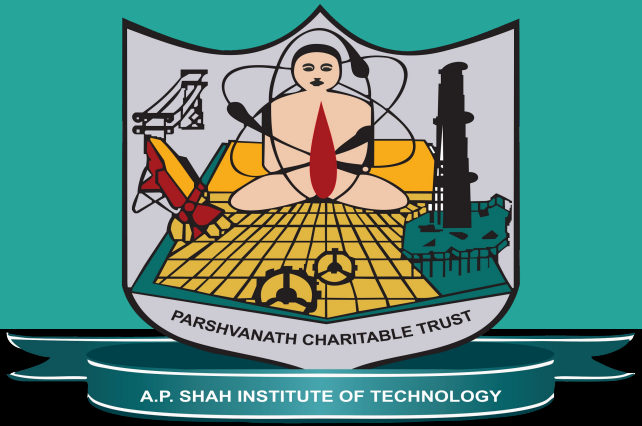



# Computer Engineering Department

A.P. Shah Institute of Technology

G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615

UNIVERSITY OF MUMBAI

Academic Year 2019-2020

A Project Report on

# Title of your project

Submitted in partial fulfillment of the degree of
Bachelor of Engineering(Sem-7)
in
**Computer Engineering**

By
Mrunal S Jadhav ( 16102030 )
Aditya G Joshi ( 16102022 )

Under the Guidance of
Prof. Sachin Malave

# 1.Project Conception and Initiation

# 1.1 Abstract


- The visual world is populated with a vast number of objects, the most appropriate labelling of which is often ambiguous, task specific, or admits multiple equally correct answers.
- A quick glance is sufficient for a human to understand and describe what is happening in the picture. The task is to transform a sentence S written in its source language, into its translation T in the target language, by maximising the probability P(T|S).
- A combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) which seeks to progress directly from image features to text can be progressed to define a single end-to-end model to maximize the likelihood of the target description sentence , given an image, instead of requiring sophisticated data preparation or a pipeline of specifically designed models.
- Thus we can develop a generative model, a probabilistic framework, based on deep recurrent architecture that combines advances in computer vision and machine translation to generate natural sentences describing an image.

# 1.2 Objectives

- Image captioning is a task that a machine learns to generate natural language sentences to describe the salient parts of an image. Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task.
- Generating complete and natural image descriptions automatically has large potential effects, such as titles attached to news images, descriptions associated with medical images, text-based image retrieval, information accessed for blind users, human-robot interaction. These applications in image captioning have important theoretical and practical research value.
- The meaningful description generation process of high level image semantics requires not only the understanding of objects or scene recognition in the image, but also the ability to analyse their states, understand the relationship among them and generate a semantically and syntactically correct sentence.

# 1.3 Literature Review

- In this project we aim to incorporate all the best methods in each stage of creating efficient deep learning model for Image captioning. Recently, a great progress in image captioning has been achieved by using semantic concepts detected from the image, which is very similar to the cognition process of humans. Researchers have proposed a multimodal Recurrent Neural Network model that creatively combines the CNN and RNN model to solve the image captioning problem. Because of the gradient disappearance and the limited memory problem of ordinary RNN, the LSTM model is a special type of structure of the RNN model that can solve the above problems.

# 1.4 Problem Definition

- In this project we hope to achieve more precise and accurate image captioning model. Here, we propose to follow this elegant recipe, replacing the encoder RNN by a deep convolution neural network (CNN).
- There will be an end-to-end system for the problem. It is a neural net which is fully trainable using stochastic gradient descent. The model combines state-of-art sub-networks for vision and language models.
- Finally, we wish to yields significantly better performance compared to state-of-the-art approaches. we propose a fully trainable attribute-based neural network founded upon the CNN+RNN architecture, that can be applied for image captioning.

# 1.5 Scope

- Translation work is achieved by using an “encoder” RNN that reads the source sentence and transforms it into a rich fixed-length vector representation, which in turn is used as the hidden state of a “decoder” RNN that generates the target sentence.
- Replacing the encoder RNN by a deep CNN can produce a rich representation of input by embedding it in a fixed-length vector, so that this representation can be used for variety of tasks.
- Developing a single end to end network to have more accurate feature extraction and efficiently generate textual description which can provide detailed information about the given image.

