ASSIGNMENT 2

SEQUENCE PROCESSING FOR DEEP CONVOLUTIONAL NEURAL NETWORKS

SEMEVAL-2021 TASK 6:
DETECTION OF PERSUASION TECHNIQUES IN TEXTS AND IMAGES

COSC2972 DEEP LEARNING
(UNDERGRADUTE LEVEL)

BY: OISIN SOL EMLYN AEONN

STUDENT ID: S3952320

TABLE OF CONTENTS

ABSTRACT

This report presents my approach to detecting ***Persuasion Techniques*** in ***Multi-Modal*** content, specifically memes combining text and images. The work is based on a recent ***Natural Language Processing (NLP)*** challenge and explores methods for integrating linguistic and visual information in ***Classification*** *tasks*.

My methodology consists of a comprehensive **Machine Learning Approach**: conducting an **EDA**, creating a **Baseline Model**, making iterative improvements through advanced **Hyperparameter Tuning**, and **Fine-Tuning** techniques to build a final optimized version.

For visual feature extraction, I utilized **MobileNetV2** and **EfficientNetB0/B1**, chosen for their optimal *performance-to-GFlops* ratio within my computational constraints. These visual features are combined with contextualized word *embeddings* and sentence representations from various sizes of **BERT**-like models, leveraging their ability to capture nuanced linguistic information.

To address the challenge of the small ***Dataset*** size and ***Class Imbalance***, I implemented ***Focal Loss, Data Augmentation, and Class Weighting***.

This approach yielded a ***F1 Micro Score*** of ***X*** on the ***Test Set***, comparable to top-performing models on this ***Data Split***. This result is particularly impressive given the constraints on time and computational resources, demonstrating the effectiveness of my approach in tackling this complex ***Multi-Modal Classification*** *task* while striving for robust *performance* on ***Unseen Data***.

PROBLEM DEFINITION AND BACKGROUND

This project addresses the **'*SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images*'** (https://aclanthology.org/2021.semeval-1.7.pdf) for *Assignment 2* of *Deep Learning* ***COSC2972*** *(undergraduate level)* at **RMIT University**. The challenge involves ***Multi-Modal Classification*** aiming to identify the ***presence*** of *22 persuasion techniques (plus a 23rd no label)* in memes, along with their corresponding sentences: integrating both linguistic and visual information.

Ethical Context:

This task has ***real-world applications*** in the complex and often ***controversial*** area of addressing online disinformation and propaganda, raising important ethical considerations about information control and freedom of expression.

Figure 1: Dataset Split

The ***Dataset*** consists of ***950 samples*** which have been ***pre-split*** as the following:

- ***Training set: 687 samples (72.32%),***
- ***Development set: 63 samples (6.63%),***
- ***Test set: 200 samples (21.05%).***

Each sample includes an ***image, pre-extracted text, and the corresponding labels,*** while the ***Enhancement Task*** added the specific sentence related to each label**.** The limited dataset size presented significant challenges for deep learning approaches, necessitating careful consideration of model architecture and training strategies to ensure ***generalizability*** to the ***Unseen Test Set***. Additionally, the *multi-modal* nature of the data (combining text and images) and the Imbalanced class distribution posed further complexities in developing an effective solution.

EVALUATION FRAMEWORK

Given the dataset's characteristics and the complexity of the task, I carefully selected an evaluation framework to assess model performance accurately overall. The primary goal was to predict multiple labels present in each meme, considering both image and text content. The framework utilized the following key metrics:

Micro F1-Score: This serves as the primary metric, providing a balanced measure of precision and recall across all instances. It is particularly suitable for multi-label classification tasks with imbalanced classes, as it gives equal weight to each sample.

Macro F1-Score: Used as a secondary metric, it offers insights into the model's performance across all classes, regardless of their frequency. This is crucial for understanding how well the model performs on less common persuasion techniques.

Position Accuracy Score: For the Enhancement Task we require a measure of the average absolute difference between predicted and actual start / end character positions of labelled text fragments.

These metrics were chosen to address the imbalanced nature of the label distribution in the dataset and to maintain consistency with existing research, and the challenge, ensuring comparability with other models see [X].

APPROACH & JUSTIFICATIONS

Data Preprocessing and Exploratory Data Analysis (EDA): The initial step involved data ingestion and reformatting. The original JSON format was converted to CSV, with labels concatenated into a single string. Image paths were appended to ensure accessibility. Extensive EDA revealed significant variations in image dimensions (200-1800 pixels), necessitating resizing. Images were predominantly square-shaped, justifying a squish methodology for resizing. Analysis of image characteristics showed normal distributions of sharpness and intensity, with positively skewed entropy. Notable issues included incorrect file extensions (JPEG renamed as PNG) and repetitive meme templates with varying labels. Text analysis examined word frequency and length distributions across labels.

