A Practical Guide to Training and Fine-Tuning Language Models

NASHVILLE DATA SCIENCE MEETUP

JULY 17TH, 2023

LIMING ZHOU

XSOLIS, FRANKLIN, TN

Agenda

Some NLP Concepts

History of Language Models

Hugging Face

- Fine – tuning pretrained Language Models
- Train a language model from scratch
- Example code

Agenda

- Some NLP Concepts
- History of Language Models
- Hugging Face
  - Fine – tuning pretrained Language Models
  - Train a language model from scratch
  - Example code

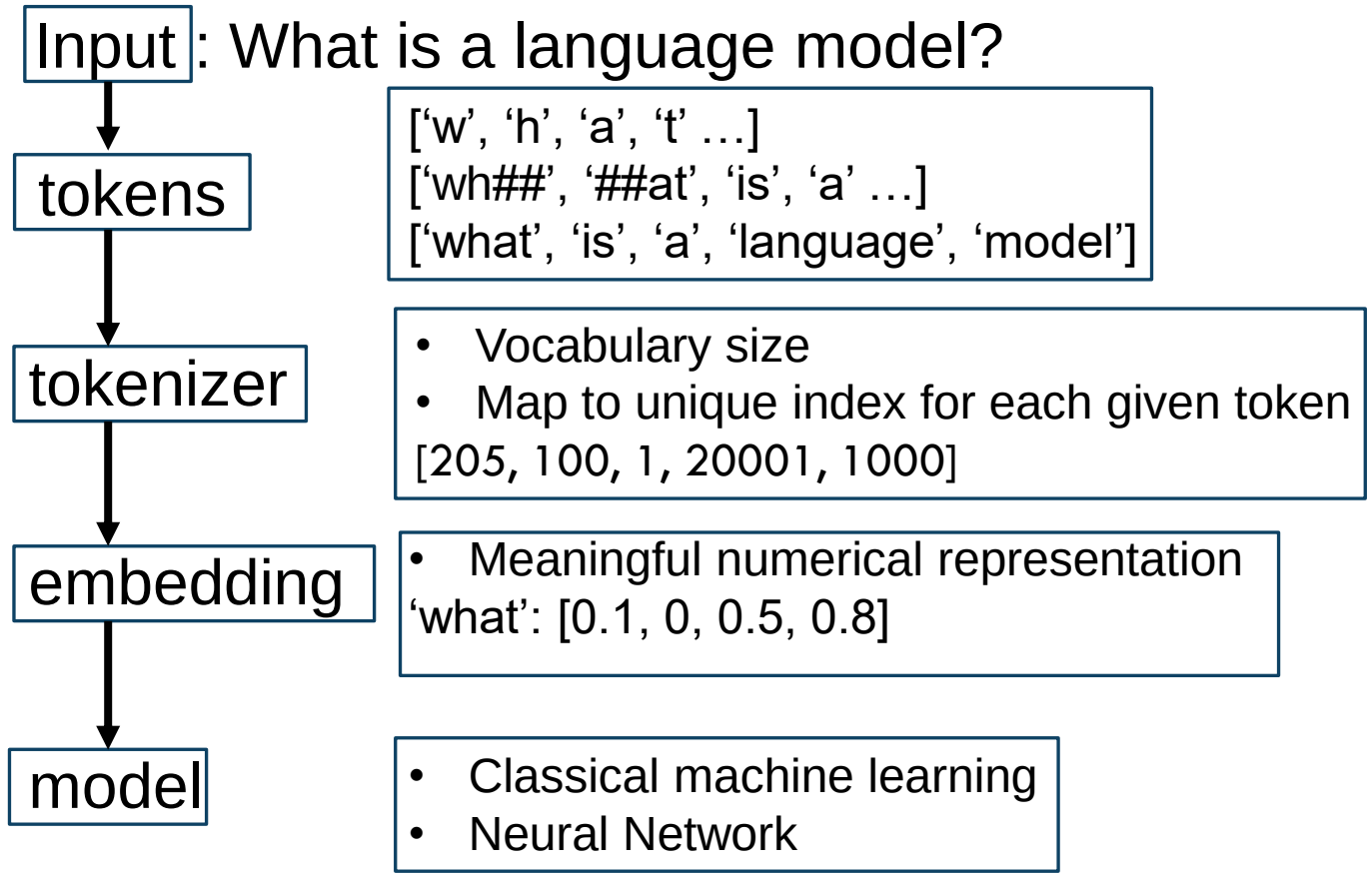
Some NLP Concepts

What is Natural Language Processing?

- Giving computers the ability to understand and generate text and spoken words in the same way human beings can