# 1.6 Technology stack

1. **Colab** - Colaboratory is a Google research project created to help disseminate machine learning education and research
2. **Pytorch** - used for applications such as computer vision and natural language processing.
3. Numpy - NumPy is the fundamental package for scientific computing with Python.
4. **Pandas** – It is a software library written for the Python programming language for data manipulation and analysis.
5. **Keras** – It is an Open Source Neural Network library written in Python and a high-level API wrapper for the low-level API that runs on top of Theano or Tensorflow
6. **Sklearn** - is a free software machine learning library for the Python programming language.
7. **Tensorflow** – It is an end-to-end open source platform for machine learning.

# 1.7 Benefits for environment & Society

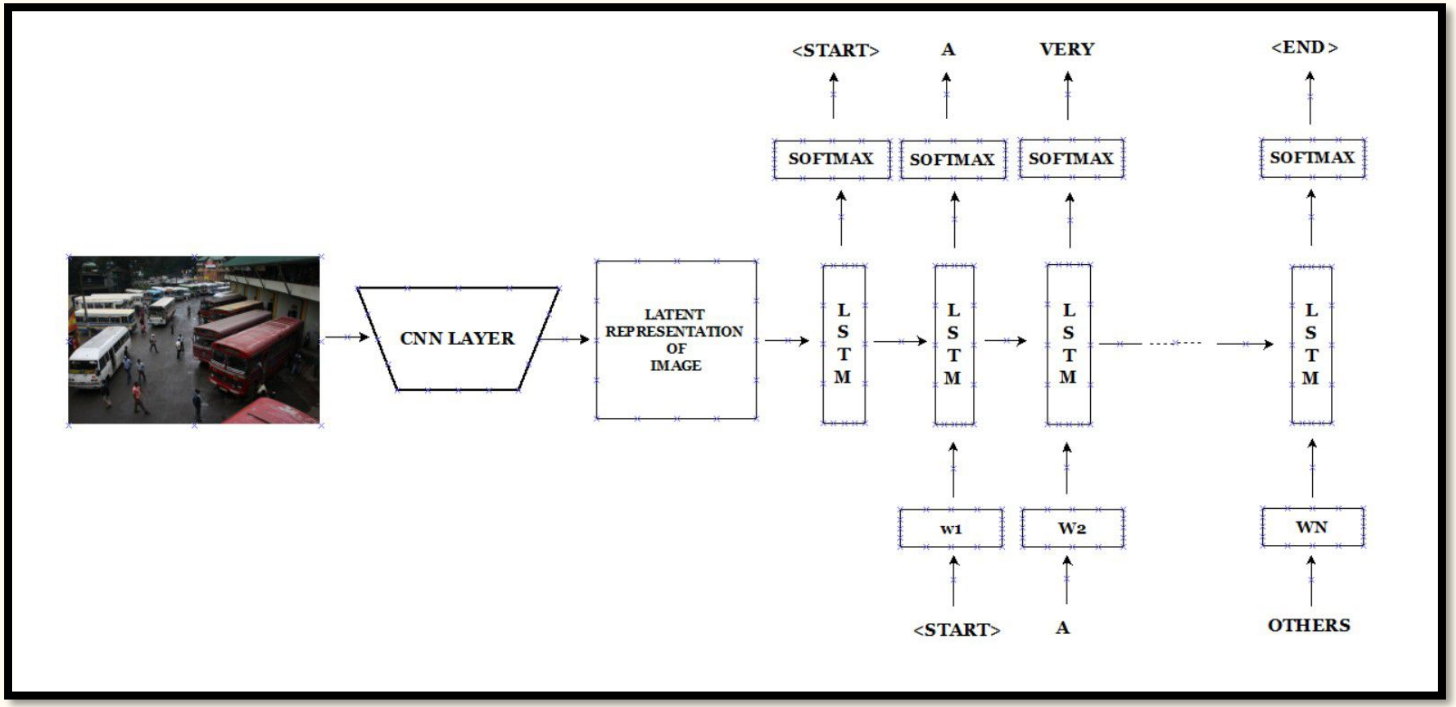
1. Helps visually impaired to understand the image by converting the captioned text into speech.
2. Helps colour blind and other vision problem patients to understand image more effectively.
3. Generate captions for images which can promote safety and protection of environment
4. Determine various pollutants present in a given image and caption them so as to reduce its generation and manage it.

