CSC-496/696: Natural Language Processing and Text as Data

Lecture 17: Using Hugging Face

Patrick Wu

Friday, November 1, 2024

Lecture Contents

1. Announcements
2. Review of Large Language Models
3. Fine-Tuning LLMs
4. Ethical Considerations and Potential Harms with LLMs
5. Hugging Face

Announcements

Assignment 4

- Released on Tuesday
- As indicated, the length of these assignments is tapering off to account for time spent on the final project
- I also know that the end of the semester is very busy
- Assignment 4 also uses the final project dataset, so it is a good chance to work with that dataset a bit if you're using it for your final project

Final Project

- Reminder: Project proposal due on November 8
- One page single spaced
- Submission is now available on Canvas
- Graded on completion

No Class on Tuesday, November 4

It’s Election Day! If you are eligible to vote, please go vote!

We’ll meet as usual one week from now

Review of Large Language Models

Three Types of LLMs

- Encoder-only models
- Decoder-only models
- Encoder-decoder models

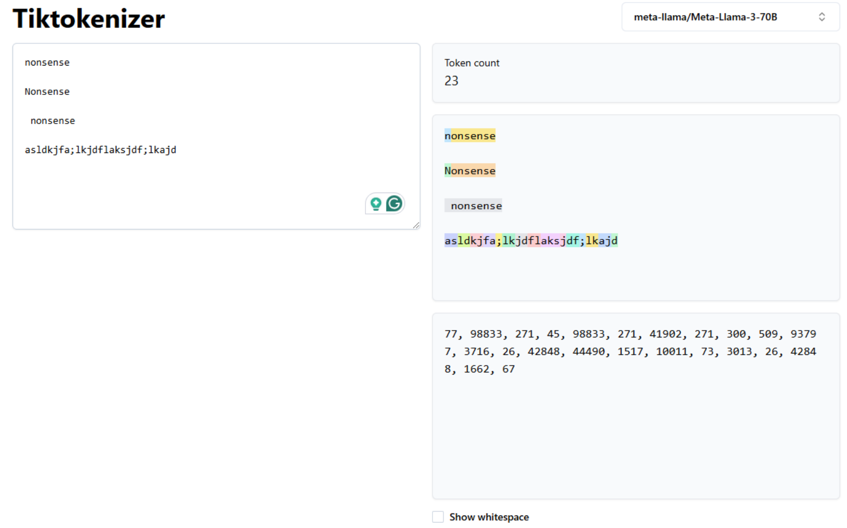
Each Type of Model is Trained Differently

- Encoder-only models: Masked language modeling
    - “The [MASK] was delicious!”
- Decoder-only models: next word prediction
    - “The food was [MASK]”

Requires a Great Deal of Data

- Common Crawl
- Colossal Clean Crawled Corpus

Tokenization Also Matters a Great Deal



Tiktokenizer

meta-llama/Meta-Llama-3-70B

```
nonsense

Nonsense

 nonsense

asldkjfa;lkjdflaksjdf;lkajd
```

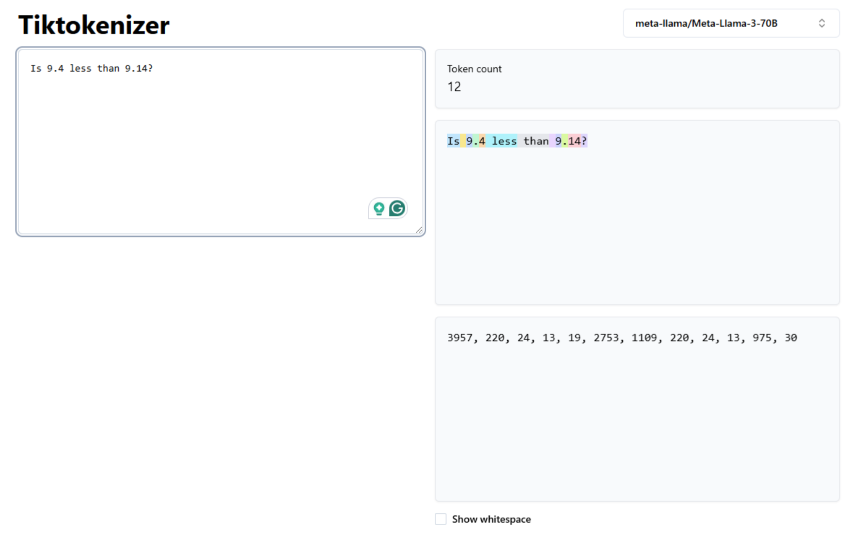
Token count
23

```
nonsense

Nonsense

 nonsense

asldkjfa;lkjdflaksjdf;lkajd
```

```
77, 98833, 271, 45, 98833, 271, 41902, 271, 300, 509, 9379
7, 3716, 26, 42848, 44490, 1517, 10011, 73, 3013, 26, 4284
8, 1662, 67
```