- NLP tasks:
  - Speech recognition
  - Sentiment analysis
  - Translation
  - Summarization
  - Question & Answer

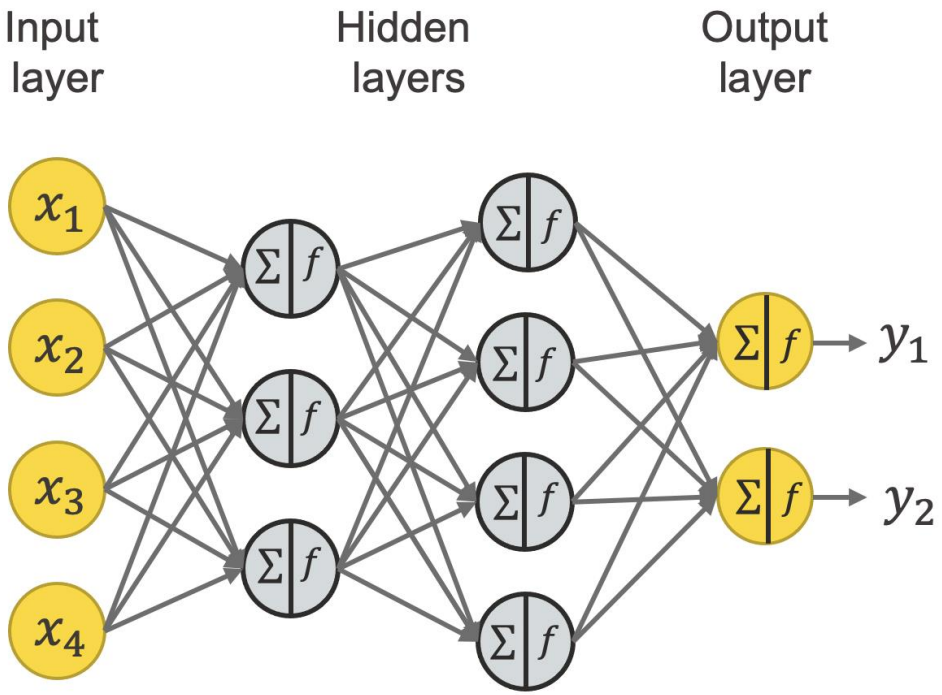
Some NLP Concepts

How to train an NLP model?



Some NLP Concepts

Neural Network



Some NLP Concepts

One-hot Encoding

categories	Sport	Politic	Science	Health
Sport	1	0	0	0
Politic	0	1	0	0
Sport	1	0	0	0
Science	0	0	1	0
Sport	1	0	0	0
Health	0	0	0	1

Agenda

Some NLP Concepts

History of Language Models

Hugging Face

- Fine – tuning pretrained Language Models
- Train a language model from scratch
- Example code

History of Language Model

What is language model?