# 2. Project Design

# 2.1 Proposed System

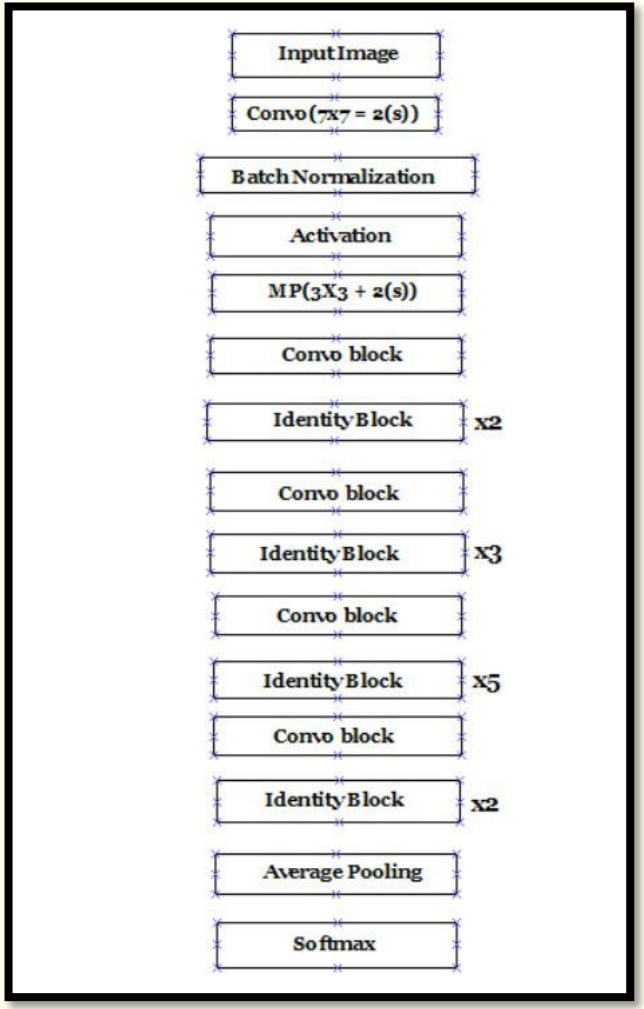
- The proposed system consists of an Encoder Convolutional Neural Network and a Decoder Recurrent Neural Network.
    - **Encoder Convolutional Neural Network** – It is a Deep Learning algorithm which can take in an input image, assign learnable weights and biases to various aspects or objects in the image and be able to differentiate one from other. The resnet50 model is pre-trained on Imagenet dataset and iswidely used in transfer learning
    - **Decoder Recurrent Neural Network** - A recurrent Neural Network (RNN) can be thought as multiple copies of same network, each passing a message to its successor. A decoder RNN uses the last hidden of CNN layer as an input to RNN decoder which uses the fixed dimensional vector representation to decode its desired output caption.

