Multimodal Deep Learning Model for Persuasion Technique Detection

This report presents a deep learning system developed to detect persuasive techniques in memes, crucial for understanding the spread of misinformation on social media. The project is divided into two tasks: the Main Task involves classifying persuasive techniques in memes, while the Enhancement Task focuses on sequence tagging to identify overlapping persuasive techniques. This report will elaborate on the system's design, evaluation, and results based on data analysis, model performance, and independent testing.

Model Design for Main Task: Multimodal Classification

The main task involves designing a multimodal deep learning model that integrates both textual and visual features to identify persuasion techniques in memes. The model leverages advanced architectures like DeBERTaV2 for text processing and ResNet50/EfficientNetB0 for image analysis, ensuring accurate classification through feature integration and attention mechanisms.

- **Text Feature Extraction**
  The model uses a pre-trained DeBERTaV2 to process the textual content of memes. The DeBERTaV2 tokenizer converts meme captions into token IDs and attention masks, ensuring sequences are padded or truncated to a consistent length of 128 tokens. These tokenized sequences are fed into the DeBERTaV2 model, generating contextual embeddings from the final hidden state. To create a fixed-length representation for each caption, the embeddings are pooled using a mean operation and then normalized using L2 normalization to maintain a consistent feature scale for training.

- **Image Feature Extraction**
  The model incorporates two pre-trained convolutional neural network (CNN) architectures: ResNet50 and EfficientNetB0. Both networks are used without their top classification layers, allowing for extraction of generalized image features. Each meme image is resized to 224x224 pixels for compatibility. ResNet50 captures detailed structural information (grid features), while EfficientNetB0 (for region features) processes the overall image context using its compound scaling technique. The outputs from both models undergo global average pooling to produce compact feature vectors representing the visual information, which are then normalized using L2 normalization.

- **Feature Integration Using Attention Mechanism**
  An attention mechanism is applied to integrate features from text and images. The text features from DeBERTaV2 serve as the query, while the image features from ResNet50 and EfficientNetB0 act as keys and values. This mechanism aligns and focuses on relevant visual elements based on textual context, enhancing the integration between modalities. The output from the attention layer is concatenated with the text features to form a unified multimodal representation.

- **Classification Layer**
  The integrated feature representation is passed through fully connected dense layers that learn the interactions between text and image features. Regularization techniques, such as dropout and batch normalization, are applied to reduce overfitting and enhance generalization. The final classification layer uses a sigmoid activation function to output probabilities for each of the 22 persuasive techniques, enabling multi-label classification. The model is trained using Focal Loss to address class imbalance, focusing on less frequent techniques, and utilizes the AdamW optimizer for efficient weight updates and improved generalization.

Model Design for Enhancement Task: Multi-Label Sequence Tagging

The model architecture is built upon a pre-trained TFBertModel (bert-base-uncased) to perform multi-label sequence tagging for identifying persuasive techniques in text spans. The core of the architecture involves leveraging BERT to generate contextual embeddings for each token in the sequence, with an option to unfreeze the last few layers (based on the trainable_layers hyperparameter) for fine-tuning specific to the task. This allows the model to adapt more effectively to the span classification problem.

Following the BERT layer, a dense layer with ReLU activation processes the token embeddings, transforming them into a form suitable for classification. The number of units in this dense layer is controlled by the dense_units hyperparameter, providing flexibility to adjust the model's capacity. To prevent overfitting, a dropout layer is applied with a rate determined by the dropout_rate hyperparameter, promoting better generalization during training. The final classification layer uses a sigmoid activation function, producing multi-label predictions for each token, which allows the model to identify multiple techniques associated with specific text spans.

The model is compiled with the AdamW optimizer, using a specified learning rate and a weight decay of to enhance generalization. The loss function employed is Focal Loss (with parameters alpha=0.5 and gamma=3), effectively handling class imbalance by focusing the training process on difficult-to-classify instances. The evaluation metrics include precision, recall, and a custom F1 score, providing a thorough assessment of the model's performance across multiple labels. This architecture, with its integration of fine-tuning and **r**egularization techniques, ensures effective multi-label span classification, making it well-suited for detecting persuasive techniques in text.

Investigation

The investigation process focused on exploring various models, configurations, and techniques to optimize the detection of persuasion techniques in memes. Initially, different architectures for image feature extraction were evaluated, including ResNet50, DenseNet121, and EfficientNetB0, to determine the most effective approach for capturing visual patterns. After extensive tuning, EfficientNetB0 and ResNet50 were selected due to their superior ability to capture both fine details and overall image context, crucial for identifying persuasive elements in memes.

