NeuraIC - Neural Image Caption Generator for Assistive Vision

Ankush Govind Chavan Pune, India  
cankush625@gmail.com

Rishabh Kumar  
Pune, India  
@gmail.comKuldeepsingh RajpurohitPune, India  
@gmail.com

MRS. Mansi BhonslePune, India  
@gmail.comAbhishek Kumar Singh  
Pune, India  
@gmail.com

Keywords—CNN, Deep Learning, image captioning, mobile application, MSCOCO

# Introduction

# Inspired by recent advances in Deep Learning based Machine Translation and Computer Vision based Object Detection have led to excellent Image Captioning models. While these models are very accurate (71.5% - 79%), these often rely on the use of expensive computation hardware making it difficult to apply these models in real-time scenarios, where actual applications can take advantage of them. In this report, we carefully follow some of the heuristic techniques and core concepts of Image Captioning and its common approaches and present our simplistic sequence to sequence based implementation with significant modifications and optimizations like using beam search instead of greedy search which enable us to run these models on low-end hardware of mobile devices. We also compare our results evaluated using various metrics with state-of-the-art models and analyze why and where our model trained on MSCOCO dataset lacks due to the trade-off between computation speed and quality. Using the state-of-the-art Flutter UI software development kit by Google, we also implement a Mobile application to demonstrate the real-time applicability and optimizations of our approach.

# Problem Definition

Automatically describing the content of an image along with their relationships or the actions being performed is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. But this could have a great impact by helping visually impaired people better understand their surroundings. These images can then be used to generate captions that can be read out loud to the visually impaired so that they can get a better sense of what is happening around them. By this project, we present a mobile application which uses a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences which describe an image captured by the mobile’s camera. The model is trained to maximize the likelihood of the target description sentence using Maximum Likelihood Estimation (MLH) given the training image. ​ What is most impressive about this method is that it is a single end-to-end model that can be defined to predict a caption, given a photo, instead of requiring sophisticated data preparation or a pipeline of specifically designed models.

Not only must the model be able to solve the computer vision challenges of identifying the objects in an image, but it must also be intelligent enough to capture and express the object’s relationships in natural language. For this reason, image caption generation has long been considered as a difficult problem. Its purpose is to mimic the human ability to comprehend and process huge amounts of visual information into a descriptive language, making it an attractive problem in the field of AI. Many major tech-companies are investing heavily in Deep Learning and AI research, as a result of which the particular problem of image captioning is being studied at several organisations by several different teams. The two main bodies of work that form the basis of this paper are [Show and Tell by Oriol Vinyals et al (2015) [1]](https://arxiv.org/abs/1411.4555) and the more advanced, attention based [Show, Attend and Tell by Kelvin Xu et al (2016) [2]](https://arxiv.org/abs/1502.03044).

Image captioning can be used for a variety of use cases such as assisting the blind using text to speech by real time responses about the surrounding environment through a camera feed, enhancing social media experience by converting captions for images in social feed as well as messages to speech. Assisting young children in recognizing objects as well as learning the English language. Major help in SEO techniques, Captions for every image on the internet can lead to faster and descriptively accurate image searches and indexing. In robotics, the perception of the environment for an agent can be given a context through natural language representation of the environment through the captions for the images in the camera feed.

## Object Detection

Object recognition is a general term to describe a collection of related computer vision tasks that involve identifying objects in digital photographs. Object detection combines two important computer vision tasks Object localization and Object classification. In Object detection we locate the presence of objects with bounding boxes and classify / label each bounding box.

## Linguistics

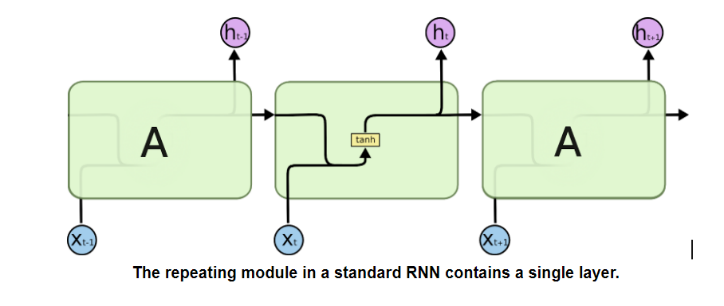
Humans don’t start their thinking from scratch every time. As we read this essay, we understand each word based on our understanding of previous words. We don’t throw everything away and start thinking from scratch again. That means our thoughts have persistence.

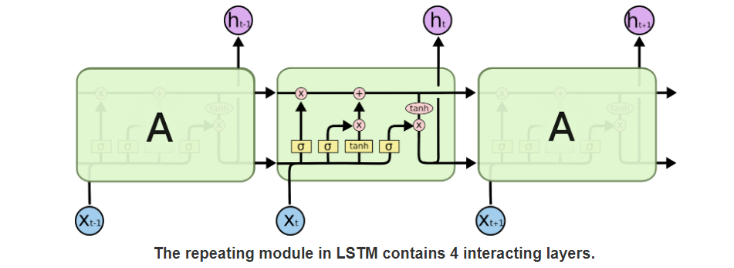
Traditional neural networks can’t do this, and it seems like a major shortcoming. Recurrent neural networks address this issue. They are networks with loops in them, allowing information to persist.

