Indian Institute of Information Technology Kottayam

MULTIMODAL SENTIMENT ANALYSIS

By

2020BCS0069 CHRISTO SOJAN
2020BCS0096 ANGATI BALA MURALI
2020BCS0184 KUMAR AMITANSHU
2020BCS0024 SYAM SIVADAS

Guided By,

Dr. Manu Madhavan

Introduction

- The project focuses on developing a machine learning model for sentiment analysis using multimodal data, such as **text, audio,** and **images**.
- The core of this project lies in dense fusion, where we **combine unimodal, bimodal, and trimodal interactions** with fusion techniques to enhance sentiment analysis by leveraging multiple modalities.
- Traditional sentiment analysis models often rely solely on text, overlooking valuable information from other modalities. By integrating multiple modalities through dense fusion, we aim to **create a more enhanced and accurate** sentiment analysis system.

Indian Institute of Information Technology Kottayam

Literature Review

Title	**Dense Fusion Network with Multimodal Residual for Sentiment Classification[1]**
Authors	Huan Deng, Peipei Kang, Zhenguo Yang, Tianyong Hao, Qing Li and Wenyin Liu
Year of Publication	2021
Journal	IEEE Xplore
Modalities	Language, Acoustic Speeches and Visual Images
Datasets	CMU-MOSI, ICT-MMMO, YouTube, and IEMOCAP
Models	The **DFMR framework** which consists of four modules-The **modality-specific module**, **dense multimodal fusion module**, **multimodal residual module**, and **sentiment classification module.**
Summary	The paper proposes the **DFMR framework for multimodal sentiment analysis**. It integrates language, speeches, and visual images using **dense fusion** and **multimodal residual modules**. The framework achieves state-of-the-art performance on benchmark datasets and provides a promising solution for sentiment analysis using multimodal data.

Results Obtained	The paper presents experimental results on four benchmark datasets. It **shows the performance of DFMR using single modalities versus multiple modalities** on the CMU-MOSI dataset. The paper demonstrates that **DFMR using multiple modalities outperforms eleven state-of-the-art baselines on these datasets**, indicating its effectiveness in multimodal sentiment analysis.
Current Status	The paper introduces the DFMR architecture for sentiment analysis, which **integrates multimodal information** and achieves state-of-the-art performance on benchmark datasets.
Future Scope	The future work may focus on **optimizing model's scalability and efficiency**, and evaluating its **performance in real-world applications**

Title	**Multi-Level Attention Map Network for Multimodal Sentiment Analysis[2]**
Authors	Xiaojun Xue, Chunxia Zhang, Zhendong Niu and Xindong Wu
YOP	2023
Journal	IEEE Transactions on Knowledge and Data Engineering journal.
Modalities	Texts and Images from user generated content
Datasets	MVSA-Single, MVSA-Multi, and Multi-ZOL.
Models used	Multi-Level Attention Map Network (MAMN), which consists of three modules: multi-granularity feature extraction, multi-level attention map generation, and attention map fusion.
Summary	This research paper presents new way to approach (MSA) as **"Multi-Level Attention Map Network"** (MAMN). The paper addresses the **challenges of noise reduction, feature extraction and correlation capture in MSA tasks**. The model utilizes techniques such as **Attention mechanisms, Gated mechanisms, and Multi-task learning.**

Results Obtained	The experimental results show that the proposed **MAMN model outperforms methods in terms of accuracy and effectiveness for both document-based and aspect-based MSA tasks**. The model achieves significant improvements in sentiment classification performance on the evaluated datasets.
Current Status	They contributed to the field "**Multimodal Sentiment Analysis**" via **novel model that addresses the challenges of noise reduction, feature extraction, and correlation capture**.However, the current status of MSA is an ongoing research area, and there is still room for further advancements and improvements in the field.
Future Scope	Paper suggest exploring **more advanced fusion methods, i**nvestigating the impact of different modalities on sentiment analysis, and exploring other application areas for MSA such as **product marketing** and **public opinion monitoring.**

Title	**Multimodal Sentiment Analysis via RNN variants[3]**
Authors	Ayush Agarwal, Ashima Yadav and Dinesh Kumar Vishwakarma
YOP	2019
Journal	IEEE Xplore
Modalities	Text, Audio and Video
Datasets	CMU-MOSI Dataset
Models used	Four variants of Recurrent Neural Networks (RNNs) for sentiment analysis (models are based on LSTM and GRU architectures) : - GRNN - LRNN - GLRNN - UGRNN
Summary	They conducted **experiments on the CMU-MOSI dataset and show that their approach achieves better sentiment classification accuracy** than existing methods on individual modalities and also after fusing the modalities using attention networks.

Results obtained	The research paper reports the accuracy achieved by the different RNN variants on the CMU-MOSI dataset. **GRNN performs best for text, GRNN and GLRNN perform best for audio, and UGRNN performs best for video**. After fusing the modalities, LRNN and GLRNN achieve the best results for multimodal sentiment analysis, with an accuracy of 78.05%.
Current Status	They contributed to the field "**Multimodal Sentiment Analysis**" via **proposing novel RNN variants and evaluating their performance on the CMU-MOSI dataset**. It demonstrates the effectiveness of using multimodal data for sentiment classification.
Future Scope	- Credibility analysis - Used for medical purposes such as detection of autism in a child - Used for predicting the sentiments, emotions, and genre of a movie by its trailer.

