CAPTION GENERATOR MODEL

Koushik Katakam
School of Computing and Engineering
University of Missouri Kansas City
Kansas city, Missouri
kkkv2@mail.umkc.edu

Saitejaswi Koppuravuri
School of Computing and Engineering
University of Missouri Kansas City
Kansas city, Missouri
sk6zb@mail.umkc.edu

Venkata Lakshmi Korrapati
School of Computing and Engineering
University of Missouri Kansas City
Kansas city, Missouri
vkvn3@mail.umkc.edu

Pavan Kumar Manchala
School of Computing and Engineering
University of Missouri Kansas City
Kansas city, Missouri
pm3zk@mail.umkc.edu

Zakari Abdulmuhaymin Ahmad H
School of Computing and Engineering
University of Missouri Kansas City
Kansas city, Missouri

aazzb4@mail.umkc.edu

***Abstract*—Characterizing a description for a picture consequently has been blasting in the field of artificial intelligence which incorporates computer vision and Natural language handling. Utilizing these two systems a deep learning model must be created so as to accomplish the ideal description for a picture. Our motto significantly centers around the model which gives more precision by training and testing different models**

Keywords—caption generation, show and tell model, NLP techniques, Deep neural networks.

I. Introduction

Caption generation is the challenging artificial intelligence problem using NLP technique and computer vision. It requires the two pictures comprehension and language show from the field of Natural language processing. For sure, a depiction must catch the articles contained in a picture, yet it additionally should express how these items identify with one another, just as their characteristics and the exercises they are associated with.

II. Related work

The issue of generating a caption from the image has almost given a tougher task in the arena of computer vision especially for streaming video. All these frameworks required for generating captions are a bit complex, brittle and fragile. A paper by Ali Farhadi [1] describes about the paper how the caption generation process has become a boom in the area and how the thought process has started. Later, after many investigations and a thorough research a new deep learning model came into existence using a Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) and called it as a Show and Tell model (STM) by Google developers [2]. Due to drawbacks with the RNN model and after the rigorous research they have come up with the Show, Attend and Tell Model by Kelvin Xu [3].

III. Project goals and objective

A. *Motivation*

Have you ever thought of generating a caption for an image? Yes, these days caption generation for an image has become an important task in the area of research of machine learning and Artificial Intelligence. No only captioning is a primary goal but predicting the objects in an image and express their relation in a process of natural language processing. The process of image captioning has been made little advanced on the advancement of neural networks.

B. *Significance/Uniqueness*

- As this arena is emerging these days, there are quite a known number of applications which provide the image captioning for an image. Similar applications include Microsoft Seeing AI, Envision AI and couple more. Recently google has come up with an idea called Google Lookout especially for the disabled. But these applications have lacked a little accuracy and facing problems especially in particular lighting conditions. Especially, these applications in particular are developed for IOS.

C. *Objectives*

Our main motto is to make a visual world into an audible one. The main objective is to create a model which generates captions by understanding the image. Deep learning models are used in order to verify the best. Our domain of interest is on “FOOD”. Various images of food are collected from the SBU Data set and train the model with these images. Finally, generating a perfect caption for an image.

D. *System Features*

- The main feature is to generate a perfect caption for the image.

- Providing user, the captions with highest point of accuracy which can be done by training the implementing it with different models and choosing the correct optimizers when needed.
- Food is the major area of focus on which we are trying to collect images and train the model using Show, Attend and Tell methodology.

IV. PROPOSED WORK

A. *Proposed Models*

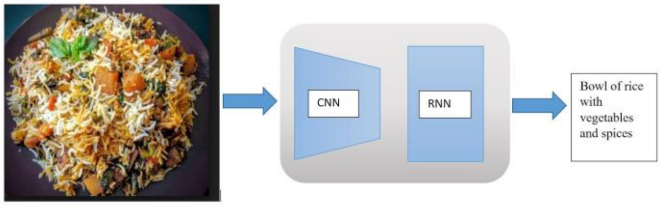


Fig: Workflow model for Image captioning

The workflow of the project is done using two different models namely CNN (Convolutional Neural Networks) and RNN (Recurrent Neural Networks) models.

