VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

PROJECT REPORT ON

IMAGE TO SPEECH USING ARTIFICIAL INTELLIGENCE

Submitted in partial fulfilment of the requirements for the following degree

B.Tech in

Computer Science and Engineering

And

CSE with Specialization in Data Science

BY

B Phaneendra Kumar Suryadevara - 20BCE0618

Manas Jati - 20BDS0223

Hrithik Ram J - 20BDS0226

Declaration

I hereby declare that the project entitled “IMAGE TO SPEECH USING ARTIFICIAL INTELLIGENCE " submitted by my team, for the award of the degree of *Bachelor of Technology in* ***CSE CORE, CSE With Specialization in Data Science*** to VIT is a record of bonafide work carried out by my team under the supervision of Mohana CM.

I further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date:

B Phaneendra Kumar Suryadevara-20BCE0618

Manas Jati -20BDS0223

Hrithik Ram J-20BDS0226

CERTIFICATE

This is to certify that the Project entitled “IMAGE TO SPEECH USING ARTIFICIAL INTELLIGENCE” submitted by **B Phaneendra Kumar Suryadevara-20BCE0618, Manas Jati -20BDS0223, Hrithik Ram J-20BDS0226** VIT, for the award of the degree of *Bachelor of Technology in CSE CORE, CSE With Specialization in Data Science*, is a record of bonafide work carried out by him / her under my supervision during the period, 24.07. 2023 to 11.11.2023, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place: Vellore

Signature of the Guide

Mohana CM

Index

S.no	Contents	Page No
1	Problem Statement	5
2	Introduction	5
3	Literature survey	6
4	Implementation	17
5	Algorithm	18
6	Complexity analysis	19
7	Important snippets	19
8	Output	21
9	Future work	22
10	References	23

Problem Statement:

Aid for the blind: we aim to create a product for the visually disabled which will help them with their daily routines. We can do this by first converting the scene into text and then the text to voice.

Introduction:

Motivation & Significance

- According to the recent census, it is said that there are 2.2 billion people suffer from visual disability all over the world, and require assistance for daily activities.
- We aim to provide a visual aid that improves daily performance, and independent living, thereby enhance the quality of life among these people.

Scope and Applications

- Our main objective is to make an ML model that will analyse the picture that the user captures on their screen and voice out the components in it.
- We will be creating a text description of captured image and then accurately convert the text into audio using Google Speech API

- The 2.2 million people that suffer from visual disability; we can say that the need for itsaid is high in demand and will increase as time goes by and as technology expands.

- Current technology allows applications to be efficiently distributed and run on mobile and handheld devices, even in cases where computational requirements are significant. Apps like; VoiceOver, Siri, Lookout, Be My Eyes, Blind Bargains etc have helped blind people do their daily activities.

Literature Survey:

Related Work So Far

Face recognition with Python and Open CV.

Image Caption generator using Computer Vision and Natural Language processing.

Image Captioning- A Deep Learning Approach.

Gaps Identified

- Accuracy of the models build was not spot on.

- Minor details in the image, like reflection of the person in the picture, sea and snow comparison etc have still not been trained to 100%.

- For some researches, the image and caption data set used is not really variant and so the accuracy of the model differs

- Even though many papers used different algorithms, like CNN& RNN, LSTM and YOLO in predicting the output description, the errors and accuracy seemed to be the same.

- We currently only have support for images which are uploaded by the user instead of a live description straight from the camera.

- Accuracy is not spot on as the dataset used is not large enough to identify various types of images.

- The time taken to train the data is huge.

- Acute details of the image is not defined by the Model.

- Model doesn't Support Face Recognition.

Drive To the Present Work

The will to make the world a better place for everyone is the drive that researchers have powered through. The need of technology is significant and vastly changing. The aim to be able to learn about new algorithms and implement them in real life and to be

able to find small details that might bring about a change in the world someday is what drove everyone into this field of science.

