CSCE 5290: Natural Language Processing

PROJECT INCREMENT 2 : Visual Captioning and Neural Machine Translation

Github :-https://github.com/PriyaEnuganti/Visual-Captioning-and-Neural-Machine-Translation
Video :- https://drive.google.com/file/d/12Y1aPzTw-NJ6IOZQgJ7OrWcEk-zzGQJt/view?usp=sharing

• Introduction >> Pankaj

The objective of our project is to create a model which can interpret the image or summarize the objects in the image and then translate it to the language of one's choice.

This could be a novel project for people who are visually challenged.

For the first increment we selected two approaches to do the visual captioning or summarisation for the objects one was the Recurrent Neural Network variant LSTM ie Long short term memory trained on COCO dataset and another was transformer learning model based on FLICKR dataset. Upon fine tuning the model and observing the results after Testing we decided to continue with the LSTM model.

For increment two, we focus on doing the Neural Machine Translation for the summarization obtained from LSTM. As it is aimed at converting one language to another language i.e. converting one sequence to another sequence will take a transformer model based on encoder decoder. The training of such model will be based on corpus from a specific language for which we are looking for output.

• Background >> All

o Related work for your topic with linked references

The paper works on low resourced Indian language translation.This poses many challenges such as structural divergence, parallel corpora, variation in word order due to syntactical divergence. Many models still use rule based machine translation.It uses Indian language dataset and does preprocessing before creating custom Tokens. Then a transformer model to train on the dataset and then uses BLEU score for output.

https://kuldeepsangwan.medium.com/machine-translation-of-indian-languages-using-transformer-bbd8c93053f7

- **Model**

o Architecture Diagram with explanation >> Aniv

The overall architecture of the system involves two modules. The input image is preprocessed and trained and the output caption is fed as input to the translator to translate the captions to the target language using transformers.

o Workflow diagram with explanation >> Sridhar

We divide the workflow of this project into two phases, the training phase and the testing phase.

TRAINING PHASE :-

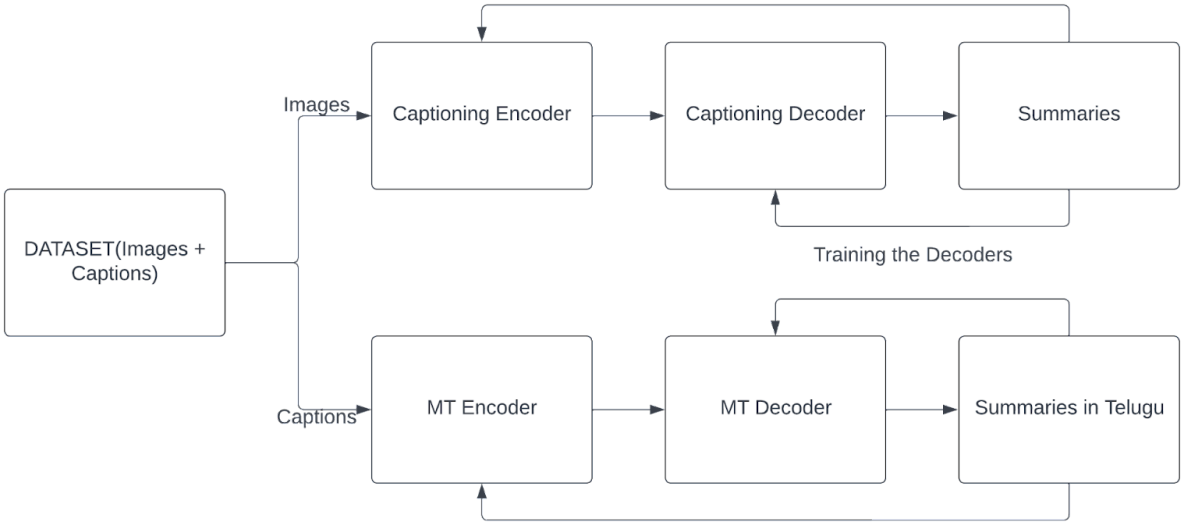


Fig. Workflow diagram - Training

As established earlier, we will be using the COCO dataset for training the models.

STEP 1 :- We start with splitting up the training network into two parts. One of the network deals with the part of image captioning, which was done in increment 1 and the other deals with translating the generated summaries into another language like Telugu.

STEP 2 :- COCO dataset comprises the images and their respective captions. We pass these into the encoder-decoder architecture of the captioning network.

STEP 3 :- The encoder-decoder architecture or the LSTM architecture forms the backbone of this network. The network trains to encode the images into suitable captions and based on its predictions on a validation set, tries to better its performance by tuning its weight matrix.

STEP 4 :- The second part of the training network deals with machine translation of the summaries generated by the captioning network. For this we have used a dataset of COCO's captions and their counterparts in Telugu language. We pass the captions and their respective translations to the machine translation network.

STEP 5 :- The encoder works on encoding these captions effectively so that the decoder makes perfect predictions on the encoded tokens.