At present, we are with a simple workflow for the project i.e., when we give an image input to the model, it should process the image and predict it, resulting a text related to the image. Generally, CNN was proposed to diagram data to a output variable. They were wound up being successful to the point, that they fit perfectly for a prediction problem including picture data as an info. RNN were intended to work with content examination. Customarily, RNN were hard to train so as to stay away from this a combination of RNN demonstrate with LSTM (Long short-term memory) is utilized.

1)Convolutional Neural Networks: Convolutional Neural Networks are intended to delineate information to an output variable which were turned out to be so viable.

2) Recurrent Neural Networks: Recurrent Neural Networks are intended to work with the arrangement stream of text information. For the most part, RNN's are generally difficult to train so inorder to conquer this trouble, Long Short-term memory asymptotically LSTM are utilized.

B. *Network Model*

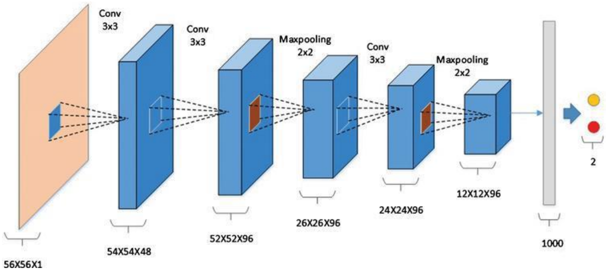


Fig: Network Diagram of the proposed project

Indeed, deep learning CNN models to train and test, each info picture will experience a movement of convolution layers with channels, pooling layers and fully connected layer pursued by applying a softmax function to group an article with probabilistic qualities 0&1.

Convolution layer is the main layer to separate features from an info picture. It safeguards the connection between pixels.

The output of the convolution layer is the multiplied estimations of info picture with channels which is given to strides and afterwards to padding stage further to ReLU activation unit.

a) Pooling Layers

So, when the images are too large, these pooling layers would reduce the number of parameters. There are three types of pooling layers namely

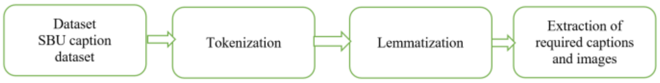
- Max pooling – largest element from the feature map
- Sum pooling – sum of all elements
- Average pooling – same as max pooling

b) Fully Connected Layer

The fully connected layer does the smoothing of our matrix into vector. It gives the output of fully connected layer to softmax activation function.

C. *Preprocessing of Image captions Dataset using NLP technique*

Preprocessing of Image captions include the concept of filtering the necessary data that is required from the entire dataset. The workflow for the preprocessing the data is as shown in figure.


Dataset SBU caption dataset
Tokenization
Lemmatization
Extraction of required captions and images

D. Feature generation using SIFT algorithm

There are many algorithms at present, but we have used Scale Invariant Feature Extraction (SIFT) feature extraction technique [3]. It has resulted us very good accuracy because the key point detectors given exactly matches the original input image.

E. Image classification using CNN model

As mentioned, CNN model is designed to map input image data to an output variable, one more advantage is that CNN has its ability to develop internal representation of a 2D image. These results are visualized using Tensorboard.

F. Image caption generation using Show and Tell Model

Show and Tell model is designed in order to generate captions for a image based on Beam search value with various probabilities. This is done by pretraining the model with certain vocabulary file. The accuracy can be tested by using various scores like BLEU, CEDAR, METER, ROUGUE etc.

G. Data Analytics using Unsupervised Learning approaches

To find the Data Analytics on the dataset we have implemented the various clustering techniques like EM clustering and KMeans clustering techniques. These techniques divide the similar data and cluster them using keywords.

