LIP READING FOR VISUAL DATA BY USING DEEP LEARNING METHODS

TEAM 4

HIMANSHEE

RICHA SHARMA

MOUNICA AYALASOMAYAJULA

MYTHRI MUCHAMARI

BACKGROUND AND PROBLEM STATEMENT

- A recent study (Easton & Basala, 1982) found that people generally have poor lip reading skills. Even on the limited subset of 30 monosyllabic words, the deaf person could only achieve 17±12% accuracy, and on 30 compound words she could only achieve 21±11% accuracy.

- It is important to automate the lip reading. Machine lip readers are practically large with applications such as enhanced hearing aids, silent dictation in public spaces, security, speech recognition in noisy environments, silent film processing and subtitling of silent films and videos etc.

- Purpose of project is to generate Text/audio for the videos where the audio is not understandable or clear, or silent videos.

LITERATURE REVIEW

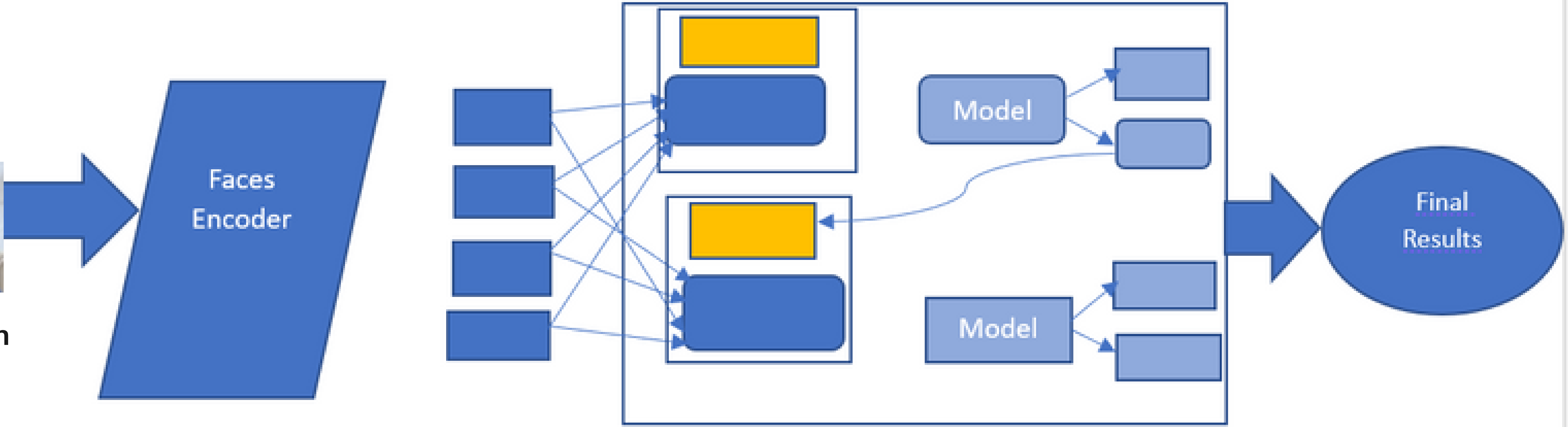
S. No	Title of the Papers (arXiv Papers)	Authors of the Paper	Summary of the Paper	Challenges and Results
1.	Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis	K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C V Jawahar	In this paper authors have used Lip2Wav models to and specifically solved the problem by focusing on individual speakers for video data. They evaluated their model with extensive quantitative metrics and human studies.	1) Human evaluation results. A) Mispronunciations with 21.5% accuracy (B) Word skips with 8.6% and (C) Homophene-based errors in the test samples with 49.9% accuracy. 2) The accuracy of their Lip2Wav model of Homophene is not that good.
2.	AuthNet: A Deep Learning based Authentication Mechanism using Temporal Facial Feature Movements	Mohit Raghavendra, Pravan Omprakash, Mukesh B R, Sowmya Kamath	This paper presents a new authentication mechanism that combines facial recognition with the unique movements of a person's face while uttering a password, known as temporal facial feature movements. This approach aims to enhance security and overcome language barriers, as users can set a password in any language.	The model achieves an accuracy of 98.1% on the MIRACL-VC1 dataset and demonstrates data efficiency by delivering good results with limited training data.
3.	Deep Learning for Visual Speech Analysis: A Survey	Changchong Sheng, Gangyao Kuang, Liang Bai, Chenping Hou, Yulan Guo, Xin Xu, Matti Pietikainen, and Li Liu	The writers discuss a range of topics related to visual speech analysis, such as speaker identification, lip reading, emotion recognition, and speech recognition. In order to extract useful information from visual speech data, the survey investigates the application of deep learning models including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their derivatives.	Different accents, speaking styles, and camera angles can all affect the lip movements that are captured during visual speech analysis. The issues these fluctuations and noise present in practical settings are covered in this study. It is challenging to successfully integrate audio and visual modalities in deep learning models for visual speech interpretation. The authors go over numerous fusion strategies and the difficulties that come with multimodal fusion.

