Model Evaluation Report with Potential Improvements

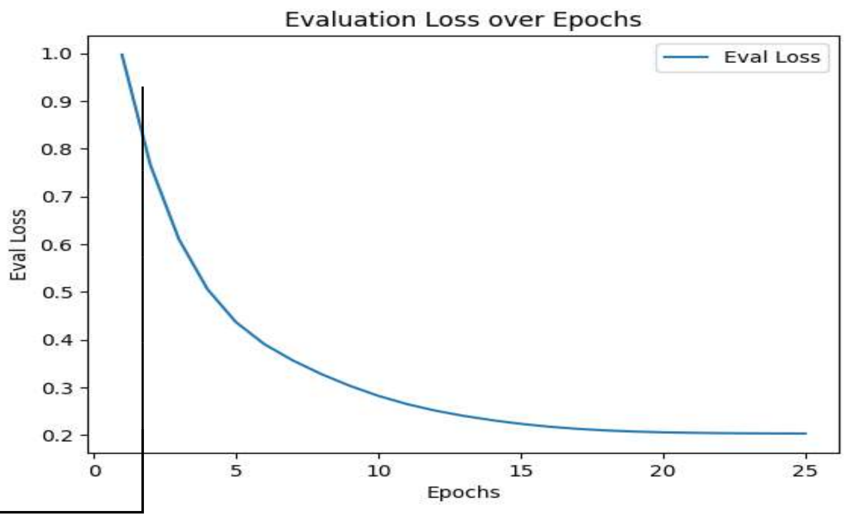
Model Overview

The model in focus is a fine-tuned version of the DistilBERT-NER model, specifically designed for Named Entity Recognition (NER) tasks related to mountains. It has been enhanced to classify tokens into 11 categories, adding two new classes: I-MOU and B-MOU for mountain entities. The model leverages the original DistilBERT weights, ensuring a solid foundation while extending its capabilities.

Key Metrics

The evaluation metrics for the model are as follows:

Metric	Score
Loss	0.2035
Precision	0.8536
Recall	0.7906
F1	0.7117
Accuracy	0.7906



Performance Analysis

1. **Precision (0.8536)**: The model exhibits high precision, indicating that when it predicts an entity, it is likely to be correct. This is particularly beneficial for applications where false positives must be minimized.
2. **Recall (0.7906)**: While recall is decent, it suggests that the model misses some true positive entities, particularly those related to mountain names.
3. **F1 Score (0.7117)**: The F1 score, which balances precision and recall, indicates that there is room for improvement, particularly in the model's ability to correctly identify mountain entities.
4. **Accuracy (0.7906)**: The overall accuracy reflects a solid understanding of the general entity classification, but further refinements are needed to enhance mountain-specific detection.

Potential Improvements

1. Data Augmentation

- **Expand the Dataset**: Currently, the model is trained on a small dataset that not cover a lot of mountain entities comprehensively is a main problem. Increasing dataset size and including more diverse sentences that represent various contexts and geographical references will significantly improve model perfomanse.
- **Use diverse dataset**: currently dataset consists only of sentences about mountains, mixing it with other datasets will allow to better distinguish different classes and prevent overtraining.

2. Model Architecture Adjustments

- **Layer Freezing**: Experimenting with freezing certain layers of the model during fine-tuning could help in retaining the general features learned from the original DistilBERT, while allowing the final layers to adapt specifically for mountain-related entities.
- **Class Weighting**: Apply class weighting during training to give more importance to the new mountain classes (B-MOU and I-MOU), which might be underrepresented in the training data.

3. Hyperparameter Tuning

- **Learning Rate Optimization**: Fine-tuning the learning rate could lead to better convergence and improved model performance. Employing techniques like learning rate scheduling may also yield positive results.
- **Batch Size Experimentation**: Testing different batch sizes can affect the model's performance. A smaller batch size might enhance the model's ability to generalize.

4. Evaluation with Additional Metrics

- **Analyze Confusion Matrix**: Creating a confusion matrix to visualize misclassifications can provide insights into specific entities that are frequently misclassified, guiding further training and data refinement.
- **Cross-validation**: Implementing k-fold cross-validation during training can provide a better assessment of model stability and robustness.

Conclusion

train_test.ipynb	Training, evaluating and demo notebook
kner_dataset	Final dataset
model_train.py	Script for model training
inference_model.py	Script for model evaluating
dataset.txt	Dataset of raw text
create_dataset.ipynb	Notebook with all dataset creation tries and methods
BIO_Dataset.txt	Dataset of raw text, labeled sentences

The fine-tuned DistilBERT-NER model demonstrates commendable performance in recognizing mountain entities. However, due to lack of time, the dataset is small and of poor quality, which is the main problem . Due to dataset refinement, it will be possible to achieve 90%+ results of the original model in all metrics. And even more with adjusting the model architecture, tuning hyperparameters, implementing post-processing techniques, and evaluating with additional metrics,.