□ Show whitespace

Tokenization Also Matters a Great Deal



Tiktokenizer

meta-llama/Meta-Llama-3-70B

```
Is 9.4 less than 9.14?
```

Token count
12

```
Is 9.4 less than 9.14?
```

```
3957, 220, 24, 13, 19, 2753, 1109, 220, 24, 13, 975, 30
```

□ Show whitespace

Fine-Tuning LLMs

Fine-Tuning a LLM

- LLMs are often good at general tasks, but maybe we want to apply the LLM to a new domain or have it behave in a specific manner
- We can do that using **fine-tuning**, which takes the parameters of a pre-trained model and updates them on a new, specialized task
- For example, we can fine-tune a model so it is very good at making political-related classifications
- Many different types of fine-tuning

Continued Pre-Training

- We can retrain all the parameters of the model on new, additional text data using the same self-supervised word prediction task
- This is as if our new data were at the tail end of the pre-training data
- This is often called “continued pretraining”

Masked Language Modeling Fine-Tuning

- Recall that BERT is an example of an encoder-only language model, which means the attention mechanism can look forward and backward
- One common approach is to fine-tune this language model to output predictions
- To do this, we can add extra neural circuitry after the last layer of the model to produce classifications
- A very efficient way to produce high-quality predictions, but you do need labeled data to fine-tune these models

Supervised Fine-Tuning

- Supervised fine-tuning, or SFT, is often used for instruction fine-tuning, where we want a pre-trained language model to learn to follow text instructions
- How ChatGPT was partially trained to follow instructions well
- We do need supervised responses to each command, so it is truly supervised (not self-supervised)

Ethical Considerations and Potential Harms with LLMs

Bias in Training Corpora

- Even if we filter out harmful content and PII from training corpora, biases still exist in these models
- Even reading a normal Twitter thread, for example, yields countless examples of biased responses
- Because LLMs are ingesting these texts and trying to predict the next word, it can learn from these biases and stereotypes
- An entire course can be taught on biases in LLMs!
- This can even affect tokenization, given that BPE tokenization is based on how frequently pairs of characters or subwords co-occur together

Biases in Training Corpora

In 2016, Microsoft released a chatbot called Tay. It did not end well. You can read about the story here.

Privacy Concerns

- Pre-training information can contain information like phone numbers and addresses
- These are often filtered out
- But you can also often reverse engineer private information

Netflix Prize Debacle

- In 2006, Netflix set up a competition called the "Netflix Prize"
- They released 100 million anonymized movie ratings, each which includes a unique subscriber ID (so you could connect movie reviews on a user-level basis)
- The goal was to predict how those users would then rate other movies, which could then be used to improve the recommendation algorithm
- Just 16 days into the competition, two University of Texas researchers said they could identify a large portion of the users in the data
- They linked together IMDb accounts (public information) with the Netflix data (private but anonymized information), which allowed them to see what other movies these users had watched without giving an IMDb rating

Privacy Concerns

- Given that the LLMs can learn highly complex relationships between words, LLMs can potentially automatically learn these relationships between different data sources, leading to private information being leaked

Hugging Face

Hugging Face

- A company based in NYC that develops computation tools for building applications using ML
- Most notable library: `transformers`
- It is compatible with PyTorch, TensorFlow, and Keras
- The industry standard now for working with LLMs

Lab

We'll work through the lab here