Dataset Challenges:

The dataset presented challenges including small size, class imbalance, and a limited validation set (63 images) lacking representation of all labels. To address these issues, class weighting was implemented instead of oversampling to mitigate overfitting on the training data, I also planned to utilise some augmentation on both the images and textual data.

Model Architecture:

The core architecture was inspired by the *Alpha Team's* approach (see X). It combines a **BERT** model for text processing with a transfer learning **CNN**. Given my computational constraints I opted for **MobileNetV2** and **EfficientNetB0/B1** for image feature extraction instead of **ResNet-50** (see X). Images were resized to *224x224* using squish methodology, reducing the input size. The multi-modal fusion was achieved through concatenation of text and image features, followed by dense layers with ReLU activation and dropout for regularization.

Training Strategy:

Initial training utilized a small version of **BERT**, specifically: **L-4_H-512_A-8** model with ***frozen weights*** for both **BERT** and **MobileNetV2** to establish a **Baseline**. This approach yielded a ***Training Micro F1-Score*** of ***0.5327***, and ***0.4983*** on the ***Development (Validation) Set*** after ***20 epochs*** with a ***batch size*** of ***32***.

Model Refinement:

Subsequent iterations incorporated techniques such as ***Gradient Accumulation***, C***ross-Entropy Loss*** (maybe swap for Focal Loss), ***Early Stopping, and Learning Rate Reduction*** to combat ***Overfitting***.

Data Augmentation techniques were applied, including ***random brightness**, **contrast**, and **saturation*** adjustments for images, and ***synonym substitution*** *and* ***back-translation*** for text. Despite a slight performance decrease, these techniques were retained to enhance ***generalizability*** beyond the ***Development Set***. Rationale: Data augmentation helps in increasing the effective size of our training set and improving model generalization.

I also *experimented* with the **BERT**-model size.

Hyperparameter Tuning:

Add limitation on epochs

A ***Grid Search*** was employed for hyperparameter optimization, focusing on ***Dropout Rates***, ***Optimizer Selection***, *and* ***Batch Size***. Exploring the following key hyperparameters:

- ~~Learning rate: Explored range 1e-5 to 5e-5~~
- ~~Batch size: Tested 4, 8, and 16~~

- ~~Focal loss parameters: α (0.75 to 0.95) and γ (1.0 to 3.0)~~

The optimal configuration was found to be: ***(dropout rate: 0.2, optimizer: Adam, batch size: 16)*** resulted in improved ***Micro F1-Scores (training: 0.7307, validation: 0.5430).***

Fine-tuning:

The final stage involved ***unfreezing*** **BERT** and **MobileNetV2** weights for ***Fine-Tuning*** with a ***low learning rate***, to optimise the model for the task. This led to a final Micro-F1 Score of: ***(training: 0., validation: 0.).***

Enhancement Task:

In addition to predicting the persuasion techniques label, I tackled the **Enhancement Task** intended for a ***High Distinction*** of identifying the specific spans of text covered by each technique. This multi-label sequence tagging task is like Named Entity Recognition but more complex due to the potential overlap of techniques within the text.

For this enhancement, I developed a model that not only predicts the labels but also identifies the start and end character positions of the text fragments associated with each label. To evaluate this aspect, I will use the previously defined ***Position Accuracy Score.***

Rationale: Transformer architectures have shown superior performance in capturing long-range dependencies and complex interactions between modalities.

3.4 Loss Function

To address the class imbalance issue, I adopted focal loss instead of standard cross-entropy loss.

Rationale: Focal loss helps in focusing the model on hard-to-classify examples, which is particularly beneficial in our imbalanced multi-label scenario.

EXPERIMENTS & TUNING

I did a lot of experimental models which even though I got more performance on some I did not implement into the machine learning flow, as they did require quite a bit more computing power.

Some I also did later on as experiments, or for the Extension Task and did not have time to integrate them fully into the Main Task as intended.

I also did do ensembling as it is very prevalent in some of the literatures like X, Y, Z but on large scale models it was impractical for me to run without a larger compute budget.

For base model I tried near 30 variants of Bert models starting with Bert small h4 going all the way up to Bert large 24h and trying other variants like Electra, Albert, deberta, etc - I spent almost a day just trying to find the largest language based transformer I could run locally on my computer and it these models would or could improve the performance on the task but counter-intuitively or in retrospect making considerable sense these models were not always the best as they do require quite a bit more training for them to understand the specific task as well as in my tests I could not dedicate more than 10 epochs per model due to time and compute constraints.

