Project Review II

NeuralC - Neural Image Caption Generator for Assistive Vision

Guide
Prof. Mansi Bhonsle

Group A9
Ankush Chavan
Kuldeepsingh Rajpurohit
Abhishek Kumar Singh
Rishabh Kumar

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 Statement

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.

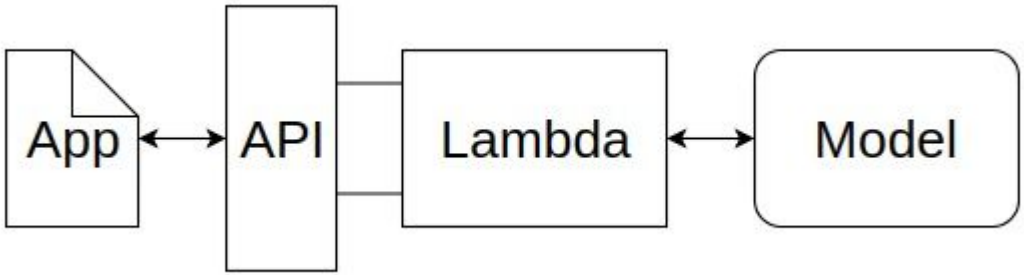
Continue...

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, asa 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] and the more advanced, attention based Show, Attend and Tell by Kelvin Xu et al (2016) [2]. 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.

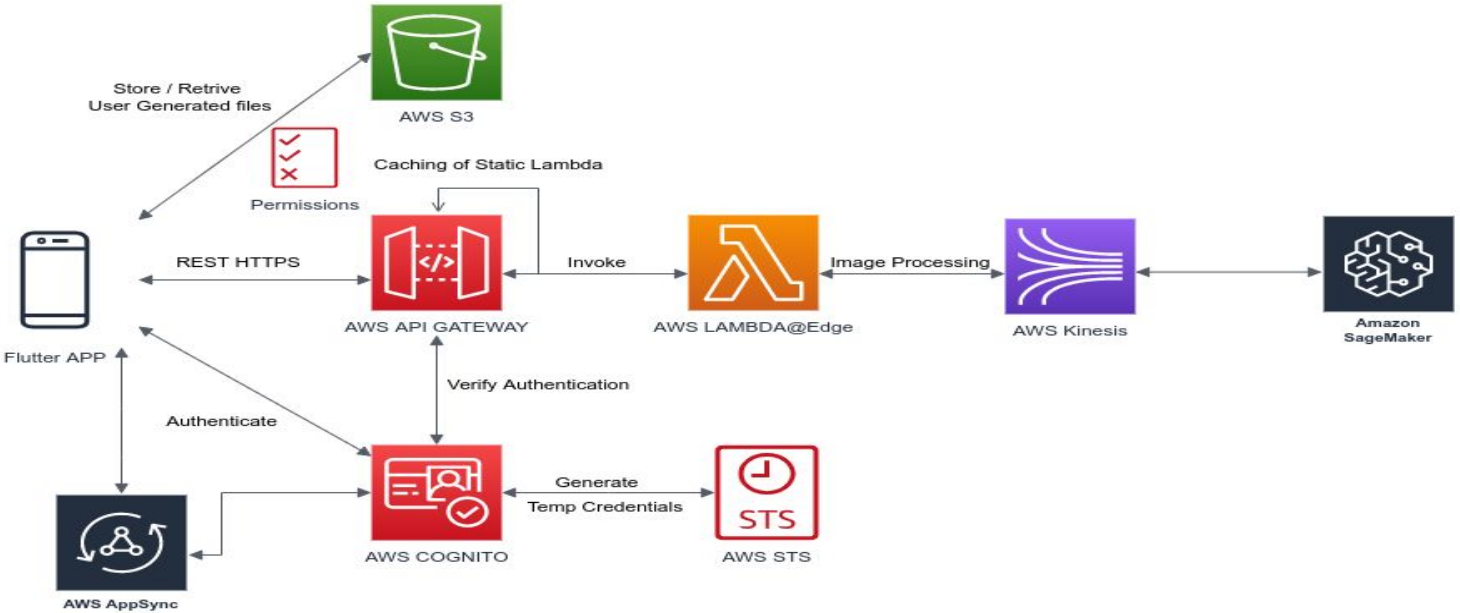
Goals and Objectives

- 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.

