

DSME 6635: Artificial Intelligence for Business Research

Deep-Learning-based NLP: Pretraining

Renyu (Philip) Zhang

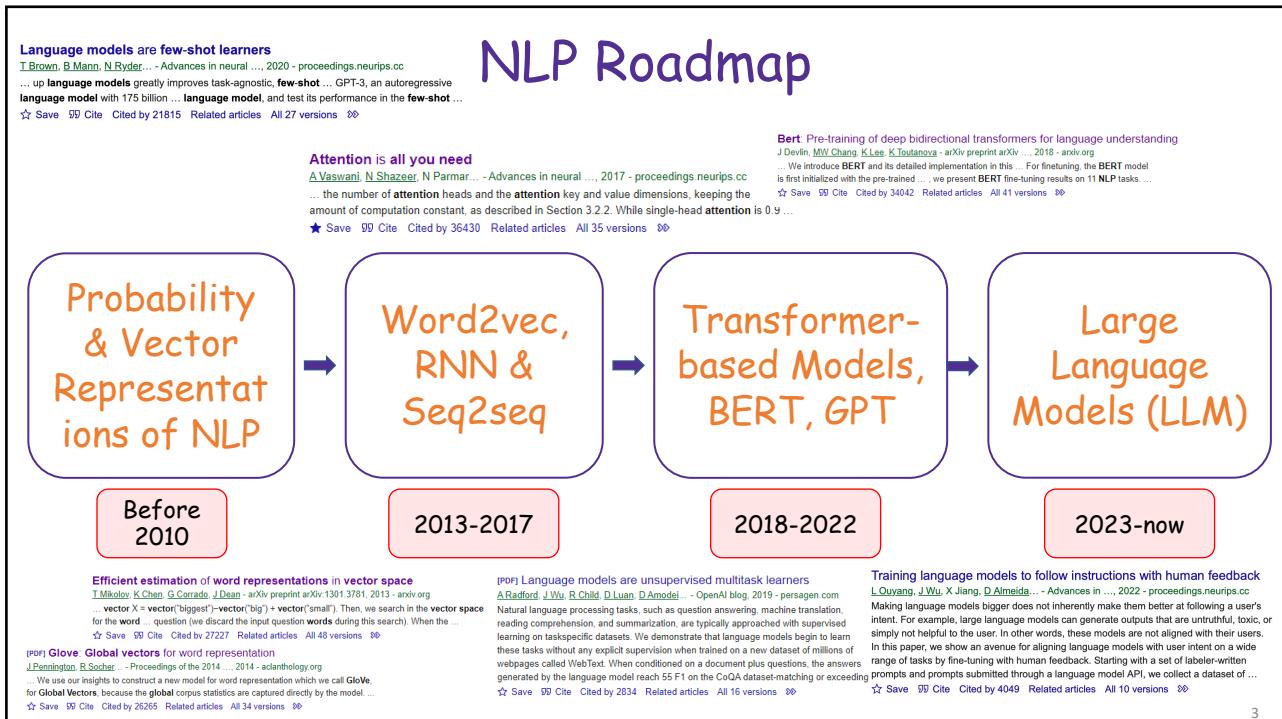
1

Agenda