PROJECT FLOW

Video

Mouth Extraction

Sequence of Faces



1.Video

2 - Mouth Extraction

3 - Feature Extraction

4 -Prepared the
Data for the Models

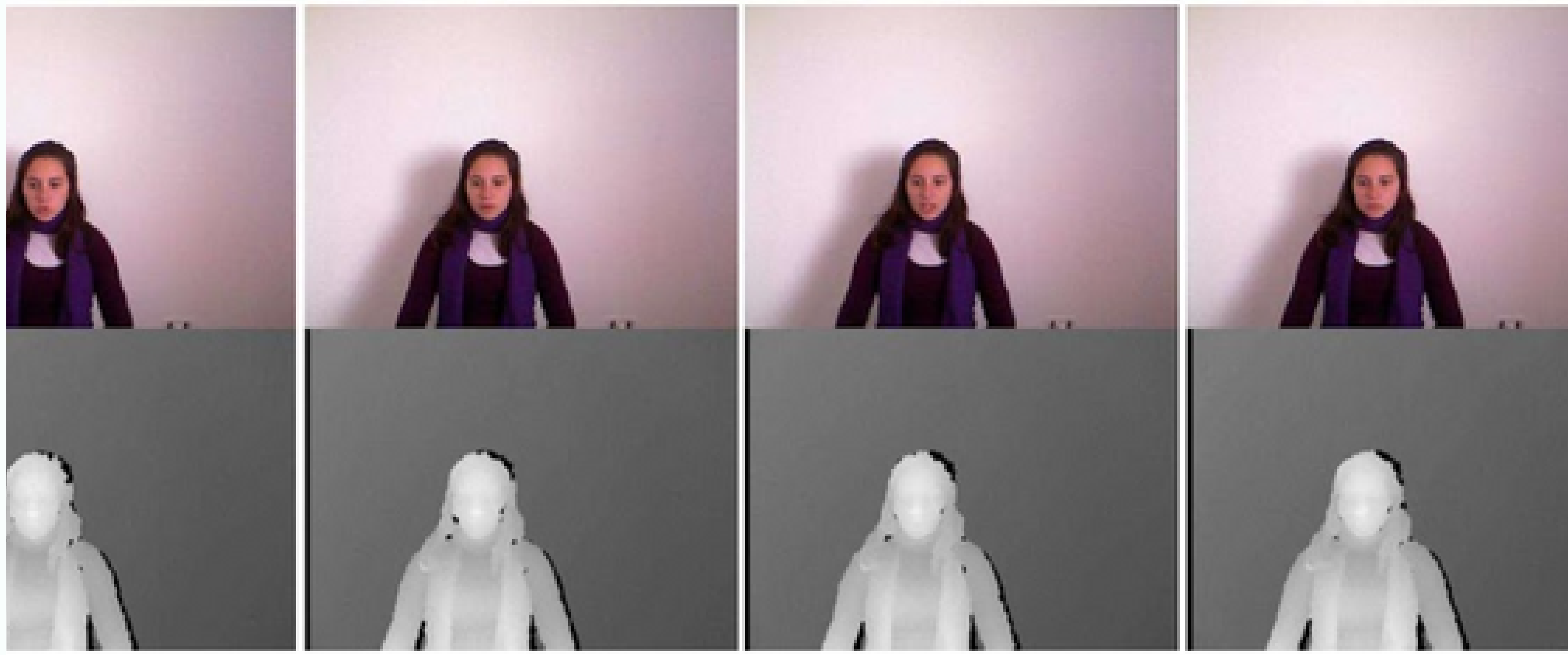
5 -Final Results
Generated Text

DATA COLLECTION

Dataset Resources

We have collected publicly available free dataset MIRACAL-VC1 https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1

MIRACL-VC1



Understanding Dataset

We tried to understand the structure of the dataset such as, the format of data, size of the data, features, dimensions of data, etc.

Dataset licenceing and policy issue

We used github data which was free and publically available.

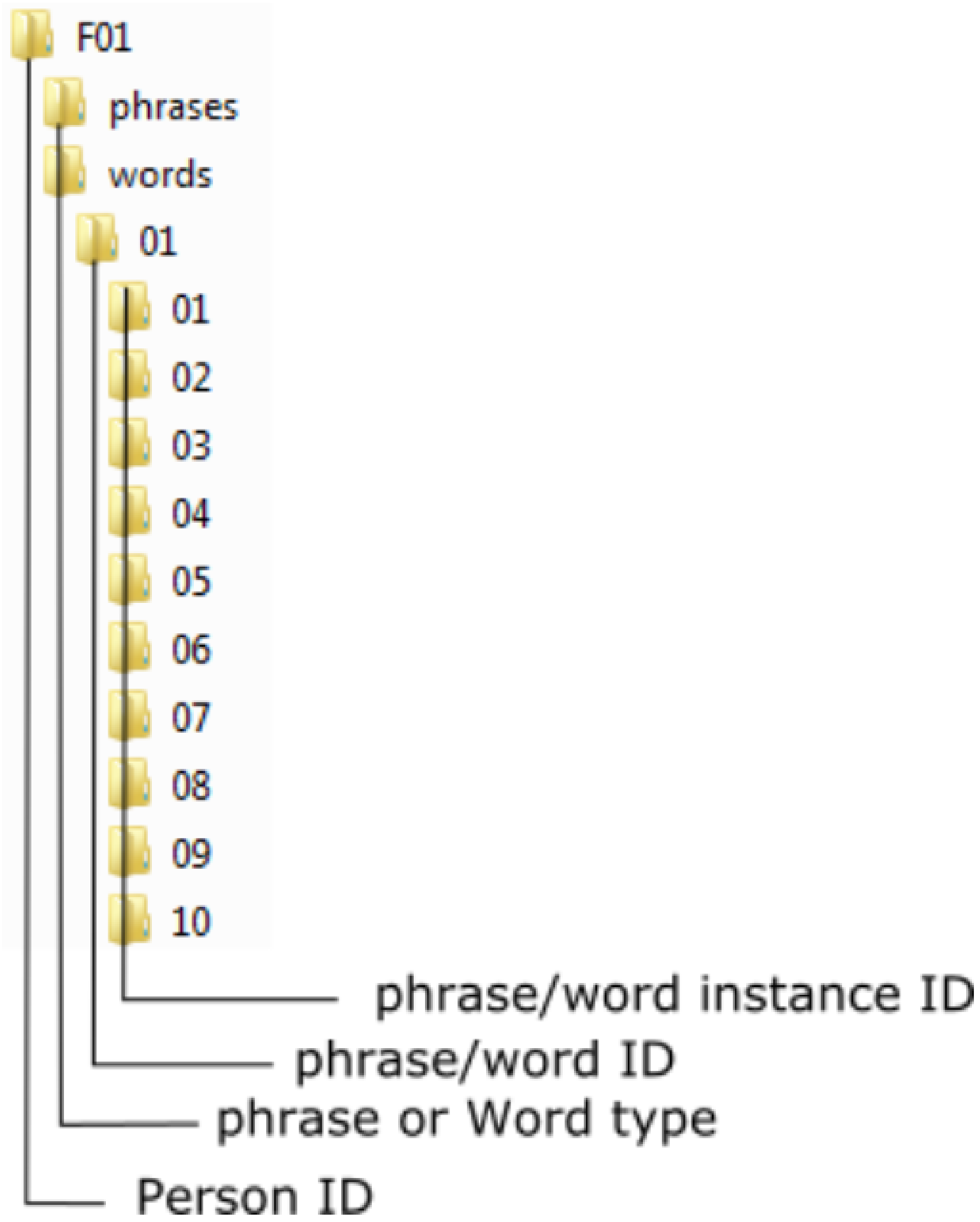
Dataset Structure

1)MIRACAL dataset has words and phases.
2)Dataset has information's about 15 speakers.

DATASET ARCHITECTURE

Folder architecture of the dataset is explained as follov

- And finally ,screenshot shows final architecture of our dataset.
- 15 folders have been created for all the 15 speakers.
- Under each speaker there are two separate folders phrases and words.
- Phrases and words components format shown in the figure.