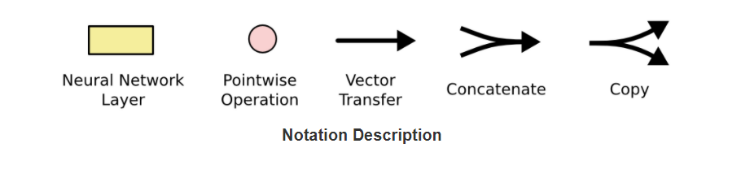
**LSTM Networks**

Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies. They work tremendously well on a large variety of problems, and are now widely used.

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn.







# ARCHITECTURE

## MODEL

In our implementation we used a sequence to sequence encoder-decoder architecture system. The encoder being pretrained InceptionV4 Convolutional Neural Network and the decoder, a deep Recurrent Neural Network with Long Short Term Memory Cells. Encoder InceptionV4 is used to transform raw images I into a fixed length embedding F which represent the convolved features for the images. These embeddings are obtained by running a forward pass till the penultimate layer i.e., the average pool layer of the InceptionV4 model. The decoder in our model has two phases, namely, training and inference. The decoder is responsible for learning the word sequences given the convolved features and original caption. The decoder’s hidden state ht is initialised using these image embeddings features F at timestep t=0. Hence the basic idea of encoder-decoder model is demonstrated by the following equations.

*F = encoder(I); Xt=0 = F; Ot = decoder(Xt:0 t)*

The training process in the RNN with LSTM Cell based decoder works on a probabilistic model in which the decoder maximizes the probability of word p in a caption given the convolved image features F and previous words Xt:0 t. To learn the whole sentence of length N corresponding to the features F the decoder uses its recurrent nature to loop over itself over a fixed number of timesteps N with the previous information (features and sampled words at timestep t) stored in its cell’s memory as a state. The decoder can alter the memory Ct as it unrolls by adding new state, updating or forgetting previous states through the LSTM forget ft, input it and output ot memory gates.

ft = σ (Wf· [ht-1, xt] + bf) (1)

it = σ (Wi· [ht-1, xt] + bi) (2)

t = σ (WC· [ht-1, xt] + bC) (3)

Ct = ft\*Ct-1 + it \* t) (4)

ot = σ (Wo· [ht-1, xt] + bo) (5)

ht = ot \* tanh(Ct) (6)

Ot =argmax( softmax(ht) ) (7)

Σ - sigmoid; Ot - Output word; tanh -hyperbolic tangent; Wo,Wf,Wi - Learnable Weight Vector;

bo,bf,bi - Learnable bias Vector;

## MOBILE APPLICATION

Flutter SDK will be used to develop a cross-platform mobile application, which will be used to capture images via device camera ,with this users can also select a pre captured image from the device.

Backend Servers will be hosted on Public Cloud e.g AWS (Amazon Web Services ), GCP(Google Cloud Platform) ,Azure etc,where these Machine Learning models will process images and also various API endpoints will be created to request and post data from and to Server. Flutter application will make an API call on these endpoints to upload images to servers or to fetch processed data from servers.

Captured or pre captured Images will be provided as an input to the backend server via API endpoint, where machine learning algorithms will further process these images and generate appropriate captions specifying what is in the image and this caption once generated it will be sent back to the Application from where the initially Images where uploaded via API call. Once these captions are successfully received in the Application, caption will be converted to speech in the application and played via the device speaker to assist visually impaired or children to learn.

Caption which is generated after processing image which is returned by server,will be stored in a local database on the device, this database will store the compressed image which will be used as a thumbnail and point to their caption, these images and caption will be listed in the app which can be used for future references.

# DATASET AND TRAINING

## Dataset

Our model will be trained on Flickr30k dataset with 31783 images having five captions each, but due to less number of training samples and every training caption beginning with “A man ... “. We will also train our model on MSCOCO (2014) [7] training dataset with 82780 images, each with five ground truth captions.

## Training

For offline evaluation, compared to other Caption Bots our implementation will use Batched data, uses TensorFlow, runs on GPU, and supports CNN finetuning. All of these together result in quite a large increase in training speed for the Language Model (~100x). Although, this split of 5000 images is not a standardized split, but it has been used by many researchers to report their results.

##### References

1. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan (2015) Show and tell: A neural image caption generator. CVPR 1, 2
2. K. Xu (2016) Show, attend and tell: Neural image caption generation with visual attention. inProc. Int. Conf. Mach. Learn.
3. Andrej Karpathy, Li Fei Fei (2015) Deep Visual-Semantic Alignments for Generating Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence (April 2017), vol 39, issue 4:664–676.
4. Ayush Yadav, Rutgers The State University of New Jersey (2017) Camera2Caption: A real-time image caption generator.
5. Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan (2016) Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge PAMI
6. [Beam Search: deeplearning.ai](https://www.youtube.com/watch?v=RLWuzLLSIgw&t=11s)