TESTING PHASE :-

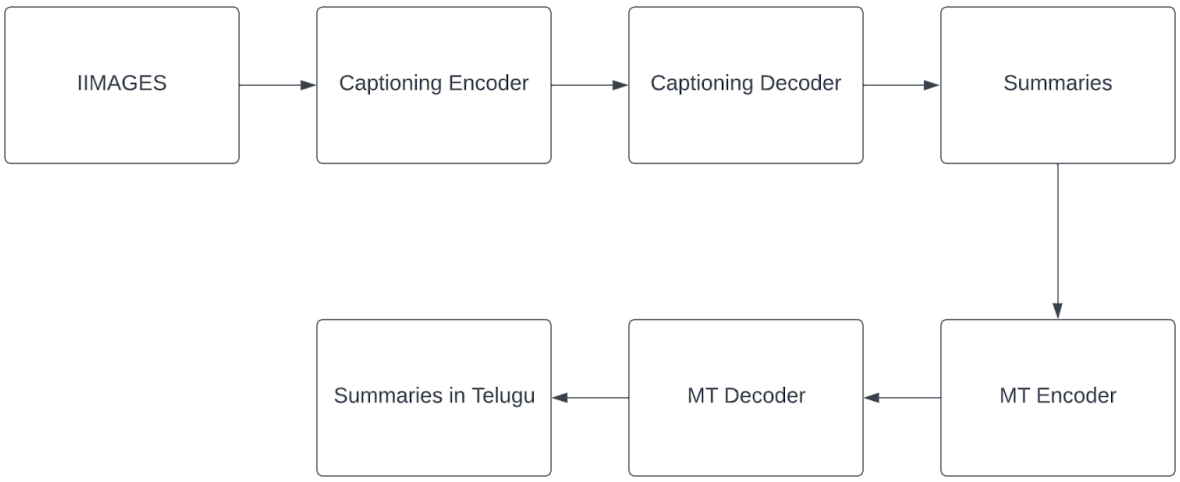


Fig. Workflow Diagram - Testing

STEP 1 :- We pass raw images as an input to the above network.

STEP 2 :- The encoder-decoder network responsible for captioning the images, generates the summaries accordingly.

STEP 3 :- These summaries are then processed using Natural Language Processing techniques.

STEP 4 :- The second encoder-decoder network which is responsible for translating these summaries, gives out a predicted summary. Encoder gives out an encrypted summary while decoder decrypts the summary and then displays it in a readable format.

• Dataset

o Detailed description of Dataset >> Pankaj

The corpus was made available from an Indian dataset which below specs

Language	**DEV**	**Valid**	**Test**
Tamil	**183451**	**2000**	**1000**
Malayalam	**548000**	**3660**	**3000**
Telugu	**75000**	**3897**	**3000**
Bengali	**658000**	**3255**	**3500**

o Detail design of Features with diagram >> Aniv

We make use of image features of size 229x229 with 3 channel RGB images with text annotations. We obtain text output in english and then translate to another language like hindi or telugu.

• Analysis of data

o Data Pre-processing >> Sridhar

Before passing the images into the network, we apply certain pre-processing techniques on the images to make the dataset more rich.

1. Color transforming:

We have applied color transformation on the pre-existing images in order to help the model generalize well to the changes in background color intensities in the image.

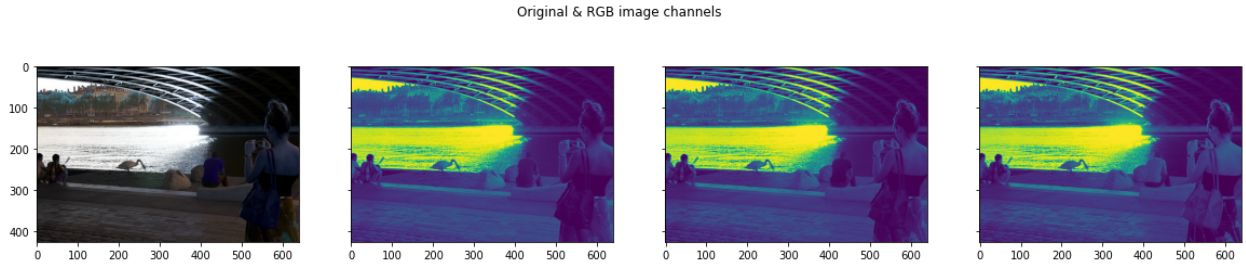


Fig. Color Transformation

2. Converting the image into grayscale:

Similar to color transformation, the captions should also be independent of the presence of RGB channels in the images to a certain extent. In order to do this, the images have been reduced down to their gray scale counterparts.

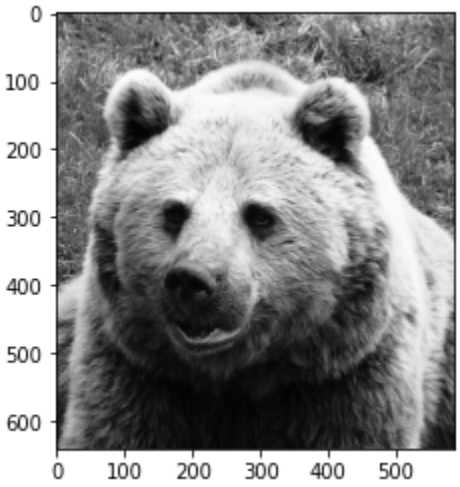


Fig. Grayscale

3. Normalization :

We restrict the range of the pixel values between 0 and 1 to avoid heavy computations.

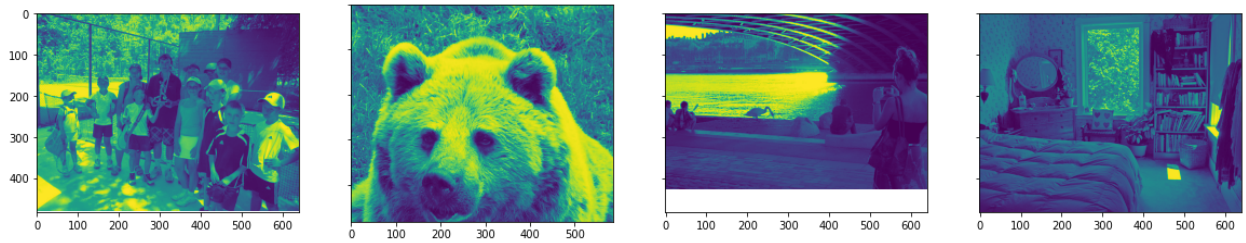


Fig. Normalized Images

4. Changing Orientation :

We have changed the orientation of certain images with the intention of data augmentation.