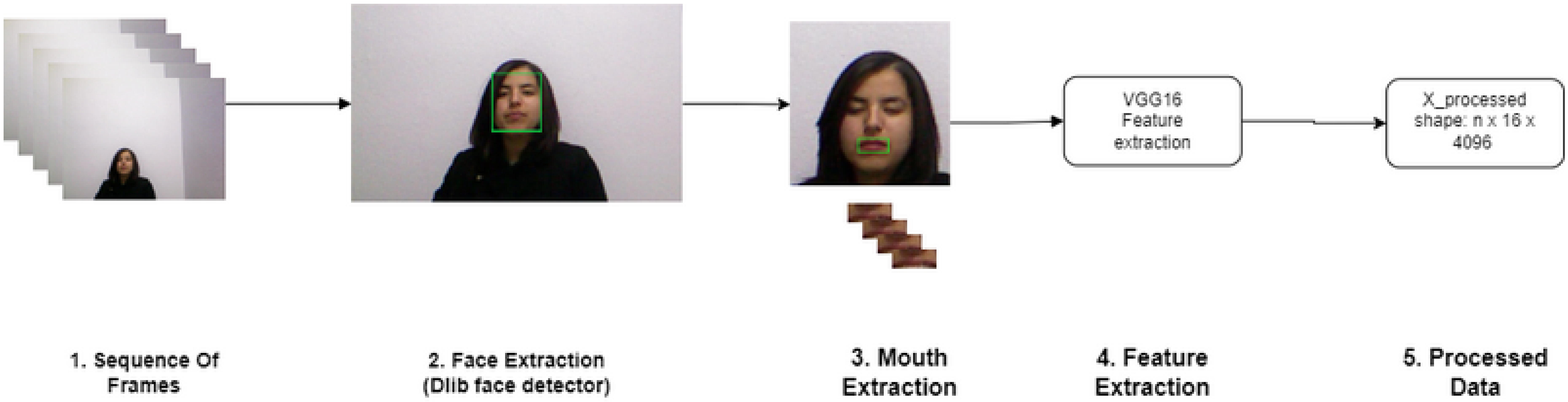


DATASET CHARACTERISTICS

- In our project, we make use of the MIRACL-VC1 dataset.
- The dataset was produced using the speech of 15 individuals, who each spoke ten words and ten sentences ten times, for a total of 3000 instances (15 x 20 x 10). Each instance consists of a series of 640 x 480 pixel color and depth images.
- To comply with the pre-trained VGGNet model, we just use the colored portion of the image and ignore the depth component.
- The words and phrases in the dataset are shown in figure.

Words	**Phrases**
Begin	Stop navigation.
Choose	Excuse me.
Connection	I am sorry.
Navigation	Thank you.
Next	Good bye.
Previous	I love this game.
Start	Nice to meet you.
Stop	You are welcome.
Hello	How are you?
Web	Have a good time.

DATA PRE-PROCESSING & PREPARATION



Face Extraction

Using MMOD+CNN to extract the Face

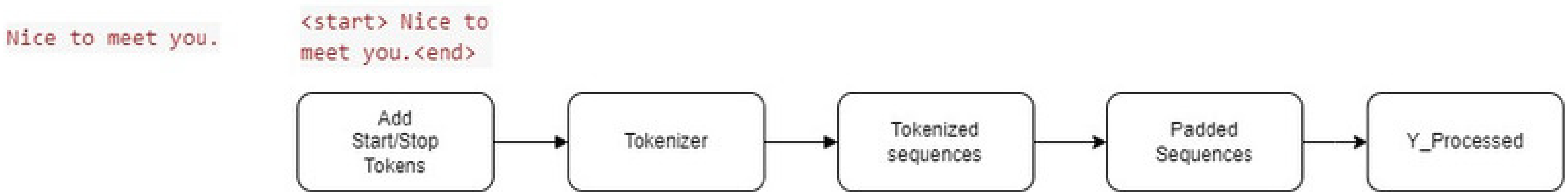
Mouth Extraction

Detected the mouth region from the face using Dlib shape predictor and then cropped that part and created image sequences.

Image Feature Extraction using VGG16

Used pretrained VGG16 model to extract features from these image sequences.

DATA PRE-PROCESSING & PREPARATION



Data preparation before Modeling

Prepared the extracted features as input features and the target text as input for the encoders and decoders.

DATA AUGUMENTATION

- Our dataset contains a total of 3000 instances only.
- Particularly for applications requiring deep learning, our dataset is limited.
- We employed data augmentation to fictitiously enhance the data size in order to address this issue.
- The original image has been changed in the following two ways as part of our data augmentation:

i) When cropping, nudged the crop area randomly in both the horizontal and vertical directions.

ii) By fluctuating the image's brightness or contrast at random.

