Automated Image Captioning

Group 3 – Cohort 18

**Team**

-Sujatha Kancharla

-Monica Nukarapu

-Suhail Pasha Kotwal

-Chaitanya Harkara

**Mentor**

-Nayan Jha

Table of Contents

[Project Description 3](#_Toc107034650)

[Objective 3](#_Toc107034651)

[Timelines 3](#_Toc107034652)

[Dataset 3](#_Toc107034653)

[Deliverables 3](#_Toc107034654)

[Technology 3](#_Toc107034655)

[Understanding of the problem 4](#_Toc107034656)

[Proposed Solution 6](#_Toc107034657)

[Option1 6](#_Toc107034658)

[Option2 11](#_Toc107034659)

[Conclusion 11](#_Toc107034660)

# Project Description

Captioning the images with proper description is a popular research area of Artificial Intelligence. A good description of an image is often said as “Visualizing a picture in the mind”. The generation of descriptions from the image is a challenging task that can help and have a great impact in various applications such as usage in virtual assistants, image indexing, a recommendation in editing applications, helping visually impaired persons, and several other natural language processing applications. In this project, we need to create a multimodal neural network that involves the concept of Computer Vision and Natural Language Process in recognizing the context of images and describing them in natural languages (English, etc). Deploy the model and evaluate the model on 10 different real-time images.

# Objective

Build an image captioning model to generate captions of an image using CNN

# Timelines

Start - 18-Jun and End (Delivery) –11-Sep

# Dataset

Flickr8k, Flickr30k & COCO

# Deliverables

* Project Technical Report
* Project presentation with desired Documents
* Summary of 3 research Papers

# Technology

* **Tools** : Natural Language Toolkit, TensorFlow, PyTorch, Keras
* **Deployments:** FastAPI, Cloud Application Platform | Heroku, Streamlit, Cloud Computing, Hosting Services, and APIs | Google Cloud

# Understanding of the problem

Automated Image captioning involves in creating an automated caption for an Image by deriving the best context of the contents of the image.

Broadly the solution should

1. Identify multiple objects within the image
2. Derive the relationship between the objects in the image based on their attributes
3. Derive the caption based on the derived context of the image in Natural language (English)

|  |  |
| --- | --- |
| A picture containing green, toy, colorful, close  Description automatically generated  Yoga for… | Trick Photography… |
| A picture containing grass  Description automatically generated  Tennis on sand … | A green frog figurine  Description automatically generated with medium confidence  Make money … |

**Key inputs**

Historically Image captioning solutions were that were developed have been template based, which were heavily hand designed and rigid in terms of Text generation.

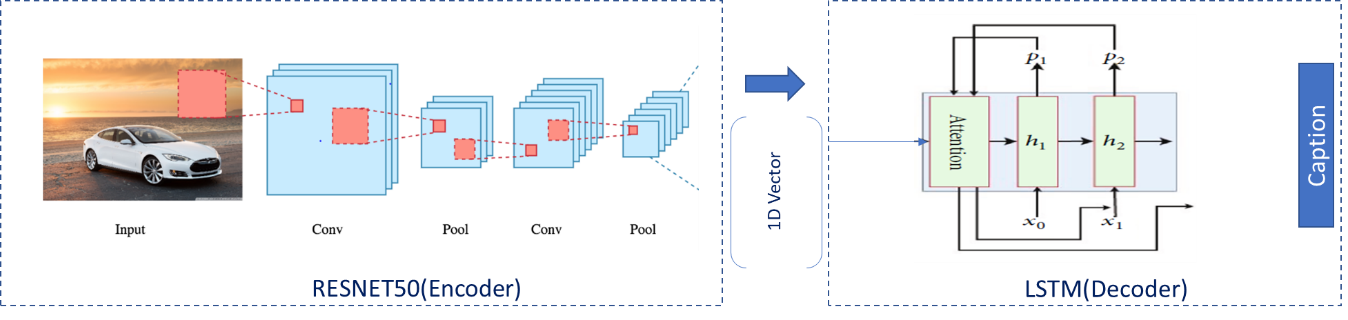
**Key considerations**

Based on the latest solutions of Text generation Using Recurring Neural Networks (RNN), there are multiple recommendations (Research papers) to develop an Image captioning solution using a combination of CNN (encoder) and RNN(Decoder). The research papers (“Show and Tell – “A Neural image caption generator” and AICRL – Automate Image Captioning Resnet50 LSTM)