Research papers

Network Compression via Mixed Precision Quantization Using a Multi-Layer Perceptron forthe Bit-Width Allocation

Deep Neural Networks (DNNs) are a powerful tool for solving complex tasks in many application domains. The high performance of DNNs demands significant computational resources, which might not always be available. Network quantization with mixed-precision across the layers can alleviate this high demand. However, determining layer-wise optimal bit-widths is non-trivial, as the search space is exponential. This article proposes a novel technique for allocating layer-wise bit-widths for a DNN using a multi-layer perceptron (MLP). The Kullback-Leibler(KL) divergence of the SoftMax outputs between the quantized and full precision network is used as the metric to quantify the quantization quality. We explore the relationship between the KL-divergence and the network size, and from our experiments observe that more aggressive quantization leads to higher divergence, and vice versa. The MLP is trained with layer-wise bit-widths as labels and their corresponding KL- divergence as the input. The MLP training set, i.e. the pairs of the layer-wise bit-widths and their corresponding KL-divergence, is collected using a Monte Carlo sampling of the exponential search space. We introduce a penalty term in the loss to ensure that the MLP learns to predict bit-widths resulting in the smallest network size. We show that the layer-wise bit-width predictions from the trained MLP result in reduced network size without degrading

accuracy while achieving better or comparable results with SOTA work but with less computational overhead. Our method achieves up to 6x, 4x, 4x compression on VGG16, ResNet50, and Google Net respectively, with no accuracy drop compared to the original full precision pretrained model, on the ImageNet dataset.

Digital modulation classification using multi-layer perceptron and time-frequency features

Considering that real communication signals corrupted by noise are generally nonstationary, and time-frequency distributions are especially suitable for the analysis of nonstationary signals, time-frequency distributions are introduced for the modulation classification of communication signals. The extracted time-frequency features have good classification information, and they are insensitive to signal to noise ratio (SNR) variation. According to good classification by the correct rate of a neural network classifier, a multilayer perceptron (MLP) classifier with better generalization, as well as addition of time-frequency features set for classifying six different modulation types has been proposed. Computer simulations show that the MLP classifier outperforms the decision-theoretic classifier at low SNRs, and the classification experiments for real MPSK signals verify engineering significance of the MLP classifier.

A Reliable Localization Algorithm Based on Grid Coding and Multi-Layer Perceptron

The traditional RSS-based fingerprint localization algorithm needs RSS values from all access points (AP) at each reference point (RP). In the large-scale indoor environment, the increasing of the number of APs will lead to establish a large-scale fingerprint database, which occupies a lot of storage space. In this paper, we propose a new reliable localization algorithm, which firstly utilizes quantized RSS to encode the monitoring region which has been divided into grids, so as to specify the grids that the interested target appears roughly. Then, we utilize Multi-Layer Perceptron (MLP) to train the grid regions in which the beacons deployment is non-isomorphic and obtain the accurate localization result. Due to the same deployment of isomorphic regions, it is imperative to train only one model to replace the others, which greatly reduces the computation of neural network. It can be concluded from the experimental results that compared with the traditional MLP-based fingerprint localization algorithm, the proposed algorithm reduces the size of fingerprint database over 80% with guarantee of localization accuracy. Moreover, our algorithm can obtain better localization accuracy compared with the other latest quantization based

Convolution in Convolution for Network in Network

Network in network (NiN) is an effective instance and an important extension of deep convolutional neural network consisting of alternating convolutional layers and pooling layers. Instead of using a linear filter for convolution, NiN utilizes shallow multilayer perceptron (MLP), a nonlinear function, to replace the linear filter. Because of the powerfulness of MLP and

1 × 1 convolutions in spatial domain, NiN has stronger ability of feature representation and hence results in better recognition performance. However, MLP itself consists of fully connected layers that give rise to a large number of parameters. In this paper, we propose to replace dense shallow MLP with sparse shallow MLP. One or more layers of the sparse shallow MLP are sparely connected in the channel dimension or channel–spatial domain. The proposed method is implemented by applying unshared convolution across the channel dimension and applying shared convolution across the spatial dimension in some com-putational layers. The proposed method is called convolution in convolution (CiC). The experimental results on the CIFAR10 data set, augmented CIFAR10 data set, and CIFAR100 data set demonstrate the effectiveness of the proposed CiC method.