MODEL IMPLEMENTATION

1.LSTM Model

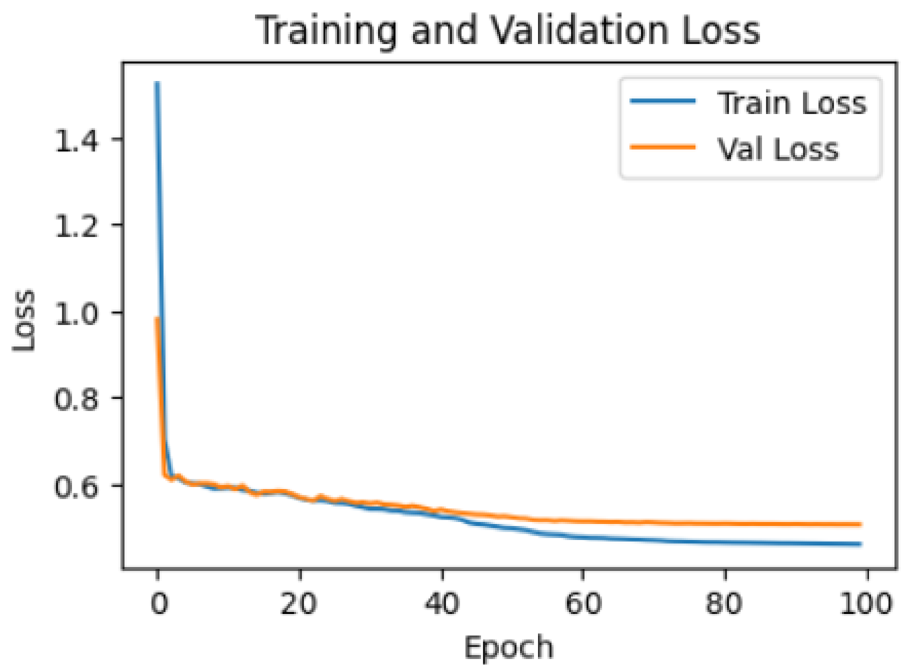
2.Stacked GRU Model

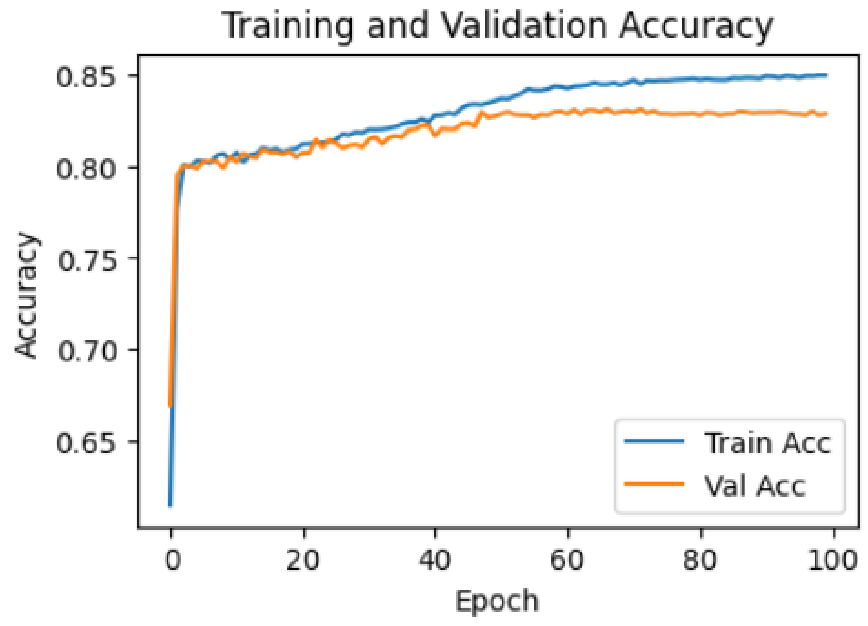
3.RNN+Attention Model

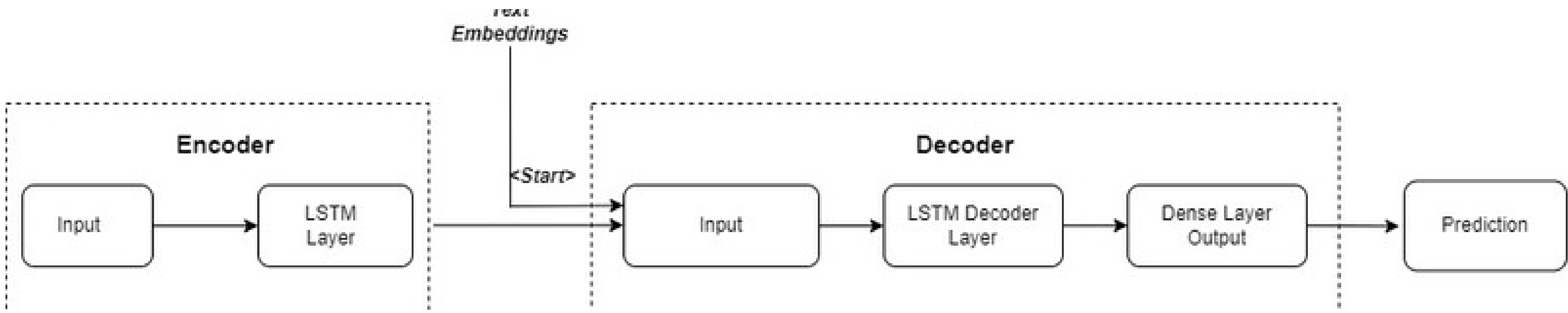
4.Multi-Head SelfAttention+FC Model

LSTM MODEL (ARCHITECTURE DIAGRAM +TABLE)

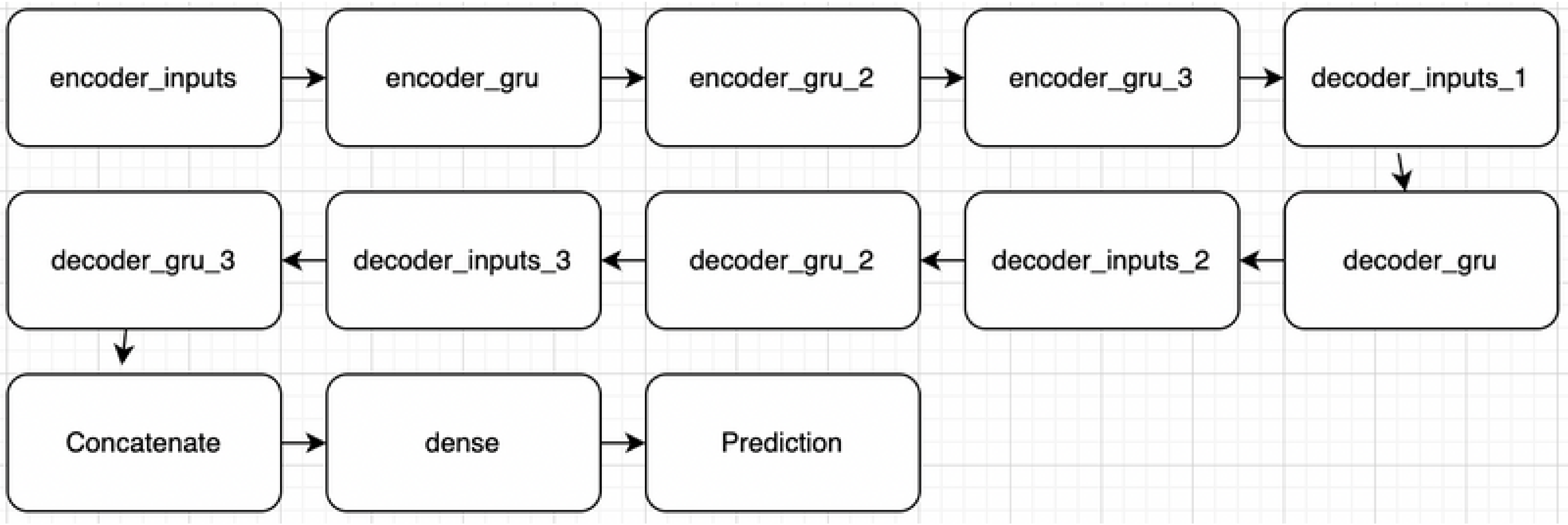
Accuracy	**84.99 %**
Loss	0.461
No. of learnable Paramters	10,617,906
Epoches	100
Learning Rate	0.01
Optimizer	adam
Btach size	64
Validation slpit	20%