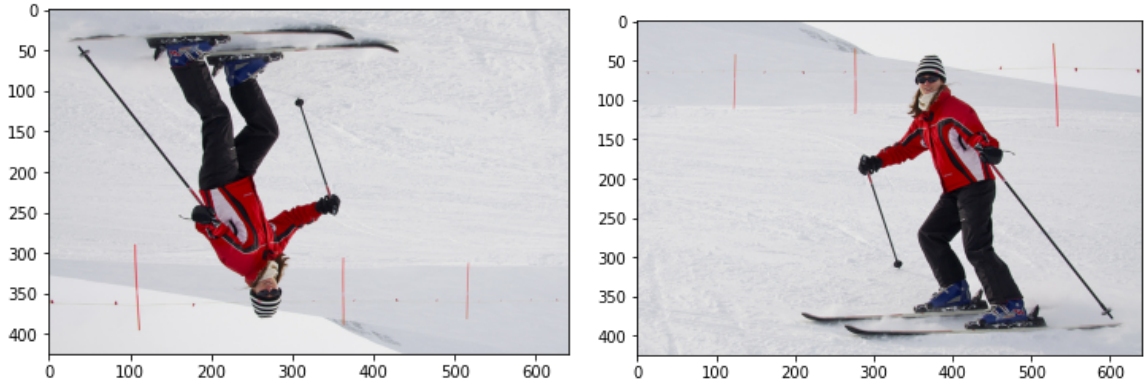


Fig. Orientation

5. Normalizing pixel brightness :

We limit the brightness of the pixels in the range 0.5 to 2.0 in order to make all the mages equally bright.

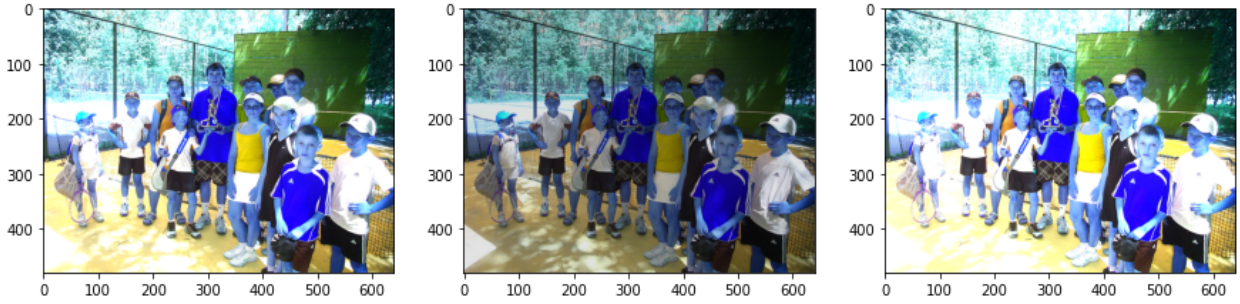


Fig. Pixel brightness

o Graph model with explanation >> Pankaj

• Implementation

o Algorithms / Pseudocode >> Aniv

- Download COCO dataset with annotations
- Train the model on inception V3 to get output english captions
- Preprocess english to telugu dataset
- train transformer to translate on dataset
- use caption as input to trained model

o Explanation of implementation >> Pankaj

- The words were tokenized with word embedding.With embedding semantic meaning was observed
- Sub word tokenization with Byte Per encoding was done to address words which will be not part of training corpus
- The dataset for English and Indian Language is cleansed by preprocess and removing apostrophe, special characters, brackets etc
- Using the dataset Vocab is created of 15000 words
- Using tensorflow tokens are generated for the vocab
- The tokens are fed to the transformer
- Positional embedding is blended with word vector to create more context
- With multi head attention layer, residual layer after embedding generates vector for encoder layer
- The decoding layer takes the feed of correct output layer and input from encoding
- The model is trained to understand the correct sequence

• Results

o Diagrams for results with detailed explanation >> Sridhar

For evaluation of our models we finalized the evaluation techniques of human evaluation of the summaries and BLEU score for the machine translation output.

Following is the training loss curve for the encoder-decoder architecture of the visual captioning model.