Efficient Convolution Neural Networks for Object Tracking Using Separable Convolution and Filter Pruning

Object tracking based on deep learning is a hot topic in computer vision with many applications. Due to high computation and memory costs, it is difficult to deploy convolutional neural networks (CNNs) for object tracking on embedded systems with limited hardware resources. This paper uses the Siamese network to construct the backbone of our tracker. The convolution layers used to extract features often have the highest costs, so more improvements should be focused on them to make the tracking more efficient. In this paper, the standard convolution is optimized by the separable convolution, which mainly includes a depth wise convolution and a pointwise convolution. To further reduce the calculation, filters in the depth wise convolution layer are pruned with filters variance. As

there are different weight distributions in convolution layers, the filter pruning is guided by a hyper-parameter designed. With the improvements, the number of parameters is decreased to 13% of the original network and the computation is reduced to 23%. On the NVIDIA Jetson TX2, the tracking speed increased to 3.65 times on the CPU and 2.08 times on the GPU, without significant degradation of tracking performance in VOT benchmark.

CSCC: Convolution Split Compression Calculation Algorithm for Deep Neural Network

Convolutional Neural Networks (CNNs) have become one of the most successful machine learning techniques for image and video processing. The most computationally intensive part of the CNN is the convolutional layers, which have the multi-channel image and multiple kernels. However, due to the network pruning operation and the application of RELU activation function operation in the training process, numerous zero values are generated in the network. This paper proposes the convolution split compression calculation (CSCC) algorithm, which improves the performance of the convolution layer by utilizing the sparse characteristic of the feature map. In the CSCC algorithm, first, the feature map is directly converted into a sparse matrix of compressed sparse row (CSR) format, which avoids expanding feature map to an intermediate matrix and reduces the memory space consumption. Second, the convolution kernel is converted into a vector. Finally, the convolution result is obtained by the sparse matrix vector multiplication (SpMV). The experimental results show that the CSCC algorithm has a good advantage in computation speed and memory consumption compared with the other convolution algorithms.

Multiscale Residual Convolution Neural Network and Sector Descriptor-Based Road Detection Method

T Road detection is a focus of research in the field of remote sensing image analysis. This task is normally difficult due to the complexity of the data, which are heterogeneous in appearance with large intra-class and lower inter-class variations that frequently lead to large numbers of gaps and errors in road extraction. In this paper, a novel road detection method is proposed that combines a multiscale deep residual convolution neural network (MDRCNN) with postprocessing. The MDRCNN is used to obtain road areas more accurately and quickly. Multiscale convolution, which provides greater accuracy, is used to acquire the hierarchical features of different dimensions. The residual connections and global average pooling are introduced to improve the efficiency of the network in the process of backpropagation and forward propagation, respectively. In the postprocessing stage, the centreline of the road can be obtained based on the road area. Geometric constraints and mathematical morphological refinement as well as leaf-to-leaf connection are used to obtain the road line. Rectangular buffer analysis and a sector descriptor tracking connection are subsequently used to improve the integrity and accuracy of the road. We experimented on two datasets with different resolutions and different scenes. Compared with other neural network methods, our method is better at connecting road fractures and eliminating errors.

A Comprehensive Review of Stability Analysis of Continuous-Time Recurrent Neural Networks

Stability problems of continuous-time recurrent neural networks have been extensively studied, and many papers have been published in the literature. The purpose of this paper is to provide a comprehensive review of the research on stability of continuous-time recurrent neural networks, including Hopfield neural networks, Cohen–Grossberg neural networks, and related models. Since time delay is inevitable in practice, stability results of recurrent neural networks with different classes of time delays are reviewed in detail. For the case of delay-dependent stability, the results on how to deal with the constant/variable delay in recurrent neural networks are summarized. The relationship among stability results in different forms, such as algebraic inequality forms, M-matrix forms, linear matrix inequality forms, and Lyapunov diagonal stability forms, is discussed and compared. Some necessary and sufficient stability conditions for recurrent neural networks without time delays are also discussed. Concluding remarks and future directions of stability analysis of recurrent neural networks are given.

Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things

A network traffic classifier (NTC) is an important part of current network monitoring systems, being its task to infer the network service that is currently used by a communication flow (e.g., HTTP and SIP). The detection is based on a number of features associated with the communication flow, for example, source and destination ports and bytes transmitted per packet. NTC is important, because much information about a current network flow can be learned and anticipated just by knowing its network service (required latency, traffic volume, and possible duration).