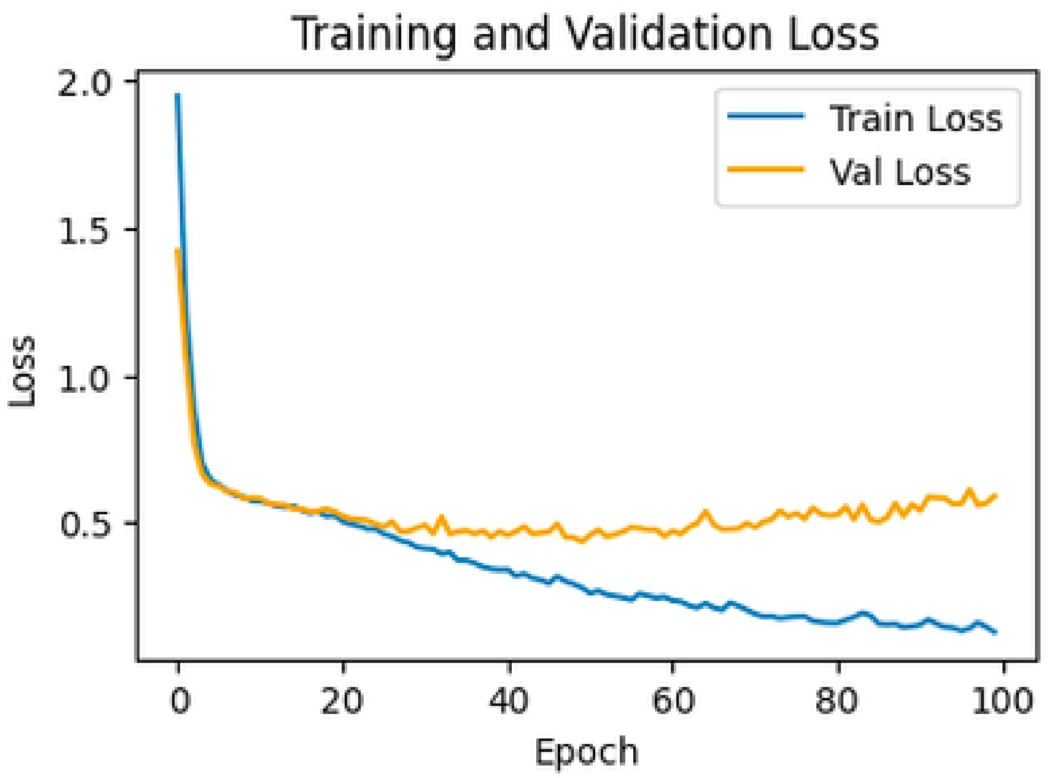


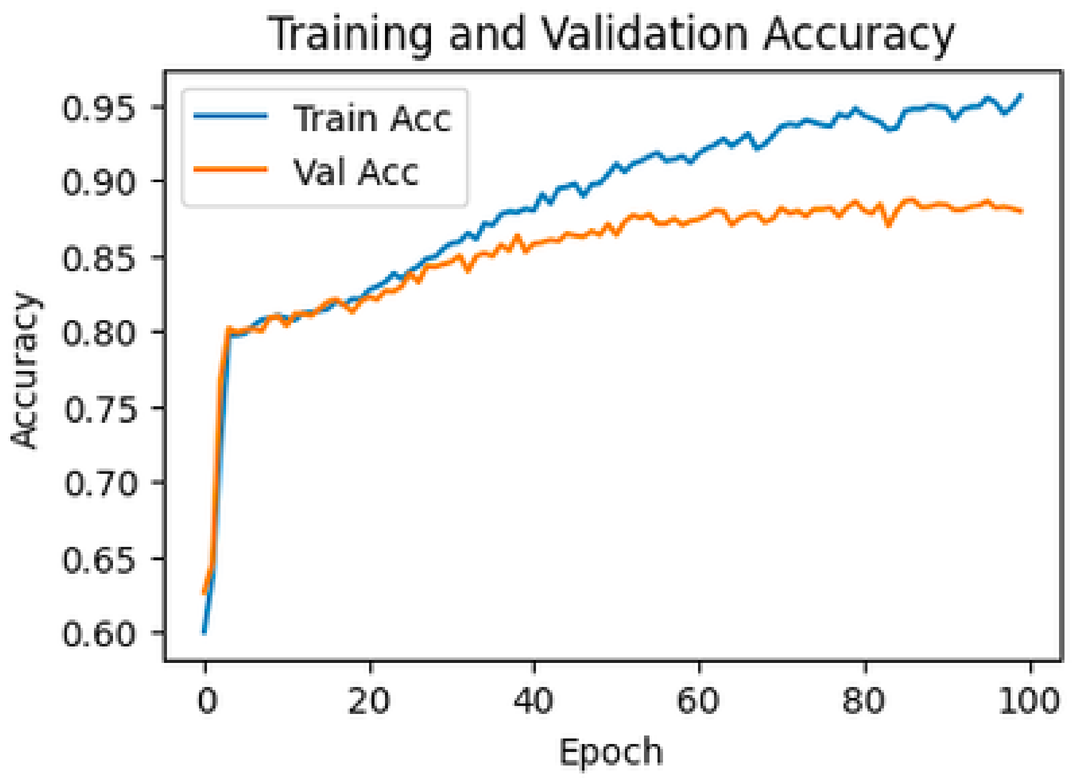




STACKED GRU MODEL (ARCHITECTURE DIAGRAM +TABLE)