V. Implementation and Evaluation

A. System Design and Implementation

a) Implementation details

The implementation details are in the Github. The link is as follows.

https://github.com/SaitejaswiK/CS5542_BigdataAnalyticsProject

b) Applications

There are many applications that are based on caption generator model for images.

- Mainly useful for visually impaired people
- Social media like Facebook can directly infer from image.
- Long term application can be describing what is happening in a video frame.

b) Evaluation and results

i. Datasets

We have used SBU dataset and Flickr 8k dataset for implementation till date.

The links for the datasets is as follows:

SBU dataset:

http://www.cs.virginia.edu/~vicente/sbucaptions/

Flickr_8kdataset:

http://nlp.cs.illinois.edu/HockenmaierGroup/Framing_Image_Description/KCCA.html

ii. System specifications

- Ram - 8GB or more
- HDD/SSD - 256 GB
- Processor - i5 or more(windows)
- GPU - 2GB or more

c)Special task:

We have implemented bottom-up attention-based caption generator model. With this our architecture has changed a little as we have used an Inception_v3 model as the image recognition model. Later necessary changes have been made in the existing model by using the features file generated by the inception model. Performing this task using the SBU dataset has become a little difficult task for us as we are stuck up with generating the vocabulary file for the dataset. So, this task we have performed on MSCOCO dataset.

Later, we trained the model using this vocabulary file and then tested the model with the SBU dataset. Later, we have generated two scores for the model namely BLUE and GLUE scores.

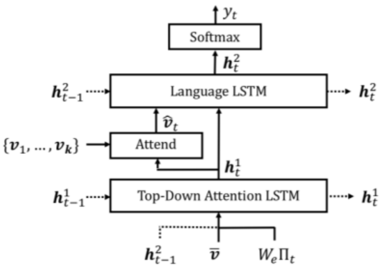


Fig: Overview of the proposed captioning model.

One major drawback for the attention model which we have developed works only with Tensorflowr1.0 version.

VI. Results

Koushik Katakam results:

The results of the processed work are expressed in the form of a table.

a) Dataset Statistics

Out of the entire dataset preprocessing of data has been done using the following keywords.

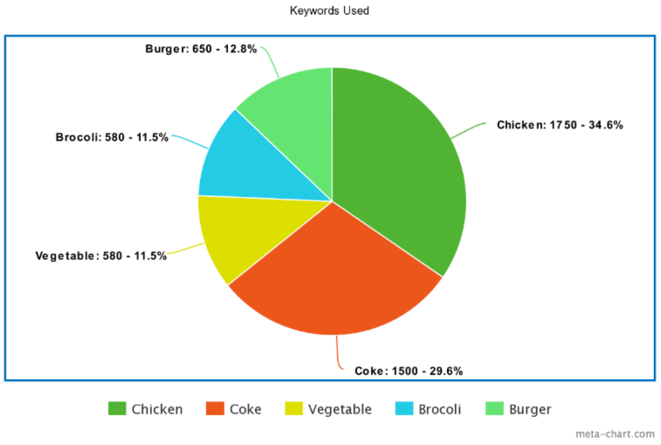


Fig: Dataset Statistics

b) Hyperparameters

- Number of Epochs - 400
- Batch Size - 60
- Activation Function - Relu
- Number of hidden layers and units - 3
- Weight Initialization - 32, 64

c) Accuracy results

Model	**Accuracy**
Dataset Used	Flickr_8k
Classification accuracy based on CNN pretrained model	92.6%
SIFT feature extraction technique	70.86%
Image classification using pretrained model	98.4%
BLEU score for Show and tell model(pretrained)	0.2222
Inception Model Classification results	98.4%
BLEU score for Show and tell model	0.16
Bottom-up attention model	CIDEr – 117.9

Pavan Kumar Manchala results:

The results of the processed work are expressed in the form of a table.

a) Dataset Statistics

Out of the entire dataset preprocessing of data has been done using the following keywords.