- Predicts the next word given a sequence of words

Stochastic Model

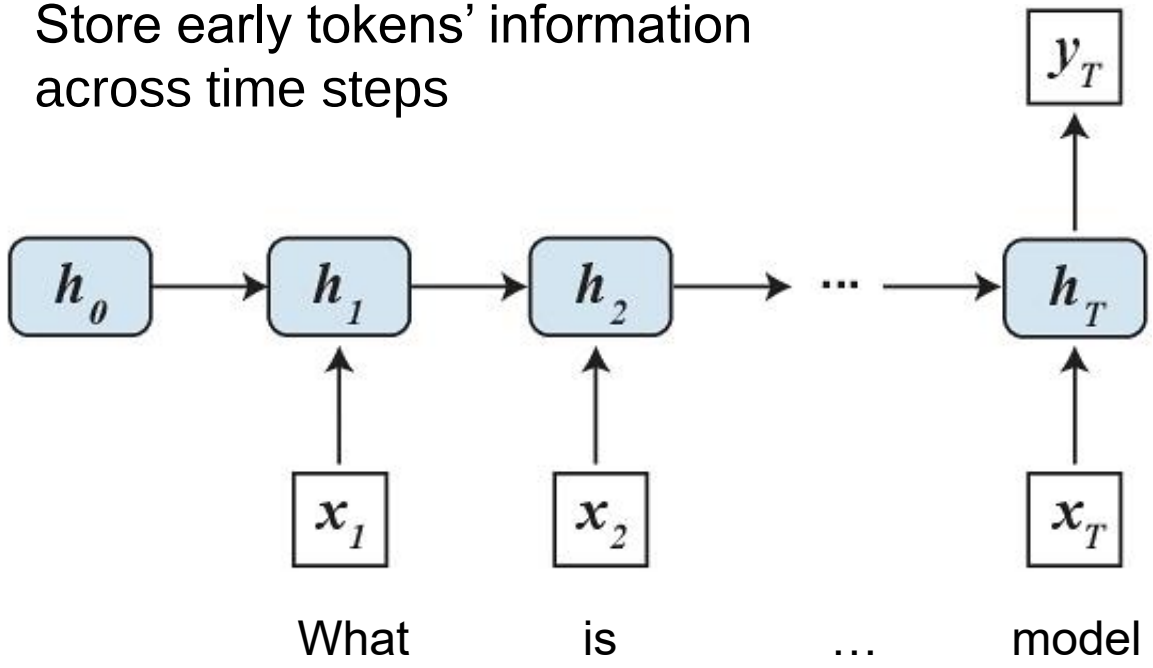
- Count number of words cooccurrence
  - P(“like”|”I”) = count(“I like”)/count(“I”)
- N-gram
  - P(“like”|”I don’t”) = count(“I don’t like”)/count(“I don’t”)