This is of particular interest for the management and monitoring of Internet of Things (IoT) networks, where NTC will help to segregate traffic and behaviour of heterogeneous devices and services. In this paper, we present a new technique for NTC based on a combination of deep learning models that can be used for IoT traffic. We show that a recurrent neural network (RNN) combined with a convolutional neural network (CNN) provides best detection results. The natural domain for a CNN, which is image processing, has been extended to NTC in an easy and natural way. We show that the proposed method provides better detection results than alternative algorithms without requiring any feature engineering, which is usual when applying other models. A complete study is presented on several architectures that integrate a CNN and an RNN, including the impact of the features chosen and the length of the network flows used for training.

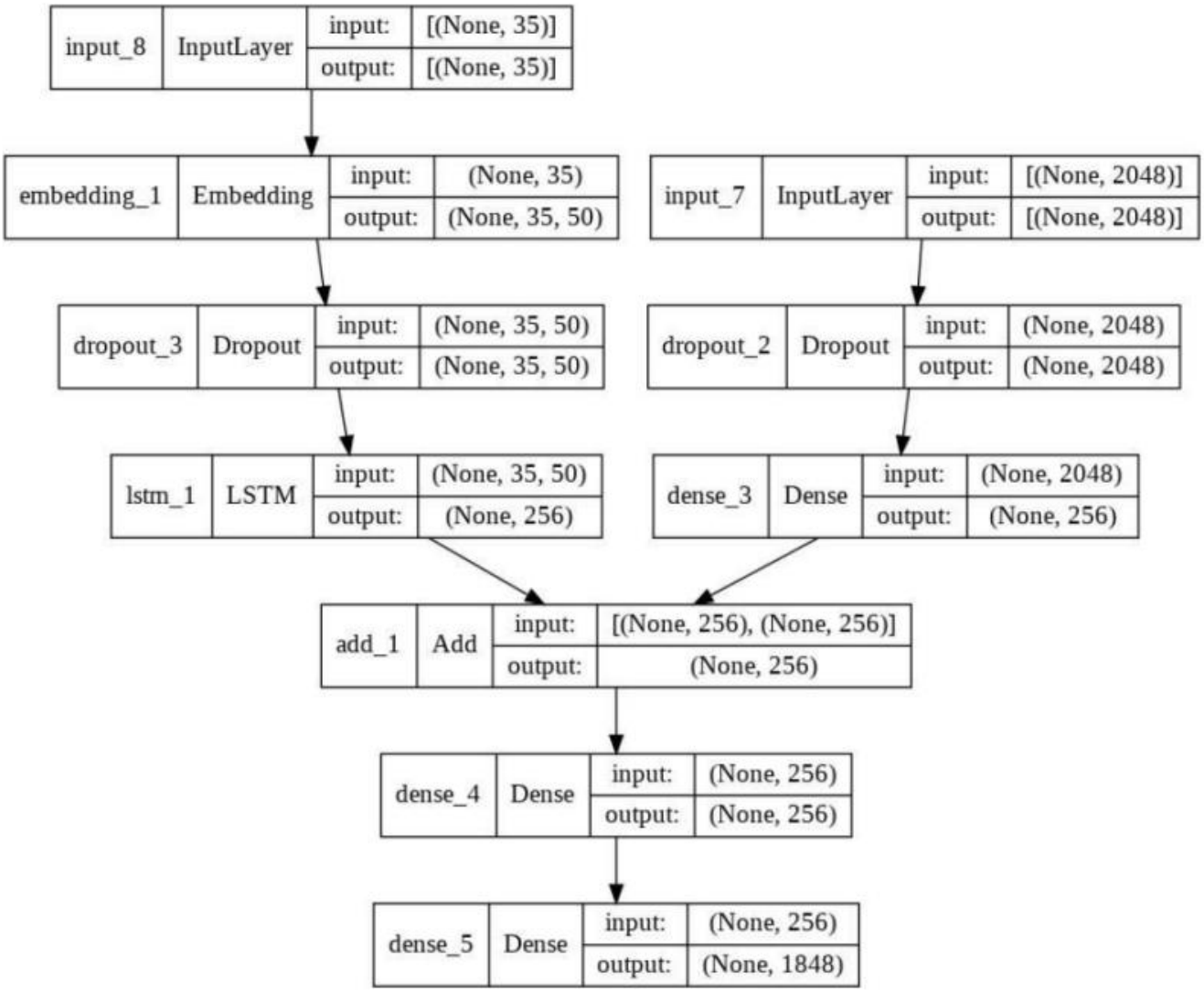
Cardiac Phase Detection in Echocardiograms With Densely Gated Recurrent Neural Networks and Global Extrema Loss