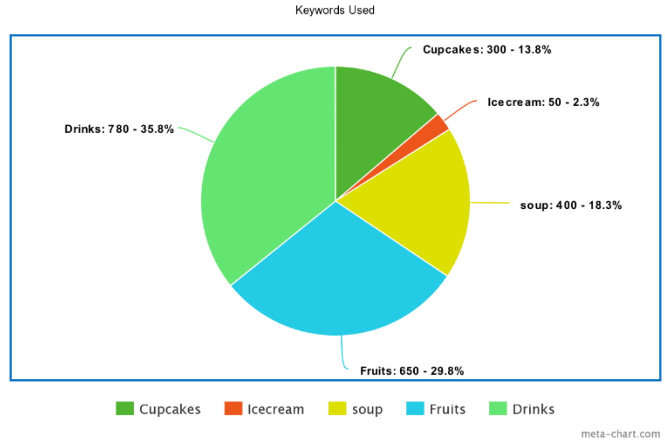


Fig: Dataset Statistics

b) Hyperparameters

- Number of Epochs - 350
- Batch Size - 60
- Activation Function - Relu
- Number of hidden layers and units - 2
- Weight Initialization - 32, 64

c) Accuracy results

Model	Accuracy
Dataset used	SBU
Classification accuracy based on CNN pretrained model	91.7%
SIFT feature extraction technique	69.7%
Image classification using pretrained model	96.2%
BLEU score for Show and tell model(pretrained)	0.48
Inception model classification results	97.6%
BLEU score for Show and tell model	0.15
Bottom-up attention model	CIDEr – 117.9

Venkata Lakshmi Korrapati results:

The results of the processed work are expressed in the form of a table.

a) Dataset Statistics

Out of the entire dataset preprocessing of data has been done using the following keywords.

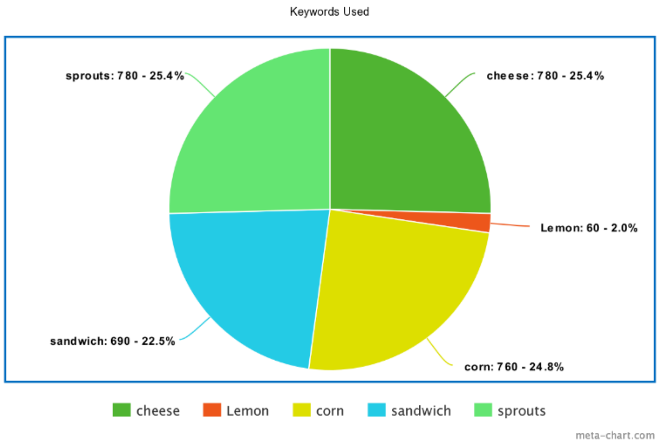


Fig: Dataset Statistics

b) Hyperparameters

- Number of Epochs - 350
- Batch Size - 50
- Activation Function - Relu
- Number of hidden layers and units - 3
- Weight Initialization - 32, 64

c) Accuracy results

Model	Accuracy
Dataset used	SBU
Classification accuracy based on CNN pretrained model	90.17%
SIFT feature extraction technique	72.3%
Image classification using pretrained model	95.4%
BLEU score for Show and tell model(pretrained)	0.24
Inception model classification results	96.2%
BLEU score for Show and tell model	0.18
Bottom-up attention model	CIDEr – 117.9

Saitejaswi Koppuravuri results:

The results of the processed work are expressed in the form of a table.

a) Dataset Statistics

Out of the entire dataset preprocessing of data has been done using the following keywords.