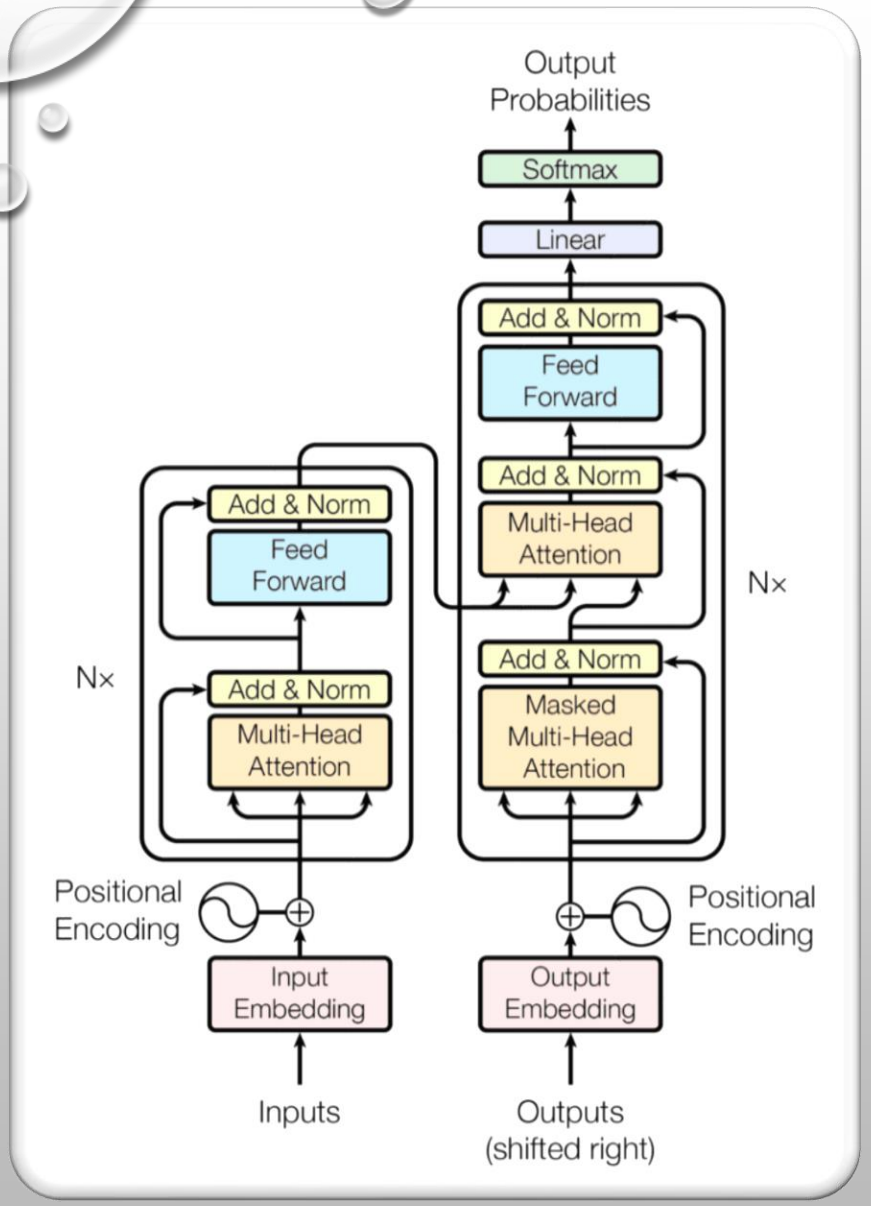
History of Language Model

Recurrent Neural Networks (LSTM, GRU)

h_i: hidden state

- Store early tokens' information across time steps





History of Language Model

Transformers (BERT, GPT, T5)

- Encoder-Decoder
  - Encoder models: understand whole sentence
  - Decoder models: text generation
- Parallel
- Attention!!!
  - Let model itself to decide which part in the sentence is important
- Hugging Face

Attention is all you need 2017 Ashish Vaswani etc.

Agenda

Some NLP Concepts

History of Language Models

Hugging Face

- Fine – tuning pretrained Language Models
- Train a masked language model from scratch
- Example code

Hugging Face

Hugging Face

Library:

- Datasets: Use Arrow
- Evaluate: Evaluate machine learning models

Classes:

- Auto
  - AutoConfig, AutoModel, AutoTokenizer,
  - AutoModelForSequenceClassification
- Trainer
  - TrainerArguments, Trainer

Hub:

