Deep Learning (COSC 2779) – Assignment 2 – 2024

Linh Thuy Do (s3927777)

1 Problem Definition and Background

The problem tackled in this project is the automatic detection of persuasion techniques in memes, a task central to addressing the growing problem of disinformation spread via social media. Memes, which combine visual and textual content, often use various propaganda techniques such as "Loaded Language," "Name Calling," and "Smears" to influence public opinion. This problem, drawn from the SemEval-2021 Task 6 dataset, is a multi-label classification challenge, where multiple persuasion techniques can appear simultaneously in each meme.

The goal is to build a deep learning system that classifies each meme by identifying which techniques it uses. This task is complex due to:

- The need to process and integrate both visual and textual modalities.
- The imbalanced dataset, where certain techniques are under-represented.
- The nuanced and context-dependent nature of memes, often requiring advanced models to capture implicit meanings.

2 Evaluation Framework

Two key performance metrics were selected for this multi-label classification task: Micro F1-Score and Macro F1-Score. The Micro F1-Score is calculated by considering the overall true positives, false negatives, and false positives across all labels. It gives more weight to frequently occurring classes and is suitable for handling imbalanced datasets. On the other hand, the Macro F1-Score calculates the F1-score for each class independently and averages them, treating all classes equally. This metric is particularly useful for evaluating the model's ability to handle minority classes. Together, these metrics ensure that both frequent and rare persuasion techniques are evaluated effectively.

The dataset is split into 70% training, 15% validation, and 15% test sets. The training set is used to fit the model, while the validation set is used to monitor the model's performance during training and fine-tune hyperparameters to prevent overfitting. The test set is kept entirely separate from the training and validation processes, and it is used only for the final evaluation to ensure unbiased assessment of the model’s generalizability to unseen data. This fixed split ensures the model is tested on data it has never seen before, mimicking real-world scenarios.

Given the imbalance in the dataset, where certain persuasion techniques appear much more frequently than others, specialized techniques were implemented to improve performance on under-represented classes. One such method is Focal Loss, a modified loss function that reduces the weight of easy-to-classify examples, enabling the model to focus more on harder examples, often those from minority classes (Lin et al. 2018). Additionally, threshold tuning was applied to each individual label. Instead of using the standard threshold of 0.5 for classifying labels, custom thresholds were set based on each label's distribution in the dataset, helping the model better distinguish between minority classes.

To prevent overfitting in this task, during the training model, I used the early stopping. The model's performance on the validation set was continuously monitored, and training was stopped once the validation performance stopped improving. This ensured that the model did not continue to train on the same data and potentially learn patterns that did not generalize to new data. Additionally, a learning rate scheduler was applied, where the learning rate was gradually reduced after 10 epochs, allowing for more refined updates to the model weights in the later stages of training. Finally, once the model was fully trained and fine-tuned, it was evaluated on the independent test set, which had been kept separate throughout the process.

3 Approach & Justifications

Given the challenges posed by a multi-label classification task, particularly with class imbalance, we adopted a robust approach to ensure both efficiency and accuracy. Our system integrated multiple components that tackled key issues such as class imbalance, complex input features, and computational limitations, allowing us to optimize performance for this task.

3.1 Dealing with Class Imbalance:

Handling class imbalance in multi-label classification is complex. Traditional techniques like oversampling and under sampling are not practical due to the risk of overfitting and inadequate representation of minority classes. Instead, we opted to modify the loss function. Specifically, Focal Loss was employed. Unlike binary cross-entropy, Focal Loss adjusts to the difficulty of classification, down-weighting easy examples while focusing on hard-to-classify instances. This adjustment directly mitigates the imbalance by emphasizing minority classes that tend to be more difficult to classify.

However, as Focal Loss alone might not fully resolve the class imbalance, particularly given the high skewness in label distribution, we further fine-tuned classification thresholds for each label. Rather than sticking with the default 0.5 threshold, custom thresholds were optimized through experiments, aiming to improve precision and recall for underrepresented classes. This decision ensured a more balanced performance across all labels, particularly improving the macro F1 score, which averages performance across all classes.