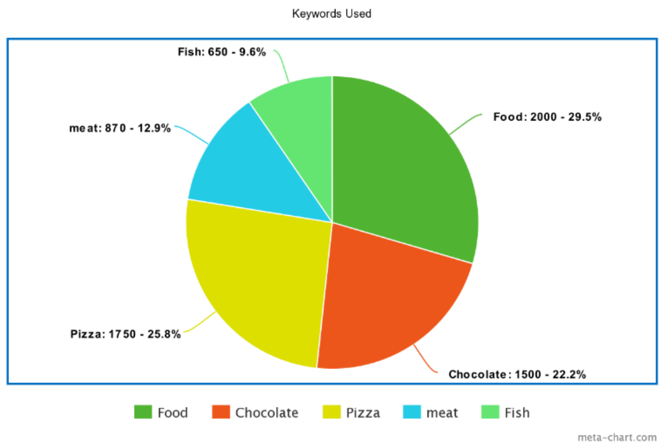


Fig: Dataset Statistics

b) Hyperparameters

- Number of Epochs - 400
- Batch Size - 50
- Activation Function - Relu
- Number of hidden layers and units - 2
- Weight Initialization - 32, 64

c) Accuracy results

Model	Accuracy
Dataset used	MSCOCO
Classification accuracy based on CNN pretrained model	89.6%
SIFT feature extraction technique	65.6%
Image classification using pretrained model	96.4%
Inception model classification results	95.4%
BLEU score for Show and tell model(pretrained)	0.182
BLEU score for Show and tell model	0.23
Bottom-up attention model	CIDEr – 117.9

Zakari,Abdulmuhaymin Ahmad H results:

The results of the processed work are expressed in the form of a table.

a) Dataset Statistics

Out of the entire dataset preprocessing of data has been done using the following keywords.

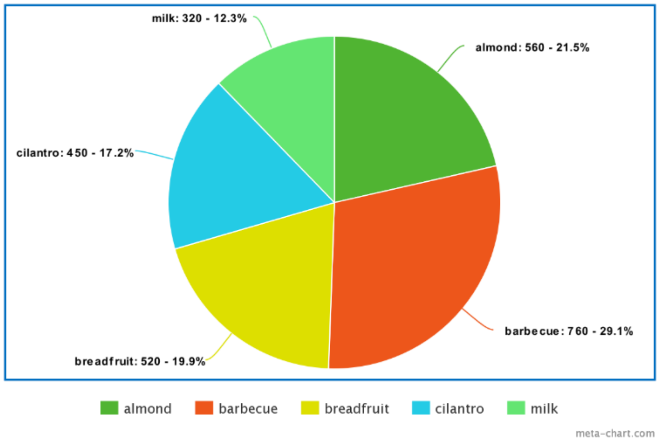


Fig: Dataset Statistics

b) Hyperparameters

- Number of Epochs - 300
- Batch Size - 40
- Activation Function - Relu
- Number of hidden layers and units - 3
- Weight Initialization - 64

c) Accuracy results

Model	Accuracy
Dataset used	SBU
Classification accuracy based on CNN pretrained model	88.2%
SIFT feature extraction technique	70.8%
Image classification using pretrained model	96.2%
BLEU score for Show and tell model(pretrained)	0.2222
Inception Model Classification results	98.4%
BLEU score for Show and tell model	0.12
Bottom-up attention model	CIDEr-117.9

CONCLUSION

- To create a caption generator model
- To understand various food images, train and test accordingly.
- Calculate the accuracy between models.

REFERENCES

[1] Every Picture Tells a Story: Generating Sentences from Images by Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi,Peter Young

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

[3] Show, Attend and Tell Model by Kelvin Xu, Jimmy Lei Ba, Ryan Kiros et al.

Paper links and websites referred:

https://www.jair.org/index.php/jair/article/view/10985

https://www.researchgate.net/publication/229051043_Feature_Extraction_Technique_Using_SIFT_Keypoint_Descriptors

https://link.springer.com/chapter/10.1007%2F978-3-642-15561-1_2

https://experts.illinois.edu/en/publications/every-picture-tells-a-story-generating-sentences-from-images

https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/

http://proceedings.mlr.press/v37/xuc15.pdf

http://homepages.inf.ed.ac.uk/keller/papers/jair16.pdf

https://arxiv.org/pdf/1810.04020

https://arxiv.org/pdf/1805.09137

www.aclweb.org/anthology/P18-1238

https://arxiv.org/abs/1707.07998

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