MULTIMODAL HATE DETECTION IN VIDEOS

Prabal Pratap Singh
(2327029)

Supervisor – Dr. Jianbo Jiao

Brief overview of Project

- My project focuses on enhancing the HateMM model, which is designed for detecting hateful speech in video content.

- The model processes multi-modal data (text, audio, and visual cues) to classify whether a video contains hateful content.

- I aim to improve the model's accuracy, efficiency, and generalizability by refining the architecture, dataset preprocessing, and training methodology.

https://ojs.aaai.org/index.php/ICWSM/article/view/22209

Problem Statement

- Hate speech in online videos is increasing, contributing to cyberbullying, misinformation, and online harassment.
- Existing hate detection models struggle with Low accuracy, High false positives and false negatives and scalability issues.
- The goal is to improve HateMM's performance by optimizing its multi-modal fusion technique and addressing preprocessing techniques.
- I have a strong background in computer vision, NLP, Neural computation and Intelligent data analysis, making this a perfect problem to apply my skills.
- Real-world impact: Improving HateMM can help social media platforms automate hate speech filtering, reducing the burden on human moderators.

Objectives & Key Goals of My Project

- From Scratch create and optimize the multi-modal fusion technique (text, audio, and visual features).
- Use advanced feature extraction techniques for better representation of hate speech.
- Reduce computational cost while maintaining high accuracy.
- Conduct qualitative analysis to understand real-world applicability (Instagram Videos)

Dataset

- The dataset used in this project is based on the dataset provided in the HateMM paper.
- To increase real-world applicability, I expanded the dataset by adding more videos collected from platforms like Instagram.
- Each new video was manually annotated to ensure accurate hate speech labeling, also enhancing dataset diversity.

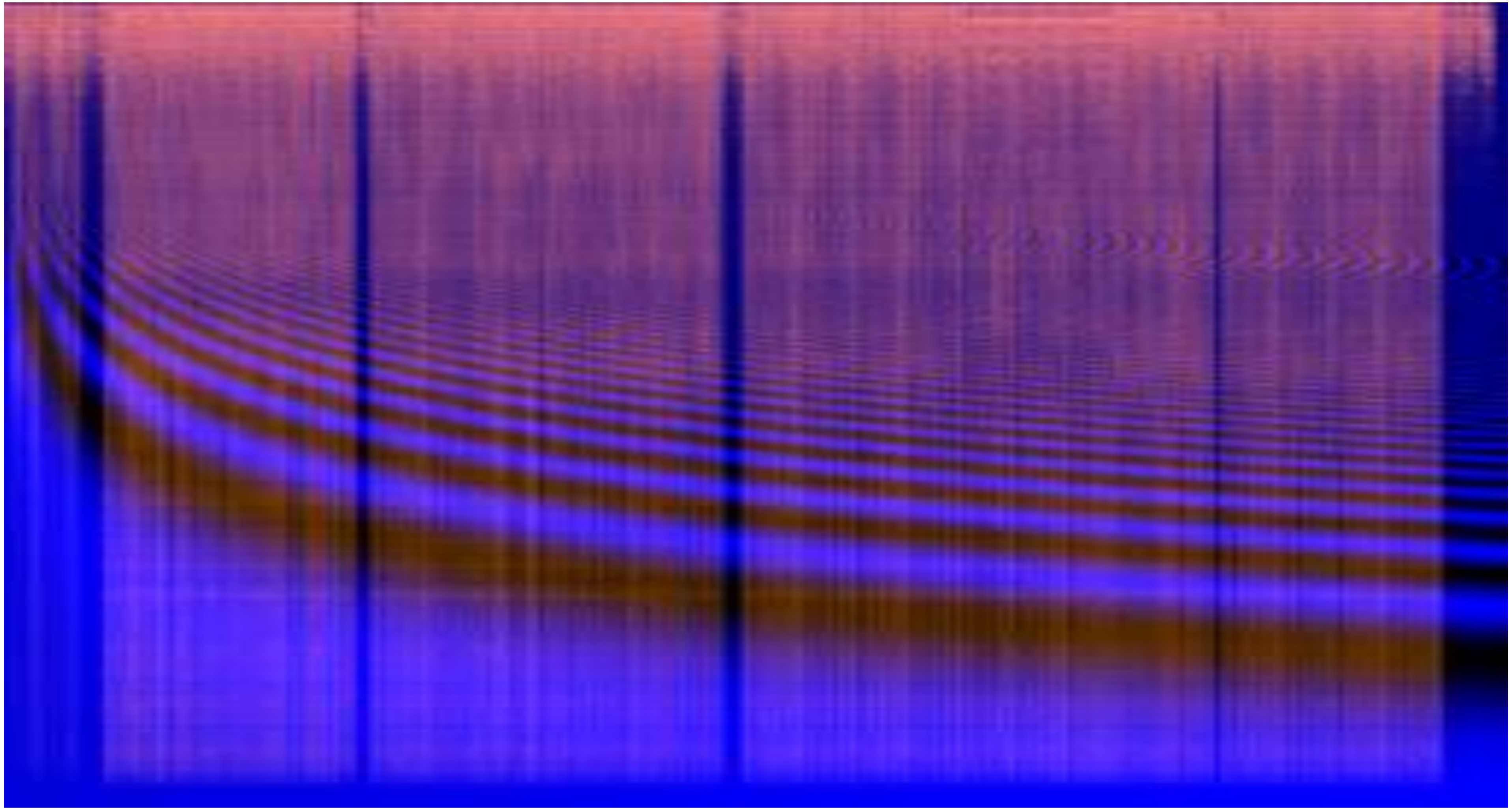
Preprocessing (Text)