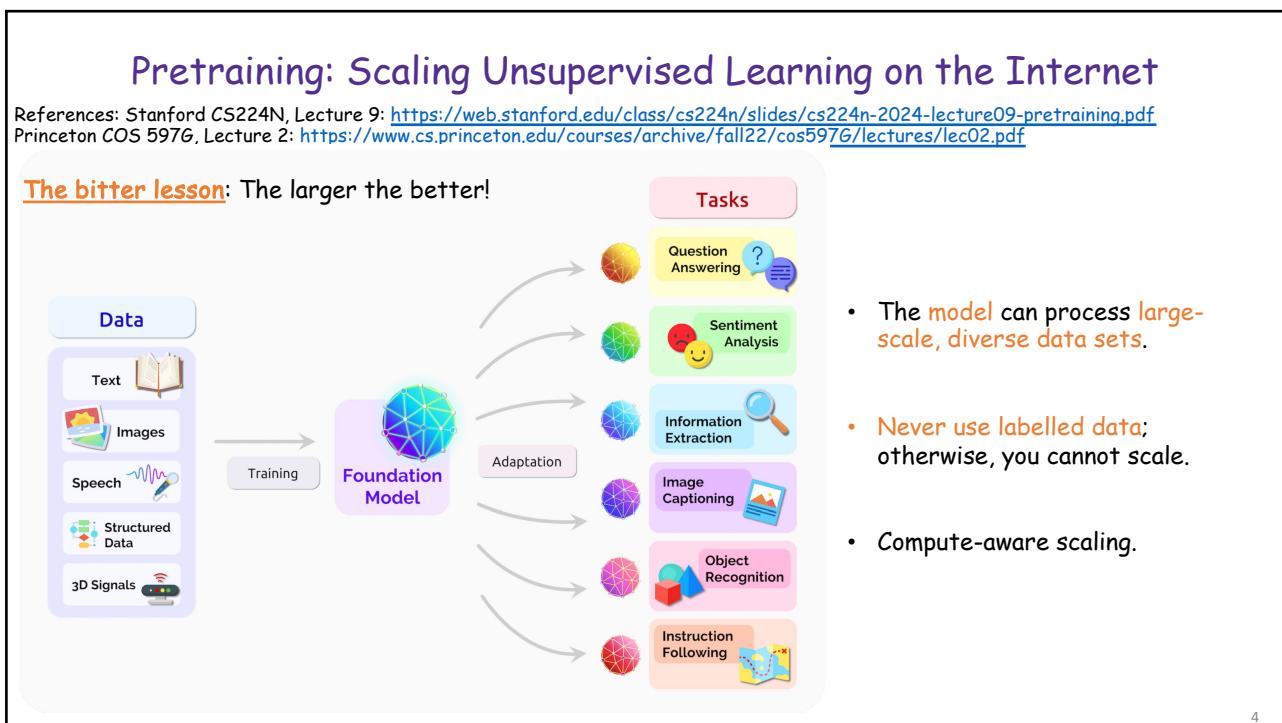
- BERT: Bidirectional Encoder Representations from Transformers
- GPT: Generative Pretrained Transformers
- Applications in Econ/Business Research