Title	**Fusion-Extraction Network for Multimodal Sentiment Analysis[4]**
Authors	Tao Jiang, Jiahai Wang, Zhiyue Liu and Yingbiao Ling
YOP	2020
Journal	Pacific-Asia Conference on Knowledge Discovery and Data Mining
Modalities	Visual and textual information.
Datasets	MVSA(Multiple Viewpoint Semantic Annotation)- Single and MVSA-Multiple
Models used	Fusion-Extraction Network (FENet) using fine grained attention and gated convolutional layers, BERT.
Summary	The document proposes a **Fusion-Extraction Network (FENet)** for multimodal sentiment analysis. The network utilizes an interactive information **fusion mechanism** to learn visual-specific textual representations and textual-specific visual representations. It also incorporates an **information extraction mechanism to filter redundant parts and extract valid information** from the multimodal representations. Experimental results on two public multimodal sentiment datasets show that FENet outperforms existing state-of-the-art methods.

Results	The experimental results show that **FENet outperforms existing state-of-the-art methods** on the two multimodal sentiment datasets. The model achieves **higher accuracy and F1 scores compared to baseline methods** such as SentiBank & SentiStrength, CoMN etc.
Current Status	The research paper contributes to the field of multimodal sentiment analysis by proposing a **novel model that effectively utilizes the relationship between visual and textual information**. It demonstrates improved performance compared to existing methods, indicating progress in the field.
Future Scope	The research opens up possibilities for further advancements in multimodal sentiment analysis. Future work could **explore more sophisticated fusion mechanisms, extraction techniques, and attention mechanisms** to enhance the understanding of multimodal data and improve sentiment analysis performance. Additionally, the proposed model could be **extended to other domains and datasets** to evaluate its generalizability.

Title	**Multimodal Sentiment Analysis: Addressing Key Issues and Setting up the Baselines[5]**
Authors	Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh and Amir Hussain
YOP	2019
Journal	55th Annual Meeting of the Association for Computational Linguistics (ACL) and the IEEE International Conference on Data Mining (ICDM)
Modalities	Text, Audio, Visual
Datasets	MOUD, MOSI, IEMOCAP
Models used	CNN, 3D-CNN, OPENSMILE, LSTM, SVM
Summary	The research paper explores multimodal sentiment analysis using deep-learning architectures for sentiment classification. It considers text, audio, and visual modalities for understanding emotions in videos. **The authors use CNN, 3D-CNN, openSMILE, and bc-LSTM models. The bc-LSTM fusion method outperforms SVM in accuracy, while audio and text modalities play crucial roles.** The paper emphasizes the need for context, different modalities, and generalizability of multimodal sentiment classifiers. Future work should focus on extracting semantics from visual features and incorporating contextual dependency learning.

Results Obtained	**The results show that the bc-LSTM fusion method consistently outperforms the SVM fusion method across all experiments**. **The audio modality performs better than the visual modality in both the MOSI and IEMOCAP datasets.** The text modality plays a crucial role in both emotion recognition and sentiment analysis, with its unimodal performance being substantially better than the other modalities.
Current Status	There are still major issues that remain mostly unaddressed in this field, such as the **consideration of context in classification, the effect of speaker-inclusive and speaker-exclusive scenarios, the impact of each modality across datasets, and the generalizability of multimodal sentiment classifiers**. The document serves as a benchmark for future research in MSA and highlights the need for further exploration and improvement in these areas.
Future Scope	Semantics Extraction uncovers deeper meaning from visual cues, Cross-Modal Feature Exploration refines fusion techniques, Contextual Dependency enhances classification, Benchmark Datasets standardize evaluations, and Generalizability ensures broader applicability, collectively driving advancements in multimodal sentiment analysis.