# 2.2 Design(Flow Of Modules)

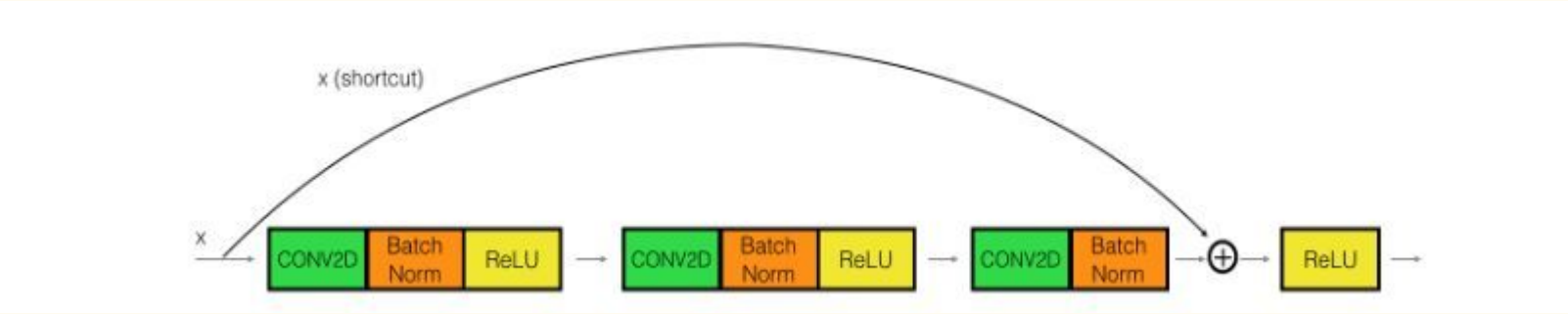


# 2.3 Description

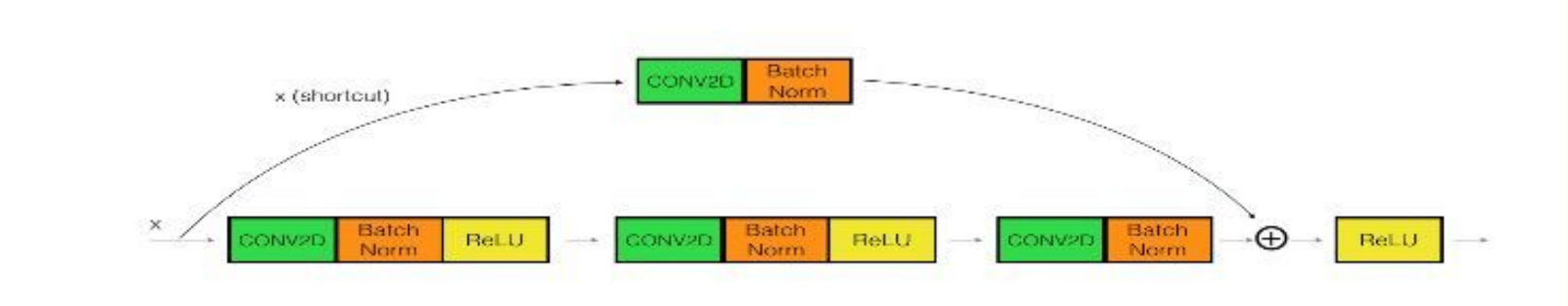
- The input image is passed on to the CNN model for feature extraction. In our project CNN model is a pre-trained Resnet50 model. The architecture of Resnet50 consists of two major blocks :
  - Identity block
  - Convolution block.



Identity Block : The identity block is the standard block used in ResNets, and corresponds to the case where the input activation (say a[l] ) has the same dimension as the output activation (say a[l+2] ).