For text feature extraction, multiple transformer-based models were tested, such as DistilBERT, RoBERTa, and DeBERTaV2. Each model's performance was assessed based on its ability to capture and represent patterns in persuasive language. DeBERTaV2 was ultimately chosen due to its enhanced performance and efficiency, offering a balanced trade-off between computational cost and accuracy.

In addition to model selection, the investigation included experimenting with different configurations of layer freezing for fine-tuning. This process involved selectively unfreezing specific layers of the pre-trained models (e.g., the last few layers of ResNet50 and EfficientNetB0) to allow for task-specific learning while retaining the benefits of pre-trained weights. Several configurations were tested, including fully freezing the initial layers and progressively unfreezing later layers across epochs. This approach aimed to strike an optimal balance between training efficiency and model performance.

Exploratory Data Analysis also highlighted the importance of addressing class imbalance within the dataset and normalisation for feature scaling. Focal Loss was applied, and its hyperparameters (e.g., alpha and gamma) were fine-tuned using Keras Tuner to focus more effectively on underrepresented classes. Additional methods, such as oversampling techniques and adjusting class weights, were explored to ensure minority classes received sufficient attention during training.

Hyperparameter optimization was another critical aspect of the investigation. Key parameters like learning rate, dropout rate, and the number of units in dense layers were tuned using Keras Tuner. By systematically searching for the best configurations, the optimization process aimed to improve performance, particularly for minority classes, while ensuring training stability and convergence.

Overall, the investigation's comprehensive approach, involving model selection, fine-tuning strategies, and hyperparameter optimization, contributed significantly to improving the model's ability to generalize across diverse and complex examples of persuasion techniques in memes.

Evaluation

The evaluation of the model was conducted separately for both the main and enhancement tasks using relevant metrics to assess performance comprehensively. For the main task, the model's effectiveness was measured through micro F1, and macro F1 scores. The primary focus was on micro F1 scores. The

micro F1 score, which aggregates true positives, false positives, and false negatives across all classes, showed moderate overall performance with a score of 0.37 on the test set. In contrast, the macro F1 score, which averages the F1 scores for each class equally, highlighted challenges with minority classes, achieving a score of 0.16 on the test set. This discrepancy indicated that the model struggled to detect less frequent techniques effectively.
For the enhancement task, the evaluation centred on the overall F1 score and accuracy specific to sequence tagging. The model achieved a very low F1 score of 0.039 on the test set, demonstrating limited effectiveness in identifying text spans associated with persuasion techniques, especially for underrepresented classes.

Both tasks were evaluated using independent test datasets to validate the model's generalization ability. Training and validation curves were monitored throughout the process to detect signs of overfitting. Divergence in validation loss from training loss in later epochs revealed some overfitting, addressed through dropout regularization and progressive layer unfreezing. The final models were selected based on the highest F1 scores observed on the test set to ensure balanced performance.

These results indicate that while the model can detect well-represented classes, it requires further improvements to enhance its effectiveness in identifying minority classes and text spans accurately.

Findings

The model evaluation and analysis yielded several key findings regarding its performance and limitations.

- **Strengths and Weaknesses**
  The model demonstrated proficiency in identifying well-represented classes, achieving a micro F1 score of 0.37 for the main task. The multimodal approach, combining text and image features, proved effective, confirming that integrating these modalities enhances meme classification accuracy. However, the model struggled with underrepresented classes, as evidenced by the macro F1 score of 0.16. The class imbalance in the dataset led to misclassification or omission of these minority classes, indicating the need for further balancing techniques.

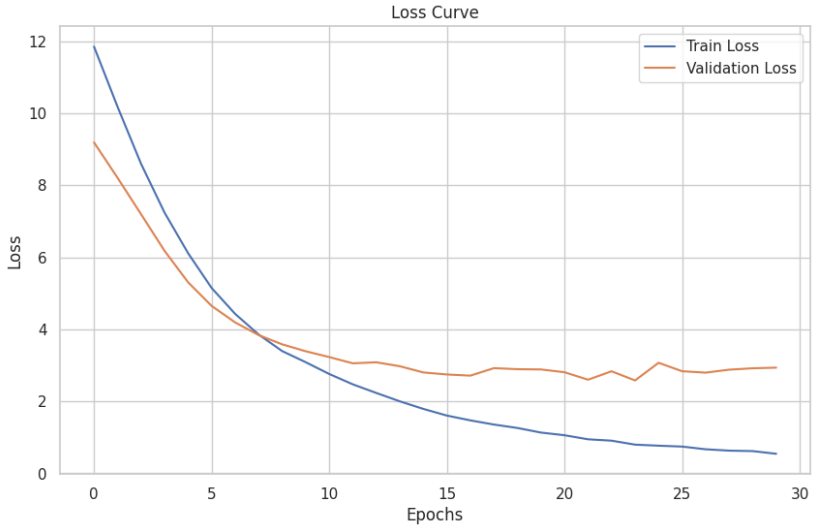
- **Overfitting**
  An analysis of the training and validation curves revealed slight overfitting, where validation loss diverged from training loss in the later epochs. This suggests that while the model learned well from the training data, it struggled to generalize, likely due to the small dataset size and insufficient representation of minority classes.