Motivation

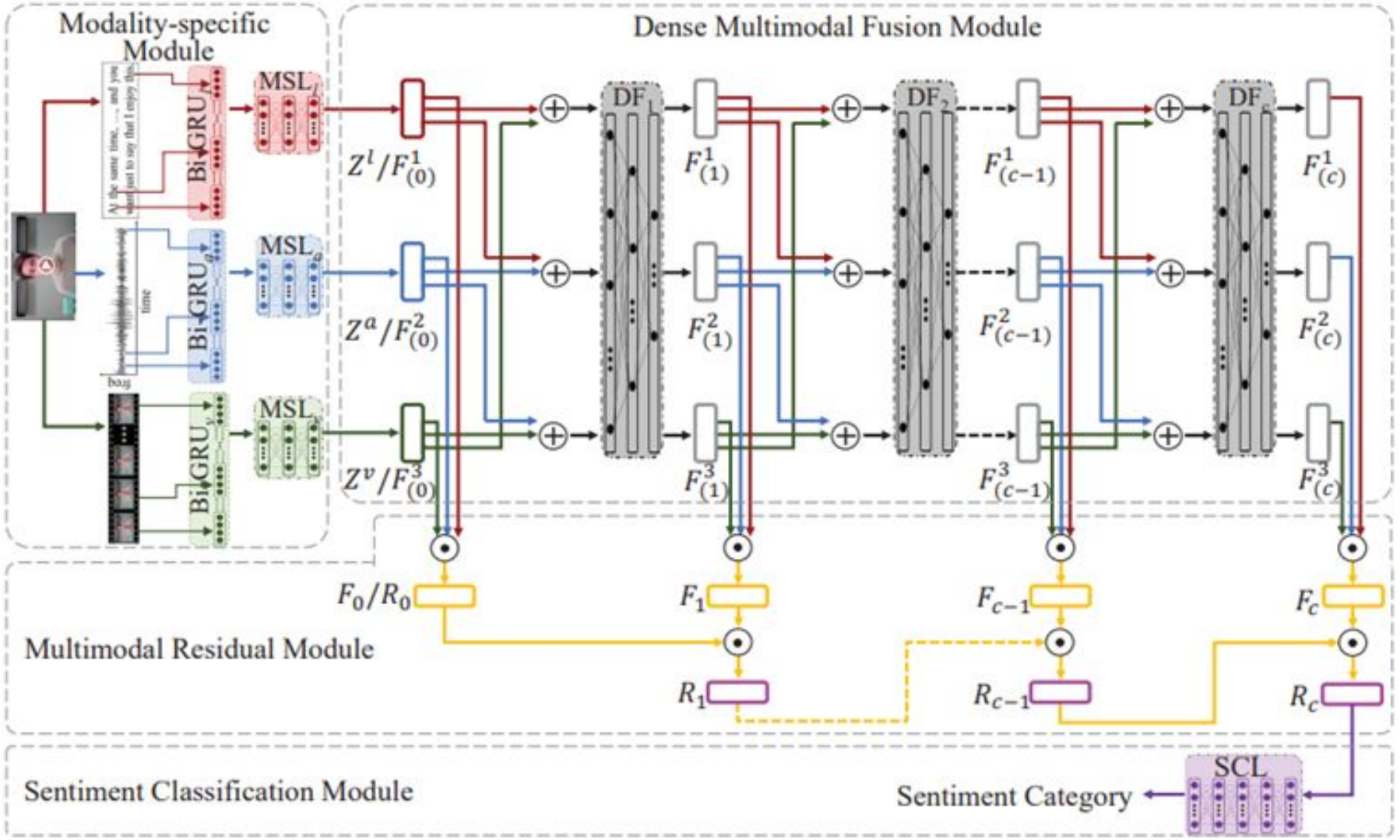
Existing approaches in multimodal sentiment analysis often neglect **cross-modal interactions**.The Dense Fusion Network with Multimodal Residual (DFMR) framework proposes an **end-to-end dense fusion network** that effectively **fuses multimodal features** in a stacking manner for sentiment analysis, providing a more accurate solution.

Additionally, the inclusion of dense fusion (DF) blocks enables the **modeling of unimodal, bimodal, and trimodal interactions** simultaneously, capturing cross-modal dynamics.

Problem Statement

- Increase in Internet usage in India
- Increase in usage of Indian languages on social platforms
- Absence of linguistic, lexical and data
- Hinders in business community and Government agencies
- Unique Needs and sentiments

Architecture



Experiment

The dataset used in for our current demonstration is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database .The IEMOCAP dataset consists of 151 videos of recorded dialogues, with 2 speakers per session for a total of 302 videos across the dataset.

Datasets Content:

videoSpeakers : There are multiple participators in one dialogue. videoSpeakers maps utterance to its speakers.

videoLabels : The emotion Labels for each utterance in a dialogue.

videoText : The text features extracts using TextCNN.

videoAudio : The video features extracts using openSMILE kitools

videoVisual : The visual features extracts using 3d-CNN.

videoSentence : The raw text info in a dialogue.

- We pass the each of the modality data into the **modality specific Bi-Directional GRU**.
- The output we get will be used to compile the model for that particular modality.
- Then the model is trained using the training dataset and evaluated using the testing dataset.
- This is performed for each of the three modalities - text, audio and visual.

- We extract all three modalities' features and pass them to the three Bi-GRUs. we then do a **simple concatenation** of the three outputs. The combined output is then fed into the output layer which is **sentiment classification module.**