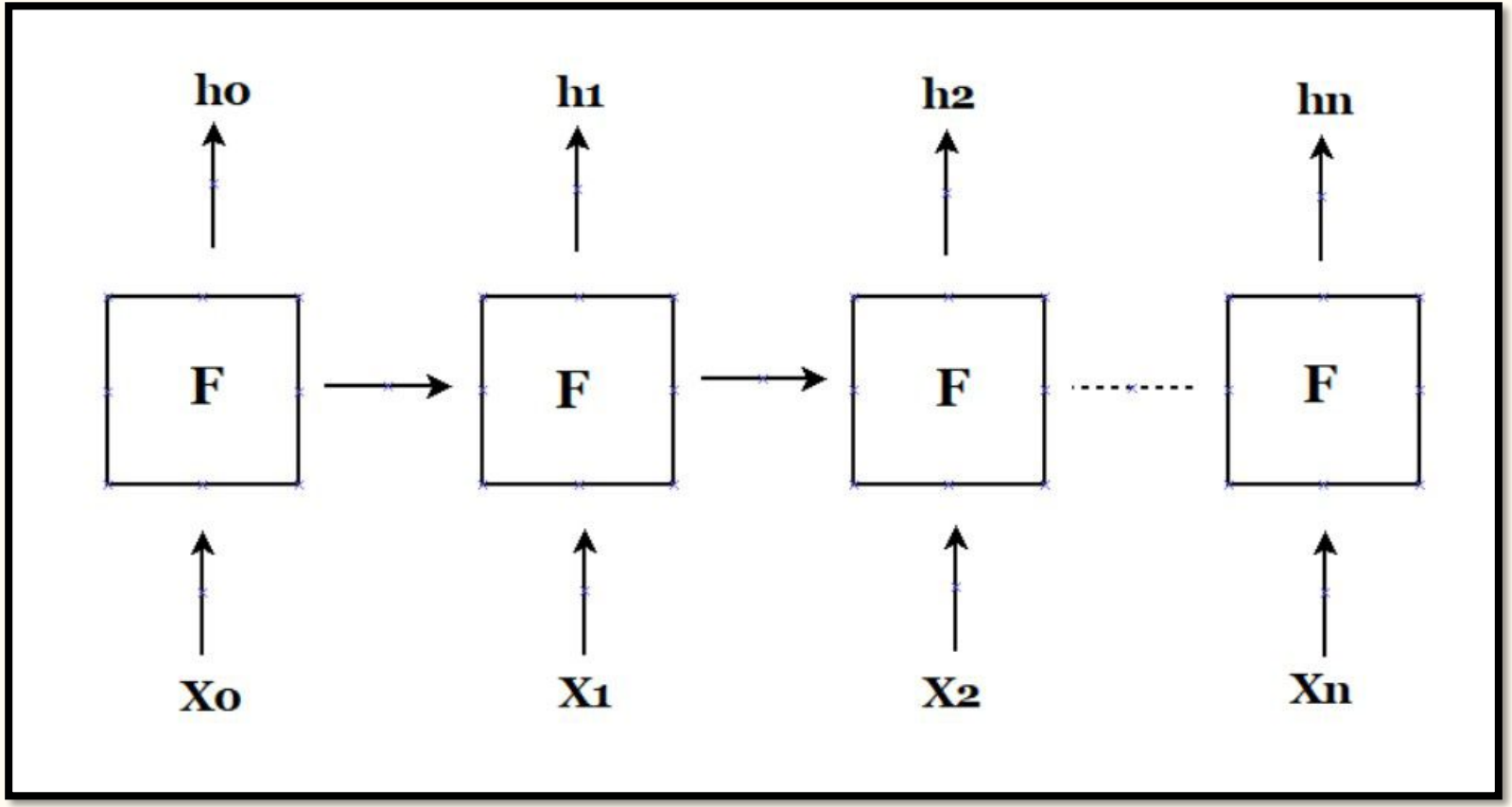



Convolution Block : This type of block is used when the input and output dimensions don't match up. The difference with the identity block is that there is a CONV2D layer in the shortcut path:



# 2.3 Description (contd..)

- The output of last hidden layer of Resnet50 is passed as input to RNN model. The first recurrent block of RNN model takes this input as well as a <START> symbol input. In the successive recurrent units of RNN the output of previous block as well as the current is passed till the last recurrent unit. The softmax activation is applied on output of LSTM block to get the desired words.