3.2 Model Architecture:

For the backbone model, we adopted a dual-branch architecture, handling both the text and image components of each meme. By employing pre-trained models (ResNet50 for image features and ALBERT for text features), we leveraged the power of transfer learning. This decision was motivated by the limited size of the dataset and the complexity of meme language and imagery, which often require substantial pre-training to capture intricate relationships.

We selected ALBERT, a lighter version of BERT, for text processing due to its efficiency and ability to handle complex language patterns with fewer resources. On the visual side, ResNet50, pre-trained on ImageNet, was chosen due to its proven ability to extract meaningful image features even from small datasets. Both branches processed their respective inputs, and the outputs were concatenated to form a unified representation, which was passed through dense layers to predict multiple labels.

This architecture enabled the model to simultaneously analyze the text and image components of each meme, providing a comprehensive understanding necessary for accurate classification of persuasion techniques.

4 Experiments & Tuning

4.1 Model Training:

To ensure the model generalized well, we froze the ResNet50 weights during training, relying on its pre-trained capabilities to extract image features efficiently. Meanwhile, ALBERT was fine-tuned to adapt to the nuances of meme language. This hybrid approach minimized overfitting while still allowing the model to learn task-specific features from the dataset.

We employed the Nadam optimizer, a blend of RMSProp and Nesterov accelerated gradient, to facilitate faster convergence, especially when fine-tuning the text branch. Additionally, a custom learning rate schedule was designed, with the learning rate remaining constant for the first 10 epochs and gradually decreasing after that. This strategy allowed the model to stabilize before refining its weights delicately in later epochs.

4.2 Hyperparameter Tuning:

To optimize model performance, we conducted a series of experiments focused on adjusting key hyperparameters, including learning rate, batch size, and threshold values for classification.

- Learning rate: We experimented with values between 0.0001 and 0.000005. The final selection of 0.000005 provided the best balance between convergence speed and generalization.
- Batch size: We tested batch sizes ranging from 8 to 32. A batch size of 16 proved optimal, balancing memory constraints and model performance.

4.3 Threshold Tuning:

A major area of tuning involved setting individualized thresholds for each label. We initially trained the model using a default 0.5 threshold for all labels, but this resulted in poor macro F1 scores. To address this, we implemented a custom threshold tuning procedure. Using validation data, we optimized the thresholds for each label by maximizing the F1 score, resulting in improved recall for minority classes.

4.4 Evaluation and Performance:

The model's performance was evaluated using both micro and macro F1 scores. While the micro F1 score provides an overall accuracy by focusing on the dominant classes, the macro F1 score evaluates the model's ability to handle underrepresented labels. By tuning the thresholds and adopting Focal Loss, we managed to achieve a reasonable balance between both metrics, avoiding bias towards majority classes.

In conclusion, the experiments demonstrated that focusing on advanced loss functions and fine-tuning, along with a multi-modal approach, can lead to substantial improvements in classifying memes, particularly under challenging conditions like class imbalance.

5 Ultimate Judgment, Analysis & Limitations

The final approach selected for this task is a dual transformer model that integrates both ResNet50 for image processing and ALBERT for text processing. This multi-modal architecture processes images and text independently before combining the features for multi-label classification. The model leverages pre-trained networks—ResNet50 for image features and ALBERT for textual understanding—before passing the processed features through transformers to handle the relationships between the two modalities. The output layer applies a sigmoid activation function, making it suitable for the multi-label classification required to detect multiple persuasion techniques in memes.

This approach was selected due to the specific nature of the problem: memes combine both textual and visual information to convey meaning, making it essential to use a model capable of handling both modalities effectively. Simpler models, which focus solely on text or images, would miss the nuances created by the interaction between these two components. By using a dual transformer model, the system can better capture these interactions (Messina et al. 2021), which are often critical for detecting complex persuasion techniques like "Loaded Language" or "Transfer."