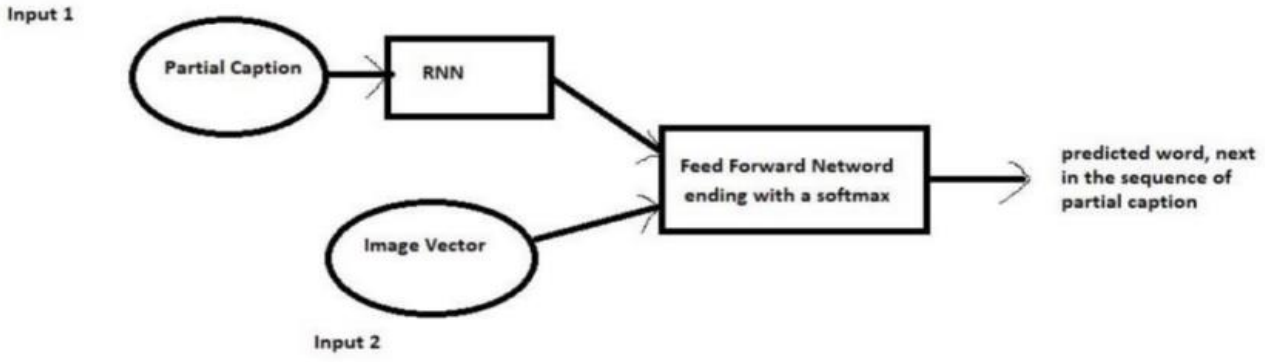
Accurate detection of end-systolic (ES) and end-diastolic (ED) frames in an echocardiographic cine series can be difficult but necessary pre-processing step for the development of automatic systems to measure cardiac parameters. The detection task is challenging due to variations in cardiac anatomy and heart rate often associated with pathological conditions. We formulate this problem as a regression problem and propose several deep learning-based architectures that minimize a novel global extrema structured loss function to localize the ED and ES frames. The proposed architectures integrate convolution neural networks (CNNs)-based image feature extraction model and recurrent neural networks (RNNs) to model temporal

dependencies between each frame in a sequence. We explore two CNN architectures: DenseNet and ResNet, and four RNN architectures: long short-term memory, bi-directional LSTM, gated recurrent unit (GRU), and Bi-GRU, and compare the performance of these models. The optimal deep learning model consists of a DenseNet and GRU trained with the proposed loss function. On average, we achieved 0.20 and 1.43 frame mismatch for the ED and ES frames, respectively, which are within reported inter-observer variability for the manual detection of these frames.

Implementation:

