Assignment 2

Lucas Waddell

Part A: Data Preparation for LLMs

Q1: Name three sources where you can find datasets for training or fine-tuning LLMs. Briefly describe the advantages and limitations of each.

Common Crawl

Advantages: Huge volume of data, covering 'real' internet use.

Disadvantages: Too much data for weak systems. No guarantees on the quality of all the data, May contain biases that are hard to detect.

Kaggle

Advantages: Fast, high quality platform. Huge quantity of datasets available across many different categories.

Disadvantages: While user submitted datasets means a diverse selection is available, it means there is no guarantee of data quality or content. Kaggle requires that users make an account for the download of datasets.

Hugging Face:

Advantages: Another fast, high quality platform. Massive breadth of datasets available.

Disadvantages: Same drawbacks as Kaggle. It requires an account, and there are no guarantees on data quality.

Q2: Explain why data preprocessing is crucial when preparing a dataset for an LLM. Mention at least three preprocessing techniques and describe their impact on model performance.

Preprocessing is crucial to avoid the phenomena of 'garbage in, garbage out'. Any model, no matter the technique or amount of compute used, is only as good as the data it is trained on. Problems such as outliers, incorrect info, or inconsistent formats can all lead to undesired effects such as increased training time, unexpected model behaviour, or poor model performance.

1. Parsing

Parsing is the process of extracting the useful information from a data source to use as input for the model. It ensures that data is in a machine-readable format and adheres to a consistent structure that the model can effectively learn from. An example would be parsing JSON datasets to remove the actual info from the JSON structure.

2. Data Augmentation

This is a technique used when the amount of data we have is limited, or it is difficult to generate more data. We can transform existing data into a similar, but different form that increases the models exposure to different varieties of information. The goal of this is of course to increase model quality.

3. Normalization

This is the process of standardizing the data to make sure consistent language is used across all the inputs. In the context of text, an example could be making everything lower case. A different example with numbers could be making sure a uniform separator, like a comma, is used for all large numbers. The standardization helps the model relate more pieces of information together.

Q3: What is tokenization in NLP? And Why is subword tokenization widely used in modern LLMs?

Tokenization is the process of taking some input data and transforming it into, generally smaller, units called tokens. A token does not necessarily mean a word. It could be any linguistic element or thing with meaning.

Subword tokenization is widely used because it does a good job of handling values that are uncommon or not present in the vocabulary of the model. The model is still able to assign some value to the unknown word by breaking it down into the subwords and then tokens. It is stated that this improves the model accuracy as it lets the model capture semantic and syntactic word relations in greater detail.

Q4: Explain what dataset bias is and how it can impact an LLM’s predictions. What steps can be taken to mitigate bias when preparing training data?

Dataset bias refers to imbalances in a dataset that lead to incorrect, inaccurate, or skewed outputs. This could lead to over/under representation of certain groups, or the propagation of misinformation in the model’s outputs.

The best tool in mitigating bias is a quality data set. Outside of this, normalization can be used to help clean the data, and data augmentation can be used to fill in underrepresented fields.

Part B : Fine-Tuning LLMs

Q1: How does fine-tuning differ from pre-training? Provide an example of a real-world use case where fine-tuning an LLM is preferable to using a pre-trained model as-is.

If pre-training is school, then pre training is law school. The example is that pre-training provides the foundational ‘understanding’ and knowledge of the model. Fine-tuning is what specializes the model, and increases its information related to a specific task.

Suppose you wanted a model that was an expert on Nova Scotia tax code. You could train a model solely on the tax code, but it is unlikely that you would have enough data to yield a very useful model. On the other hand, you could take a strong foundational model such as LLaMA or deepseek, and fine tune it on the Nova Scotia tax code. Fine tuning lets the model retain its general understanding and law knowledge while gaining a deep expertise on the NS tax code.

Q2: What are three common evaluation metrics for fine-tuned LLMs?

1. Perplexity: Measure how well the model predicts the next token in some sequence.
2. Human Evaluation: Use people to assess how the text sounds and reads.
3. ROUGE: Measure the similarity of the model output with human generated reference text.

Q3: Why is parameter-efficient fine-tuning useful when working with large-scale Models?

It is useful because it reduces costs. Since we are not adjusting every component, not everything needs to be retrained. For huge models where training costs are a significant barrier, this decreases the expenses that must be incurred to increase model quality.