Deep Learning Task Report

Introduction

This report outlines the experimentation with different architectures and strategies to solve a sentiment classification task using the SST-2 dataset. The goal was to progressively enhance the model's ability to classify sentences as positive or negative. Starting with simpler architectures and progressing toward more advanced setups, the experiments aimed to evaluate the effectiveness of different strategies. Finally, a zero-shot classification approach was implemented using a pre-trained language model.

Steps in the Experiment

Step 1: Importing a Decoder-Only Model

- The initial step involved loading a pre-trained decoder-only language model (GPT-2) and its tokenizer.
- The SST-2 dataset was prepared and tokenized to create input data compatible with the model.

Step 2: Generating Outputs from the Model

- The model generated textual outputs for various inputs using different decoding configurations. These included hyperparameters like `temperature`, `top_k`, and `top_p`.
- While this step demonstrated the generative capabilities of GPT-2, it was not directly tied to the classification task.

Step 3: Loading and Preparing the SST-2 Dataset

- The SST-2 dataset, part of the GLUE benchmark, was loaded and tokenized. Padding and truncation were applied to ensure uniform input lengths, enabling efficient processing during training.

Step 4: Using the CLS Token for Classification

- The decoder's final layer was removed, and a classification head was added. The embedding of the first token (CLS token) was used as input for the classification layer.
- **Results**:
    - This setup struggled with generalization, with modest training accuracy and a clear bias toward the positive class.
    - Validation performance indicated limited capacity to distinguish between the classes effectively.
    - The confusion matrix showed a high number of false negatives, suggesting the model's inability to correctly predict negative samples.

Step 5: Adding a Linear Layer for Aggregation

- Instead of relying solely on the CLS token, a linear layer aggregated embeddings from the entire input sequence.
- **Results**:
    - While the training performance slightly improved, validation accuracy declined.
    - The model remained heavily biased toward the positive class, as indicated by the confusion matrix.
    - The aggregation of embeddings diluted critical information necessary for accurate classification, leading to poor generalization.

Step 6: Introducing Bidirectional Attention

- A bidirectional multi-head attention layer replaced the linear layer. This mechanism allowed the model to learn relationships between all tokens in the input sequence, improving contextual understanding.
- **Results**:
    - This architecture significantly enhanced performance. Both training and validation accuracy improved.
    - The confusion matrix showed better balance between classes, with a notable reduction in false positives and false negatives.
    - The attention mechanism's ability to model bidirectional relationships between tokens proved critical for capturing nuanced sentiment cues.

Step 7: Implementing Unidirectional Attention

- Two variants of unidirectional attention were tested:
    - **Left-to-Right Attention**: The model only considered preceding tokens for each token in the sequence.
    - **Right-to-Left Attention**: The model only considered succeeding tokens for each token in the sequence.
- **Results**:
    - **Left-to-Right Attention**: Performed well, with improved training and validation accuracy over earlier setups. However, it exhibited slightly more class imbalance, as indicated by higher false positive rates.
    - **Right-to-Left Attention**: Outperformed all other fine-tuned models, achieving the highest validation accuracy. The confusion matrix showed a well-balanced performance with significant reductions in false positives and false negatives, making it the best among fine-tuned models.

Step 8: Zero-Shot Classification

- A pre-trained zero-shot classification pipeline was employed using the `textattack/bert-base-uncased-SST-2` model.
- **Results**:
    - This approach achieved the highest overall accuracy without requiring any fine-tuning, showcasing the power of pre-trained models for general-purpose tasks.
    - The classification metrics were well-balanced, with high precision and recall for both classes.

- While the zero-shot model slightly favored the negative class, its overall performance exceeded all fine-tuned models. The confusion matrix highlighted its effectiveness in correctly classifying both negative and positive samples.

Key Observations

1. **Performance Improvements**:
   - Each architectural enhancement improved the model's ability to generalize.
   - Bidirectional and Right-to-Left attention mechanisms stood out as the most effective among the fine-tuned models.
2. **Class Balance**:
   - Simpler architectures (Steps 4 and 5) struggled with class imbalance, heavily favoring the positive class.
   - Attention-based models (Steps 6 and 7) and the zero-shot classifier (Step 8) addressed this issue effectively, with Right-to-Left Attention providing the best balance among the fine-tuned models.
3. **Zero-Shot Strength**:
   - The pre-trained zero-shot classifier outperformed all other methods without requiring additional training. This highlights the potential of pre-trained models for achieving strong performance on unseen tasks.
4. **Training Efficiency**:
   - Fine-tuning required significant computational effort compared to the zero-shot approach, which achieved high performance with minimal setup.
5. **Insights from Confusion Matrices**:
   - Confusion matrices revealed the evolution of class balance across setups. Simpler models showed high false negatives, particularly for the negative class, while advanced attention-based architectures reduced these errors significantly.
   - The Right-to-Left Attention model and the zero-shot classifier displayed the best balance, with a lower tendency to favor one class over the other.

Conclusion

- **Bidirectional and Right-to-Left Attention** models demonstrated excellent fine-tuning performance, with the latter achieving the best results among the trained models.
- The **zero-shot classifier** achieved the highest overall accuracy with no training effort, showcasing the power of large pre-trained models.
- This task highlights the trade-off between fine-tuning complexity and zero-shot simplicity, with attention mechanisms providing robust intermediate solutions for sentiment classification.

Recommendations

1. **Use Zero-Shot Models**:
   - For quick deployment with minimal effort, zero-shot classification models like `textattack/bert-base-uncased-SST-2` should be the go-to choice.
2. **Leverage Attention Mechanisms**:
   - When fine-tuning is necessary, attention-based architectures such as Bidirectional or Right-to-Left setups should be prioritized.
3. **Further Exploration**:
   - Explore hybrid models that combine fine-tuned attention mechanisms with pre-trained zero-shot models for enhanced performance.
4. **Address Class Imbalance**:
   - Use weighted loss functions or oversampling techniques to mitigate class imbalance in future experiments.