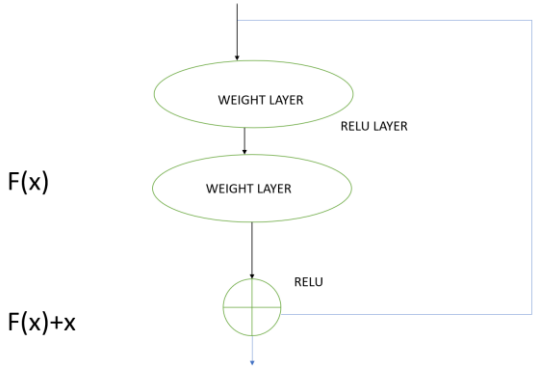

Fig 3:Residual Neural Network Block

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

The traditional CNN consists of Convolutional Layer, ReLU (Rectified Linear Unit)Layer and Pooling Layer. After passing the input through the traditional CNN the output is as follows:
H(x)=f(wx+b) or H(x)=f(x) where H(x) is the output value and x is the input and w is the weights that are multiplied and b is the bias that is added and f() is the activation function. We see that input is not equal to output in case of traditional CNN .So if we apply this to extract image features or classify the images there will be error in the result and the accuracy is low.
When it comes to ResNet model the skip connections are the core of this model. This skip connections are the short cut path that is followed by the gradient to reach the output layer. When this skip connections are applied the output is equal to input .i.e; H(x)=x+f(x) where f(x)=0 as represented by the above figure. So we can observe that when the images are passed through ResNet model the output will be equal to input without any bias or weights added .Thus when ResNet is used for image feature extraction there is no much loss of data or image features. Hence ResNet is better in extracting image features than traditional CNN model.

The Residual Neural Network has various layers in it like Convolution layer, ReLU(Rectified Liner Unit),Batch Normalization, Pooling layer, and Flatten .The description and working of various layers of Residual Neural Network is given below:
Convolution Layer, When the image is passed through the convolution layer ,then the image is converted to pixel values. Image filters(feature map) are applied on the image and convolution operation is performed. The output of this convolution layer is passed through the ReLU layer. When the convoluted image matrix is passed through the Rectified Liner Unit layer. It applies ReLU activation function and modifies the pixel values. ReLU Activation Function output={input when input>=0,0 when input<0(i.e.; negative values).
The output from the ReLU layer is sent to Batch Normalization layer where it performs normalization and standardization operation by adding extra layers to the network to scale the input to common size ultimately making the network faster and stable. Then the output of this layer is passed to pooling layer.
Pooling layer is generally used to decrease the size of the image. Max Pooling layer generally strides a window of small matrix across its input. It only selects the maximum value from each submatrix. Thus it decreases the size of the input matrix without much loss.
Flatten Layer, Generally the purpose of this flatten layer is it will convert the image featured matrix to single layered vector in this ResNet. Basically after applying the max pool function on the matrix we still want to still decrease the size of matrix and convert it to a single layered feature vector which is considered to have image features for that purpose flattening is done right after max pooling in ResNet. After the matrix is converted to single layer then this is said to consist of image features and is passed to Long short term memory unit for generating each term of the caption sequence using vocabulary that we have built.

Thus we are extracting the image features from the ReNet model in the form of single layered vector and these are passed to the LSTM Networks to generate captions .

4.5.2.LSTM

The output of ResNet(Image feature vector) and vocabulary built by using training data set captions are passed to Long Short Term Memory Networks to generate captions. When we pass image feature vector and vocabulary as input to first layer of LSTM, it generates the first word of the caption using training knowledge. The next words of a caption are generated with the help of image feature vector and previously generated words. Finally, all these words are concatenated to generate the caption for the given image.
Long short term memory cells are the advanced RNN's which can remember data from long periods. This Long Short Term Memory Networks can overcome the problem of vanishing gradient which exists in Recurrent Neural Networks. In traditional RNN's they cannot remember long sequence of data due to vanishing gradient problem. So in the case of caption generation RNN's cannot remember important words that are generated previously and which are required for generation of future words. For example in the case of predicting the last word of this sentence,"I am from France. I speak very fluently in French". It is important to remember the starting word France which is not possible in case of traditional RNN but Long Short Term Memory Networks do not have this issue. So LSTM's are preferred for caption generation compared to traditional RNN's .
Long Short Term Memory Networks have cells which consists of various gates like input gate, forget gate and output gate.[figure 4]

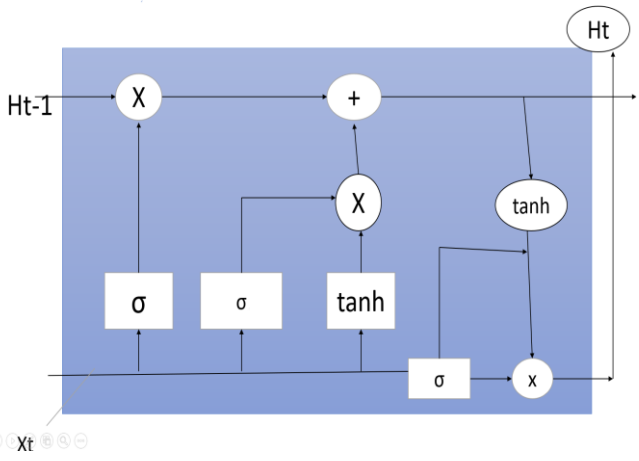

Fig 4:LSTM Cell Architecture

In the figure xt is the input to the cell and ht-1 is the output that is remebered from the previous layer and ht is the output of the present cell.The first step in LSTM is deciding what we have to forget this is decided by sigmoid function. It takes Ht-1 and Xt as inputs and give value 1(keep it as it is don't forget or 0 throw away all the matter).It is represented by the given equation where f(t) is the forget gate,f(t)=σ(Wf.[Ht-1,Xt]+bf).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

