Foundations of NLP

Lecture-11

BERT, Masked Language Modeling and LLMs

Mahindra™ University

Global Thinkers. Engaged Leaders.

Recap

- Sequence to Sequence learning
- Attention
- Transformers
- BERT
- Word-piece tokenization Algorithm
- Domain-specific BERT
- Implementation of BERT

Word-piece tokenization Algorithm

https://huggingface.co/learn/nlp-course/chapter6/6?fw=pt%E2%80%8B

Clinical BERT embeddings

Look at the Dataset

https://aclanthology.org/W19-1909.pdf

Masked Language Modeling

Masked Language Modeling (MLM) (Devlin et al., 2019).

MLM uses unannotated text from a large corpus. Here, the model is presented with a series of sentences from the training corpus where a random sample of tokens from each training sequence is selected for use in the learning task.

Once chosen, a token is used in one of three ways:

- **It is replaced with the unique vocabulary token [MASK].**
- **It is replaced with another token from the vocabulary, randomly sampled based on token unigram probabilities.**
- **It is left unchanged.**

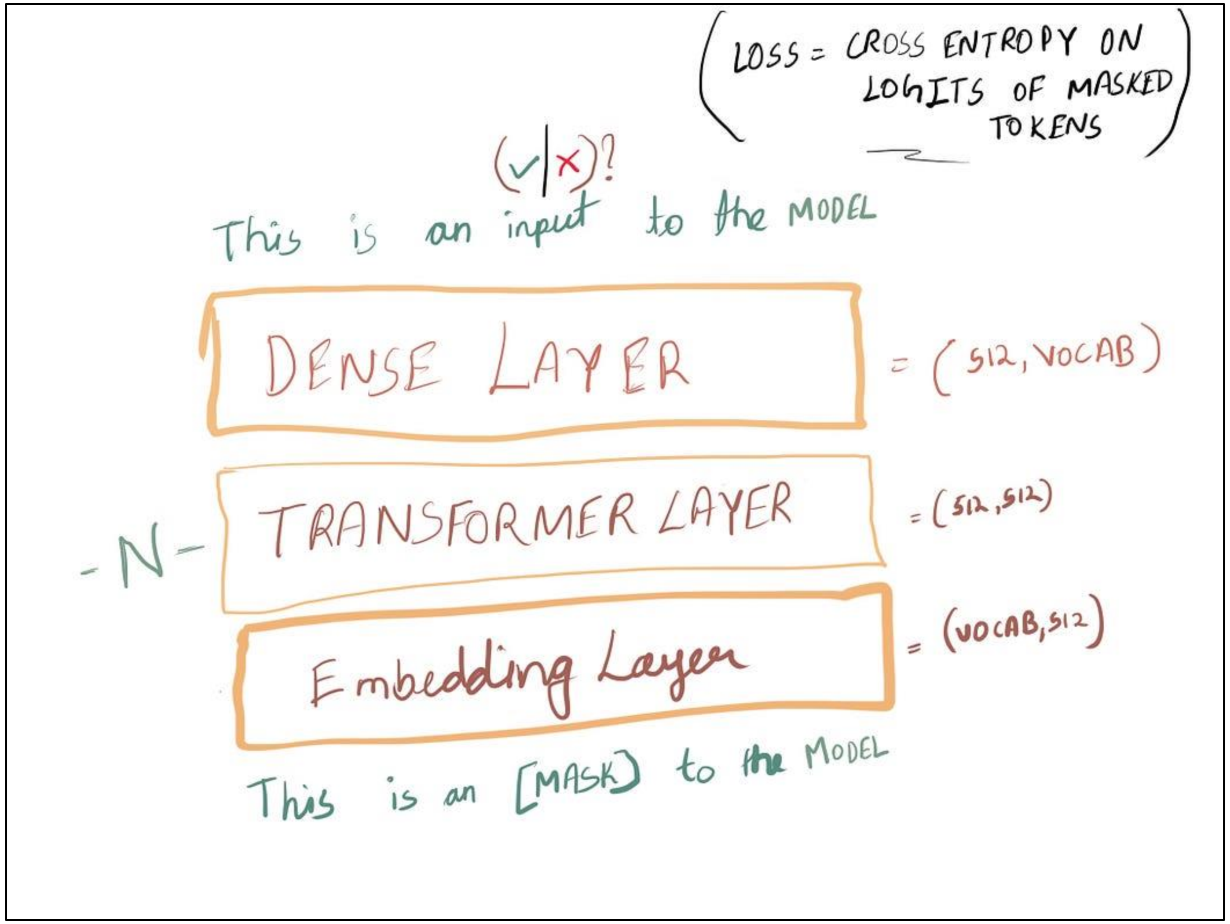
https://web.stanford.edu/~jurafsky/slp3/11.pdf