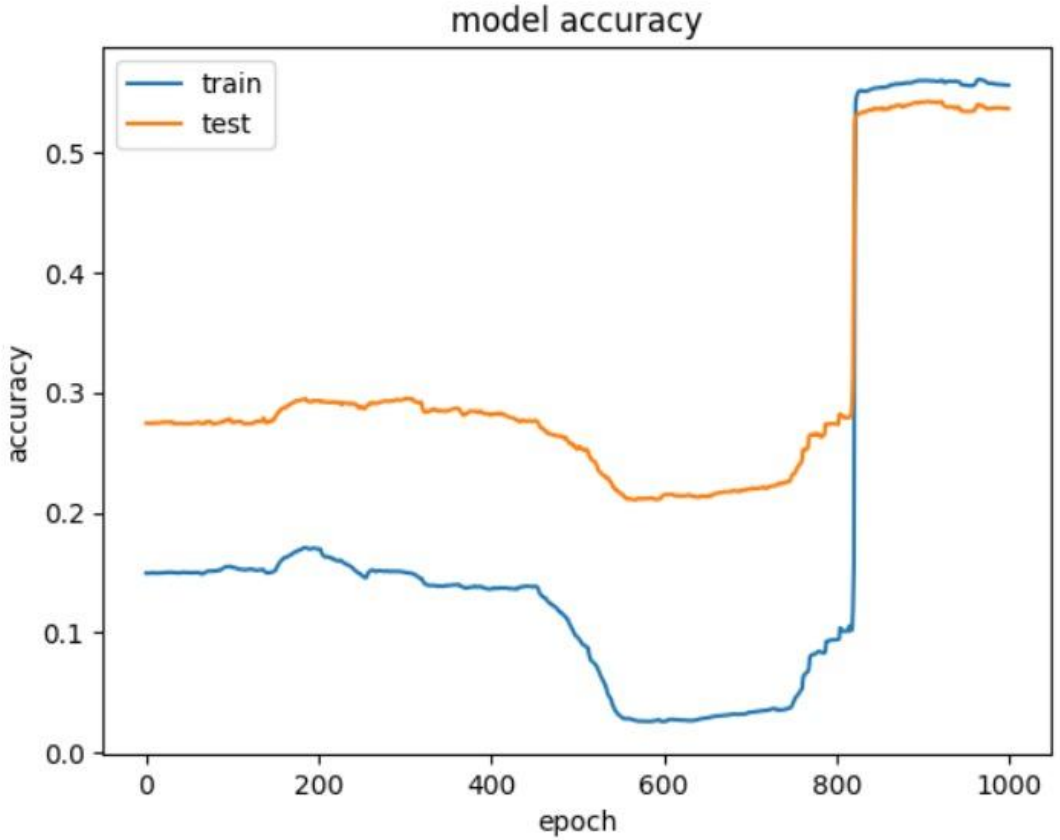
- The classification labels are:
  - **0 - Happy**
  - **1 - Sad**
  - **2 - Neutral**
  - **3 - Angry**
  - **4 - Excited**
  - **5 - Frustrated**

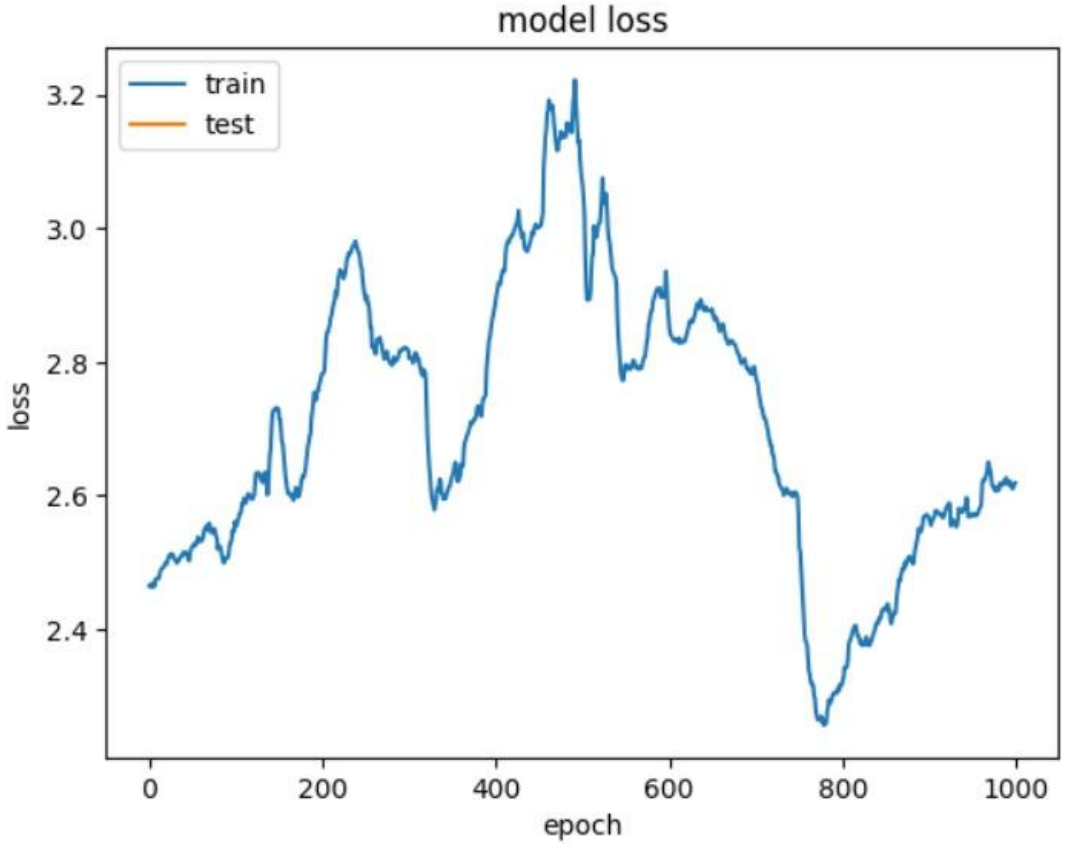
Comparison

Accuracy of different modalities compared with the accuracy of MSA

Audio Analysis	Text Analysis	Visual Analysis	MSA Analysis
0.1839	0.0412	0.1464	0.5980