The decision to use pre-trained models (ResNet50 and ALBERT) was based on their demonstrated success in their respective domains—image recognition and language understanding. These models allow for a reduction in training time while benefiting from the extensive knowledge they have gained from large datasets (ImageNet for ResNet50 and large text corpora for ALBERT). Fine-tuning these models on the meme dataset allows them to adapt to the specific challenges posed by meme content, such as sarcasm or irony in text and implicit meaning in images.

Several configurations were explored before finalizing the model:

- **Focal Loss** was implemented to handle the severe class imbalance observed in the dataset, where some persuasion techniques are significantly under-represented. Focal Loss proved effective by focusing the model's learning on harder-to-classify, minority classes.
- **Learning rate** was initially set to 0.00001, with a **learning rate scheduler** applied after 10 epochs, gradually reducing it to improve convergence during fine-tuning.
- **Early stopping** was implemented to prevent overfitting, halting training once the validation performance stopped improving, which helped maintain the model's generalizability.
- The model used a **batch size of 16**, balancing computational efficiency and performance during training.

Other approaches were tested, including using different architectures like **EfficientNet** for image processing and **BERT** for text, but these models either resulted in slower convergence or did not provide significant improvements over ResNet50 and ALBERT. Attempts to use **simple CNNs or LSTM models** for either modality also resulted in lower performance, particularly in capturing complex relationships between text and images.

Despite the promising results, the model does have some limitations that must be addressed for real-world deployment:

1. **Performance on Minority Classes**: Although Focal Loss improved the model's focus on under-represented classes, the **Macro F1-Score** remained relatively low. This indicates that minority persuasion techniques, such as "Repetition" or "Causal Oversimplification," are still harder for the model to classify accurately. For real-world use, increasing the dataset size, particularly for these minority classes, would likely improve performance. Additionally, using **data augmentation** for the text and image components could help further address the imbalance.
2. **Domain-Specific Fine-Tuning**: The reliance on pre-trained models like ALBERT and ResNet50, while beneficial for general understanding, may not fully capture the specific nuances of memes, which often involve humor, irony, or evolving cultural references. A possible improvement for real-world implementation would be fine-tuning on a larger, more diverse meme dataset that includes more varied and complex persuasion techniques.
3. **Model Efficiency**: The model's use of transformers and pre-trained networks requires considerable computational resources, which may limit its scalability. To address this, a **lighter version** of the model or **knowledge distillation** could be applied, where a smaller model is trained to mimic the performance of the larger model, allowing for faster and more efficient inference without sacrificing much accuracy.
4. **Evolving Nature of Memes**: In real-world applications, memes and the language used within them constantly evolve. This dynamic nature presents a challenge to static models. To overcome this, the system would require **periodic retraining** on new meme datasets to keep up with emerging trends and new techniques of persuasion, ensuring the model remains effective over time.