Masked Language Models

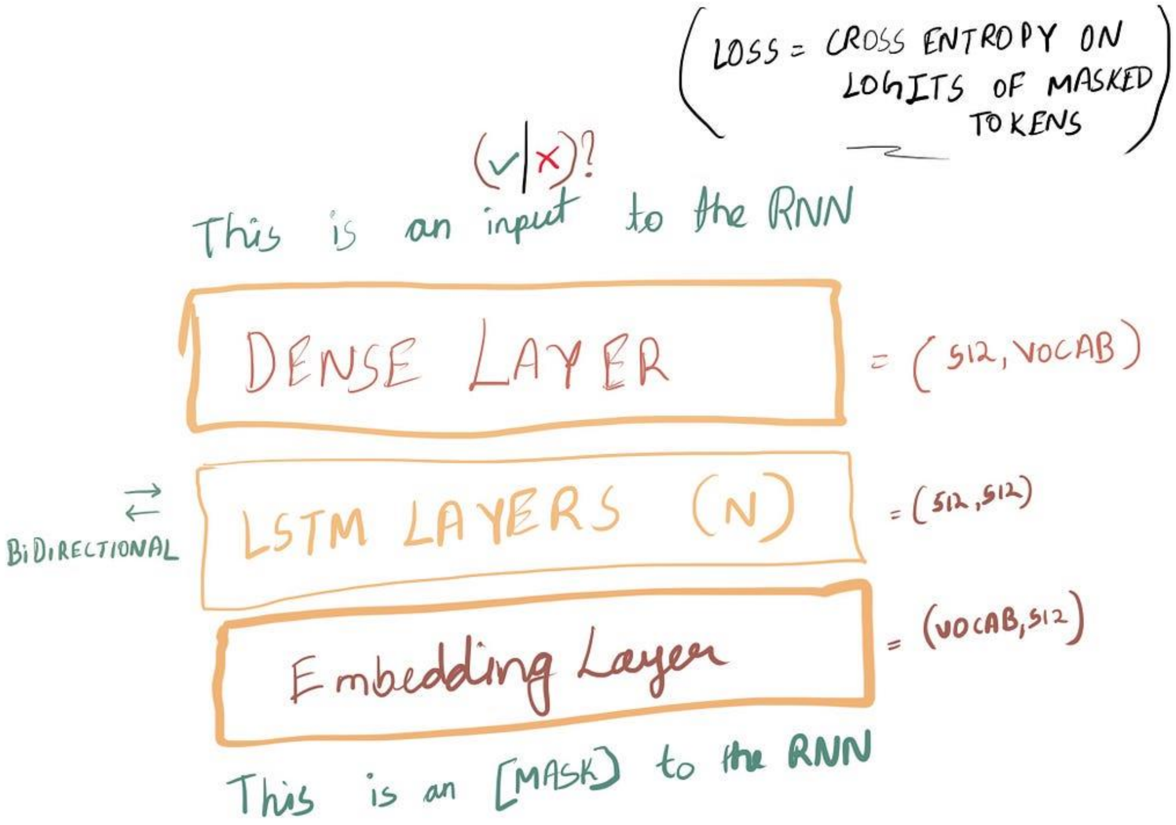
BERT:

- This model is trained via masked language modeling, where instead of predicting the **following word**, we **mask** a word in the middle and **ask the model** to guess the word given the words on both sides.
- Sometimes called **encoder-only,** because they produce an encoding for each input token but generally aren’t used to produce running text by decoding/sampling.
- **Masked language models are not used for generation**
- They are generally instead used for **interpretative** tasks

MLM Objective


LOSS = CROSS ENTROPY ON LOGITS OF MASKED TOKENS
(✓|✗)?
This is an input to the MODEL
DENSE LAYER
= (512, VOCAB)
-N-
TRANSFORMER LAYER
= (512, 512)
Embedding Layer
= (VOCAB, 512)
This is an [MASK] to the MODEL

MLM Objective


LOSS = CROSS ENTROPY ON LOGITS OF MASKED TOKENS
(✓|✗)?
This is an input to the RNN
DENSE LAYER
= (512, VOCAB)
⇄
BiDIRECTIONAL
LSTM LAYERS (N)
= (512, 512)
Embedding Layer
= (VOCAB, 512)
This is an [MASK] to the RNN

Why MLM is useful

- **Contextualized Representations:** By predicting masked words based on their surrounding words, the **model learns rich, context-sensitive representations of words**. This helps the model understand nuanced meanings and word relationships, making it more effective at tasks like question answering, sentiment analysis, and text classification.
- **Pretraining for Transfer Learning: MLM is often used as a pretraining objective in models like BERT**. After pretraining on a large corpus using MLM, the model can be fine-tuned for specific downstream tasks (like text classification or named entity recognition) with a relatively smaller amount of task-specific data.

