MULTIMODAL SENTIMENT ANALYSIS

In partial fulfilment of BITS F312, Neural Networks and Fuzzy Logic

Prepared By:-

Mayur Arvind - 2016B1A70603G
Aneesh Garg - 2016B3A70340G
Aayush Soni - 2016A7PS0720G

INTRODUCTION

We experience the world in a multimodal fashion - through sight, sound, textures, smells, and tastes. A problem is referred to as multimodal when it includes multiple such modalities. To progress towards a general-purpose AI system, we need to develop methods to interpret such multimodal signals together.

For example, images are usually associated with explanations and captions, and texts contain images to augment the main idea of the article. These different modalities are characterized by very different statistical properties, and learning from such multimodal data is a challenge.

Though combining different modalities or types of information for improving performance seems intuitively appealing task, but in practice, it is challenging to combine the varying level of noise and conflicts between modalities.

This project involves a multimodal analysis of memes from both the picture and the associated text and attempting to develop models that generate accurate predictions given a new meme.

THE TASKS

Task A - Classify a meme as being *motivational* or *non-motivational*.

Task B - Identify the type of sentiment expressed in the meme - *positive*, *negative*, *very_positive*, *very_negative* or *neutral*

Task C - Quantify the degree of offensiveness of a meme

DATA CLEANING / PREPROCESSING

- Entries with erroneous/undefined values were dropped.
- The images were resized to 256 x 256.
- For the text, we performed standard NLP preprocessing techniques like removing numbers, single characters, punctuation symbols, extra spaces and lemmatization. In addition, the names of websites, which don't add any meaningful context, were removed.

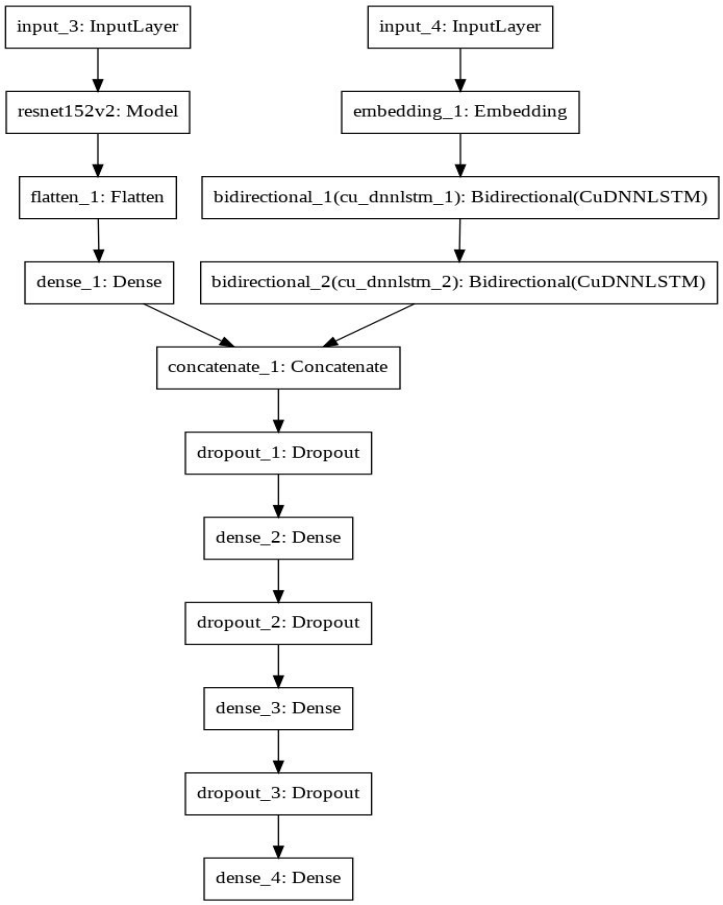
THE MODEL

For the images, we experimented with various pretrained models, including VGG, Inception and settled on ResNetV2, with 152 layers. We also tried out various pretrained embeddings for the text modality, such as GloVe and ELMo. While ELMo is in general, a superior approach, the unique nature of the dataset, the limited number of samples, and the large (1024 dimensional) size of the ELMo vectors, which necessitated a large number of epochs and many times got stuck at local optima, led us to choose GloVe.