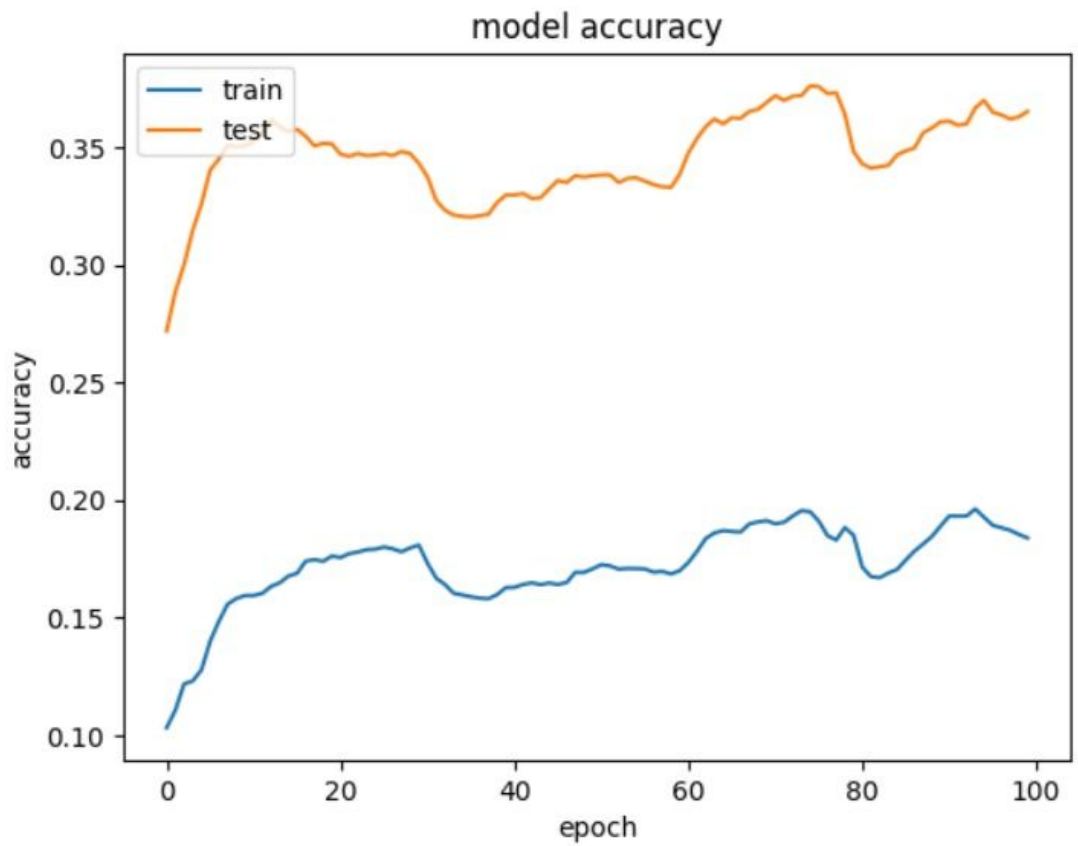
Demonstration

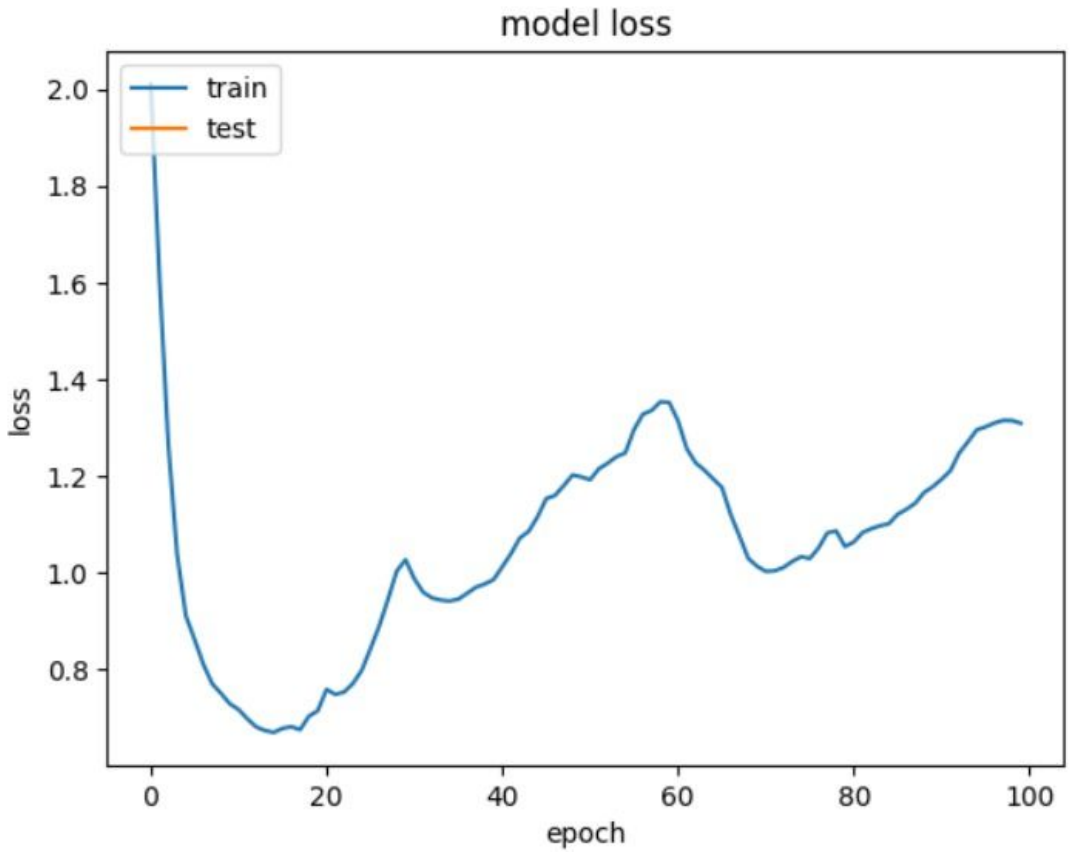