ho
h1
h2
hn
F
F
F
F
Xo
X1
X2
Xn

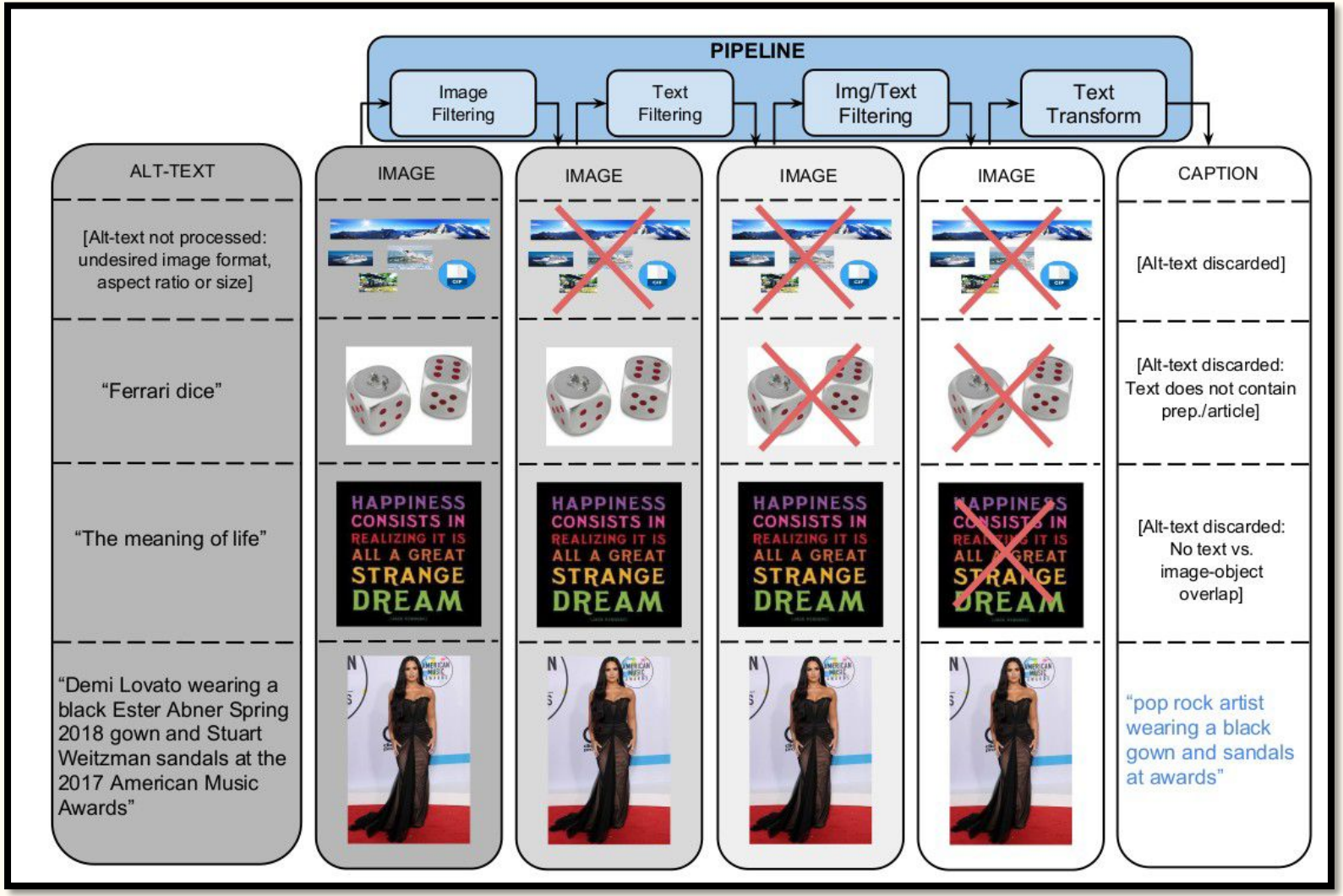
# 2.6 Module-1 : Dataset

- **Motivation** - A large amount of work has been done on image caption generation task. Most of the significant work in solving computer vision tasks involve following prominent datasets.

    1. MSCOCO Dataset
    2. Flickr30K Dataset
    3. Pinterest Image Dataset

# Contd..

- **Google’s Conceptual Caption Dataset**
  - Conceptual Captions, contains an order of magnitude more images than the MS-COCO dataset and represents a wider variety of both images and image caption styles.
  - An automatic pipeline extracts, filters, and transforms candidate <image, caption pairs>,with the goal of achieving a balance of cleanliness, informativeness, fluency, and learnability of the resulting captions. This pipeline is known as Flume pipeline which processes billions of web pages parallely.


PIPELINE
Image Filtering
Text Filtering
Img/Text Filtering
Text Transform
ALT-TEXT
IMAGE
IMAGE
IMAGE
IMAGE
CAPTION
[Alt-text not processed: undesired image format, aspect ratio or size]
[Alt-text discarded]
“Ferrari dice”
[Alt-text discarded: Text does not contain prep./article]
“The meaning of life”
HAPPINESS CONSISTS IN REALIZING IT IS ALL A GREAT STRANGE DREAM
[Alt-text discarded: No text vs. image-object overlap]
“Demi Lovato wearing a black Ester Abner Spring 2018 gown and Stuart Weitzman sandals at the 2017 American Music Awards”
“pop rock artist wearing a black gown and sandals at awards”

# Module-2 : ENCODING THE IMAGE

- The encoding part of the network that compresses the input into a latent-space representation. It can be represented by an encoding function h=f(x). To learn this method, we propose to study and evaluate the following two methods to learn the representation and find the best fit for the dataset.

