Assignment 8:

“Final Kaggle Submission – Text Generation with Llama”

Shobhit Pachauri

sxp230024

University of Texas at Dallas

MS Business Analytics and AI

BUAN 6342.S01 - Applied Natural Language Processing - F24

Prof. Harpreet Singh

December 13, 2024

Homework 5: Encoder-Only Models for Tweet Emotion Detection

Models and Theory

In Homework 5, encoder-only models were employed to classify emotions in tweets. The primary models explored were:

- **DistilBERT**: A lightweight version of BERT, offering 97% of its language understanding capabilities but with reduced computational costs, making it 40% smaller and 60% faster.
- **RoBERTa Base**: A robustly optimized version of BERT, featuring improved pretraining techniques such as dynamic masking and larger batch sizes.
- **DistilRoBERTa**: A distilled variant of RoBERTa, which strikes a balance between computational efficiency and performance.

These models use self-attention mechanisms to capture contextual information effectively and were pre-trained on extensive corpora, enabling them to generalize well across various tasks.

Key Parameters

```
config = AutoConfig.from_pretrained(model_name)
config.num_labels = num_labels
model = AutoModelForSequenceClassification.from_pretrained(
    model_name,
    config=config,
)
```

Results and Analysis

The models were evaluated on F1 Macro scores and accuracy:

- **DistilBERT**: F1 Macro of 50.63%.
- **RoBERTa Base**: F1 Macro of 52.56%, with the highest accuracy of 6.99%.
- **DistilRoBERTa**: F1 Macro of 51.01%.

Summary: RoBERTa Base demonstrated the highest performance in accuracy and F1 Macro, showcasing the benefits of its advanced training methodology. DistilBERT, while slightly behind, provided a faster and more resource-efficient alternative.

Homework 6: Fine-Tuning on Emotion Datasets

Key Improvements

Building on Homework 5, Homework 6 focused on fine-tuning the models on specific emotion datasets. Key enhancements included:

- Adoption of consistent evaluation metrics, such as the F1 score, to provide reliable comparisons.
- Implementation of hyperparameter optimization techniques to refine model performance.

Results and Analysis

Fine-tuning yielded improved F1 scores for all models:

- DistilBERT emerged as the best compromise between computational efficiency and accuracy, offering competitive performance while maintaining its lightweight design.

Summary: Homework 6 demonstrated that fine-tuning encoder-only models on domain-specific datasets significantly improves their ability to detect emotions accurately.

Homework 7: Multilingual and Domain-Specific Classification

Models and Theory

Homework 7 explored advanced models for multilingual and domain-specific classification tasks:

- **Llemma**: Achieved 80% accuracy in cross-lingual tasks, showcasing its robustness across languages.
- **MTEB (WhereIsAI UAE Large V1)**: Achieved 83% accuracy in domain-specific classification, highlighting its specialization.
- **Gemma2**: Another advanced model investigated for these applications.

Key Parameters

```
model = AutoModelForSequenceClassification.from_pretrained(
    model_name,
    num_labels=len(label_columns),
    problem_type="multi_label_classification"
)
```

Summary: The exploration of multilingual and domain-specific models revealed that specialized pretraining and architectures contribute significantly to performance in niche tasks, with MTEB leading the results.

Homework 8: Zero-Shot Classification with LLAMA Models

Models and Theory

Homework 8 introduced decoder-based models for zero-shot classification tasks:

- **LLAMA Base Model**: A decoder-only transformer with 7 billion parameters, designed for generalization without fine-tuning.
- **LLAMA Instruct Model**: A fine-tuned variant optimized for instruction-following tasks.

These models employ rotary positional embeddings and grouped-query attention to enhance efficiency and excel in few-shot and zero-shot learning scenarios.

Key Parameters

```
model = AutoPeftModelForCausalLM.from_pretrained(
    checkpoint,
    quantization_config=bnb_config,
    torch_dtype=torch_data_type,
    trust_remote_code=True,
    device_map='auto'
)
```

Results and Analysis

The models were evaluated on accuracy and F1 Macro metrics:

- **LLAMA Base Model**: Accuracy of 71.79% and F1 Macro of 0.1721.
- **LLAMA Instruct Model**: Accuracy of 70.07% and F1 Macro of 0.1615.

Summary: While the LLAMA Base Model slightly outperformed the Instruct Model, both models demonstrated robust zero-shot learning capabilities, emphasizing the potential of decoder-only transformers in such tasks.

Conclusion

Throughout this series of assignments (HW5 to HW8), we explored a range of advanced natural language processing techniques, progressing from encoder-only models to large language models capable of zero-shot classification. This journey provided valuable insights into the evolving landscape of NLP and emotion detection in text. Key Learnings and Developments:

1. Encoder-Only Models (HW5):

   We began with traditional encoder-only models like DistilBERT, RoBERTa Base, and DistilRoBERTa for tweet emotion detection. These models demonstrated the effectiveness of self-attention mechanisms in capturing contextual information. RoBERTa Base emerged as the top performer with an F1 Macro score of 52.56%, highlighting the benefits of its robust optimization techniques.
2. Fine-Tuning Strategies (HW6):

   Building on HW5, we explored fine-tuning techniques on specific emotion datasets. This process significantly improved model performance, with DistilBERT emerging as a strong contender balancing efficiency and accuracy. This assignment underscored the importance of domain-specific fine-tuning in enhancing model performance.
3. Multilingual and Domain-Specific Models (HW7):

   We ventured into more specialized models like Llemma, MTEB, and Gemma2 for multilingual and domain-specific tasks. MTEB (WhereIsAI UAE Large V1) achieved an impressive 83% accuracy in domain-specific classification, demonstrating the power of specialized pretraining for niche tasks.

4. Zero-Shot Classification with LLAMA Models (HW8):

   The final assignment introduced us to decoder-based models for zero-shot classification. We experimented with the LLAMA Base Model and LLAMA Instruct Model, both 7B parameter models. The Base Model slightly outperformed the Instruct Model, achieving 71.79% accuracy and an F1 Macro of 0.1721. This showcased the potential of large language models in tackling classification tasks without task-specific fine-tuning.

Key Takeaways:

1. Model Evolution: We observed a clear progression from task-specific encoder models to more versatile decoder-based models capable of zero-shot learning.
2. Trade-offs: Throughout the assignments, we encountered trade-offs between model size, computational efficiency, and performance, learning to balance these factors based on specific use cases.
3. Importance of Fine-Tuning: The significant improvements seen in HW6 emphasized the critical role of domain-specific fine-tuning in enhancing model performance.
4. Versatility of Large Language Models: The performance of LLAMA models in zero-shot classification tasks demonstrated the potential of large language models to generalize across various NLP tasks without extensive task-specific training.

Future Directions:

Based on our experiences, future work could explore:

- Investigating the impact of different prompting strategies in zero-shot and few-shot learning scenarios

- Exploring multi-task learning approaches to leverage the versatility of large language models across various NLP tasks
- Developing more robust evaluation metrics for multi-label emotion classification tasks

In conclusion, this series of assignments provided a comprehensive journey through the current state-of-the-art in NLP, from specialized encoder models to versatile large language models. The progression highlighted the rapid advancements in the field and the increasing potential of AI to understand and classify human emotions in text, paving the way for more sophisticated and nuanced applications in sentiment analysis and beyond.

Wandb link: https://wandb.ai/shobhit-pachauri-university-of-texas-at-dallas/multilabel_tweet_classification/workspace