2

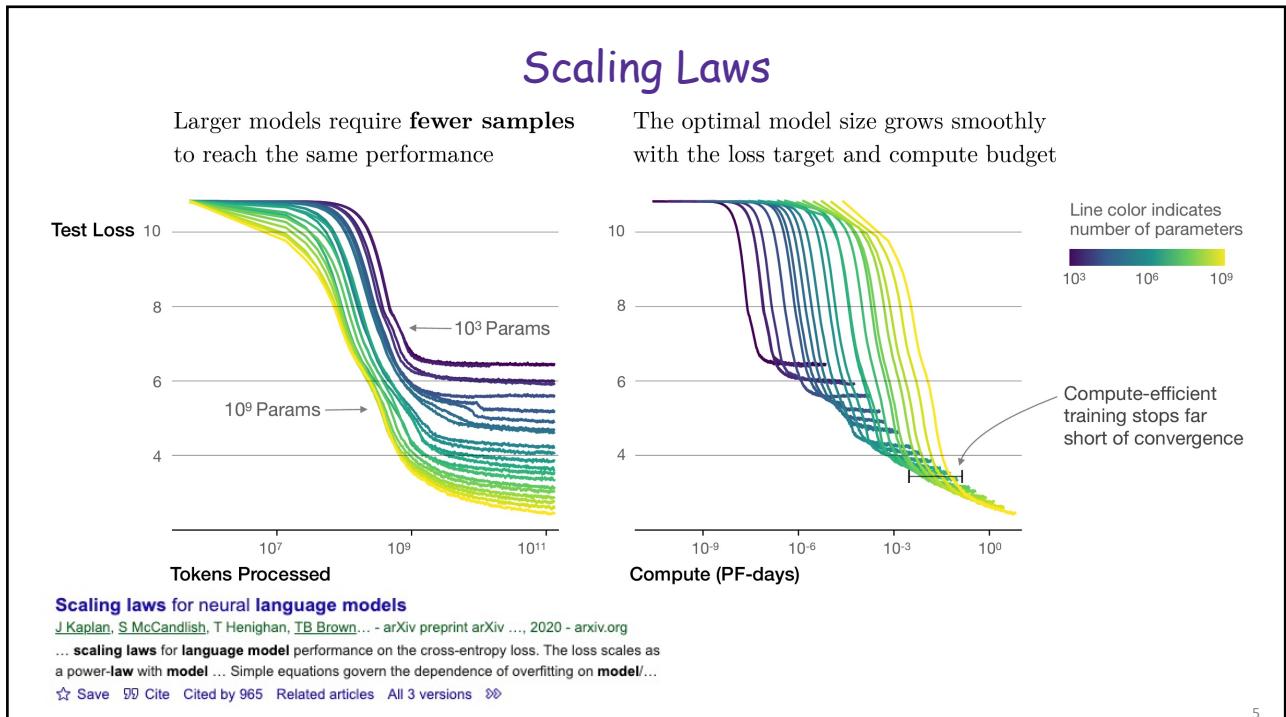
2



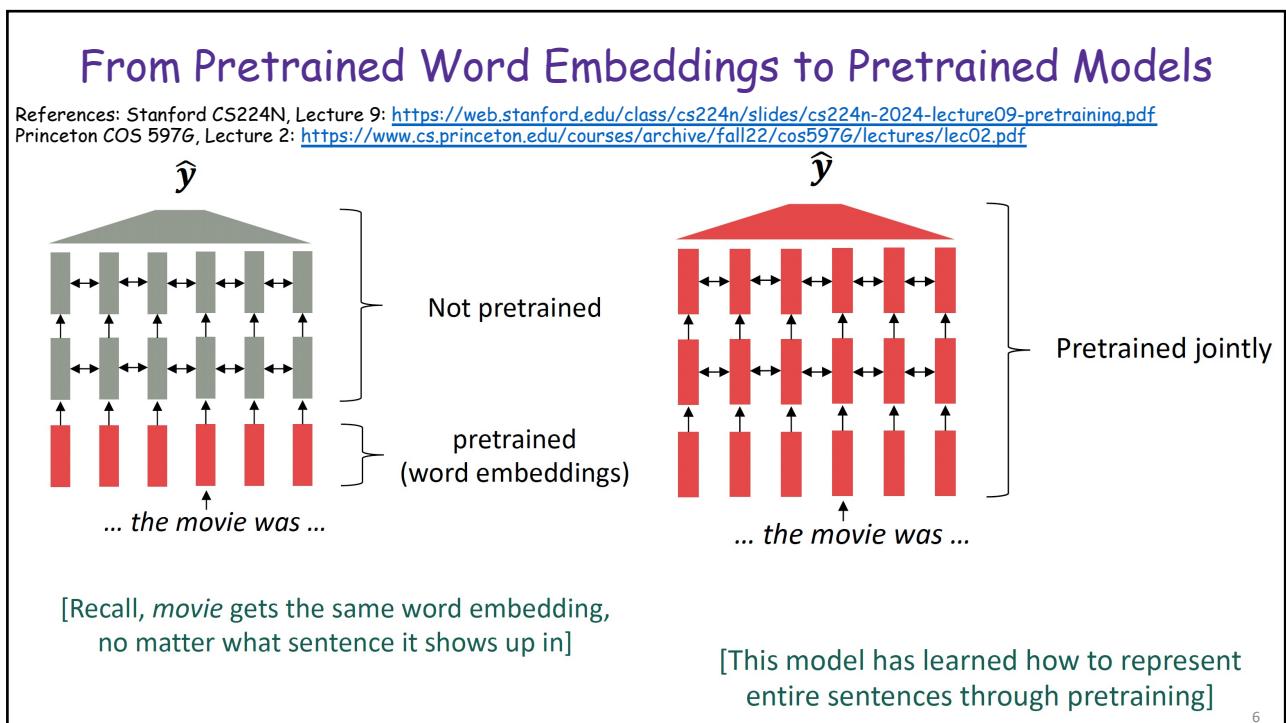
3



4



5



6

6

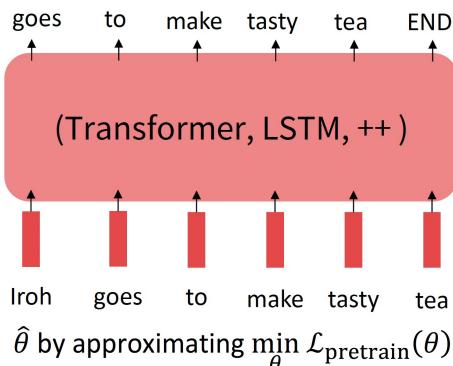
From Pretrained Word Embeddings to Pretrained Models

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>
Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>

- Pretraining can improve downstream NLP applications by serving as **parameter initialization**.

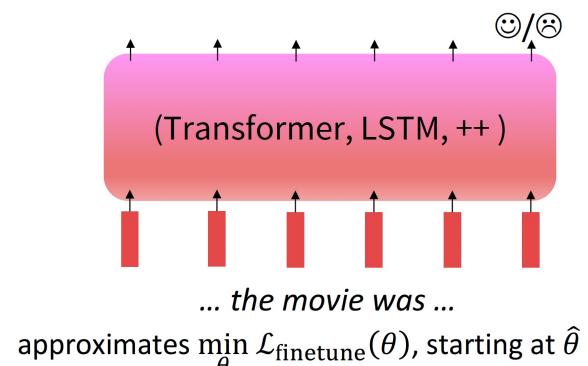
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



Step 2: Finetune (on your task)

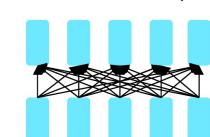
Not many labels; adapt to the task!



7

Three Pretraining Architectures

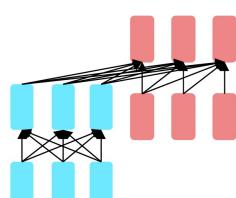
References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>
Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



Encoders

- Can condition on future.

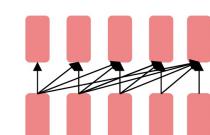