Pretraining and Fine-tuning

- **Details of BERT Architecture in the below Research Paper [Must Read!!]**

https://arxiv.org/pdf/1810.04805

Transfer Learning

This aspect of the pretrain-finetune paradigm is an instance of what is called **transfer learning** in machine learning: the method of acquiring knowledge from one task or domain, and then applying it (transferring it) to solve a new task.

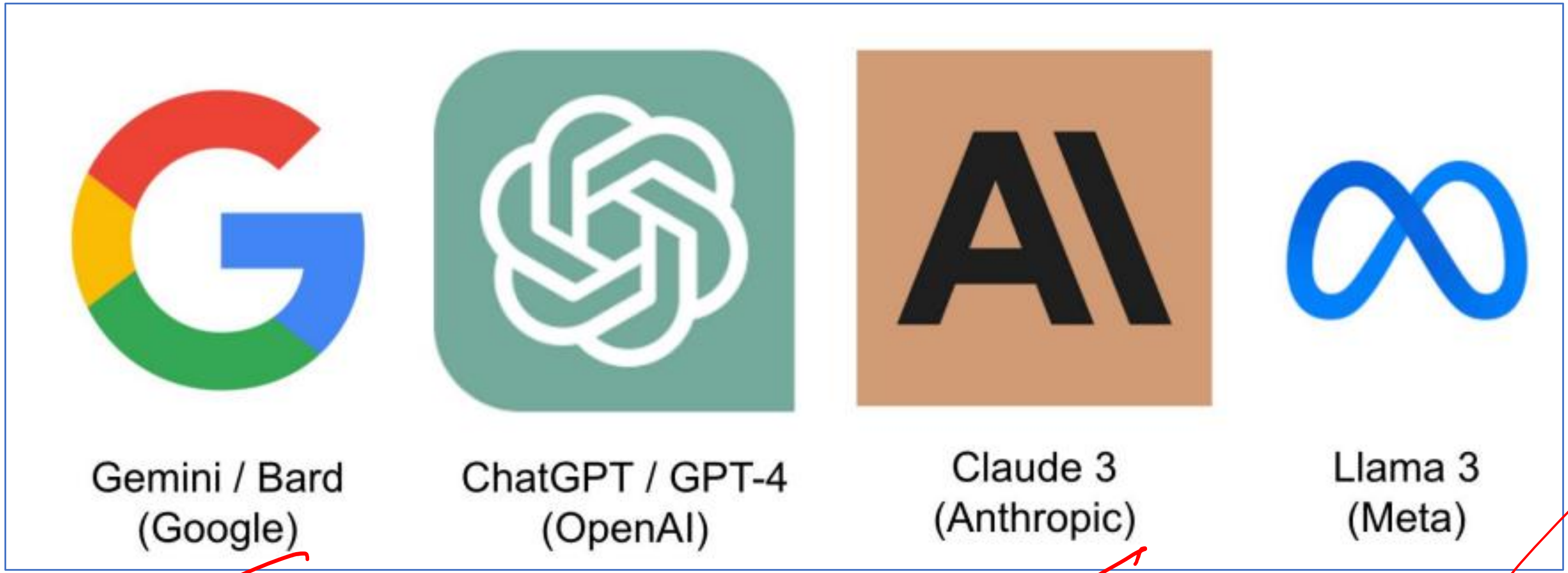
Limitations of BERT

- Devlin et al. (2019:§5): admirably detailed but still partial ablation studies and optimization studies.
- 2. Devlin et al. (2019): “The first [downside] is that we are creating a mismatch between pre-training and fine-tuning, since the [MASK] token is never seen during fine-tuning.”
- 3. Devlin et al. (2019): “The second downside of using an MLM is that only 15% of tokens are predicted in each batch”
- 4. Yang et al. (2019): “BERT assumes the predicted tokens are independent of each other given the unmasked tokens, which is oversimplified as high-order, long-range dependency is prevalent in natural language”

Domain-specific use-cases for Fine-tuned LLMs

- Support issue prioritization
- Fraud detection
- Blog writing
- Lead qualification
- Text classification
- Question answering
- NER