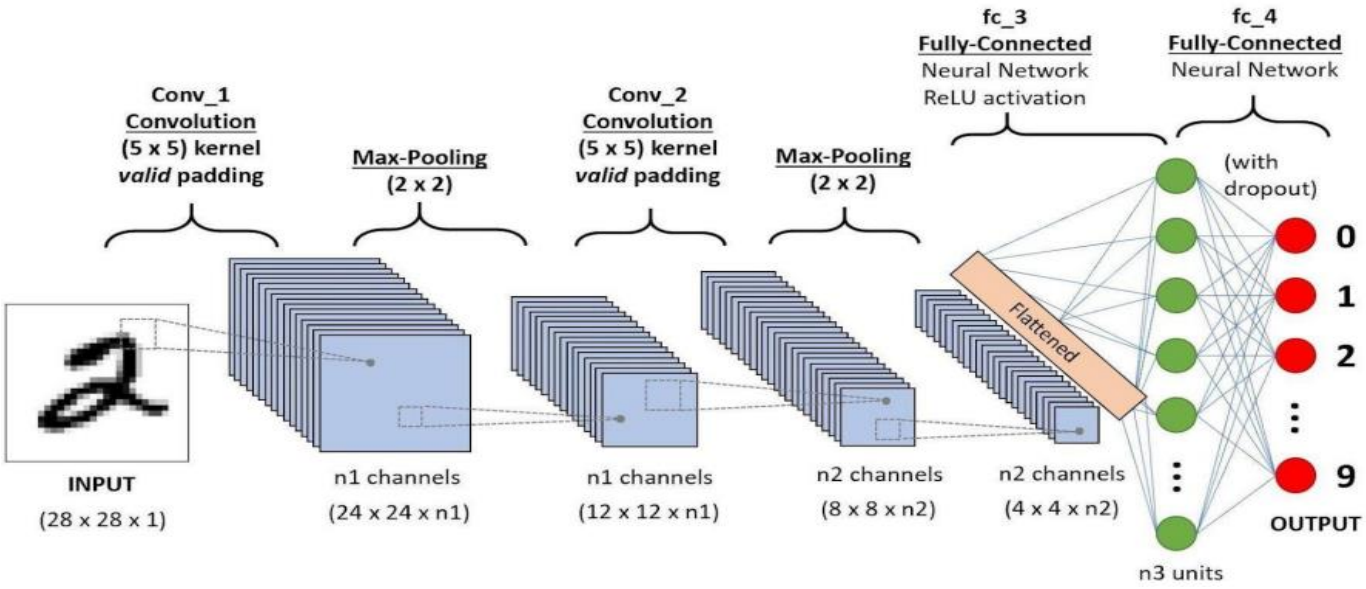
Architecture/Flowchart





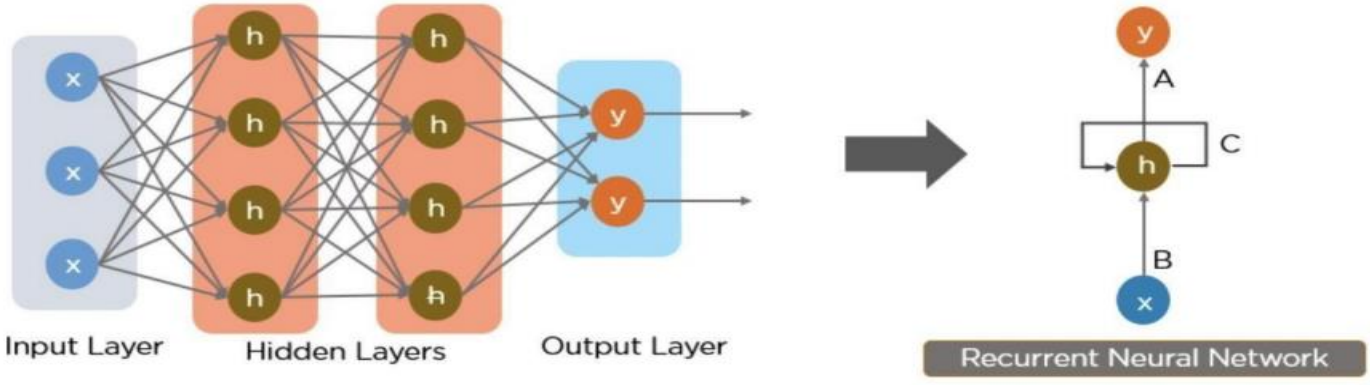
Algorithm:

CNN-Convolutional Neural Network



The layers in this stage helps in analysing the input images via a ResNet50 Model, that categorises images via gradients, edges and colour in the starting layers and the further layers train to understand shapes like wheels and people.

RNN-Recurrent Neural Network



The layers in this stage helps in analysing the input images via a ResNet50 Model, that categorises images via gradients, edges and colour in the starting layers and the further layers train to understand shapes like wheels and people.