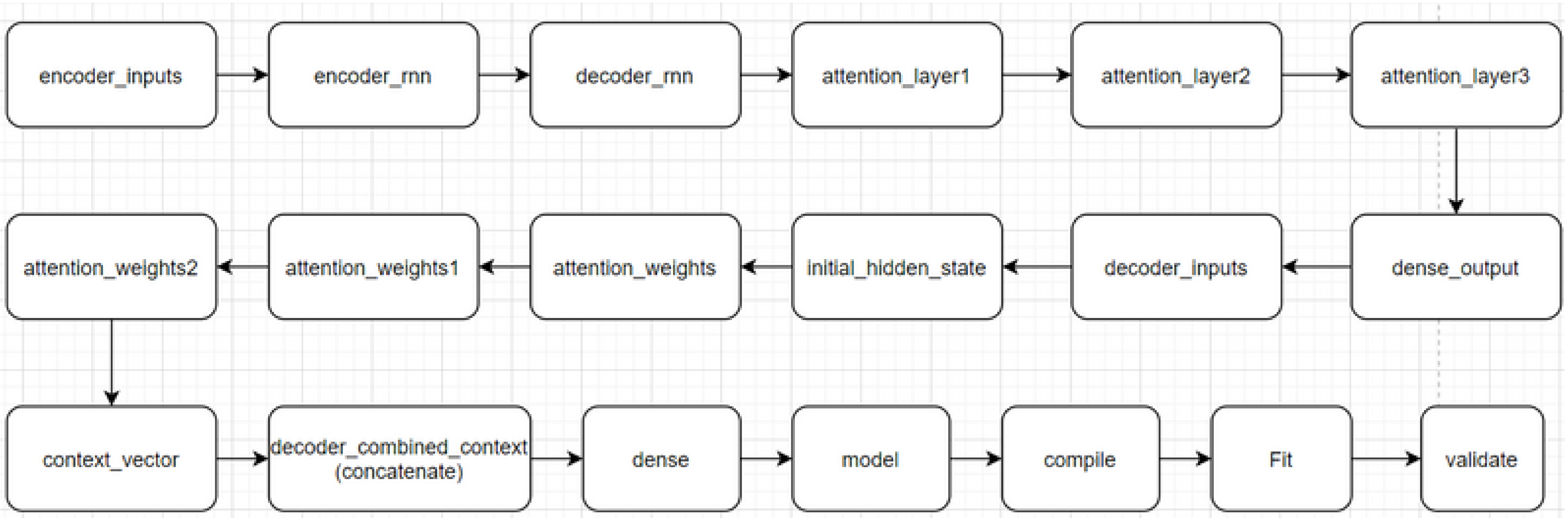


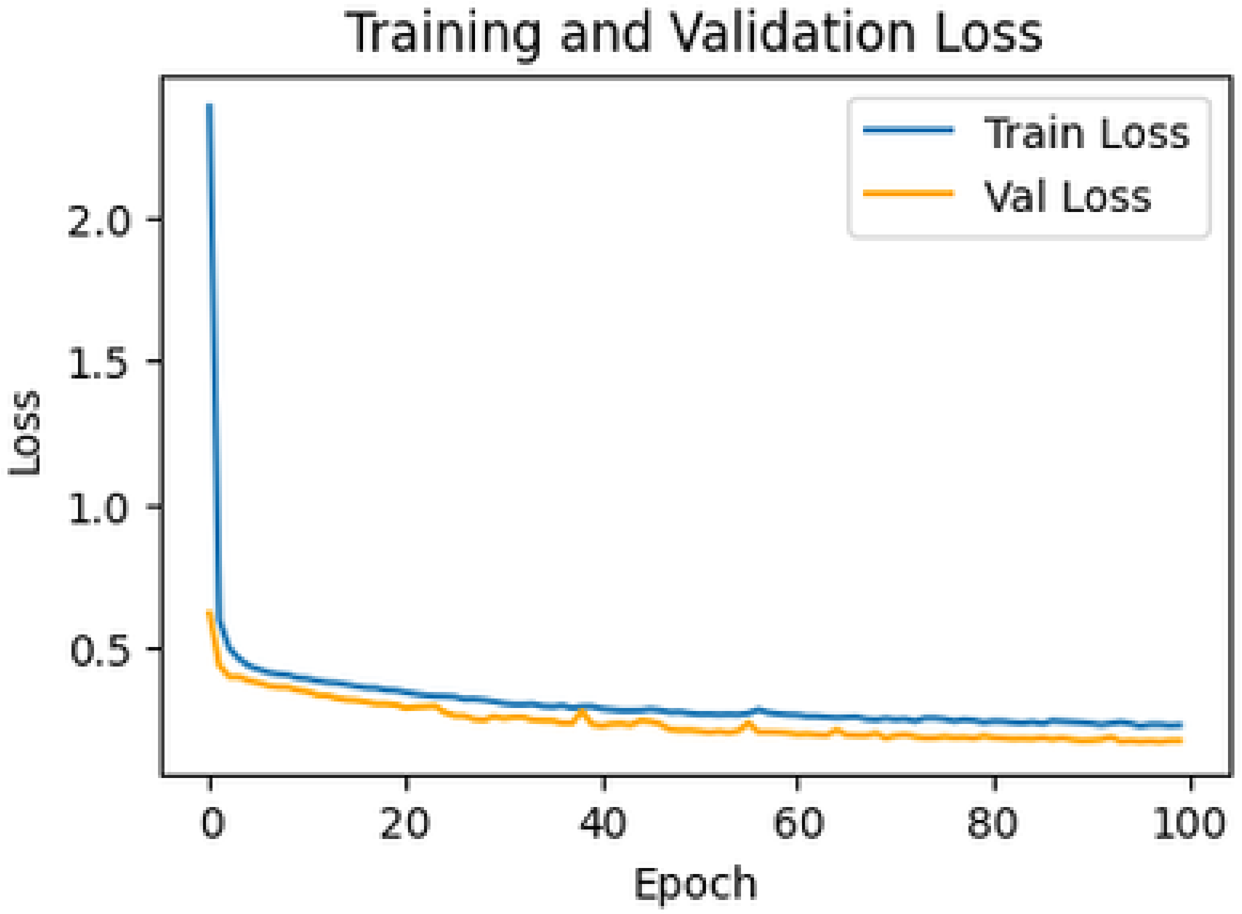


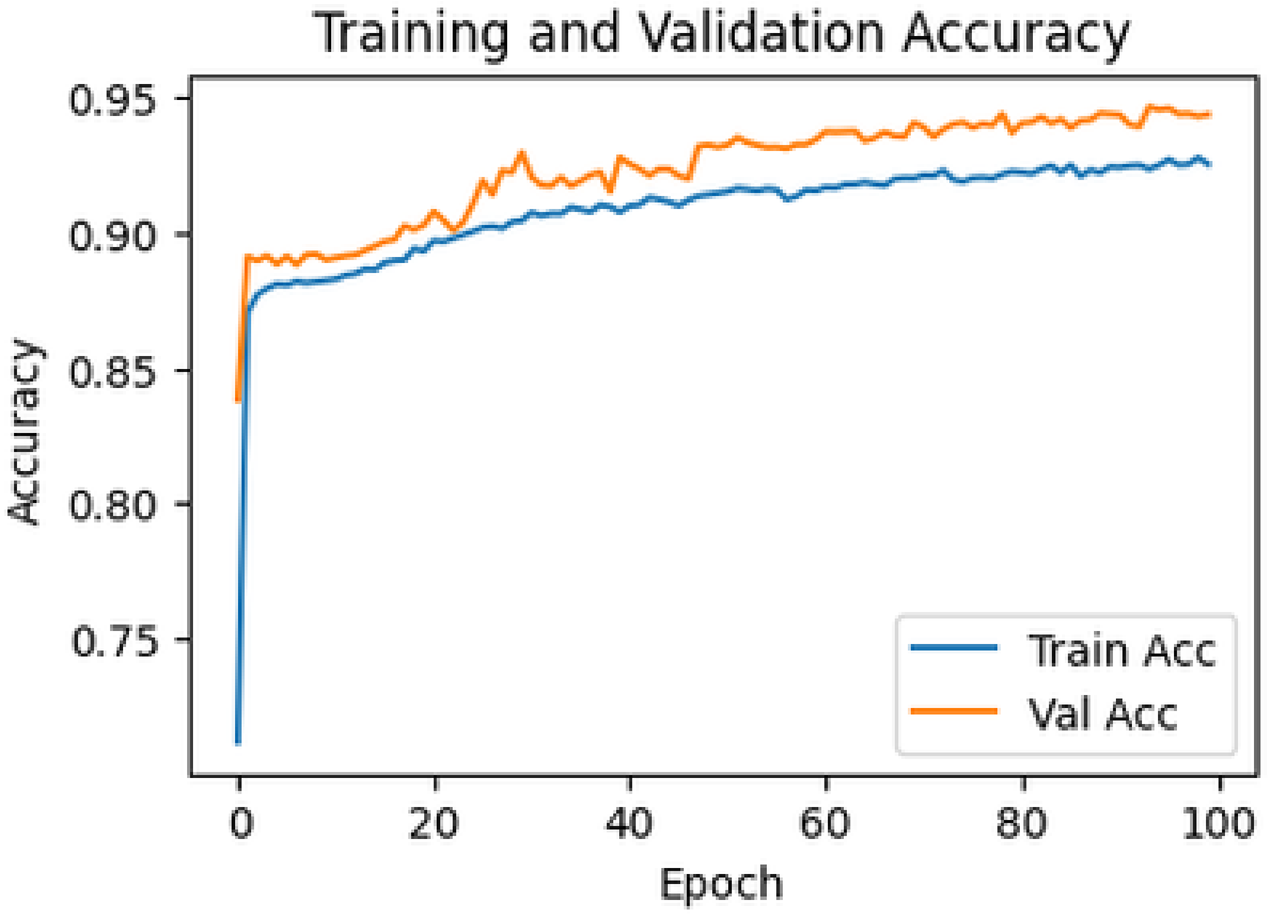


Accuracy	**87.95 %**
Loss	0.6136
No. of learnable Paramters	4,881,458
Epoches	100
Learning Rate	0.01
Optimizer	adam
Btach size	64
Validation slpit	20%