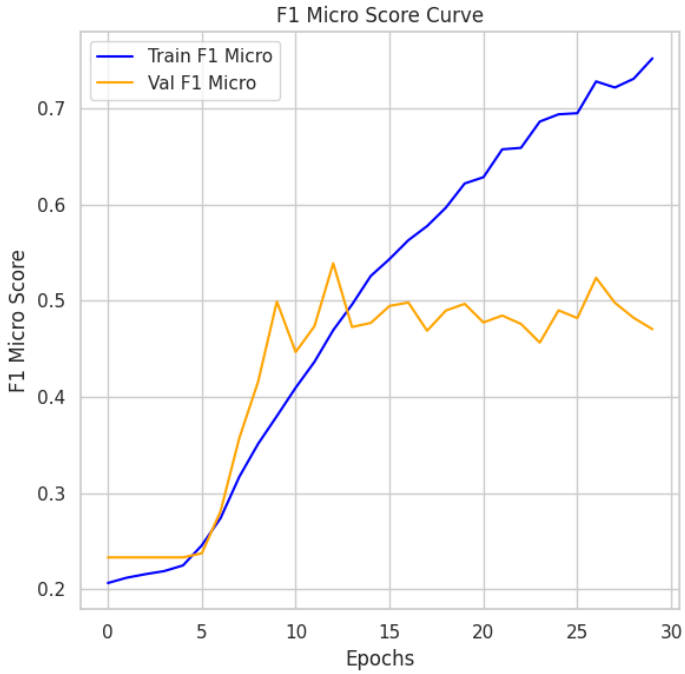
- **Ethical Considerations**
  The class imbalance resulted in biased predictions, with majority classes being more accurately identified. This imbalance risked reinforcing stereotypes and underrepresenting important minority persuasive strategies. Additionally, privacy concerns were noted, as memes often contain sensitive content. Careful handling and anonymization are necessary to protect individuals' rights.

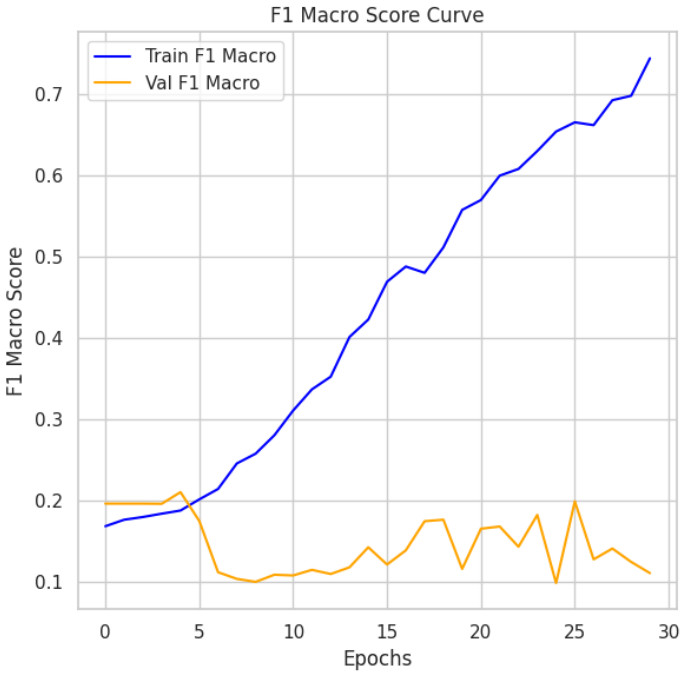
- **Future Improvements**
  To address these challenges, employing data augmentation and oversampling techniques like SMOTE could help balance the dataset, particularly for minority classes. Incorporating advanced architectures such as transformers with cross-attention mechanisms may also improve feature fusion between text and images for more precise classifications. Expanding the dataset size and using greater computational resources would enable deeper fine-tuning, enhancing the model's generalization capabilities across all classes.

These findings highlight the model's potential and the areas requiring improvement, particularly in handling diverse and imbalanced data to achieve more comprehensive and unbiased predictions.

Plots for Main Task:







Plots for Enhancement Task: