Multimodal Sentiment Analysis Project

Luca Alfano

Introduction

This report summarizes the methodology, results, and observations from the implementation of a multimodal sentiment analysis model using text and image data from the enriched MVSA Twitter dataset. The project integrates natural language processing (NLP) and computer vision techniques to classify sentiment based on textual and visual content.

Data Preprocessing Steps

Text Preprocessing

1. **Tokenization**: The text data is tokenized using a pre-trained tokenizer from Hugging Face's ***AutoTokenizer***. It converts raw text into token IDs suitable for input to a transformer-based NLP model.
2. **Padding and Truncation**: Text inputs are padded to ensure uniform length (128 tokens) and truncated if longer than the maximum length.
3. **Embedding Extraction**: Using the pre-trained ***AutoModel***, text embeddings are generated by averaging the last hidden states of the transformer model.

Image Preprocessing

1. **Conversion to RGB**: Images are converted to RGB format to ensure compatibility with the MobileNetV2 model.
2. **Feature Extraction**: Images are processed using ***MobileNetV2ImageProcessor*** and passed through a modified MobileNetV2 model (classifier layers removed). The output feature maps are averaged spatially to create embeddings.
3. **Normalization**: Image pixel values are normalized as required by MobileNetV2.

Model Architectures

NLP Model

The text embeddings are extracted using a transformer-based model, leveraging its contextual understanding of language. First, the model is fine-tuned for the task as a stand-alone model; then, the embeddings are generated in the Multi-modal model.

Computer Vision Model

Similar to the NLP model the MobileNetV2 architecture is used for image feature extraction, fine-tuning the stand-alone model and later, in the Multi-modal model, the classifier layers are removed to focus on generating embeddings.

Multi-modal Fusion Model

The ***MultimodalModel*** freezes the layers of the NLP and CV models that were fine-tuned separately, and combines text and image embeddings using:

- **Linear Projections**: Text embeddings are projected to 128 dimensions, while image embeddings are projected to 1280 dimensions.
- **Multi-Head Attention Mechanism**: A multi-head attention layer fuses the textual and visual features into a unified representation. This mechanism allows the model to focus on relevant parts of both modalities for classification tasks.
- **Classifier**: A feed-forward neural network with ReLU activation and dropout is used for sentiment classification.

Attention Mechanism

The attention mechanism plays a crucial role in multimodal fusion:

1. How It Works:
    - Multi-head attention computes weighted relationships between text and image embeddings, enabling the model to prioritize certain features over others based on context.
    - The attention layer outputs a combined representation that captures inter-modality dependencies.
2. Relevance:
    - Attention is essential for identifying correlations between textual sentiment cues (e.g., words like "happy") and visual sentiment indicators (e.g., smiling faces).
    - It enhances interpretability by focusing on salient features in both modalities.

Results and Observations

Performance Metrics

- Since the problem is framed as a multi-class classification the model was trained using a weighted F1 metric for optimization, utilizing the ***evaluate*** library from Hugging Face. The final evaluation and conclusions on the model also consider accuracy and recall scores, leveraging Confusion Matrix and Classification Report from ***scikit-learn***.
- Results indicate a balanced performance in sentiment classification across the three classes: positive, negative, neutral. The model achieves an overall accuracy of 72%, which is reasonable but suggests that there is still potential for enhancement, particularly in the negative and neutral classes where some confusion is evident both from the Matrix and from the metrics (0.81 f1 for the positive label, 0.68 for the neutral and 0.67 for the negative) More details about the model performance are available in the notebook.

Challenges Faced

1. Dataset Complexity:
   - While working separately on the text or image data was straight-forward, generating the emebddings in a way that could potentially be reused on future data points presented quite the challenge, requiring the implementation of a custom ***MultimodalPreprocessor*** (able to handle the preprocessing steps for both text and image data, including tokenization, feature extraction, and model preparation) and a ***MultimodalDataset*** (expanding the Torch Dataset class for the task at hand). Both classes are pickled and available for further use.
   - Variability in image quality and text length posed challenges during preprocessing.
2. Computational Constraints:
   - Training large models on multi-modal data was resource-intensive, requiring GPU acceleration and careful management of the resources available.

Conclusion

The multimodal sentiment analysis model successfully integrates NLP and computer vision techniques using attention mechanisms for sentiment classification. While challenges such as dataset variability and computational demands were encountered, the project demonstrates the effectiveness of multimodal approaches in understanding complex social media data.
To further improve the performance of the model hyper-parameters tuning should be considered:

- Learning Rate
- Weight Decay
- The number of heads for the attention mechanisms
- The number of training Epochs
- The Batch size of the training data loader
- The Loss Criterion
- The Optimizer

Are all examples of hyper parameters that are independent from the actual model optimization and that can be optimized, for a better performance using libraries like ***optuna*** and additional training effort.

Appendix

Resources mentioned in the report

MultimodalPreprocessor

MultimodalDataset

Multimodal Model

GitHub repository

Google Colab Notebook