- Datasets
- Pre-trained models

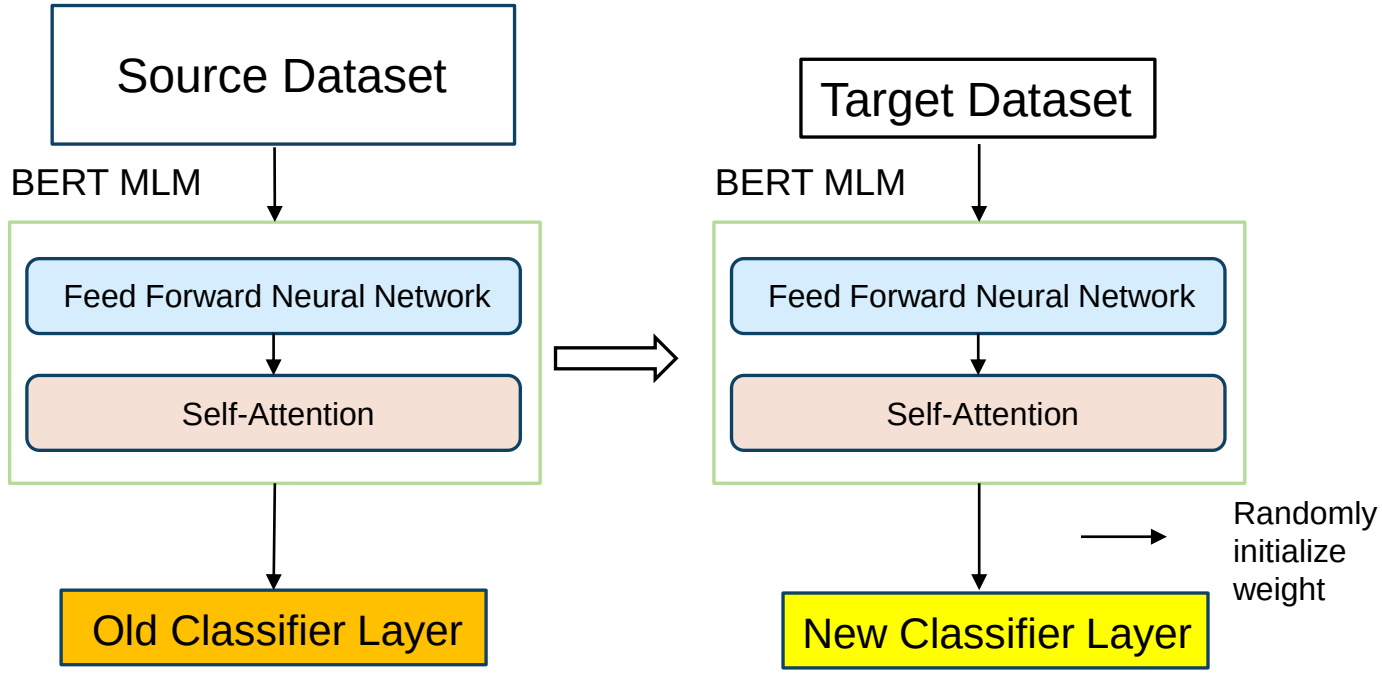
```
tokenizer("Welcome to the 🤗 Tokenizers library. This is a library from Hugging Face.")
```

```
{'input_ids': [101, 6160, 2000, 1996, 100, 19204, 17629, 2015, 3075, 1012, 2023, 2003, 1037, 3075, 2013, 17662, 2227, 1012, 102], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]}
```