Recent Large Language Models

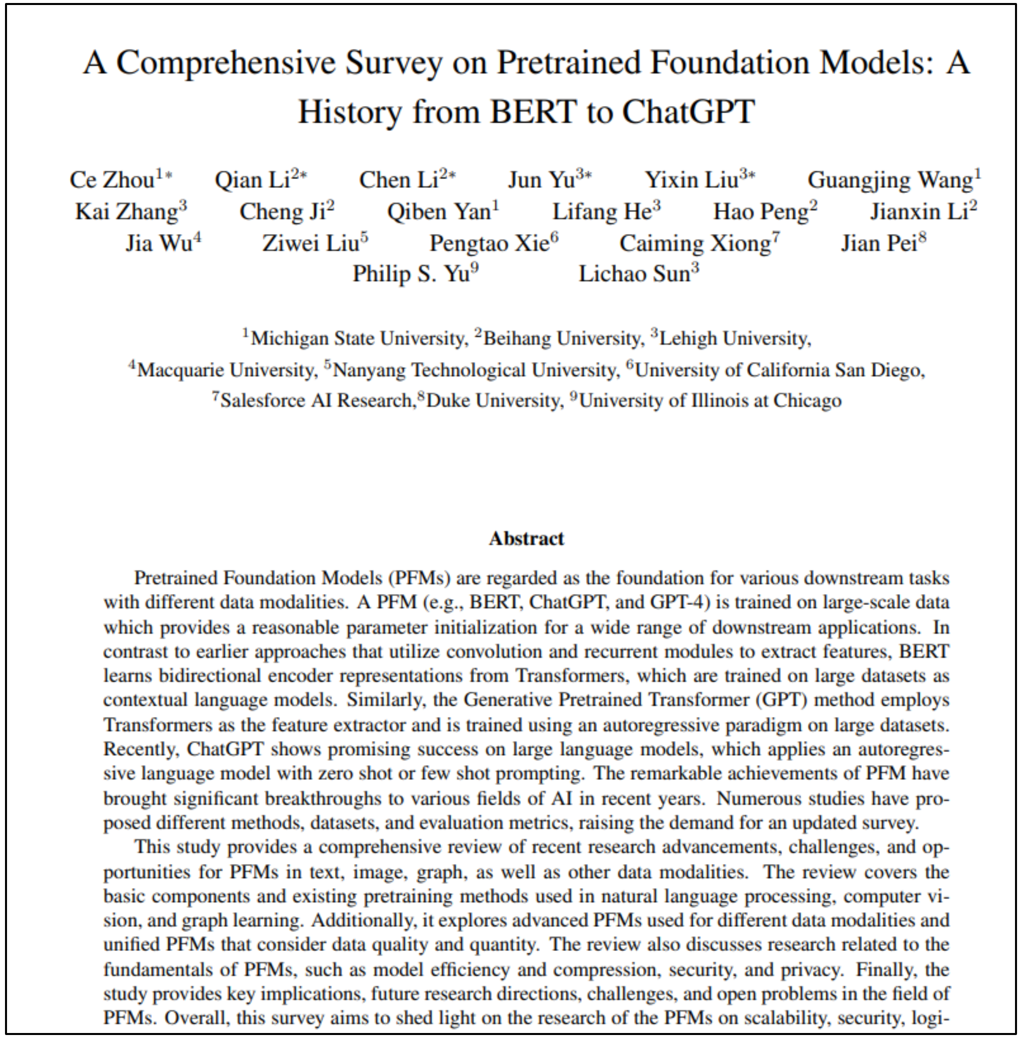


Leaderboard performance: SuperGlue

Super glue is the leaderboard over the range of challenging NLP tasks suchh as question answering, machine translation, CoRef, etc.

https://super.gluebenchmark.com/leaderboard/

History from BERT to ChatGPT



A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT

Ce Zhou[1*] Qian Li[2*] Chen Li[2*] Jun Yu[3*] Yixin Liu[3*] Guangjing Wang[1]
Kai Zhang[3] Cheng Ji[2] Qiben Yan[1] Lifang He[3] Hao Peng[2] Jianxin Li[2]
Jia Wu[4] Ziwei Liu[5] Pengtao Xie[6] Caiming Xiong[7] Jian Pei[8]
Philip S. Yu[9] Lichao Sun[3]

[1]Michigan State University, [2]Beihang University, [3]Lehigh University,
[4]Macquarie University, [5]Nanyang Technological University, [6]University of California San Diego,
[7]Salesforce AI Research,[8]Duke University, [9]University of Illinois at Chicago