After deciding about what data we have to forget using the forget gate now we should decide what information have to be stored in Ht of the cell's state for long time series data processing. It is divided into two sub parts where sigmoid (σ) Neural network layer is the layer which decides what values need to be modified. And the second step is the tan h layer that creates vector of modifies new values that are added to cell state. The cell state carries the information from one cell to other cell. These steps are represented by the given formulas:

It=σ(Wi.[Ht-1,Xt]+bi),Ct=tanh(Wc.[Ht-1,Xt]+bc).
Then we update the cell by using the given formula :Ctt=f(t)*Ct-1+It*Ct
Finally our output is updated by the given equations: Ot=σ(Wo[Ht-1,Xt]+bo) and Ht=Ot*tanh(Ctt).In this way during the process of training the captions are processed like this in the Long Short Term Memory and the words generated at each cell state are passed into the next cell states finally LSTM's concatenate all the words and generate the caption for the given images.

5.RESULT ANALYSIS

After defining and fitting the model. We trained our model for 50 epochs. It is observed that during the initial epochs of training the accuracy is very low and the captions generated are not much related to given test images. If we train the model for atleast 20 epochs then we have observed that the captions generated are some what related to the given test images. If the model is trained for 50 epochs we observe that the accuracy of the model increases and the captions generated are much related to the given test images as follows in the following figures.[figure 5] [figure 6]

Fig 5: Caption generated for given test image

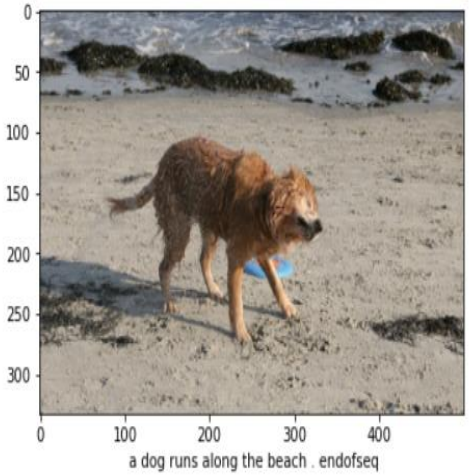

Fig 6: Caption generated for given test image

6.CONCLUSION

Image captioning deep learning model is proposed in this paper. We have used RESNET-LSTM model to generate captions for each of the given image. The Flickr 8k data set has been used for the purpose of training the model. RESNET is the architecture of convolution layer. This RESNET architecture is used for extracting the image features and this image features are given as input to Long Short Term Memory units and captions are generated with the help of vocabulary generated during the training process. We can conclude that this ResNet-LSTM model has higher accuracy compared to **CNN-RNN and VGG Model. This model works efficiently** when we run the model with the help of Graphic Processing Unit. This Image Captioning deep learning model is very much useful for analyzing the large amounts of unstructured and unlabeled data to find the patterns in those images for guiding the Self driving cars, for building the software to guide blind people.

FUTURE SCOPE

In our paper we have explained about generating captions for the images. Even though deep learning is advanced upto now exact caption generation is not possible due to many reasons like hard ware requirements problem, no proper programming logic or model to generate the exact captions because machines cannot think or make decisions as accurately as human do. So in future with the advancement of hardware and deep learning models we hope to generate captions with higher accuracy.It is also thought to extend this model and build complete Image-Speech conversion by converting captions of images to speech.This is very much helpful for blind people.

REFERENCES

[1] Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran. (2018). Image Captioning Based on Deep Neural Networks. MATEC Web of Conferences. 232. 01052. 10.1051/matecconf/201823201052.
[2] A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture," 2019 International Arab Conference on Information Technology (ACIT), 2019, pp. 246-251, doi: 10.1109/ACIT47987.2019.8990998.
[3] S. Das, L. Jain and A. Das, "Deep Learning for Military Image Captioning," 2018 21st International Conference on Information

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Fusion (FUSION), 2018, pp. 2165-2171, doi: 10.23919/ICIF.2018.8455321.

[4] GGeetha,T.Kirthigadevi,G GODWIN Ponsam,T.Karthik,M.Safa," Image Captioning Using Deep Convolutional Neural Networks(CNNs)" Published under licence by IOP Publishing Ltd in Journal of Physics :Conference Series ,Volume 1712, International Conference On Computational Physics in Emerging Technologies(ICCPET) 2020 August 2020,Manglore India in 2015.

[5] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[6] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recogni-tion and description." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[7] Lu, Jiasen, et al. "Knowing when to look: Adaptive attention via a visual sentinel for im-age captioning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 6. 2017.

[8] Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. "Im2text: Describing images us-ing 1 million captioned photographs." Advances in neural information processing systems. 2011.

[9] Chen, Xinlei, and C. Lawrence Zitnick. "Mind's eye: A recurrent visual representation for image caption generation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[10] Feng, Yansong, and Mirella Lapata. "How many words is a picture worth? automatic caption generation for news images." Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

[11] Rashtchian, Cyrus, et al. "Collecting image annotations using Amazon's Mechanical Turk." Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics, 2010