Operation Flow



High Level System Design



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.

R-CNN Model Family

The R-CNN family of methods refers to the R-CNN, which may stand for "*Regions with CNN Features*" or "*Region-Based Convolutional Neural Network*," developed by Ross Girshick, et al. This includes the techniques R-CNN, Fast R-CNN, and Faster-RCNN designed and demonstrated for object localization and object recognition.

a. **R-CNN:** It is a relatively simple and straightforward application of CNNs used to propose candidate regions or bounding boxes of potential objects in the image called "*selective search*," to the problem of object localization and recognition. A downside of the approach is that it is slow, requiring a CNN-based feature extraction pass on each of the candidate regions generated by the region proposal algorithm.

Continue...

b. **Fast R-CNN:** It is proposed as a single model instead of a pipeline to learn and output regions and classifications directly. The architecture of the model takes the photograph as a set of region proposals as input that are passed through a deep convolutional neural network. A pre-trained CNN, is used for feature extraction. The end of the deep CNN is a custom layer called a Region of Interest Pooling Layer, that extracts features specific for a given input candidate region. The output of the CNN is then interpreted by a fully connected layer then the model bifurcates into two outputs, one for the class prediction via a softmax layer, and another with a linear output for the bounding box. This process is then repeated multiple times for each region of interest in a given image.

c. **Faster R-CNN:** Since the Fast R-CNN was also not fast enough to be used in real time systems because it takes approximately 2 sec to generate the output on an input image. Instead of using selective search algorithms Faster R-CNN uses a region proposal generation algorithm called Region Proposal Network.

YOLO Model Family

Another popular family of object recognition models is referred to collectively as YOLO or "*You Only Look Once*," developed by Joseph Redmon, et al. The R-CNN models may be generally more accurate, yet the YOLO family of models are fast, much faster than R-CNN, achieving object detection in real-time.

a. **YOLO:** The approach involves a single neural network trained end to end that takes a photograph as input and predicts bounding boxes and class labels for each bounding box directly. The technique offers lower predictive accuracy (e.g. more localization errors), although operates at 45 frames per second and up to 155 frames per second for a speed-optimized version of the model.

Continue...

b. **YOLOv2 (YOLO9000):** Although this variation of the model is referred to as YOLO v2, an instance of the model is described that was trained on two object recognition datasets in parallel, capable of predicting 9,000 object classes, hence given the name "*YOLO9000*". A number of training and architectural changes were made to the model, such as the use of batch normalization and high-resolution input images. Like Faster R-CNN, YOLOv2 model makes use of anchor boxes, pre-defined bounding boxes with useful shapes and sizes that are tailored during training. The choice of bounding boxes for the image is pre-processed using a k-means analysis on the training dataset.

c. **YOLOv3:** The improvements were reasonably minor, including a deeper feature detector network and minor representational changes.

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.

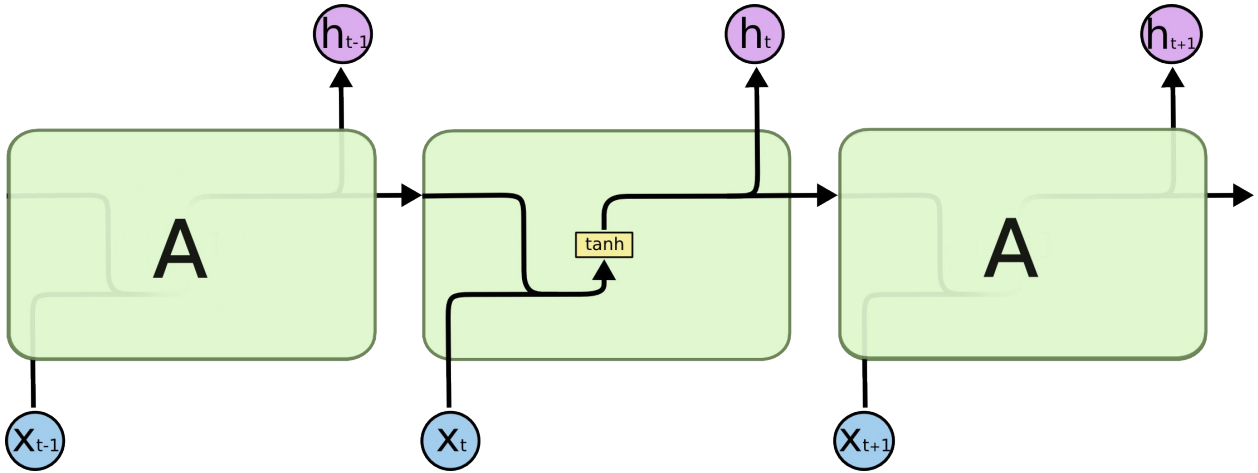
Traditional neural networks can’t do this, and it seems like a major shortcoming. **Recurrent neural networks** address this issue. They are networks with loops in them, allowing information to persist.