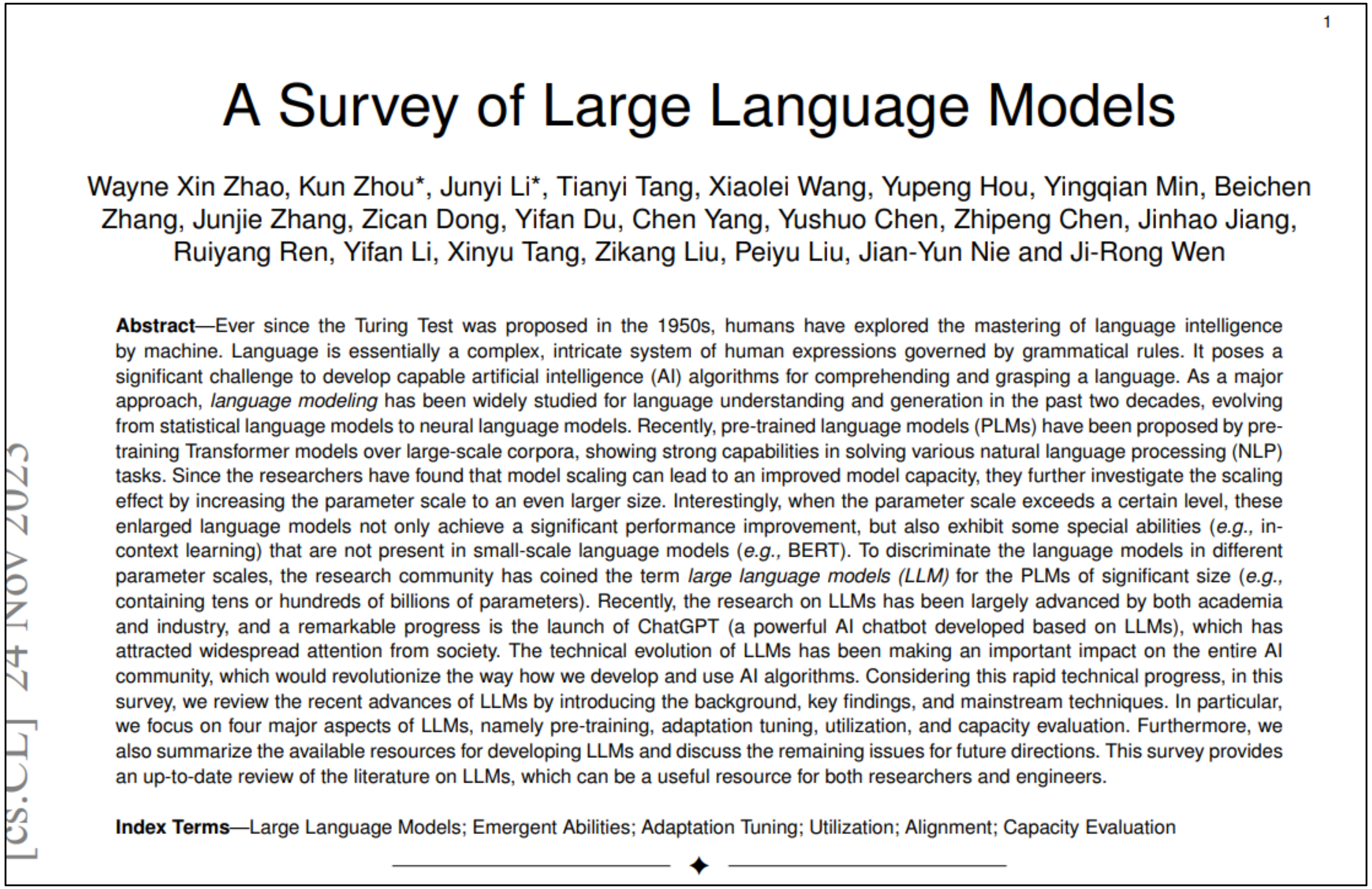
Abstract

Pretrained Foundation Models (PFMs) are regarded as the foundation for various downstream tasks with different data modalities. A PFM (e.g., BERT, ChatGPT, and GPT-4) is trained on large-scale data which provides a reasonable parameter initialization for a wide range of downstream applications. In contrast to earlier approaches that utilize convolution and recurrent modules to extract features, BERT learns bidirectional encoder representations from Transformers, which are trained on large datasets as contextual language models. Similarly, the Generative Pretrained Transformer (GPT) method employs Transformers as the feature extractor and is trained using an autoregressive paradigm on large datasets. Recently, ChatGPT shows promising success on large language models, which applies an autoregressive language model with zero shot or few shot prompting. The remarkable achievements of PFM have brought significant breakthroughs to various fields of AI in recent years. Numerous studies have proposed different methods, datasets, and evaluation metrics, raising the demand for an updated survey.

This study provides a comprehensive review of recent research advancements, challenges, and opportunities for PFMs in text, image, graph, as well as other data modalities. The review covers the basic components and existing pretraining methods used in natural language processing, computer vision, and graph learning. Additionally, it explores advanced PFMs used for different data modalities and unified PFMs that consider data quality and quantity. The review also discusses research related to the fundamentals of PFMs, such as model efficiency and compression, security, and privacy. Finally, the study provides key implications, future research directions, challenges, and open problems in the field of PFMs. Overall, this survey aims to shed light on the research of the PFMs on scalability, security, logi-

https://arxiv.org/pdf/2302.09419

Large Language models are on Hype!!!!!