The general structure of the model is as follows, with only the final layer differing for each task.

This can be considered a form of early fusion, with the ResNet and LSTM layers playing the role of feature extractors.

The Dense layers, interspersed with Dropout layers, perform the final classification.



In order to give equal weight to both the textual and visual modalities, both the streams are downsampled to a 128 dimensional vector before concatenation (fusion) and further processing.

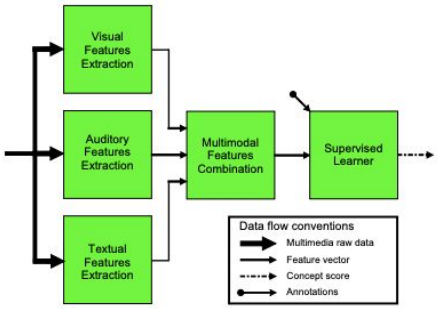
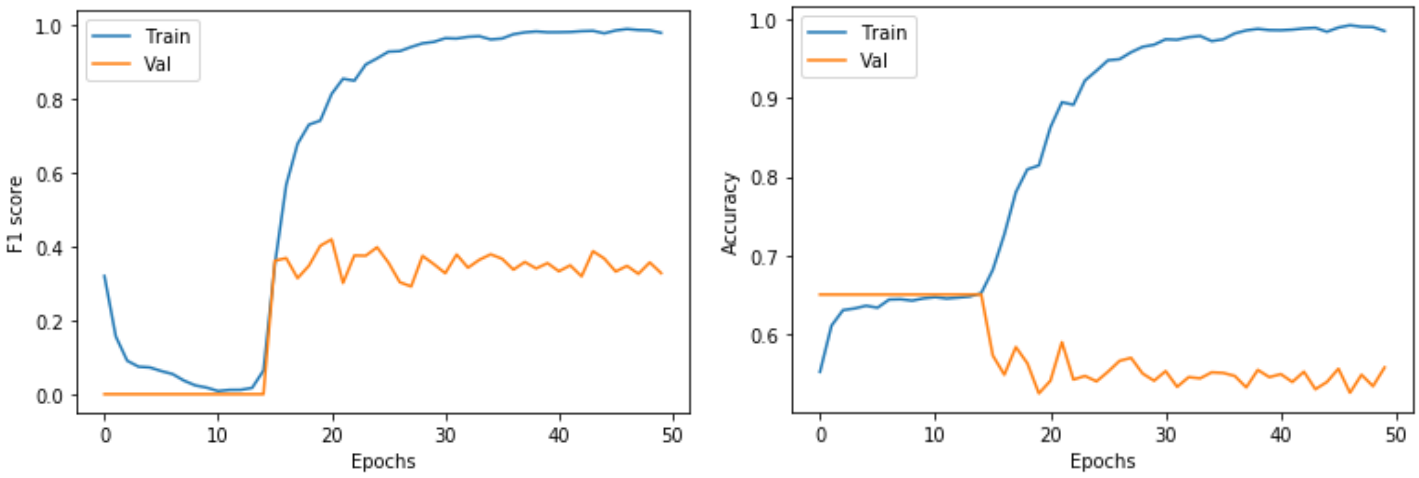


Figure 1: General scheme for early fusion. Output of unimodal analysis is fused before a concept is learned.