RNN+ATTENTION MODEL

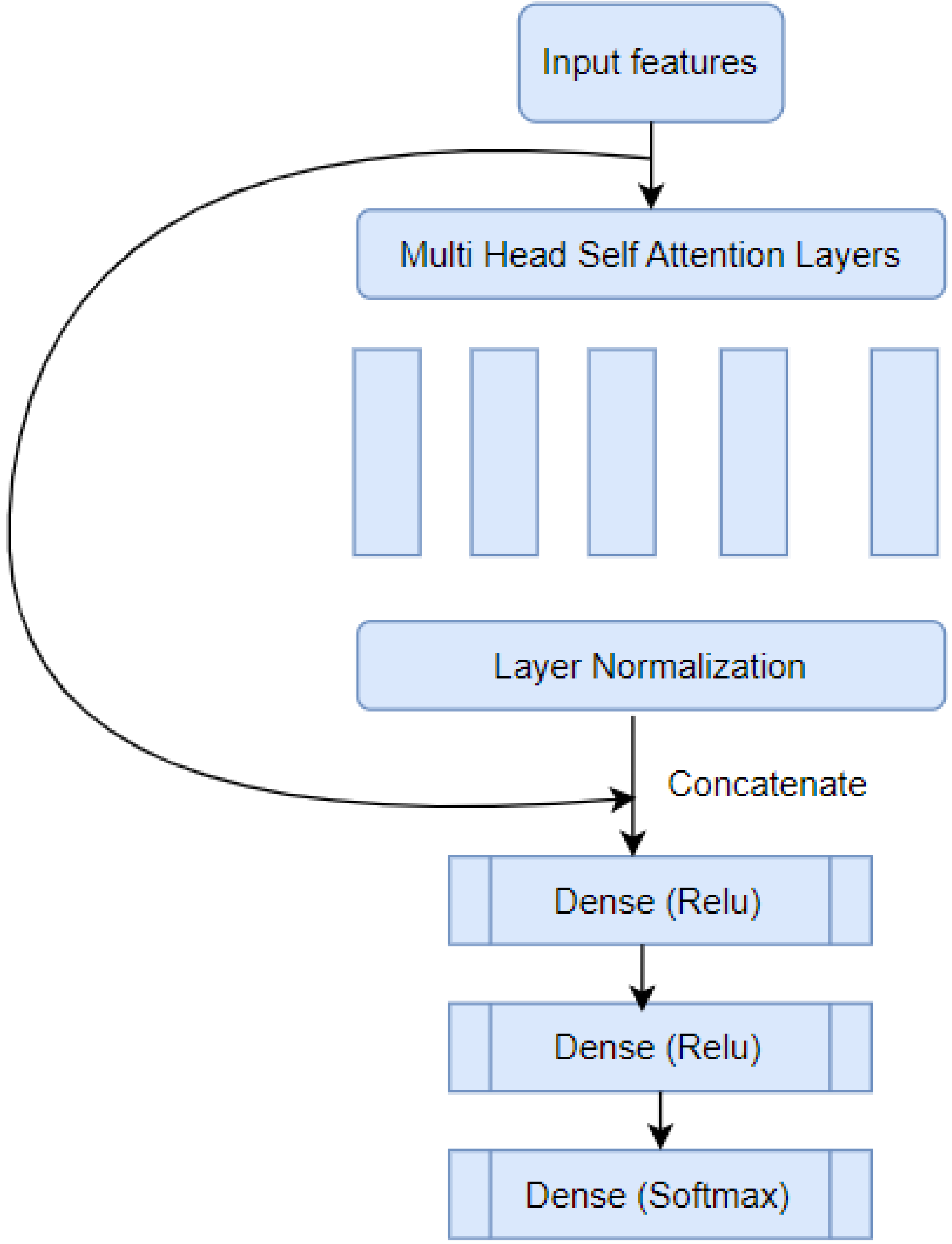






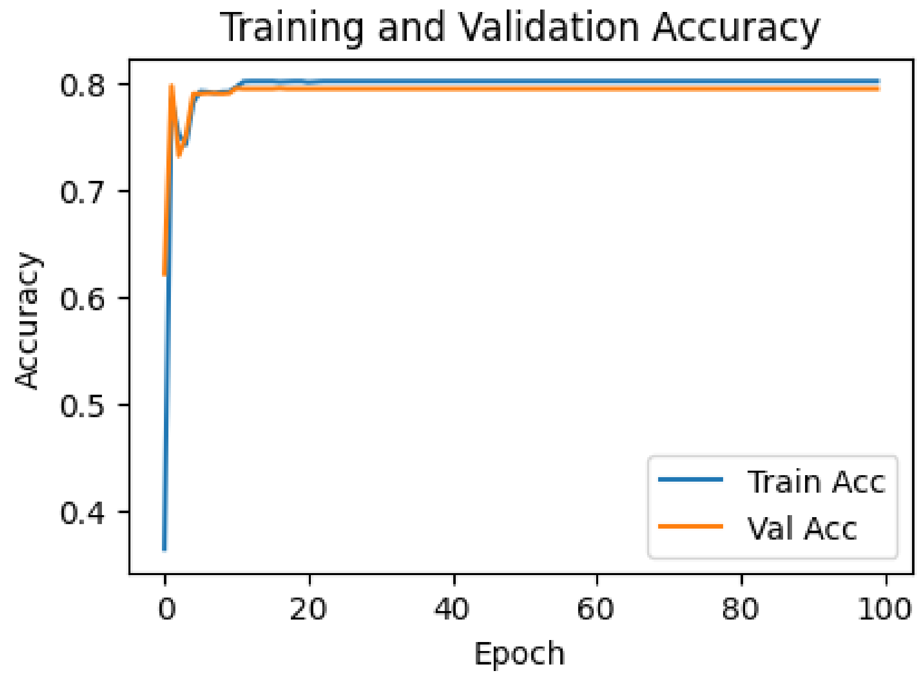
Accuracy	**94.3 %**
Loss	0.178
No. of learnable Paramters	1,223,266
Dropout	0.5
Epoches	100
Learning Rate	0.01
Optimizer	adam
Btach size	64
Validation slpit	20%

MULTI-HEAD SELF ATTENTION +FC MODEL



Accuracy	**80.02%**
Loss	0.6096
No. of learnable Paramters	44,808,754
Epoches	100
Learning Rate	0.01
Optimizer	adam
Btach size	64
Validation slpit	20%