Text analysis

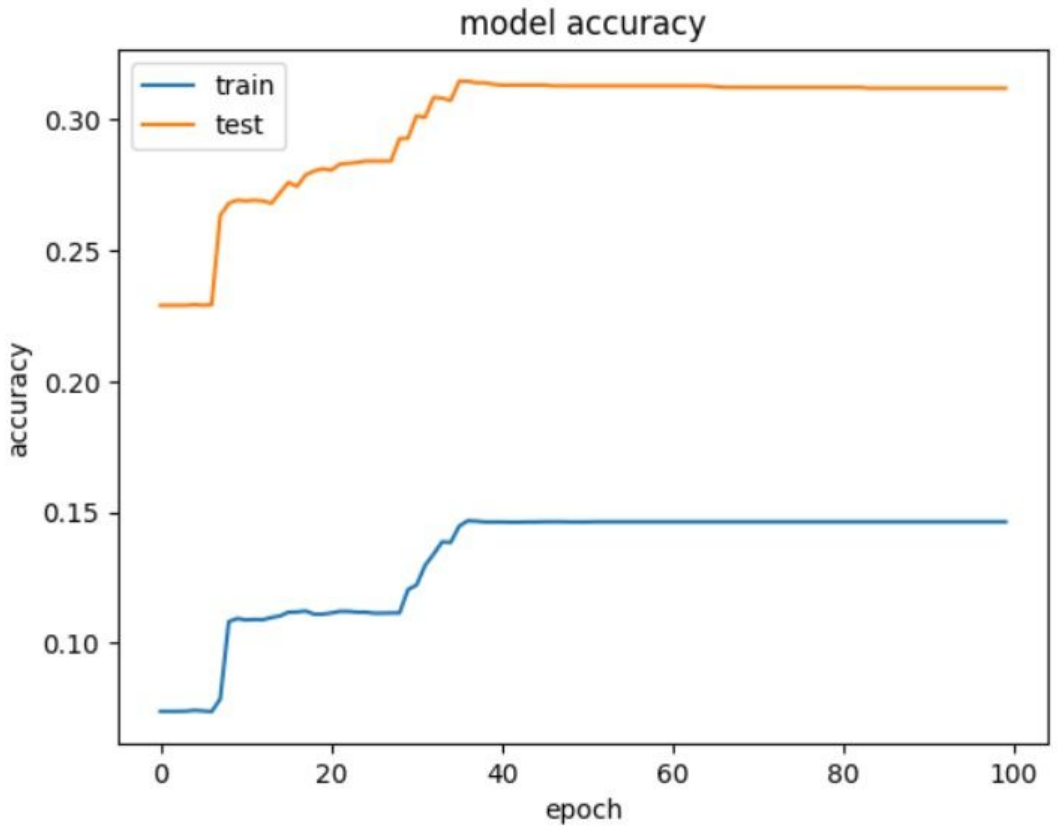
Indian Institute of Information Technology Kottayam