Hugging Face

Fine Tuning

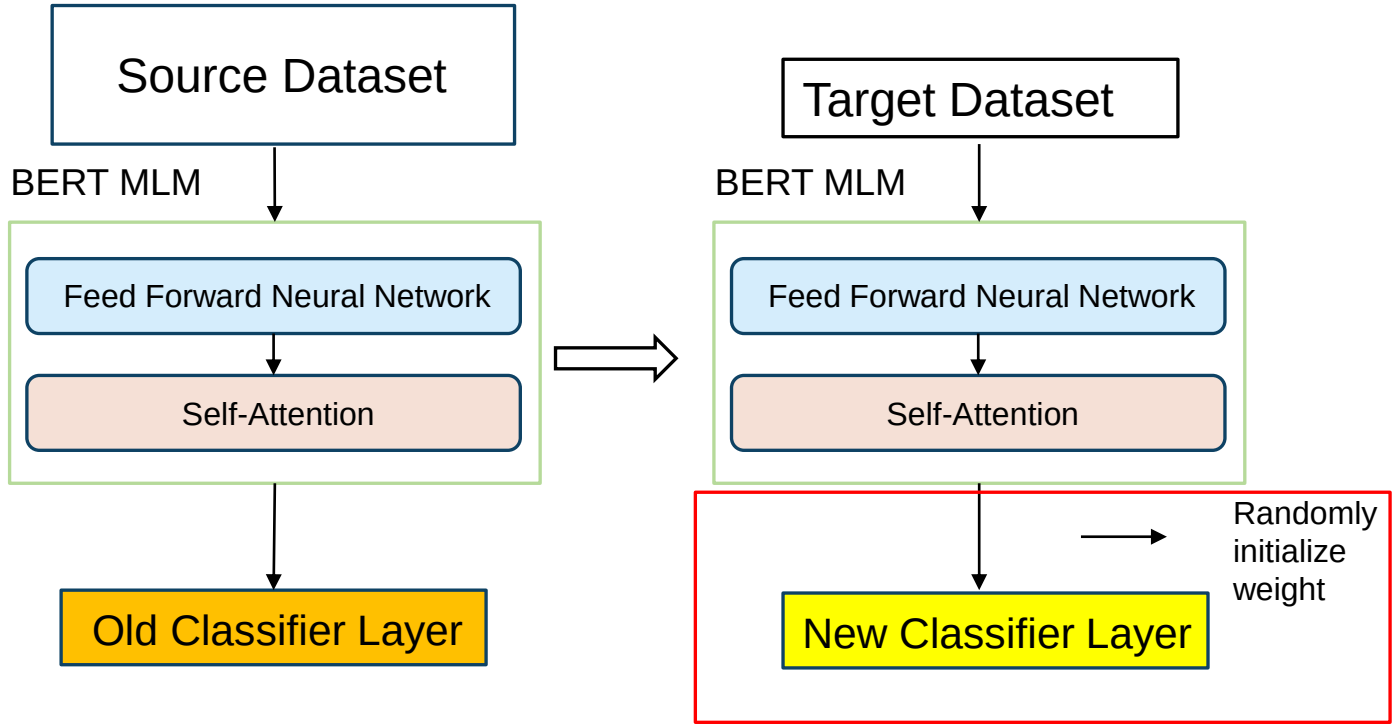
- Transfer learned knowledge from source dataset to target dataset
  - If you have small target dataset
  - If you can’t train a model from scratch
  - If you work in a unique domain (specialized vocab)