Complexity Analysis

CNN and RNN:

Model	FLOPs	PN (million)	Depth	Stream #
AlexNet	$\mathbf{7.25 \times 10^8}$	58.3	7	1
VGG16	1.55×10^{10}	134.2	16	1
ResNet50	3.80×10^9	23.5	**50**	1
GoogLeNet	1.57×10^9	**6.0**	22	1
JLML-ResNet39	1.54×10^9	7.2	39	**5**

(V) Comparisons of Model Size and Complexity. We compared the proposed JLML-ResNet39 model with four seminal classification CNN architectures (Alexnet [Krizhevsky

	Ops	Activations
Attention (dot-prod)	$n^2 \cdot d$	$n^2 + n \cdot d$
Attention (additive)	$n^2 \cdot d$	$n^2 \cdot d$
Recurrent	$n \cdot d^2$	$n \cdot d$
Convolutional	$n \cdot d^2$	$n \cdot d$
Multi-Head Attention with linear transformations. For each of the h heads, $d_q = d_k = d_v = d/h$	$n^2 \cdot d + n \cdot d^2$	$n^2 \cdot h + n \cdot d$
Recurrent	$n \cdot d^2$	$n \cdot d$
Convolutional	$n \cdot d^2$	$n \cdot d$

n = sequence length d = depth k = kernel size

Program :

Training of Model

```
epochs = 20
batch_size = 3
steps = len(train_descriptions)#number_pics_per_batch
```

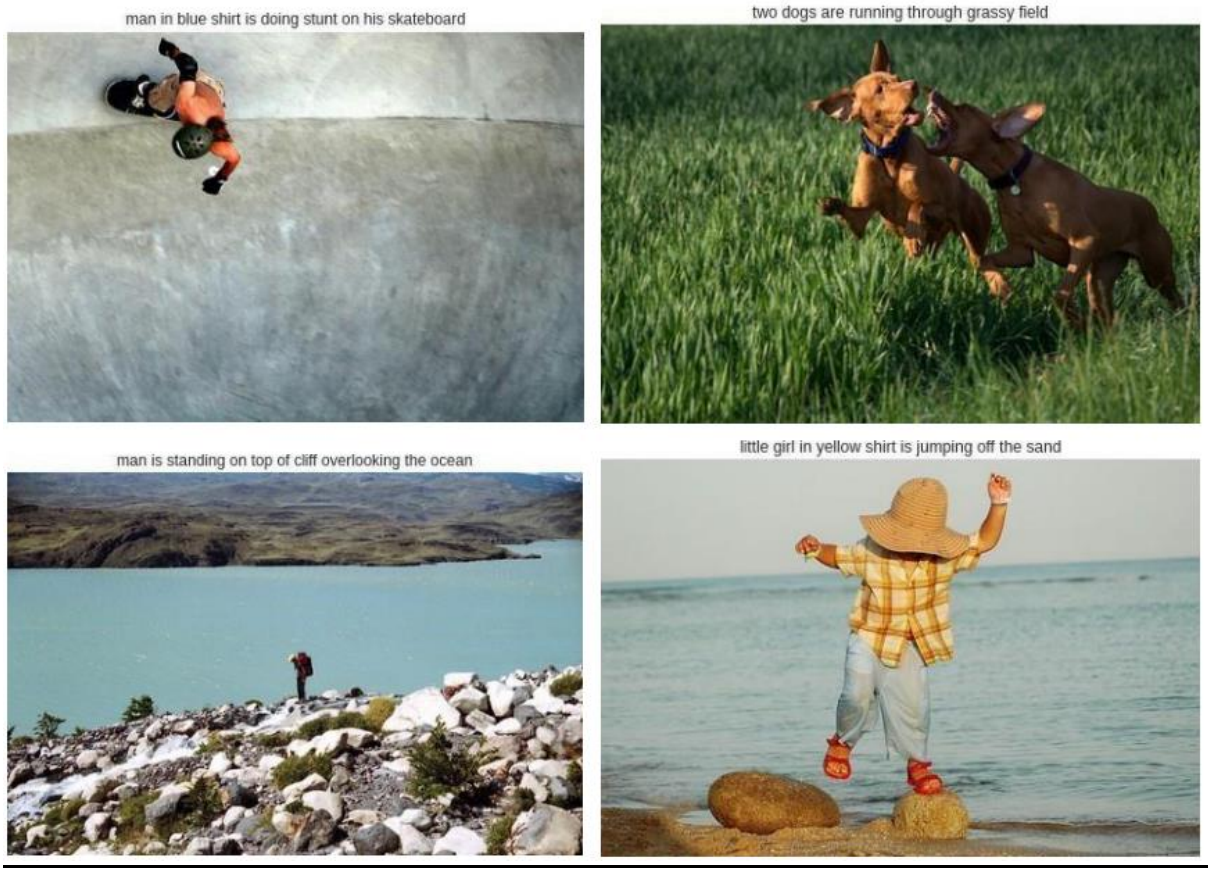
```
def train():

    for i in range(epochs):
        generator = data_generator(train_descriptions,encoding_train,word_to_idx,max_len,batch_size)
        model.fit_generator(generator,epochs=1,steps_per_epoch=steps,verbose=1)
        tf.keras.models.save_model(to_file='/content/model_'+str(i)+'.h5')
        model.compile(optimizer='sgd',
            loss='mse',
              metrics=[tf.keras.metrics.Accuracy()])
```

```
model = load_model('/content/model_9.h5')
```

```
def predict_caption(photo):

in_text = "startseq"

for i in range(max_len):

sequence = [word_to_idx[w] for w in in_text.split() if w in

word_to_idx]sequence =
pad_sequences([sequence],maxlen=max_len,padding='post')

ypred = model.predict([photo,sequence])

ypred = ypred.argmax() #Word with max prob always - Greedy

Samplingword = idx_to_word[ypred]

in_text += (' ' + word)

if word == "endseq":break

final_caption = in_text.split()[1:-1]

final_caption = ' '.join(final_caption)return

final_caption

# Pick Some Random Images and See

Resultsplt.style.use("seaborn")

for i in range(15):

    idx = np.random.randint(0,1000)

    all_img_names = list(encoding_test.keys())img_name =
    all_img_names[idx] photo_2048 =
    encoding_test[img_name].reshape((1,2048))

    i = plt.imread("/content/Images/"+img_name+".jpg")caption =
    predict_caption(photo_2048)

#print(caption)

plt.title(caption)

plt.imshow(i)

plt.axis("off")
```

```
plt.show()
text_to_say = caption
language = "en"
gtts_object = gTTS(text = text_to_say,lang = language,slow = False)
gtts_object.save("/content/test.wav")
from IPython.display import Audio
Audio("/content/test.wav")
```