    1. **Transfer Learning on ResNet50**
    2. **Using Autoencoders**

# Module-2 : ENCODING THE IMAGE (contd..)

- Neural Style Transfer (NST) uses a previously trained convolutional network, and builds on top of that.
- The implementation of ResNet50 model trained on more than million images from the ImageNet database was a major breakthrough as it allowed training deep neural networks.
- Two main types of blocks are used in a ResNet, depending mainly on whether the input/output dimensions are same or different.
    - The Identity Block
    - The Convolution Block

# MODULE 3: DECODER

- The target variable is the captions that our model is learning to predict. The output of the previous module (encoder) is fed as an input to the decoder RNN which would generate the sentences.

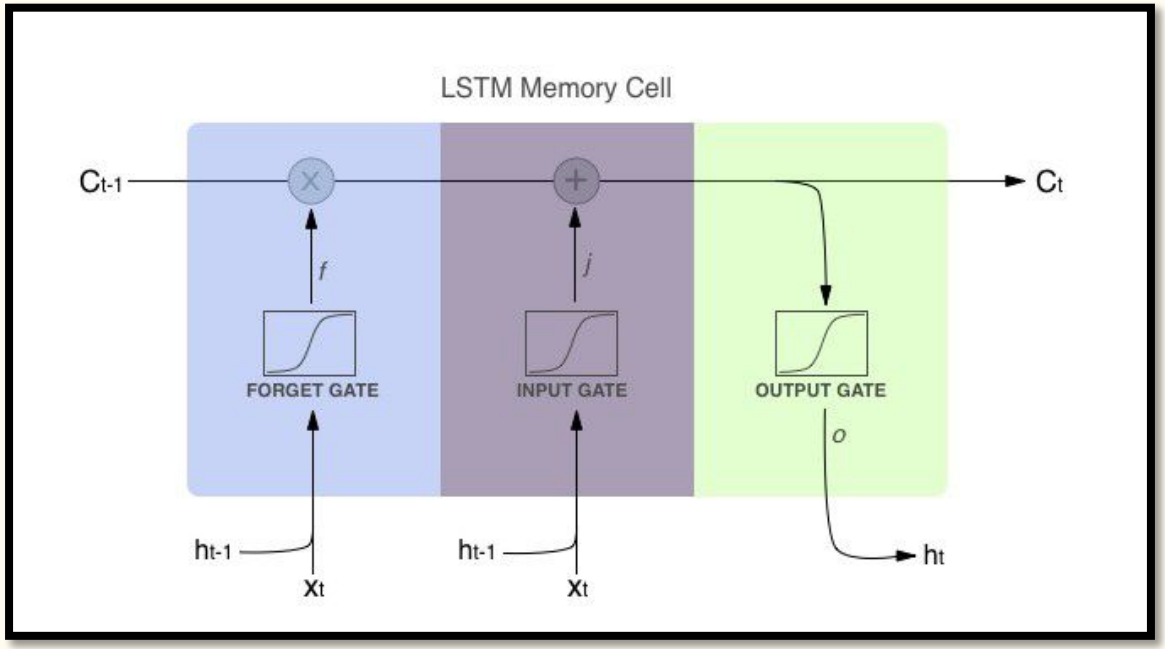
$$\theta^{\star} = \arg\max_{\theta} \sum_{(I,S)} \log\ p(S|I;\theta),$$