MULTI-HEAD SELF ATTENTION +FC MODEL ACCURACY



EXPERIMENTS AND RESULTS

STACKED GRU

```
Precision score:  0.7218788400867777
F1-Score:  0.7057046769480033
Confusion Matrix:
 [[1803    0   12 ...    0    0    0]
 [   0  108    1 ...    0    0    0]
 [   3    0   37 ...    0    0    0]
 ...
 [   0    4    0 ...   13    0    0]
 [   0    0    0 ...    0   34    0]
 [   0    0    0 ...    0    0   34]]
```

RNN+ATTENTION

```
Validation loss of RNN+ATTENTION model: 0.1728864312171936
Validation accuracy of RNN+ATTENTION model: 0.9437500238418579
Precision score of RNN+ATTENTION model:  0.5739718579796977
F1-Score of RNN+ATTENTION model:  0.5365455024833097
Confusion Matrix:
 [[8416    3    0 ...    0    0   10]
 [   2   63    2 ...    0    0    0]
 [  17    0    9 ...    0    0    0]
 ...
 [   0    0    0 ...    8    0    0]
 [   0    0    0 ...    0   30    0]
 [   0    0    0 ...    0    0   32]]
```

SELFATTENTION+FC

```
Precision score:  2.8851307437599583
F1-Score:  2.8740810127263208
Confusion Matrix:
 [[9144    0    6 ...    0    0    0]
 [   0  565    0 ...    3    0    0]
 [   9    2  246 ...    1    0    0]
 ...
 [   0    4    1 ...  124    0    0]
 [   1    0    0 ...    0  149    0]
 [   1    0    0 ...    0    0  149]]
```

LSTM

```
Precision score:  0.8613882162263837
F1-Score:  0.8504055507657193
Confusion Matrix:
 [[9148    0    2 ...    0    0    0]
 [   1  566    0 ...    7    0    0]
 [  23    0  225 ...    6    0    0]
 ...
 [   0    4    2 ...  113    0    0]
 [   0    0    0 ...    0  150    0]
 [   0    0    0 ...    0    0  150]]
```

CHALLENGES AND KEY LEARNING

- Data preprocessing: Processing image frames and reshaping the data
- Sequence to Sequence: Working with temporal sequence data and working with self attention models.
- Handling overfitting and underfitting issues.
- Real time Lip Reading Model: Working with real time data / Unseen data.

DISCUSSION AND FUTURE IMPROVEMENT

Performance — RNN with attention model has given the best accuracy as compared to other models.

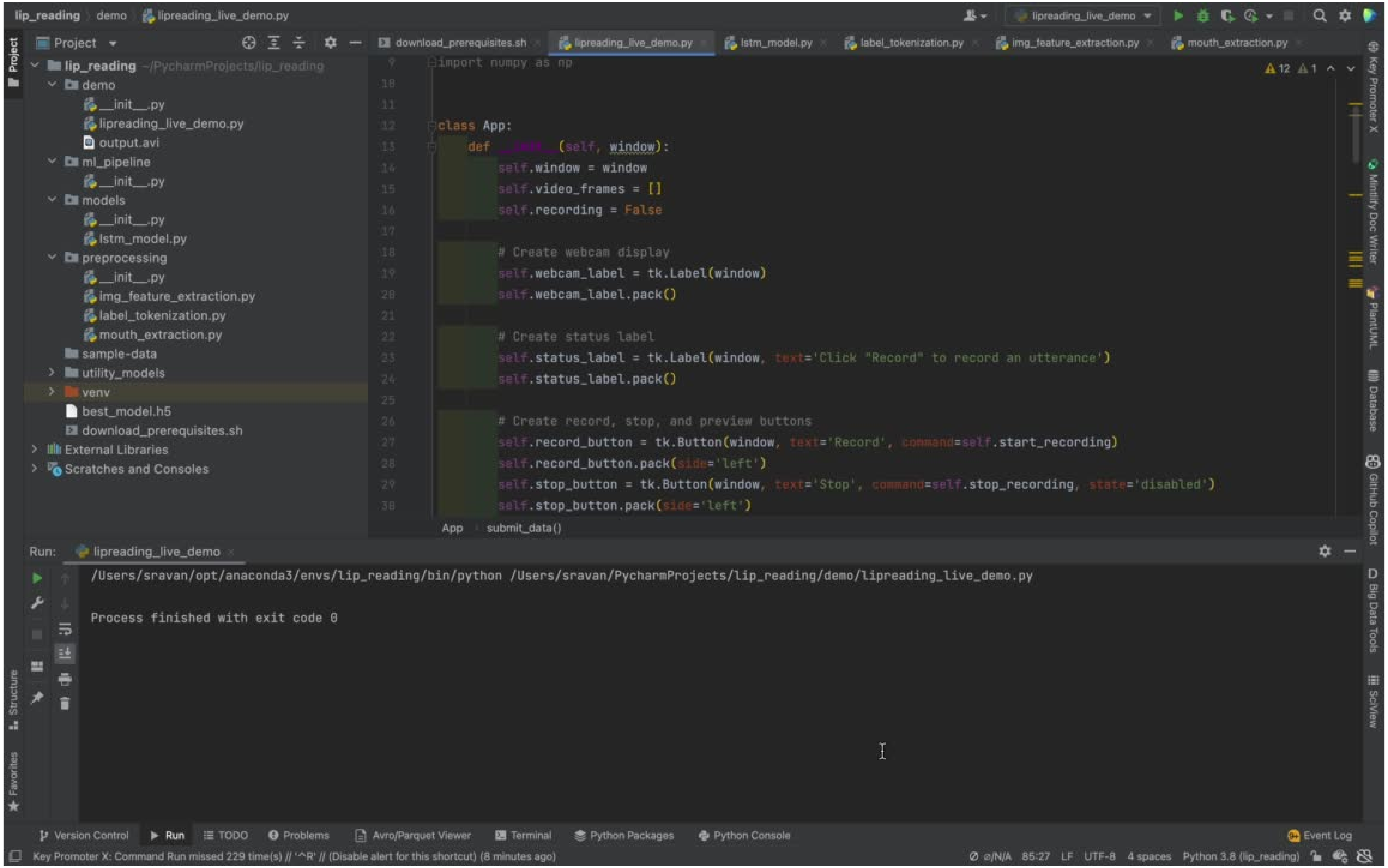
Memory — RNN with attention model has the least number of parameters

Future improvements

Exploring more models — Working with transformers , GCN etc.

Increasing training — Training the model for more epochs and with more data. Increasing the vocabulory.

PROTOTYPE/POC



TEAM CONTRIBUTION

LITERATURE SURVEY	3-5 Papers by each team members.
DATA COLLECTION	All Members
DATA PRE-PROCESSING	Face extraction & Mouth extraction-Mythri, Preparing phrases and words and combining them in training dataset- Himanshee, Image feature extraction and Data Augumentation-Mounica, Data preparation for modeling-Richa
MODEL IMPLEMENTATION	LSTM- Mythri, Stacked GRU- Mounica, RNN+Attention- Richa, Self-Attention+FC- Himanshee
DOCUMENTATION AND REPORT	All members
REAL-TIME LIPREADING MODEL	Himanshee, Mythri, Richa , Mounica (All members)

Thank You