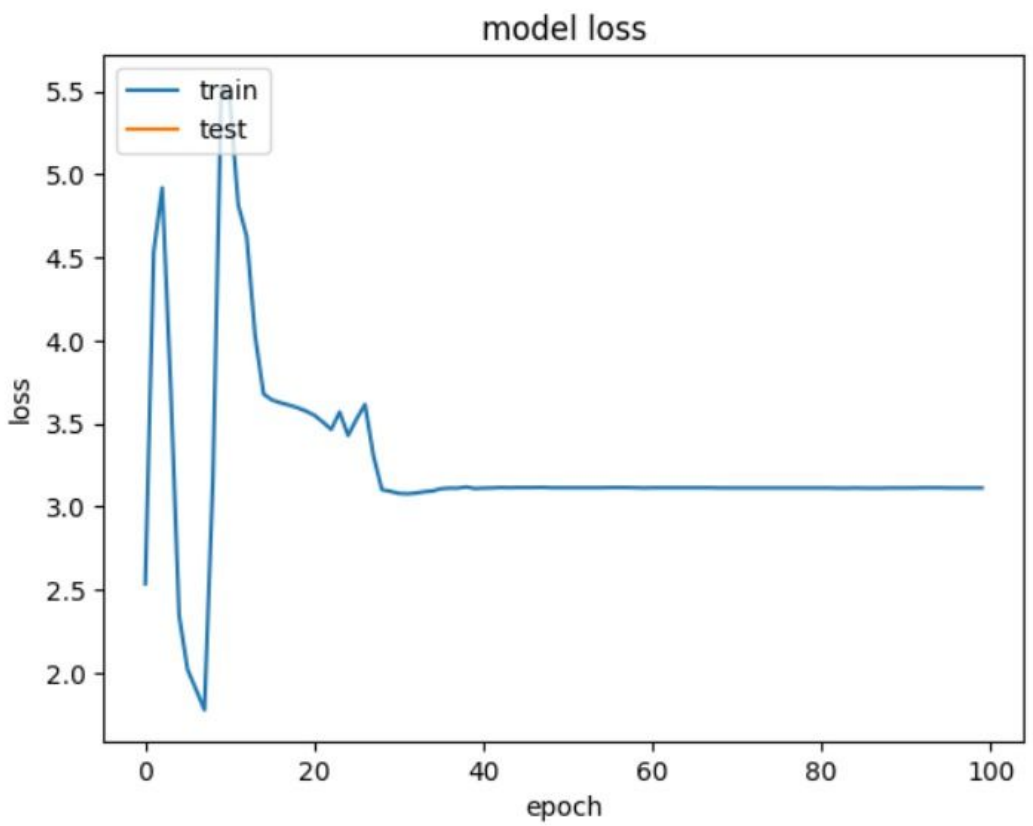




Audio sentiment analysis

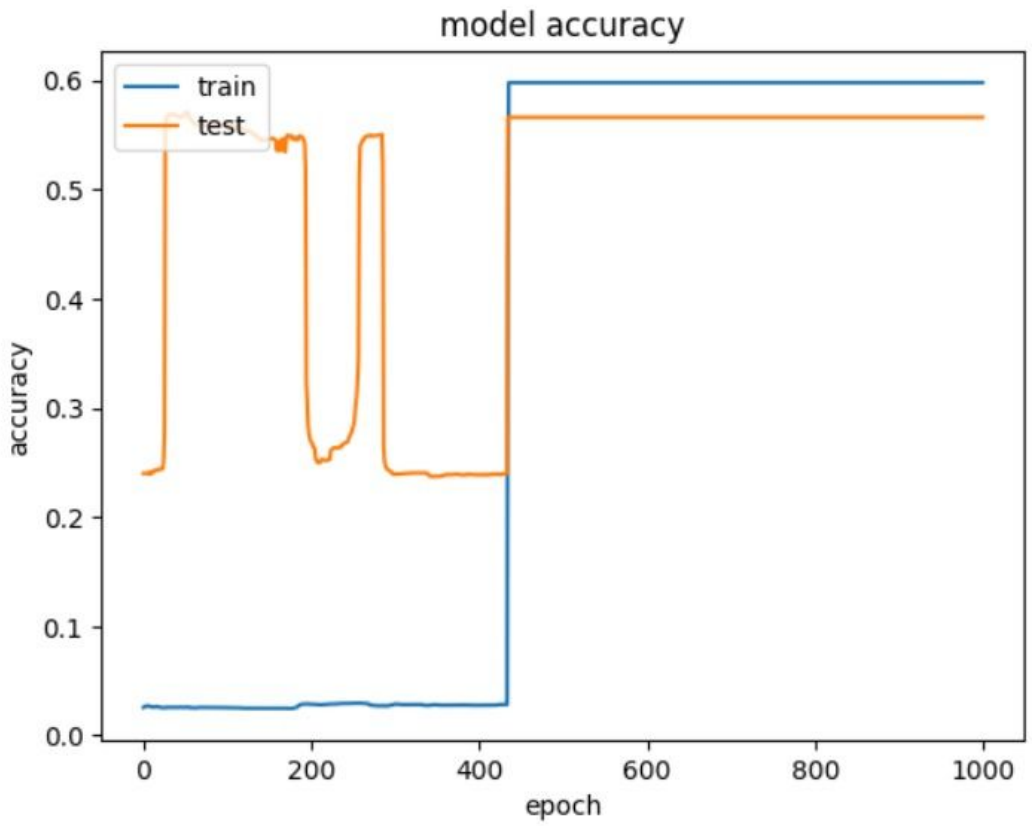
Indian Institute of Information Technology Kottayam