- To extract text from video content, I used OpenAI Whisper, a state-of-the-art automatic speech recognition model.
- Whisper provided a highly accurate transcript of what was spoken in the video, ensuring that all text-based hate speech was captured effectively.
- Although I intended to use large model but my computer didn’t had enough memory to accommodate the processing
- To convert text to tensor I used BERT pretrained tokeniser [1,1024]

Preprocessing (Audio)

- Instead of using raw 1D audio signals, I converted them into 2D STFT spectrograms as they have been proven to be more effective for hate speech detection (based on literature review).
- To introduce a level of spatial knowledge, I incorporated sine encoding in the spectrograms.
- A [3 × 256 × 256] tensor, where each channel captures different features of the spectrogram.

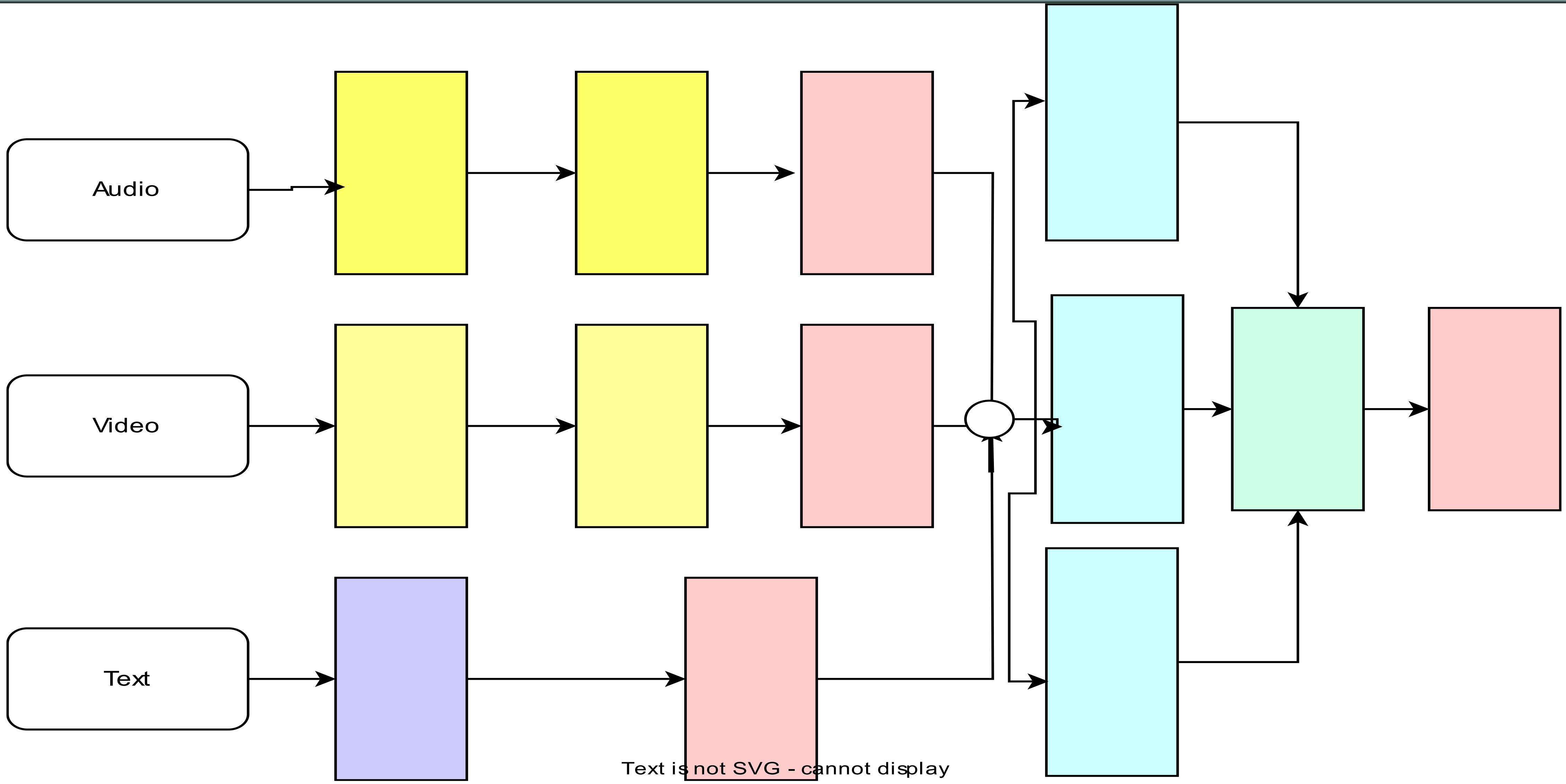
Preprocessing (Audio)



Preprocessing (Video)

- Frames were selected based on the highest sound intensity, as hateful speech is more likely to occur in loud speech segments.
- From each video, 50 frames were extracted to maintain computational efficiency while capturing key visual context
- Final Video Representation: A [3 × 64 × 64] tensor, optimized for further model processing.

Architecture of Best Model (82% Acc)



Implementation Details

- This model is trained using PyTorch, efficiency while capturing key visual context
- Training involved careful hyperparameter tuning to achieve optimal performance.
- BceLoss was used as this a binary classification problem
- Learning rate was adjusted dynamically
- Used dropout layers and others techniques (regularisation) to handle generalizability and training momentum