- it is common to applythe chain rule to model the joint probability over S0 ; . . . ; SN , where N is the length of this particular example as

$$\log p\ (S|I) = \sum_{t=0}^{N} \log\ p(S_t|I, S_0, \ldots, S_{t-1}),$$

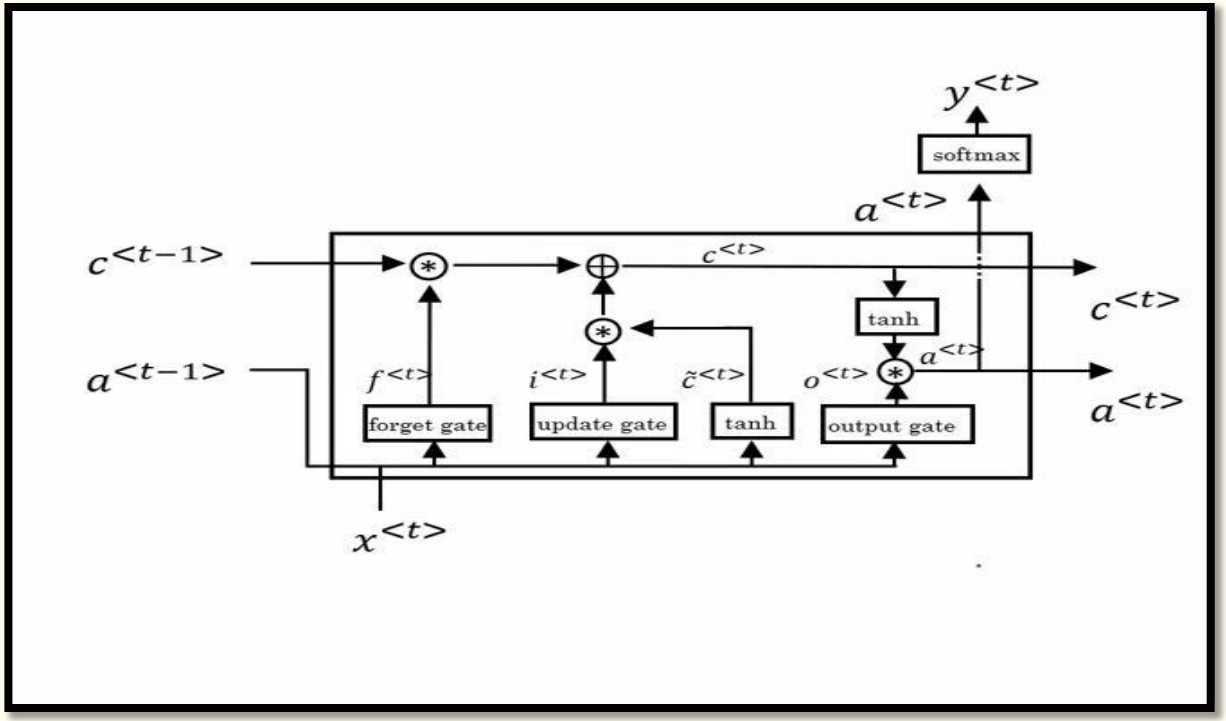
# MODULE 3: DECODER (contd ..)

- To model this probability we use Long Short Term Memory (LSTM) which is a special type of RNN.



# MODULE 3: DECODER (contd ..)

- Lastly, we filter the cell state and then it is passed through the activation function which predicts what portion should appear as the output of current LSTM unit at timestamp t.



# 2.7 References


- Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge --- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan
- Conceptual Captions: A Cleaned, Hypernym, Image Alt-text Dataset For Automatic Image Captioning --- Piyush Sharma, Nan Ding, Sebastian Goodman, Radu Soricut.
- Image Captioning Based on Deep Neural Networks --- Shuang Liu, Liang Bai, Yanli Hu and Haoran Wang
- Transforming Auto-encoders--- G. E. Hinton, A. Krizhevsky & S. D. Wang
- Farhadi A. et al. (2010) Every Picture Tells a Story: Generating Sentences from Images. In: Daniilidis K., Maragos P.,Paragios N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6314. Springer, Berlin, Heidelberg
- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention — Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio ; Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2048-2057, 2015

# 3.Planning for next semester
—

# Planning

- Deploy a single end to end network to develop more accurate feature extraction and efficiently generate textual description which can provide detailed information about the given image.
- To do a comparative analysis of Autoencoders and Transfer Learning on generate an accurate latent representation of an image.
- To weight the spatial locations according to their importance and implement Attention Mechanism for the achieved Image Captioning Model.

# Thank You