# Proposed Solution

## Option1

* **Solution Architecture**



**Inspiration**

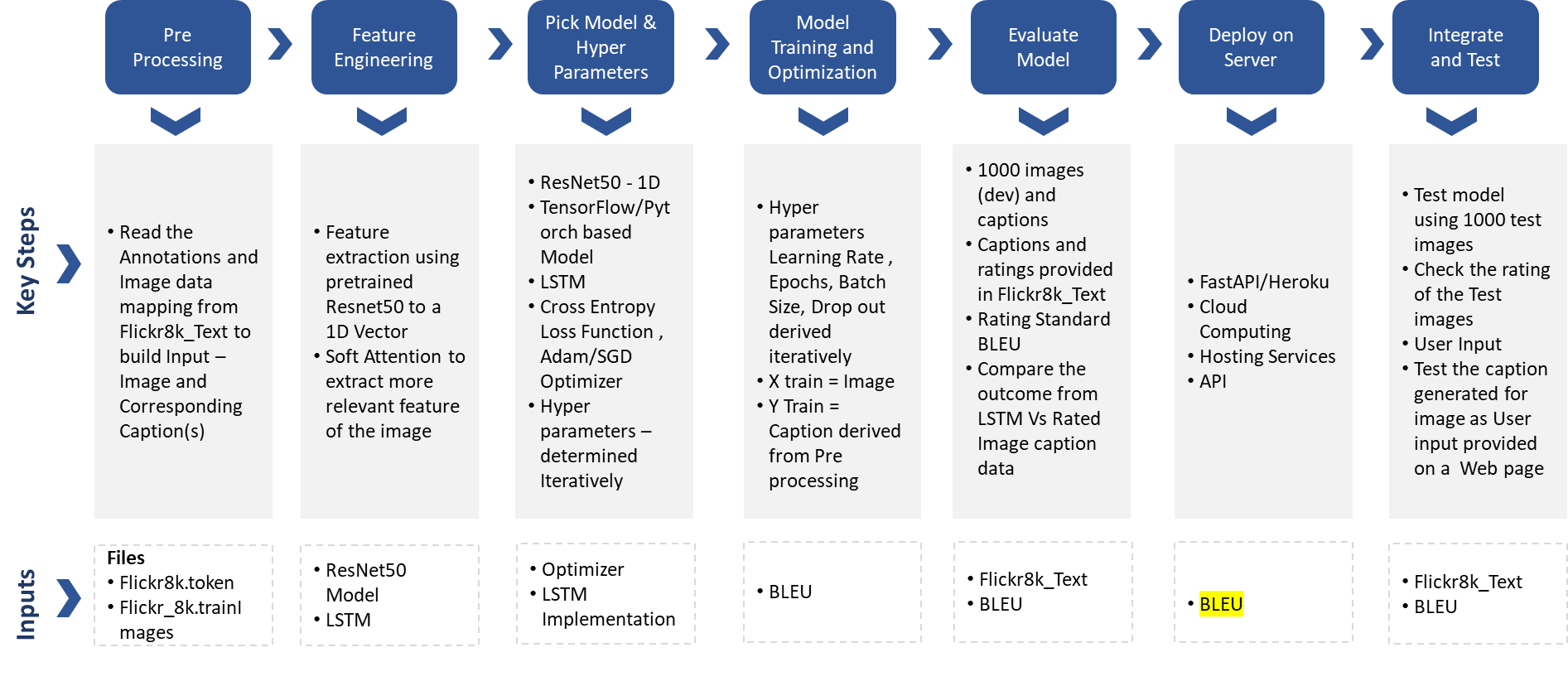
Automatic image captioning based on ResNet50 and LSTM with soft attention *( Reference – Technical paper - AICRL – Automate Image Captioning Resnet50 LSTM*) . The model was designed with one encoder-decoder architecture where ResNet50, a convolutional neural network was adopted as the encoder to encode an image into a compact representation as the graphical features and then, a language model LSTM was selected as the decoder to generate the description sentence. A soft attention model was integrated with LSTM such that the learning can be focused on a particular part of the image to improve the performance.