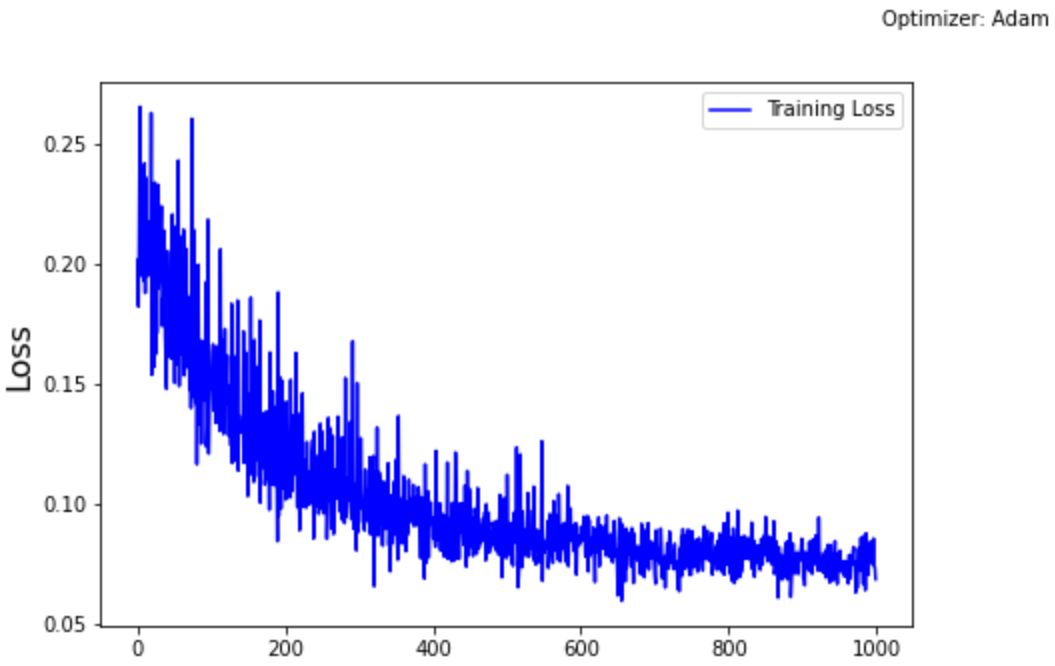


Fig. Loss Curve for image captioning model

From the above loss curve, we can observe that the loss goes on decreasing significantly with the number of epochs.

To evaluate the generated summaries, we use a well defined method called Cosine similarity which measure the extent of similarity between the generated captions and the ground truth values of those captions.

Following is the result obtained for one of the summaries,

Fig. Similarity score

Following are some of the summaries generated by the captioning model,

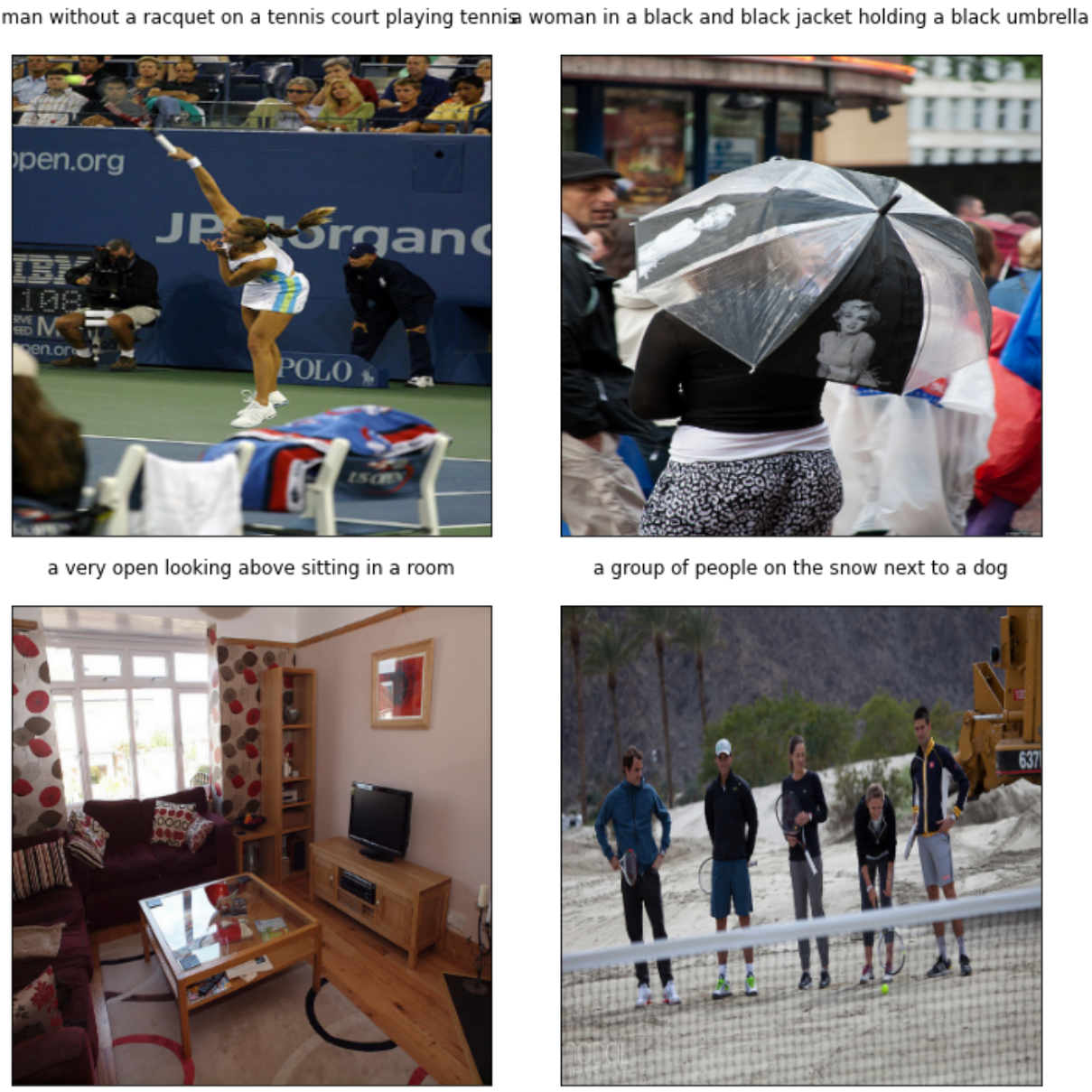


Fig. Out of captioning model

Next, we turn to the machine translation outputs. Following are some of the outputs obtained from the captioning model that were passed as an input into the network.