- Example: BERT.



Encoder-Decoders

- Combining encoder and decoder.

- Example: T5



Decoders

- Cannot condition on future.

- Example: GPT

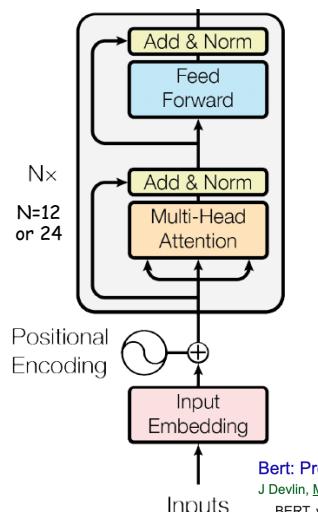
- All (very) large language models are decoders.

8

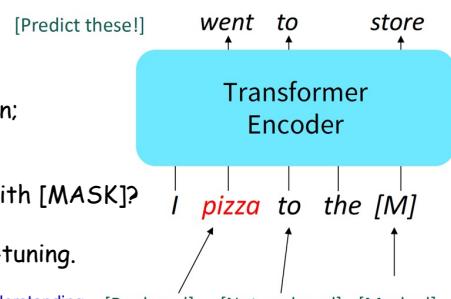
8

BERT: Bidirectional Encoder Representations from Transformers

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



- Key idea: Learn representations based on **bidirectional context**.
 - We went to the river **bank**. vs. I need to go to the **bank** to make a deposit.
- Pretraining objectives: **masked language modeling + next sentence prediction**
- 15% of tokens are randomly masked.
- The masked tokens in the inputs:
 - 80% replaced with [MASK];
 - 10% replaced with a random token;
 - 10% no change.
- Why not all masked tokens replaced with [MASK]?
 - [MASK] tokens are never seen in fine-tuning.