ULTIMATE JUDGMENT, ANALYSIS & LIMITATIONS

Model Performance for Main Task:

<TABLE FIGURE X>

Comparison of our performance to others.

4.
5. Context understanding: The model struggles with memes that require broader contextual or cultural knowledge.
6. Computational resources: High-performing models like ERNIE-ViL are computationally intensive, potentially limiting real-time applications. EPOCHS ON

-discuss how little the multi-modal part actually matters in our scenario and I and others found that the combination with images only increased our F1 score by a few % while

Show a table of my models performance verses benchmarks as well as what I would do next time given more time, less constraints around compute & accessibility of weights .

After extensive experimentation, the ERNIE-ViL model emerged as the best performer, achieving an F1 score of 57.14 on the test set. This outperformed both DeBERTa+ResNet50 (55.96) and DeBERTa+BUTD (56.21) configurations.

KEY FINDINGS:

1. ~~Multimodal pre-training advantage: ERNIE-ViL's superior performance can be attributed to its pre-training on large-scale image-caption data, allowing it to learn more general and robust multimodal representations.~~
2. ~~Visual feature impact: The BUTD object detection features slightly outperformed ResNet50 grid features, suggesting the importance of salient region information in meme analysis.~~
3. ~~Focal loss effectiveness: Switching from cross-entropy to focal loss improved F1 scores by approximately 4 points, demonstrating its efficacy in handling class imbalance.~~

Error Analysis:

-Go through case examples and explain why e.g. why loaded language classified as smth else or smth

Examining misclassifications revealed several patterns:

- Confusion between closely related techniques (e.g., "Loaded Language" vs "Name Calling/Labeling")
- Difficulty in detecting subtle visual cues
- Challenges with memes requiring external context or cultural knowledge – use of slogans, or wording in complex contexts or under used ways.

Limitations and Future Work:

1. Dataset size: The small dataset remains a significant limitation, potentially hindering the model's ability to generalize to diverse real-world memes.
2. Visual-textual alignment: Current approaches may not fully capture the intricate relationships between text and image elements in memes.
3. Talk about the literature and what we could have done differently and other teams did - particularly around ensembles and larger models of Bert which require more compute but also more time to train. More images would have been the best thing we could have been offered as the less than 1000 was not merciful in variety or anything. I would have loved to see this as then this model could be taken outside the realm of the US election - liberal vs republicans and coronavirus memes as the real internet has many more

   HYPERPARAMS, NUMBER OF HYPERPARAMS, LENGTH AND SIZE OF MODEL AND RETRAINING.

   One particular thing I would consider next time is using a bur stable cluster for gpu compute to parallelise training of language based models like Bert as this would increase the speed at which a model could be trained, and increase the amount of hyperparams I can turn at once.

Future work directions:

- Explore more sophisticated visual-textual alignment techniques
- Investigate ways to incorporate external knowledge bases
- Develop more efficient model architectures for real-time processing

Collect and annotate a larger, more diverse dataset of memes USING LLMS THAT ARE MULTI-MODAL – THIS WILL ALLOW US TO CREATE A VERY GOOD TASK SPECIFIC IF THESE ARE AUTHENTICATED BY A HUMAN Another thing is having access to more data for training would be very useful as we had under 700 samples for training and under 300 for validation and test. This presented quite a considerable challenge and I think one very smart way that we can easily increase the size of the dataset would be to utilise existing LLM with vision capability to 1. Ocr the text from the image and 2. Define the labels associated with said meme. These could tehn be fed to a human who could confirm and validate the labels before increasing our datasets.

-

Model Performance for Enhancement Task:

<TABLE FIGURE Y>

CONCLUSION

This project demonstrated the effectiveness of transfer learning and multimodal fusion techniques in tackling the challenging task of persuasion technique detection in memes. The ERNIE-ViL model, combined with focal loss and careful data augmentation, proved most effective in navigating the constraints of a small dataset.

While the achieved performance is promising, there's significant room for improvement, particularly in handling subtle persuasion techniques and memes requiring deep contextual understanding. Future work should focus on more sophisticated multimodal integration, larger datasets, and incorporation of external knowledge.

The developed system shows potential for real-world applications in content moderation and digital literacy education, though further refinement is needed for practical deployment.

APPENDIX

1

2

3

REFERENCES

Dimitrov, D., et al. (2021). SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. Proceedings of the 15th International Workshop on Semantic Evaluation, 70-98.

Yu, F., et al. (2020). ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph.

Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

He, K., et al. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

Lin, T.Y., et al. (2017). Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision, 2980-2988.