Hugging Face

Fine Tuning

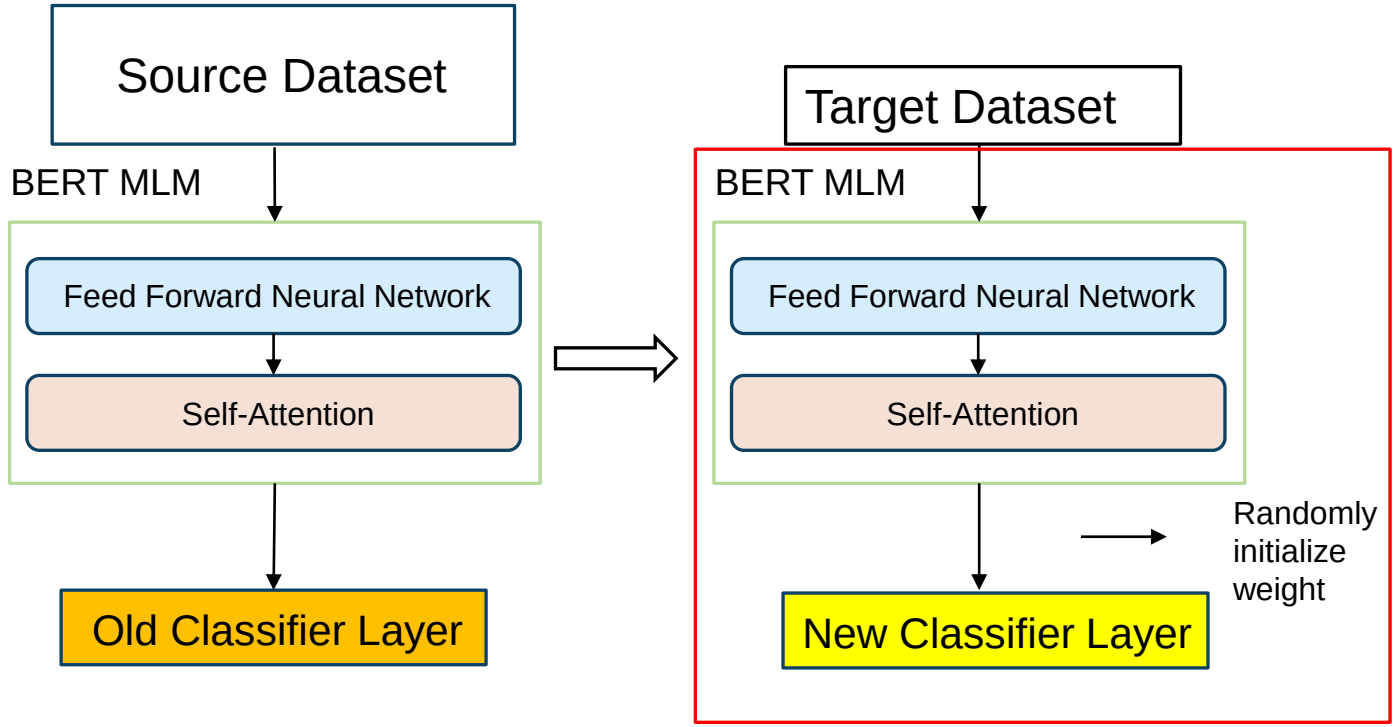
- Transfer learned knowledge from source dataset to target dataset
  - If you have small target dataset
  - If you can’t train a model from scratch
  - If you work in a unique domain (specialized vocab)



Hugging Face

Fine Tuning

- Transfer learned knowledge from source dataset to target dataset
  - If you have small target dataset
  - If you can’t train a model from scratch
  - If you work in a unique domain (specialized vocab)