1

A Survey of Large Language Models

Wayne Xin Zhao, Kun Zhou*, Junyi Li*, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie and Ji-Rong Wen

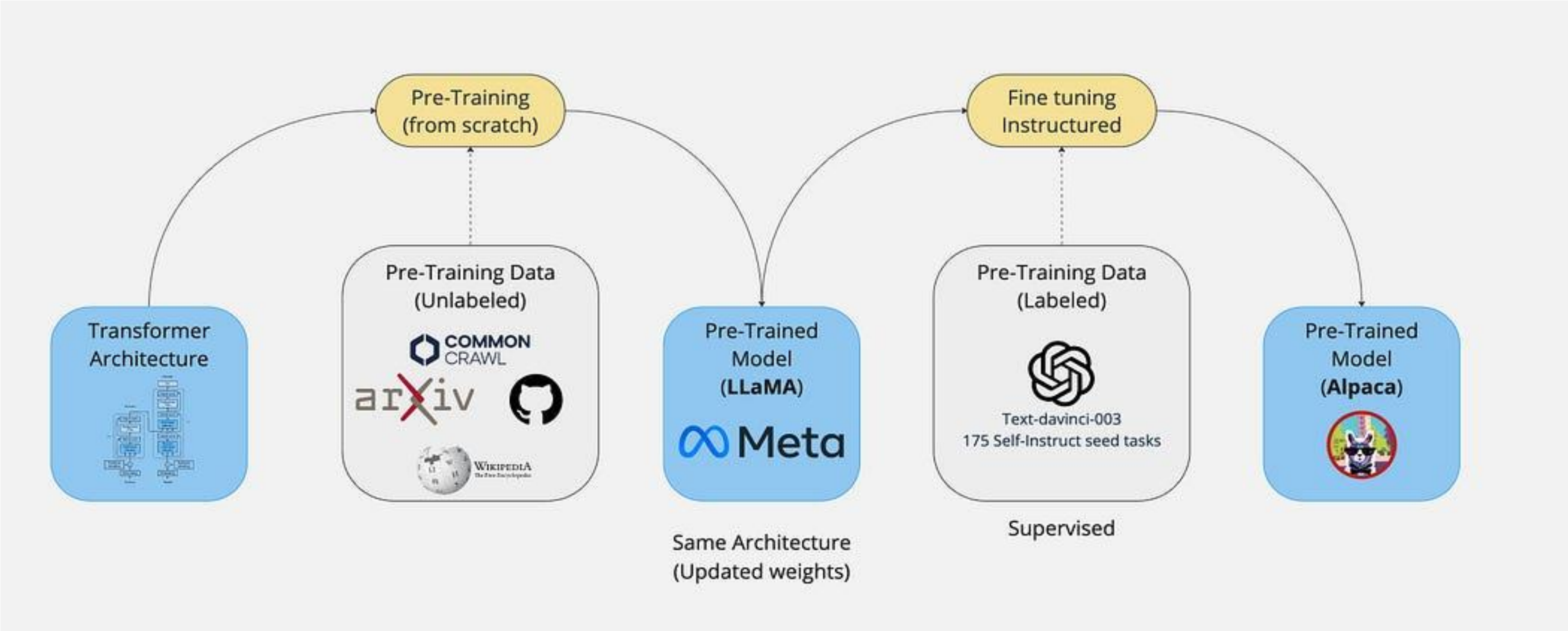
Abstract—Ever since the Turing Test was proposed in the 1950s, humans have explored the mastering of language intelligence by machine. Language is essentially a complex, intricate system of human expressions governed by grammatical rules. It poses a significant challenge to develop capable artificial intelligence (AI) algorithms for comprehending and grasping a language. As a major approach, *language modeling* has been widely studied for language understanding and generation in the past two decades, evolving from statistical language models to neural language models. Recently, pre-trained language models (PLMs) have been proposed by pre-training Transformer models over large-scale corpora, showing strong capabilities in solving various natural language processing (NLP) tasks. Since the researchers have found that model scaling can lead to an improved model capacity, they further investigate the scaling effect by increasing the parameter scale to an even larger size. Interestingly, when the parameter scale exceeds a certain level, these enlarged language models not only achieve a significant performance improvement, but also exhibit some special abilities (*e.g.*, in-context learning) that are not present in small-scale language models (*e.g.*, BERT). To discriminate the language models in different parameter scales, the research community has coined the term *large language models (LLM)* for the PLMs of significant size (*e.g.*, containing tens or hundreds of billions of parameters). Recently, the research on LLMs has been largely advanced by both academia and industry, and a remarkable progress is the launch of ChatGPT (a powerful AI chatbot developed based on LLMs), which has attracted widespread attention from society. The technical evolution of LLMs has been making an important impact on the entire AI community, which would revolutionize the way how we develop and use AI algorithms. Considering this rapid technical progress, in this survey, we review the recent advances of LLMs by introducing the background, key findings, and mainstream techniques. In particular, we focus on four major aspects of LLMs, namely pre-training, adaptation tuning, utilization, and capacity evaluation. Furthermore, we also summarize the available resources for developing LLMs and discuss the remaining issues for future directions. This survey provides an up-to-date review of the literature on LLMs, which can be a useful resource for both researchers and engineers.