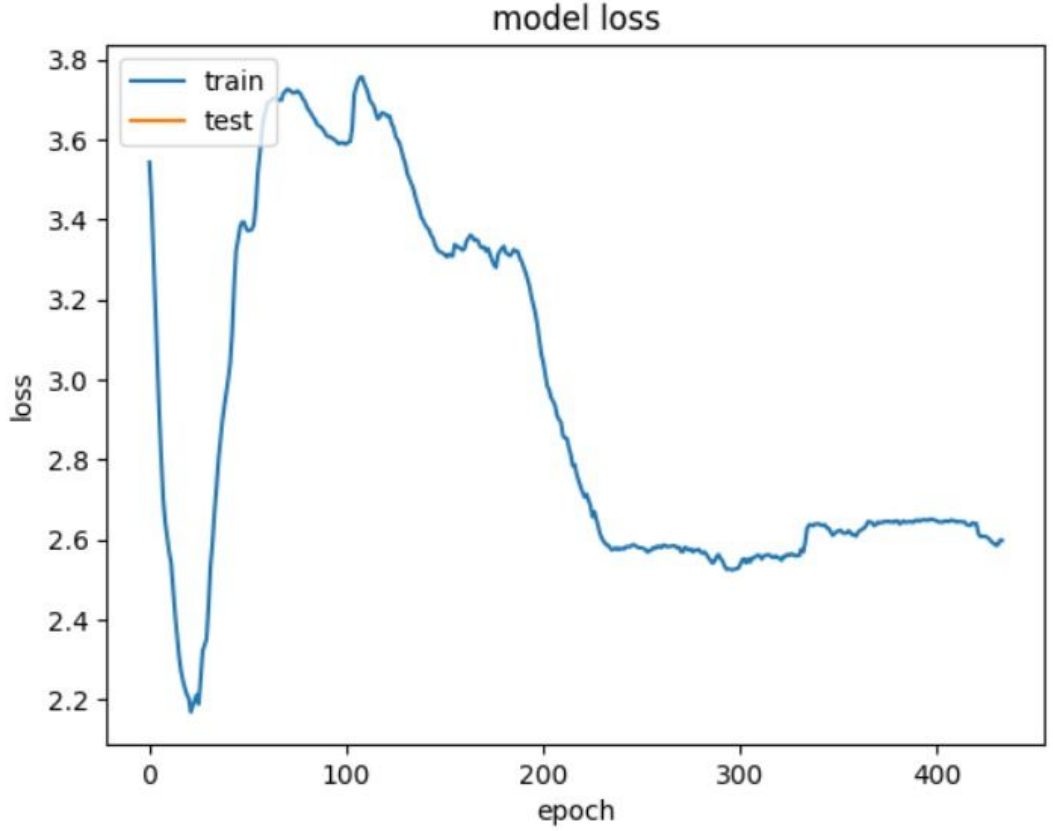




Visual sentiment analysis

Indian Institute of Information Technology Kottayam





MSA Result

Future Scope and Conclusion

- We are currently limited by our resources. We cannot implement the larger dataset of CMU MOSEI using the resources we have at hand. In future, we plan to implement our model using the CMU MOSEI dataset which provides more accurate results efficiently.

- We have currently used simple concatenation to combine our multiple modalities to produce a result. The actual architecture we are going to implement consists of the dense fusion extraction module in place of simple concatenation which will greatly increase the accuracy of the model.

- At present we have implemented the multimodal sentiment analysis only in the English language. Our next step would be to create a dataset for implementing it on the regional languages in India, and to carry out the implementation maintaining the same accuracy as in the previous scenario.

REFERENCES

- [1] Jiang, T., Wang, J., Liu, Z., Ling, Y. (2020). Fusion-Extraction Network for Multimodal Sentiment Analysis. In: Lauw, H., Wong, RW., Ntoulas, A., Lim, EP., Ng, SK., Pan, S. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2020. Lecture Notes in Computer Science(), vol 12085.
- [2] X. Xue, C. Zhang, Z. Niu and X. Wu, "Multi-Level Attention Map Network for Multimodal Sentiment Analysis," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 5105-5118, 1 May 2023, doi: 10.1109/TKDE.2022.3155290.
- [3] A. Agarwal, A. Yadav and D. K. Vishwakarma, "Multimodal Sentiment Analysis via RNN variants," *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, Honolulu, HI, USA, 2019, pp. 19-23, doi: 10.1109/BCD.2019.8885108.
- [4]S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh and A. Hussain, "Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines" in *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17-25, Nov.-Dec. 2018, doi: 10.1109/MIS.2018.2882362.Springer, Cham.
- [5] H. Deng, P. Kang, Z. Yang, T. Hao, Q. Li and W. Liu, "Dense Fusion Network with Multimodal Residual for Sentiment Classification," *2021 IEEE International Conference on Multimedia and Expo (ICME)*, Shenzhen, China, 2021, pp. 1-6, doi: 10.1109/ICME51207.2021.9428321.