INPUT:- 'a man without a racquet on a tennis court playing tennis', 'a woman in a black and black jacket holding a black umbrella'

OUTPUT :- टेनिस खेलने के लिए टेनिस कोर्ट में बिना राकेट के एक आदमी', 'एक नारंगी और नारंगी जैकेट में एक नारंगी छत पकड़े हुए एक महिला'

Fig. Translated captions

We also perform a second set of translations from English to Telugu, following are some outputs for the same.

- **Project Management**

    - **o Implementation status report**

    Work completed

        - **Description >> Pankaj**

        As mentioned in Increment-1 , following task were completed with Increment-2.

        - Build a transformer model for machine translation
        - Summary text from Image is feed as input to model
        - Model will provide output in the language which is lowly resourced like Indian regional language
        - Researched multiple articles based on encoder-decoder and pre trained like GPT-3
        - Selected Telugu language and preprocess both english & Telugu dataset
        - Transformer model was trained on 15K+ vocab
        - BLEU score was published

- **Responsibility (Task, Person) >> Priya**

TASK	PERSON
Processing the captions generated by the captioning model using various Natural language processing techniques. Working on the data analysis and project management part in the project	Priya
Training the machine translation model to translate the summaries obtained from the captioning model. Working on research for different possible ways to perform machine translation of the summaries.	Aniv
Research on possible evaluation methods for the context of this project. Implementing evaluation of models and results using similarity score for captions and BLEU score for translations.	Pankaj
Integrating the image captioning network and machine translation network and fine-tuning the hyperparameters to reduce loss in both the networks	Shridhar

- **Contributions (members/percentage) >> Priya**

Priya : Applied various pre-processing techniques on the images for the image captioning model and natural language processing techniques for the machine translation model.

Aniv : Built the encoder-decoder architecture for both the models and trained the network. Researched on the possible options for the machine translation architecture.

Pankaj : Evaluated the models using different evaluation techniques like loss curve for captioning and its human evaluation, BLEU score for the translated summaries.

Shridhar : Tuning the hyper-parameters for both the models and integrating them. Reducing the loss involved in training both the models.

- **Issues/Concerns >> Pankaj**
    - The dataset required to trained models are huge
    - Run into multiple issue while executing in Colab and have to pay license fee to buy more RAM
    - We need more dataset to train our object summary model as its accuracy can still be improved
    - Need more resource on other language of worlds apart from English, Spanish, French etc
    - Pre trained model like GPT3 are very good option for better accuracy in short period

- **References/Bibliography >> All**

https://prateekjoshi.medium.com/english-to-hindi-translation-made-simple-with-transformers-library-33f64f745552

https://kuldeepsangwan.medium.com/machine-translation-of-indian-languages-using-transformer-bbd8c93053f7

https://www.sciencedirect.com/science/article/pii/S2666651020300024

https://towardsdatascience.com/neural-machine-translation-15ecf6b0b

https://www.tensorflow.org/text/tutorials/nmt_with_attention

PROJECT INCREMENT 1 : Visual Captioning and Neural Machine Translation

Name	Email Id	Student Id
Aniv Chakravarty	anivchakravarty@my.unt.edu	11424113
Priya Enuganti	priyaenuganti@my.unt.edu	11508112
Shridhar Kshirsagar	ShridharKshirsagar@my.unt.edu	11541383
Pankaj Yadav	pankajyadav@my.unt.edu	11486346

Github link
https://github.com/PriyaEnuganti/Visual-Captioning-and-Neural-Machine-Translation

Video link
https://drive.google.com/file/d/1xw-xNAdGoXGHy-HIi16hEHby4zIJvEaE/view?usp=share_link

• **Related Work (Background)** :-

Image captioning is the process of generating textual description for an image.It uses a combination of Natural Language Processing and Computer vision to generate.
The solution consists of two modules – one is an image based model to extract features and nuances out of image, second one is language based model – translates features and objects given by our image based model to natural sentence.
The dataset used is either from COCO or google open images dataset for images.
The features can be extracted using the CNN and feature vector is linearly transformed to have the same dimension as input for Recurrent Neural Network.
Recurrent Neural Network is trained with label and target text.This follows encoder and decoder model where CNN act as an encoder and RNN is a decoder where it does language modeling up to the world level.
ref : https://towardsdatascience.com/image-captioning-in-deep-learning

Neural Machine Translation is the process of translating from one language to another language using a single large neural network. It is achieved by combining Neural networks and Natural Language processing. Neural Machine translation uses the Probabilistic approach. While working on NMT its important to look out for

model building i.e. how to design its architecture, inference which aims on ways of generating translated output for given input and finally learning which is about how to make the NMT model learn the parameters from input data.

Modeling a translation model can be done in different levels such as for whole document, for each paragraph or on each sentence. For this application the focus is on sentence-level as we expect a sentence from visual captioning.
Ref : https://arxiv.org/pdf/2012.15515.pdf

• **Dataset : -**
COCO has large scale object detection and image caption. It has several features:Flickr Dataset:

- Object identification with instance annotations
- Context recognition
- Dense Pixel to identify vectors
- High number of labeled images
- 1.5 Mio object instances
- Multiple categories of object
- Many stuff categories which has boundary to build context
- pre-trained key points

Flickr Dataset:
- Has been selected from various flickr groups to form 30000 images.
- Manually selected to have distinct scenes and situations.
- Each image has five different captions describing the significant features and entities from the image.

