Potential improvements

Data

The first thing I would like to emphasize is to improve the dataset. For example, I could have taken a ready-made dataset from hugging face or kaggle that was already labeled, but I decided to show how you can create your own dataset using some tricks.

An improvement would be to make/take a bigger dataset, to label it very well, to use data augmentation - paraphrasing, masking words. It would also be necessary to make it balanced (although I have it balanced by mountain names) but not by classes. So another idea was to change loss function to handle imbalance in classes. I also needed to add more mountain names, as the model could not always distinguish the local mountain. For training, it would be better to use cross validation, especially when the dataset does not have a lot of data (1100 in mine). Additionally, improvement can be to add mountain names to tokenizer vocabulary so that tokenizer won't split unknown words into sub-tokens. It will improve the efficiency of the model. Another improvement can be to make model able to distinguish different abbreviated names of mountains.

Model

I was using default bert-based-uncased model, and potential improvement could be to test another models, like: RoBERTa, T5, BART, DistilBERT – which are pretrained on larger data and have wider vocabulary and other improved features in training.

Parameters

Use techniques like grid search or random search with Optuna to find the best combination of hyperparameters (e.g., learning rate, batch size, number of epochs) that will optimize my model. Due to computational resources and limited time, I have to manually select best parameters which I've discovered. But using more detailed analysis of parameters optimization can improve model efficiency.