Index Terms—Large Language Models; Emergent Abilities; Adaptation Tuning; Utilization; Alignment; Capacity Evaluation

[cs.CL] 24 Nov 2023

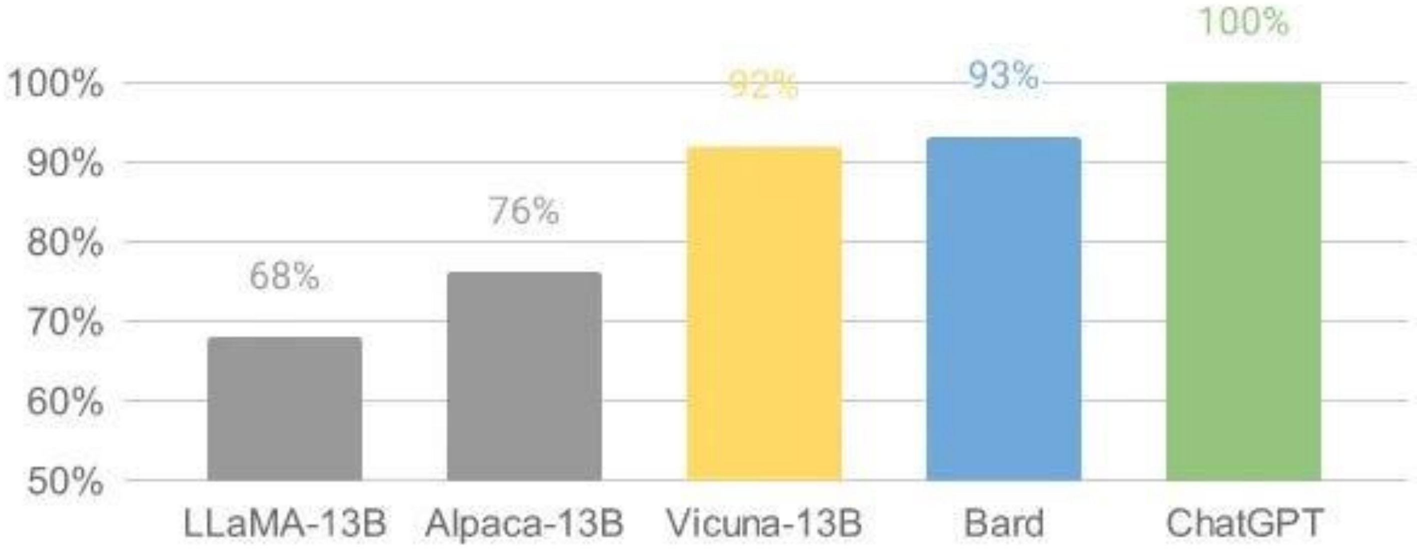
https://arxiv.org/abs/2303.18223

Example of fine-tuning a LLaMA-based model



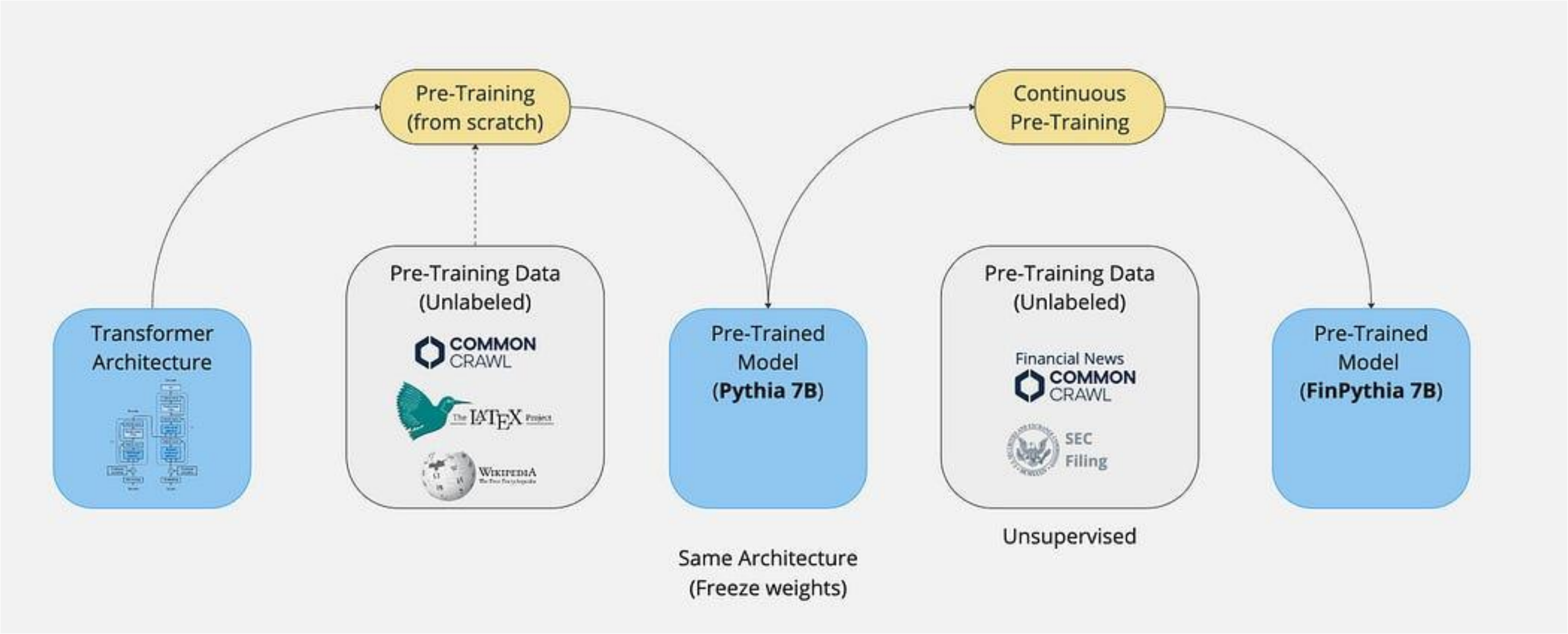
Example of fine-tuning a LLaMA-based model