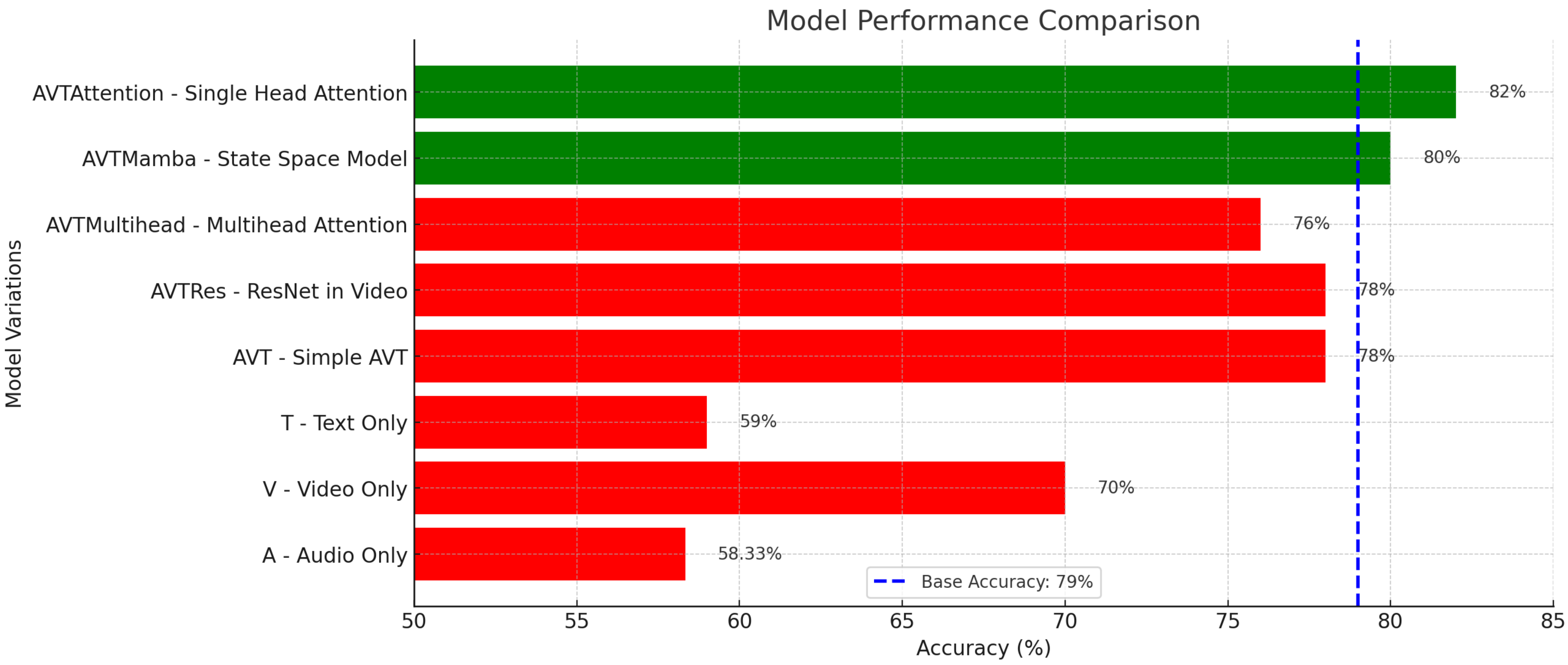
Implementation Details

- This model is trained using PyTorch, efficiency while capturing key visual context
- Training involved careful hyperparameter tuning to achieve optimal performance.
- BceLoss was used as this a binary classification problem
- Learning rate was adjusted dynamically
- Used dropout layers and others techniques (regularisation) to handle generalizability and training momentum
- Batch was set to 2 because of high memory requirements to train the model and my computer didn’t had enough GPU memory

Other Models and their Results

- A – Audio Only 58.33% Accuracy
- V- Video Only 70% Accuracy
- T- Text only 59% Accuracy
- AVT – Simple Audio + Video + Text 78% Accuracy
- AVTRes – Resnet implementation in Video 78% Accuracy
- AVTMultihead – used Multihead attention 76% Accuracy
- AVTMamba – used State Space Model (with gates) 80% Accuracy
- AVTAttention – used single head attention 82% Accuracy

Other Models and their Results



Interpretation

- AVTAttention performed the best in video classification even better that the original model in the paper!!!
- Attention mechanisms effectively focus on important features in multimodal inputs (audio, video, and text).
- But to My surprise Multihead attention performs worse that single head attention.
- Multimodal Fusion is superior to Single Modality models
- AVTMamba is a State-space model with gating mechanisms are excellent at capturing long-range dependencies.
- This likely allows the model to integrate context better across audio, video, and text modalities.
- Single channel models does not perform very well

Potential Improvements and Future work

- The machine used for training had only 16GB RAM and did not have a high-end GPU, which may have limited batch size and precision settings during training.
- AMP was enabled, without it the model could have allowed much higher precision
- The current model was trained using 64×64 images, which lack fine-grained details.
- Fusion of multiple architectures like Mamba, Attention, and LSTMs could further enhance sequential understanding.
- Experimenting with cross-modal transformers or contrastive learning techniques could also yield better results.

- Since the model performer better than the one listed in paper it could be published as told by my supervisor

Thank you!