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.

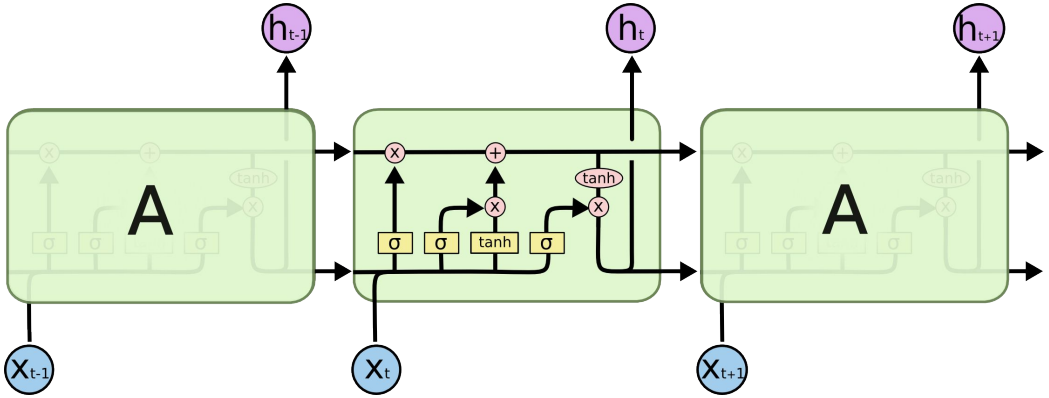
LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn.

Continue...

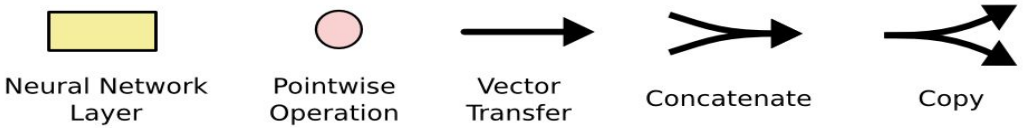


The repeating module in a standard RNN contains a single layer.

Continue...



The repeating module in LSTM contains 4 interacting layers.



Notation Description

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.

Continue...

$$F = encoder(I); \quad Xt=0 = F; \quad Ot = decoder(Xt:0 \ t)$$

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.

Continue...

$$ft = \sigma\,(Wf \cdot [ht\text{-}1, xt] + bf) \quad (1)$$

$$it = \sigma\,(Wi \cdot [ht\text{-}1, xt] + bi) \quad (2)$$

$$t = \sigma\,(WC \cdot [ht\text{-}1, xt] + bC) \quad (3)$$

$$Ct = ft * Ct\text{-}1 + it * t) \quad (4)$$

$$ot = \sigma\,(Wo \cdot [ht\text{-}1, xt] + bo) \quad (5)$$

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

$$Ot = \operatorname{argmax}(\operatorname{softmax}(ht)) \quad (7)$$

σsigmoid; OtOutput word; tanh hyperbolic tangent; Wo,Wf,WiLearnable Weight Vector; bo,bf,biLearnable bias Vector;

Cross Platform 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.

Continue...

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.

Technologys

- Machine Learning
- R-CNN
- Fast R-CNN
- YOLO
- LSTM Networks
- Deep Learning
- Neural Networks
- Object Detection
- Recurrent Neural Networks
- Flutter
- Cross Platform Development
- API
- Cloud Computing