Output:



Future Work:

When we improve the program the accuracy and efficiency of the model will increase, and we can auto play the audio and let the user click a picture via their device and even connect their contacts to their families for face recognition and other AI features. This can be used for other social media accounts where the user doesn't need to click a picture, x`but the AI just does an immersive reader not only of the text but also of the pictures that are displayed on the screen. Improvements can be made as much as possible for the model and testing and training using large datasets to improve accuracy can be done. When spoken about accuracy of the model, there are many minute details to be taken care of and when the image dataset and vocabulary expands such details can be noticed and can be rectified. A lot of modifications can be made to improve this solution like: a lot of modifications can be made to improve this solution like:

- Using a larger dataset
- Changing the model architecture, e.g. include an attention module.
- Doing more hyper parameter tuning (learning rate, batch size, number of layers, number of units, dropout rate, batch normalization etc.).
- Use the cross validation set to understand overfitting.
- Using Beam Search instead of Greedy Search during Inference.
- Using BLEU Score to evaluate and measure the performance of the model.

Reference:

1. James C. Bezdek, James Keller, Raghu Krisnapuram, Nikhil Pal (1990) 'Fuzzy Models and Algorithms for Pattern Recognition and Image Processing', The handbooks of fuzzy sets series, 455-457
2. Roberto Manduchi, Sri Kurniawan (2013) 'Assistive Technology for Blindness and LowVision' (1) 23-48
3. Andrej Karpathy, Li Fei-Fei () 'Deep Visual-Semantic Alignments for Generating ImageDescription' [online] 5-7
4. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning toalign and International Journal of Advanced Science and Technology Vol. 29, No. 3s, (2020), pp. 975
5. K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares,H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machinetranslation.In EMNLP, 2014.
6. D. Elliott and F. Keller. Image description using visual dependency representations. InEMNLP, 2013.
7. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young,C. Rashtchian, J. Hockenmaier, and D.Forsyth. Every picture tells a story: Generating sentences from images. In ECCV, 2010.
8. R. Gerber and H. Nagel. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. In ICIP. IEEE, 1996.