* **Key Highlights**

**Encoder**

* Represent the image, using pretrained convolutional neural network (CNN), ResNet50, which is a very deep network that has 50 layers
* Extract visual features, which use ResNet50 network as the encoder to generate a **1D vector** representation of the input images

**Decoder**

* Soft attention is implemented by adding an additional input of attention gate into LSTM that helps to concentrate selective attention
* LSTM networks are used to accomplish the tasks of machine translation and sequence generation
* Execution Plan
* **Solution Design and Implementation**
* **CNN Design**

Using Pretrained ResNet50 for creating 1D vector from Image input. The size has to matched to the input to Attention and LSTM (Size of Dictionary)

* **Attention Design**

The attention gate can be represented as an addition input for LSTM. The soft attention depends on the previous output of LSTM and extracted features of input image

* **LSTM Design**

Input for LSTM is an Attention vector which is of same dimensions as One hot representation of the words (dimension of Dictionary)

* **Key Functions and Hyper Parameters (Sample)**

These are representative details which will be updated as per our results in future

|  |  |
| --- | --- |
| Function/Parameter | Value |
| Loss Function | CrossEntropy |
| Optimizer | Adam |
| Learning Rate | 0.003 |
| Epochs | 50(Training) |
| Batch Size | 100 |
| Droup out | 0.2 |

* **Training and Testing**
* **Training Dataset** 
  + 6000 Images from Flickr\_8k.trainImages.txt
  + Caption Mapping from Flickr8k.token.txt
* **Validation Dataset**
  + 1000 Images from Flickr\_8k.devImages.txt
  + Caption Mapping from CrowdFlowerAnnotations.txt
* **Testing Dataset**
  + 1000 Images from Flickr\_8k.testImages.txt
  + Caption Mapping from ExpertAnnotations.txt
  + Additional Noisy Dataset – To be determined on preparation of the data

We propose to use Standard rating Metrics Models to derive the efficiency of Model during Validation and Testing. The Expert annotations and Crowd Annotations will be used for this purpose

We intend to use **BLEU** rating model to come up with standard rating for Validation and Testing of the Model

* **Results**

To be determined post the training and Testing of the Solutions

* **Key Highlights**

**Pros**

* Feature extraction through Pre trained model like Restnet50 will help reduce the time to build and optimum CNN
* LSTM will be help reduce the vanishing gradient descent problem for RNN
* Implementation of Attention layer would reduce the processing of entire image vector in every LSTM
* BLEU rating could help us already existing public successful models

**Cons**

* The accuracy validated so far are not very high based on the Technical papers
* If Accuracy is not high it might have limitations of implementation for Visually challenged

Additional highlights will be to be determined post the training and Testing of the Solutions

* **Key Application in real life**
  + Recommendations in editing applications,
  + Usage in virtual assistants,
  + For image indexing,
  + For visually impaired persons,
  + For social media
  + Natural language processing applications.
* **Key Challenges and Learnings**

To be determined post the training and Testing of the Solutions

* **References**

Technology Papers - “Show and Tell – “A Neural image caption generator” and AICRL – Automate Image Captioning Resnet50 LSTM)

## Other Options

Based on the results of the solution we would like try the following models (time permitted)

**Other Pre-Trained CNN ( Encoder) Models**

* Inception V3, Xception models
* GoogleNet

**Other Decoder Models (Word Embeddings)**

* Word2Vec
* GloVe
* BERT

**Other Decoder(s)**

* Transformers
* LSTM Without Attention

**Datasets**

* Flickr30
* COCO

# Conclusion

To be determined based on the final solution and results