Named Entity Recognition for Mountain Names

Hospodarchuk Dmytro

October 24, 2024

1 Introduction

The goal of this project was to develop a named entity recognition (NER) model for identifying mountain names within text. This report outlines the process of dataset creation, model selection, fine-tuning, and evaluation. The project involved training two models: one fine-tuning a pre-trained transformer-based model (DistilRoBERTa) using LoRA PEFT (Parameter-Efficient Fine-Tuning), and the other training a custom transformer model from scratch.

2 Dataset Creation

The dataset for this task was obtained from Hugging Face, specifically from the dataset named ”NER Mountains” by Telord (link). It consists of labeled text data where mountain names are annotated for NER tasks.

To address the data imbalance in the dataset (where many tokens do not belong to any named entity), a weighted cross-entropy loss function was used to assign higher importance to the minority classes.

3 Model Selection

For this NER task, two different approaches were employed:

- **Fine-Tuning DistilRoBERTa with LoRA PEFT:** DistilRoBERTa, a lightweight version of RoBERTa, was selected due to its balance between performance and computational efficiency. The model was fine-tuned using LoRA (Low-Rank Adaptation).
- **Training a Custom Transformer Model from Scratch:** A smaller transformer model was trained from scratch to explore the trade-offs between model size and accuracy. This approach aimed at creating a more lightweight solution.

4 Training Details

4.1 Fine-Tuning DistilRoBERTa

The fine-tuning process utilized a weighted cross-entropy loss function to handle class imbalance. Hyperparameters such as learning rate, batch size, and number of epochs were tuned to achieve optimal performance. LoRA was applied to fine-tune the model's weights, which allowed for efficient adaptation to the NER task.

4.2 Custom Transformer Training

A smaller transformer model was trained using the same dataset. Despite being less accurate than the fine-tuned DistilRoBERTa, this model demonstrated decent performance while having a significantly lower computational cost.

5 Results

The performance of the fine-tuned DistilRoBERTa model surpassed that of the custom transformer model. The use of LoRA PEFT and a pre-trained language model provided a significant advantage in terms of accuracy. The results indicate that using a pre-trained model is more effective for this specific NER task.

Model	Acc	Precision (TP)
DistilRoBERTa (Fine-tuned)	0.96	0.94
Custom Transformer	0.88	0.85

Table 1: Performance comparison of NER models.

6 Future Improvements

To further enhance the NER performance, the following steps can be considered:

- **Larger Dataset:** Increasing the size of the dataset by adding more labeled examples would help improve the model's generalization and accuracy.
- **Training Larger Models:** Using larger models such as RoBERTa-base or RoBERTa-large can improve performance, though it would require more computational resources.
- **One-Shot or Few-Shot Learning:** Implementing one-shot or few-shot learning approaches could help make the model generalize better to unseen examples.

- **Data Augmentation:** Augmenting the training data using techniques such as back-translation or synonym replacement can increase the variety of training examples and improve model robustness.
- **Hyperparameter Tuning:** Further tuning of hyperparameters like learning rate, batch size, and regularization terms can yield better performance.

7 Conclusion

This project demonstrated the effectiveness of fine-tuning pre-trained language models for NER tasks. The use of LoRA PEFT for fine-tuning DistilRoBERTa showed significant performance benefits over training a model from scratch. Future work can focus on expanding the dataset, exploring larger models, and implementing advanced training techniques to improve results.