Summary

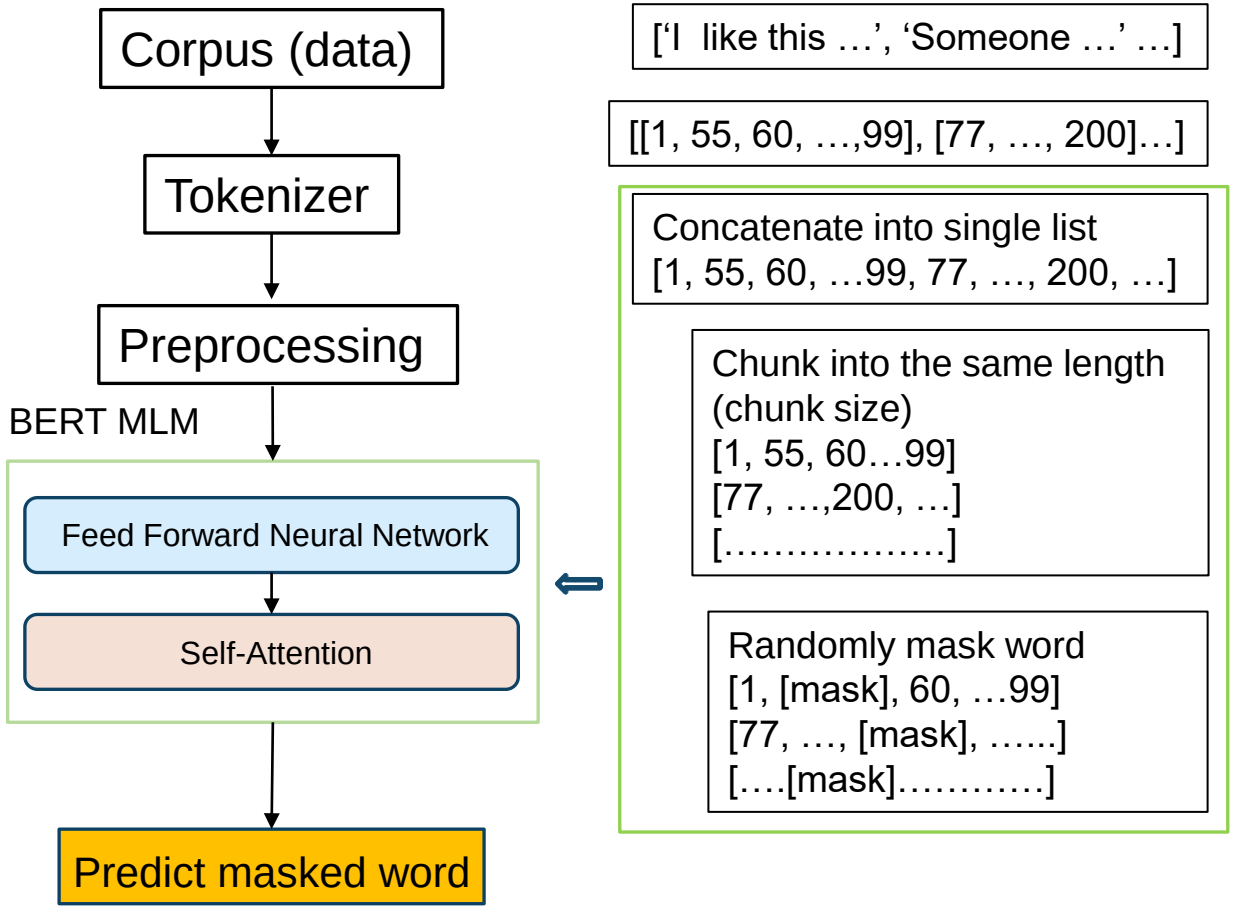
Fine Tuning

- Total size of source dataset:
  - BooksCorpus (800M words)
  - English Wikipedia (2,500 M words)
  - Totally ~26GB
- Fine tuning uncased base BERT model for CNN-news dataset,
  - 6 different categories: business, entertainment, health, news, politics, sport
  - Total size of target dataset:
    - ~38k cases
    - ~24MB
- Fine tuning with classifier layer only:
  - # of trainable parameters = 4,614
  - Converge after 6th epoch
  - Training time for each epoch: 3.5 mins
  - F1: 0.46, acc: 0.66
- Fine tuning the whole model:
  - # of trainable parameters = 109,486,854
  - Converge after 9th epoch
  - Training time for each epoch: 8.8 mins
  - F1: 0.85, acc: 0.96

Hugging Face

Masked Language Model

- Predict masked word based on given context
- Try to learn the language rules



Summary

Masked Language Model

- Uncased BERT masked language Model
- OpenWebText: ~16GB
- Vocab_size = 30k
- Used based BERT config
- Chunk_size = 128
- Batch_size =256
- 10K steps/checkpoint
- 132 mins / checkpoint
- CPU: AMD Ryzen Threadripper 3960X 24-Core Processor
- Memory: 256GB
- GPU: 2 X NVIDIA RTX A5000 (24GB)

Reference

Tokenization in NLP:
https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/

Word embedding:
https://www.turing.com/kb/guide-on-word-embeddings-in-nlp

Machine Learning course (Stanford)
https://www.youtube.com/watch?v=jGwO_UgTS7I

Hugging Face NLP course:
https://huggingface.co/learn/nlp-course/chapter1/1

Attention is all you need:
https://arxiv.org/abs/1706.03762

Bert:
https://arxiv.org/abs/1810.04805

Q&A