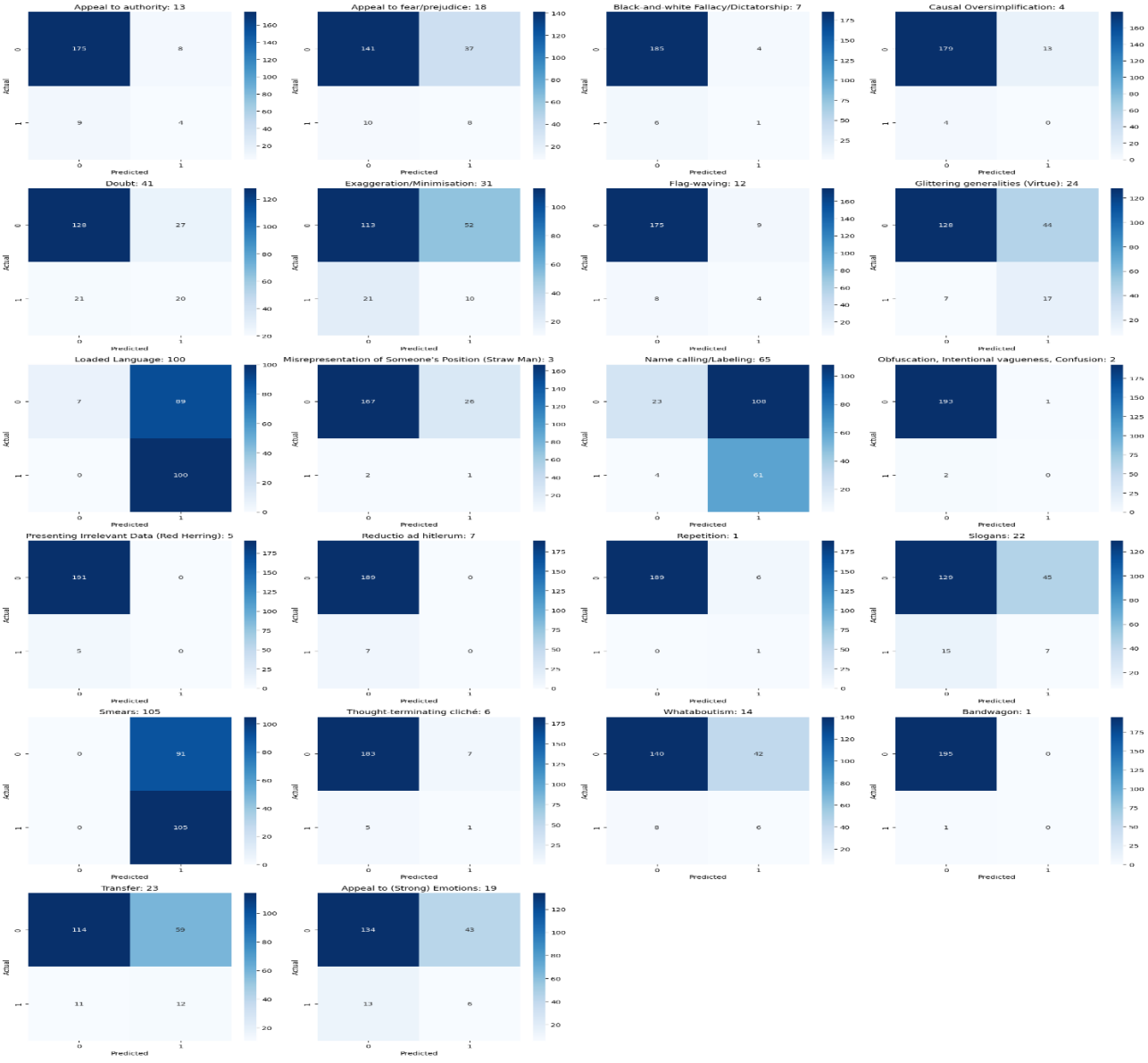
In conclusion, the selected approach proves to be highly effective for detecting persuasion techniques in memes. The choice of architecture, parameter settings, and handling of class imbalance all contributed to the model's strong performance. However, further improvements in handling minority classes, efficiency, and adaptability will be necessary for real-world deployment, where the system must maintain high accuracy across a constantly evolving dataset.

6 References:

Lin T, Goyal P, Girshick R, He K and Dollár P (7 February 2018) 'Focal Loss for Dense Object Detection', *Cornell University*, accessed 20 September 2024. https://arxiv.org/abs/1708.02002

Messina N, Falchi F, Gennaro C and Amato G (5 August 2021) 'AIMHatSemEval-2021 Task 6: Multimodal Classification Using an Ensemble of Transformer Models', *the 15th International Workshop on Semantic Evaluation*, accessed 13 September 2024. https://aclanthology.org/2021.semeval-1.140.pdf

7 Appendix



Appendix 1: Confusion Matrix of the Final Model