• **Detail design of Features**
The image is RGB three channels of pixel value between 0 and 255. The captions and annotations of images are in text format in the form of a sentence. We build two models based on long short term memory neural network models along with transfer learning on different pre-trained models like Inception v3 in order to use the input image feature in order to generate text output as a sentence. The training preprocess of the input images of model-1 generates 2048 features from the image. We then make use of tf-idf word frequencies to build a vocabulary along with special tags for start, end, padding and UNK or unknown for any out of vocabulary words encountered. After tokenizing and encoding the training samples we perform training by setting hyperparameters of the number of hidden states (300), word embed size (128), batch size (512), number of epochs (500) and learning rate of (1e-3). The transfer learning

model consists of a two-layer encoder-decoder architecture with 2 input layers, 1 embedding layer, 2 LSTM layers and a dense layer at the end. The activation function used is tanh to better account for image captioning. The loss is calculated using a sparse entropy matrix.

• **Analysis**

After a detailed analysis of the dataset, the available processing capacity and the time in hand, two datasets were shortlisted. One being the COCO (Common Objects in Context) and the other being the Flickr dataset. COCO being a massive dataset, only 5000 instances from the data were picked to work upon for this project. The images come with a list of 4-5 annotations attached to them. The images and their captions constitute the training data of the model. The preprocessing part involves cleaning the annotations of any tags or punctuations and then are passed on to the model.

• **Implementation**

Implementation has two parts. First step is building a visual captioning model using neural networks and Natural Language Processing. The next step is developing a neural machine translation model that would translate the output of the visual captioning model from English to Spanish, which is the second most popular language spoken in the United States.

The visual captioning model building is done in two ways. The first method is to do transfer learning on a COCO dataset using LSTM. The second method is to build a CNN model with encoder decoder architecture using flickr dataset. The accuracies of the two models are then compared to choose the best among them.

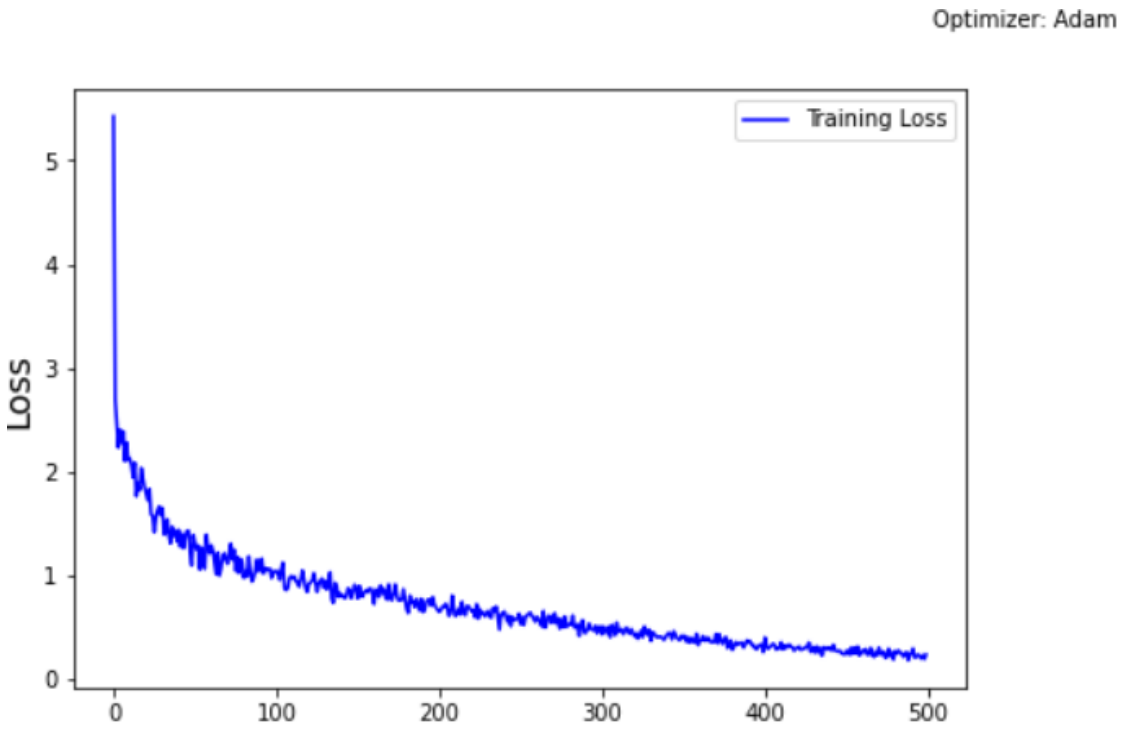
Neural Machine translation is the second module for the project. It will be implemented using RNN with encoder-decoder architecture. Encoder here aims to focus on the order of words and complex dependencies among phrases in the original language. The decoder is simply a language model that leans from the encoder output. Will compare the results with google translation for evaluating model performance and accuracy measures.

• **Preliminary Results**

Model 1: Training results

# of epoch	loss function	# of batch
40	1.3	4.5

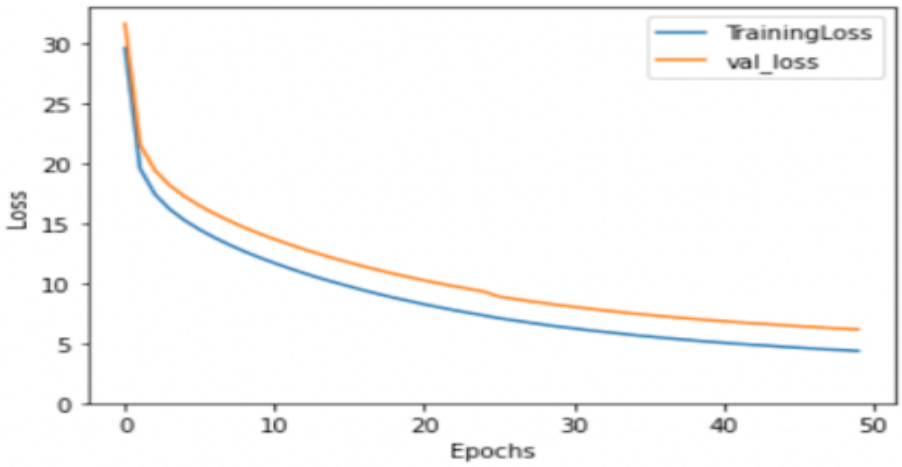
100	1.03	4.5
200	0.65	4.5
500	0.22	4.5



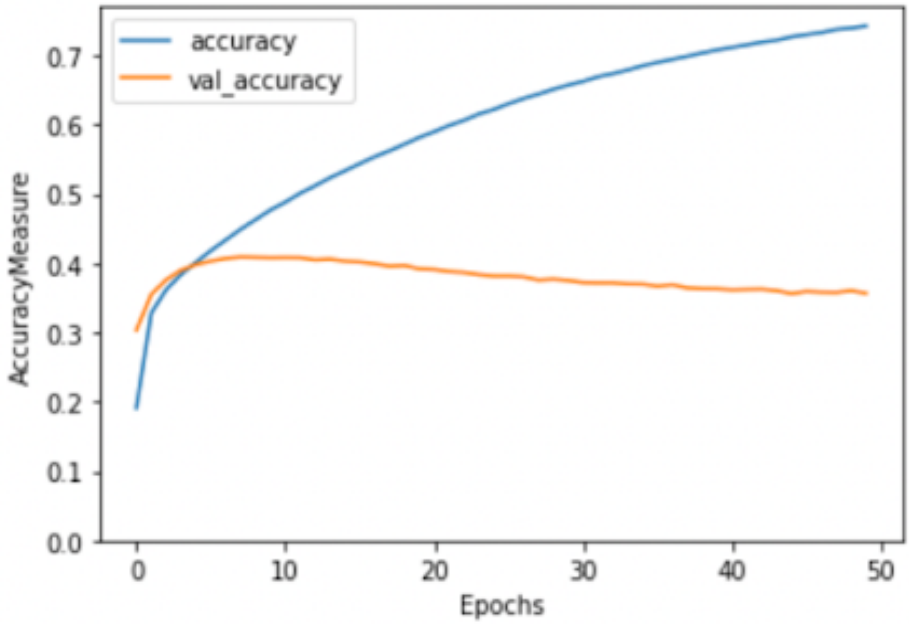
Model2: Training results

After tuning hyper parameters here epochs and batch size, a batch size of 64 and 50 epochs has given decent results so far and need some improvement further.

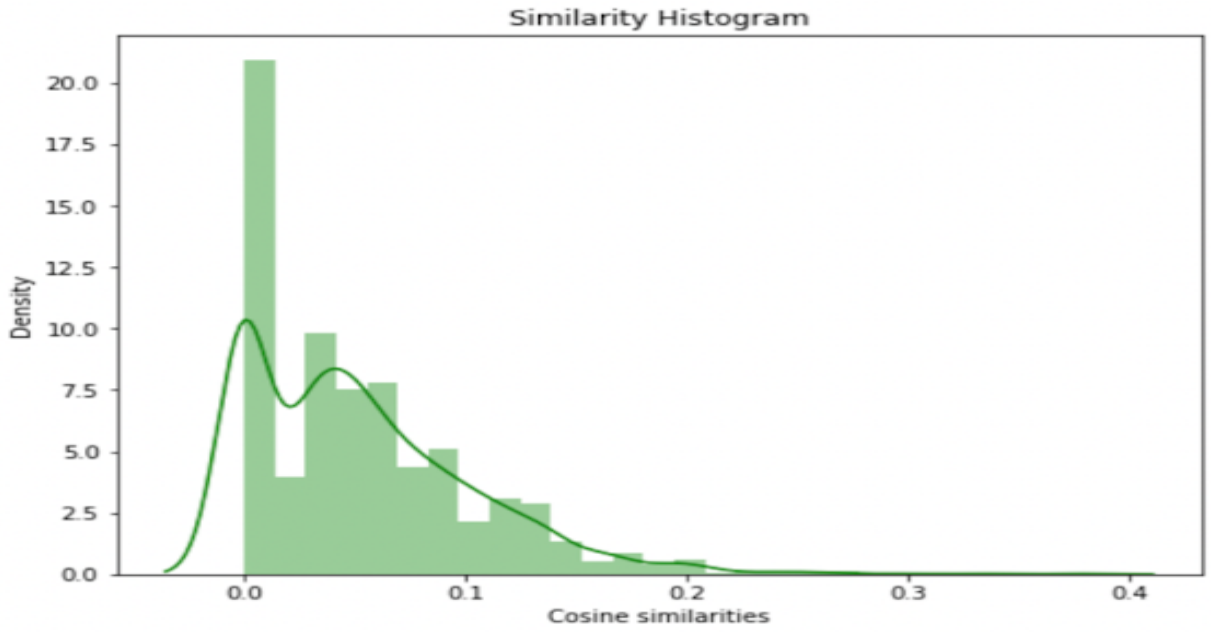
The Training Loss and validation loss for 50 epochs can be seen below. It started high at around 30 and has seen a decreasing trend as the epochs increases and has seen a steady decrease between 40 to 50 epochs.



The accuracy plot between training and validation accuracy says that model is learning better on training data with increasing epochs and invalidation it sees a slight decreasing trend. This could be further improved by adding more training data and increasing epochs.



Cosine Similarity Measure is used to evaluate model performance by comparing the similarity between expected caption vs model generated caption. The resulting distribution for over 1500 testing images can be visualized below.



• **Project Management**

We make use of Whatsapp group chat for quick communication and zoom for weekly meetings. All code and documentations are updated through google groups and github. In terms of keeping track of timelines and milestones we have each been able to complete our tasks set up.

- Implementation status report
  - Literature Survey: Completed on 10/14/22
  - Captioning model group Priya and Shridhar Completed on 10/26/22
  - Captioning model group Aniv and Pankaj Completed on 10/27/22
  - Captioning model evaluation: in progress
  - translation model: to be started for increment 2
  - translation model evaluation: to be started for increment 2
  - report of final results, documentations and demo: in progress

IMPLEMENTATION STATUS REPORT

Work completed

Description

The project was laid out with two major objectives of Visual captioning, i.e. generating a caption for an RGB image and Neural Machine Translation, i.e. translation of caption corpus into another language. Increment 1 marks the completion of the first half of the project, i.e Visual Captioning. Given below is the summary of how the tasks were split up and integrated together.

• **Responsibility (Task, Person)**

Task	Description(Priya)	Person
Research and Background	Reading and getting insights from previous work done on the similar project idea. This would he used to enhance the existing idea adding or changing implementation.	Everyone
Data Analysis and preprocessing	Analyzing the images and annotated captions to see any patterns or check for necessary processing to be done. Then preprocessing is applied on images to vary features such as intensity or orientation when necessary.	Aniv, Priya
Designing and building the architecture	Desigining the basic architecture is key which includes the number of layers in Neural networks then the encoder-decoder which does help in image captioning	Pankaj, Shridhar
Training and Tuning	Model training is where it learns the corelation between image and its caption by updating weights. Tuning is done by	Aniv, Priya

	varying hyper parameters like epochs, batch size to achieve good performance on unseen data.	
Performance Measures and Testing	Measuring the model performance on unseen data using techniques like cosine similarity here to check for sentence matching between expected and predicted caption.	Pankaj, Shridhar

• Contributions :

Aniv Chakravarty: Went over the literature survey of Attention to Attention Image captioning and Enhanced Descriptive Image Captioning and proposed these papers for the initial meeting to determine possible methods.
Began to work on the implementation of image captioning model with Pankaj as part of sub team model 1. Built transfer learning model using subset of COCO dataset and Inception V3 on google colab notebook and debugging any issues. Performing preliminary hyperparameter tuning to get some relevant output captions before letting Pankaj perform in detail tuning and evaluation. Update Increment 1 report for design of features and project management. Building translator in Hindi and telugu.

Pankaj Yadav: Did online research to identify the related project and approaches to accomplish this Task. Coordinated within the team for approach and will continue now to work on a Model based on the LSTM.

Priya Enuganti:

Did background research work focusing on neural Machine Translation to understand how it outperforms other translation techniques and its architecture. Discussed the same with the team on how to further implement it. Analyzed the flickr dataset and worked on image processing, building the encoder-decoder architecture, training CNN model for visual captioning.
Also worked partially on testing and tuning the model.

Shridhar Kshirsagar :

Studied research related to model training and better evaluation methods for the captions generated. Came across a term called ‘confidence score’ for the generated captions. Did the hyper parameter tuning and performed model evaluation using Cosine Similarity technique in Natural Language Processing.

Work to be completed

Description

- Need to build a model for machine translation
- Summary text from Image will be feed as input to model
- Model will provide output in the language of choice
- Research have to be conducted for appropriate features
- Model training
- Testing from the output of one the successful image captioning model

• **Responsibility (Task, Person)** :: Priya

Task	Description(Shridhar)	Person
Data Analysis and preprocessing	Analyzing the generated captions and removing unnecessary punctuations, stopwords from the captions. Additionally, using lemmatization techniques to extract the lemma out of the sentence.	Aniv
Designing and building the architecture	The architecture of the neural machine translation block will consist of two neural networks. An encoder trying to encode the obtained captions and a decoder trying the decode the content and give an output in the context of another language.	Shridhar
Training and Tuning	Combinations of	Priya

	hyperparameters such as number of neurons, number of epochs, dropouts, etc. will be tried and tested out.	
Accuracy Measures and Testing	The generated text will be compared against Google translator using an accuracy measure like cosine similarity,etc.	Pankaj

• **Issues/Concerns**

- Model selected needs huge amount of training dataset
- even 5000 images were not sufficient
- Colab has memory issues despite using GPU
- Captioning tokenisation needs more fine tuning to get accurate results
- Only COCO and google open image are readily available
- Flickr dataset is diverse and needed more images to improve performance
- Model performance need to be further enchanced
- Finding corpus for the language of translation
- Identifying algorithm to convert language with appropriate stopwords/grammer

• References/Bibliography :

1. https://ieeexplore.ieee.org/document/9197977
2. https://github.com/PriyaEnuganti/Visual-Captioning-and-Neural-Machine-Translation
3. https://vision.cornell.edu/se3/wp-content/uploads/2018/03/1501.pdf
4. https://openaccess.thecvf.com/content_ICCV_2019/papers/Huang_Attention_on_Attention_for_Image_Captioning_ICCV_2019_paper.pdf
5. https://github.com/husthuaan/AoANet
6. https://aclanthology.org/2021.acl-short.36.pdf
7. https://github.com/Gitsamshi/Nli-image-caption
8. https://github.com/susantabiswas/image-summarizer/blob/master/Image_Captioning.ipynb
9. https://arxiv.org/pdf/2012.15515.pdf
10. https://github.com/nageshsinghc4/Neural-machine-translation-NMT