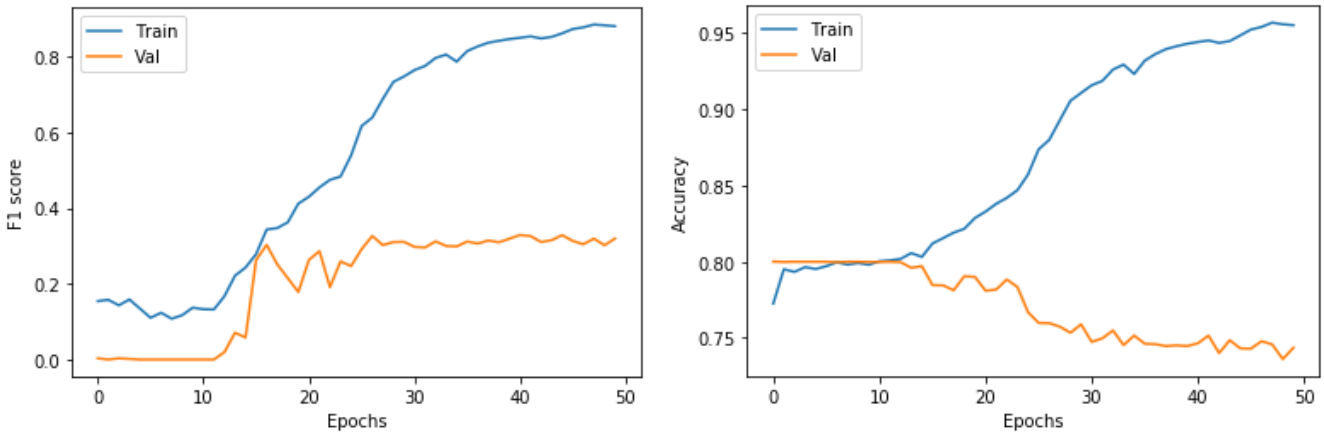
RESULTS

Task 1 -



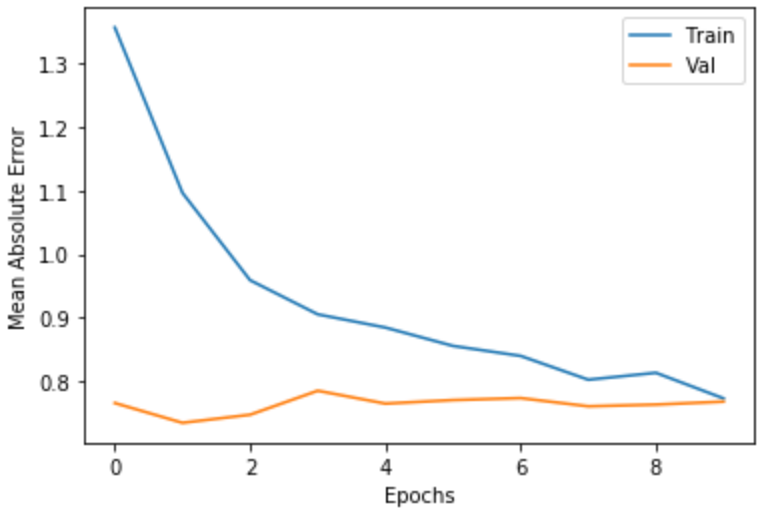
A validation F1 score of ~0.37 was obtained after 50 epochs, while the accuracy was ~56%.

Task 2 -



A validation F1 score of ~0.33 and an accuracy was ~74% were achieved.

Task 3 -



The mean absolute error on the validation set was ~0.76

CONCLUSION

In the coming decades, multimodal learning will be of the utmost importance, for example, with the advent of self-driving cars, whose systems take visual and radar signal inputs and IoT based home systems, which have temperature and speech-based inputs. This project was immensely enjoyable, and gave us a lot of insights into the entire machine learning pipeline, right from data cleaning and preprocessing to building suitable models for selecting optimal hyperparameters and generating accurate outputs. While the model's learning capability was limited by the small size of the dataset and its unique nature, the general approach followed can be successfully applied to a range of other tasks.

REFERENCES

1. **Early versus Late Fusion in Semantic Video Analysis -** http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.5928&rep=rep1&type=pdf
2. **Deep contextualized word representations - https://arxiv.org/abs/1802.05365**
3. **Learn to Combine Modalities in Multimodal Deep Learning - https://arxiv.org/abs/1805.11730**