Relative Response Quality Assessed by GPT-4



Relative Response Quality Assessed by GPT-4

Pre-train a Pythia based model



Example of further pre-train a Pythia based model

Acknowledgments

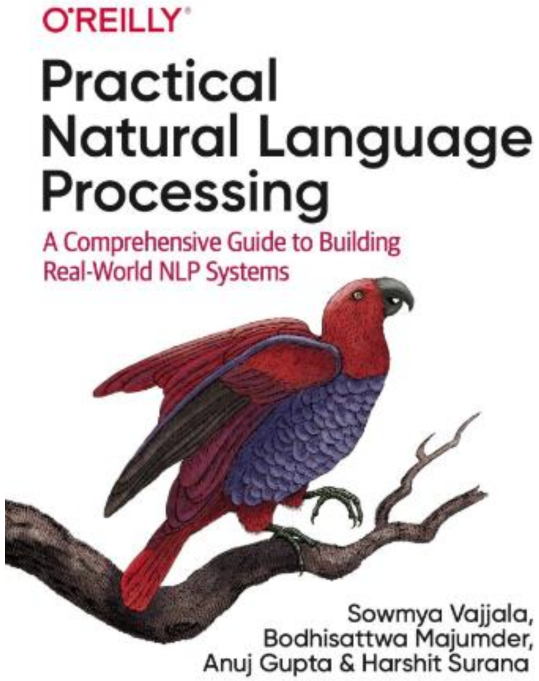
- These slides were adapted from the book SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
- Some modifications from CS224N presentations and resources found in the WEB by several scholars.

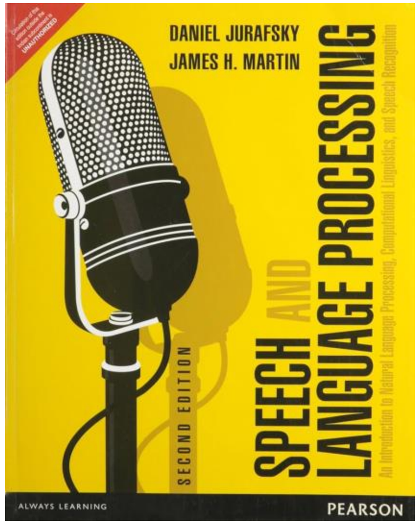
Reference materials

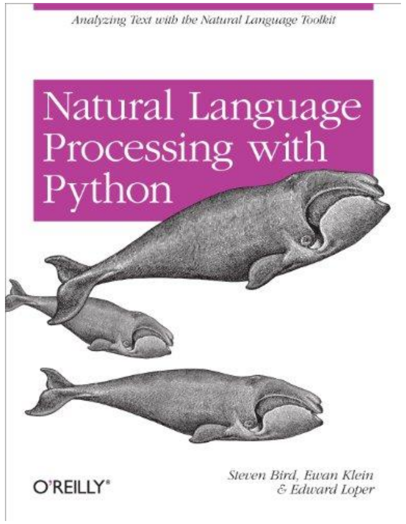
- https://lab.vlanc.co.in/teachings/fall2024-AI-schedule.html

- Lecture notes

- (A) Speech and Language Processing by Daniel Jurafsky and James H. Martin
- (B) Natural Language Processing with Python. (updated edition based on Python 3 and NLTK 3) Steven Bird et al. O'Reilly Media


O'REILLY
Practical Natural Language Processing
A Comprehensive Guide to Building Real-World NLP Systems
Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta & Harshit Surana


DANIEL JURAFSKY
JAMES H. MARTIN
SECOND EDITION
SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
ALWAYS LEARNING
PEARSON


Analyzing Text with the Natural Language Toolkit
Natural Language Processing with Python
O'REILLY
Steven Bird, Ewan Klein & Edward Loper