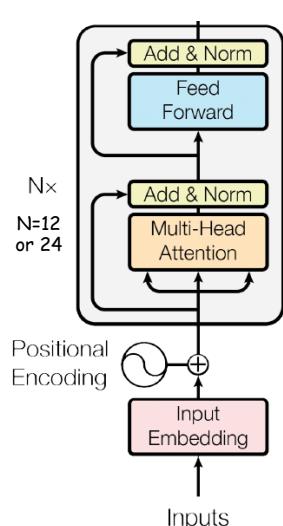
Bert: Pre-training of deep **bidirectional** transformers for language understanding [Replaced] [Not replaced] [Masked]
 J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv ..., 2018 - arxiv.org
 ... BERT, which stands for **Bidirectional Encoder Representations** from Transformers. Unlike ...
 2018, BERT is designed to pretrain deep **bidirectional representations** from unlabeled text by ...
[☆ Save](#) [⤓ Cite](#) [Cited by 93230](#) [Related articles](#) [All 46 versions](#) [⤔](#)

9

9

Next Sentence Prediction (NSP)

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



- Understanding the **relationships between two sentences** are also important.
- Reduce the gap between pretraining and finetuning.

[CLS]: a special token always at the beginning

[SEP]: a special token used to separate two segments

Input = [CLS] the man went to [MASK] store [SEP]

he bought a gallon [MASK] milk [SEP]

Label = IsNext

They sample two contiguous segments for 50% of the time and another random segment from the corpus for 50% of the time

Input = [CLS] the man [MASK] to the store [SEP]

penguin [MASK] are flight ##less birds [SEP]

Label = NotNext

10

10

Subwords and Input Embeddings

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>
 Princeton COS 597G, Lecture 2: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



- To make sure training and testing vocabularies are consistent, uncommon words are split into components.

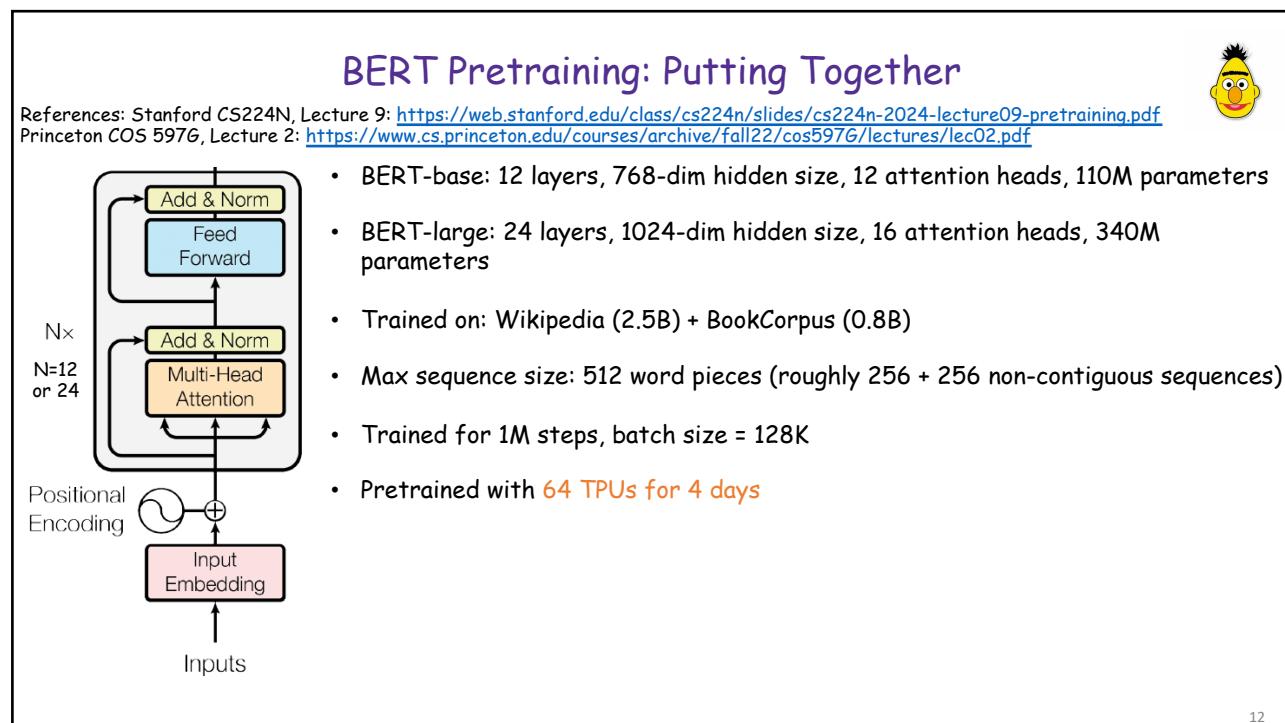
word	vocab mapping	embedding
Common words	hat → hat	█
Variations	learn → learn	█
misspellings	taaaaasty → taa## aaa## sty	█
novel items	laern → la## ern##	█
	Transformerify → Transformer## ify	█

Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	#ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{#ing}$	$E_{[SEP]}$
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B	E_B
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

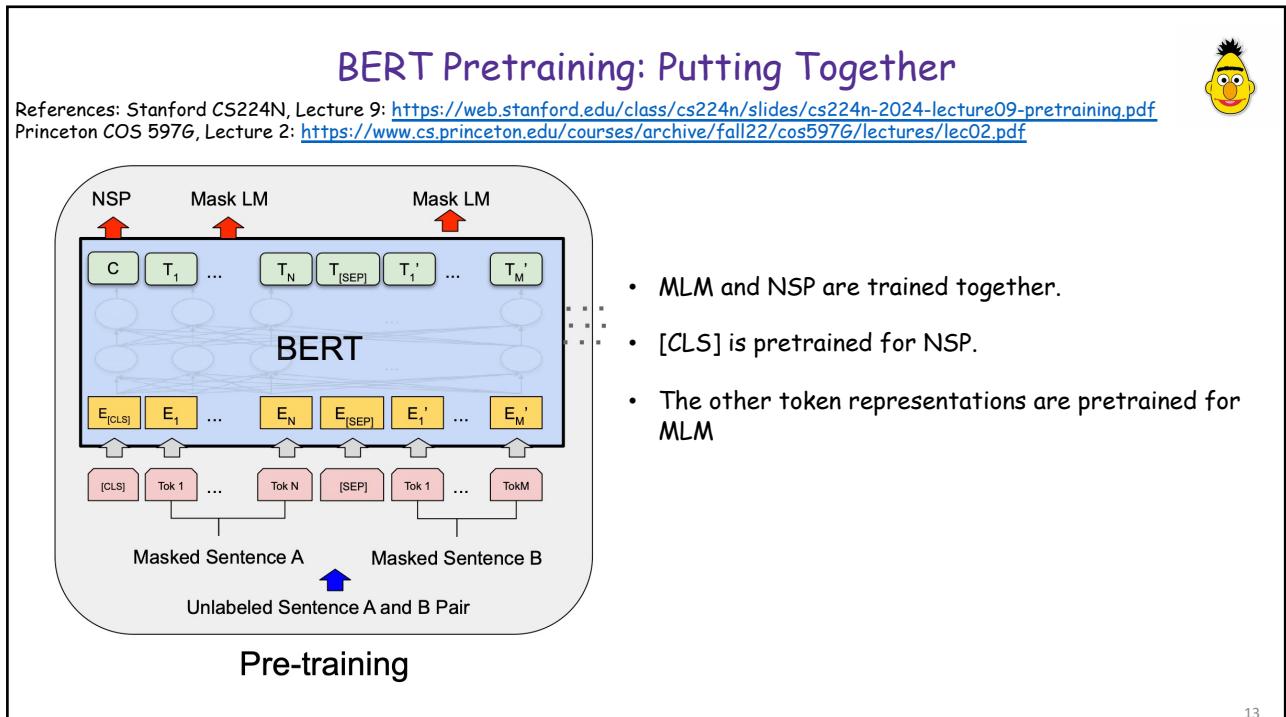
Which of the two segments?

11

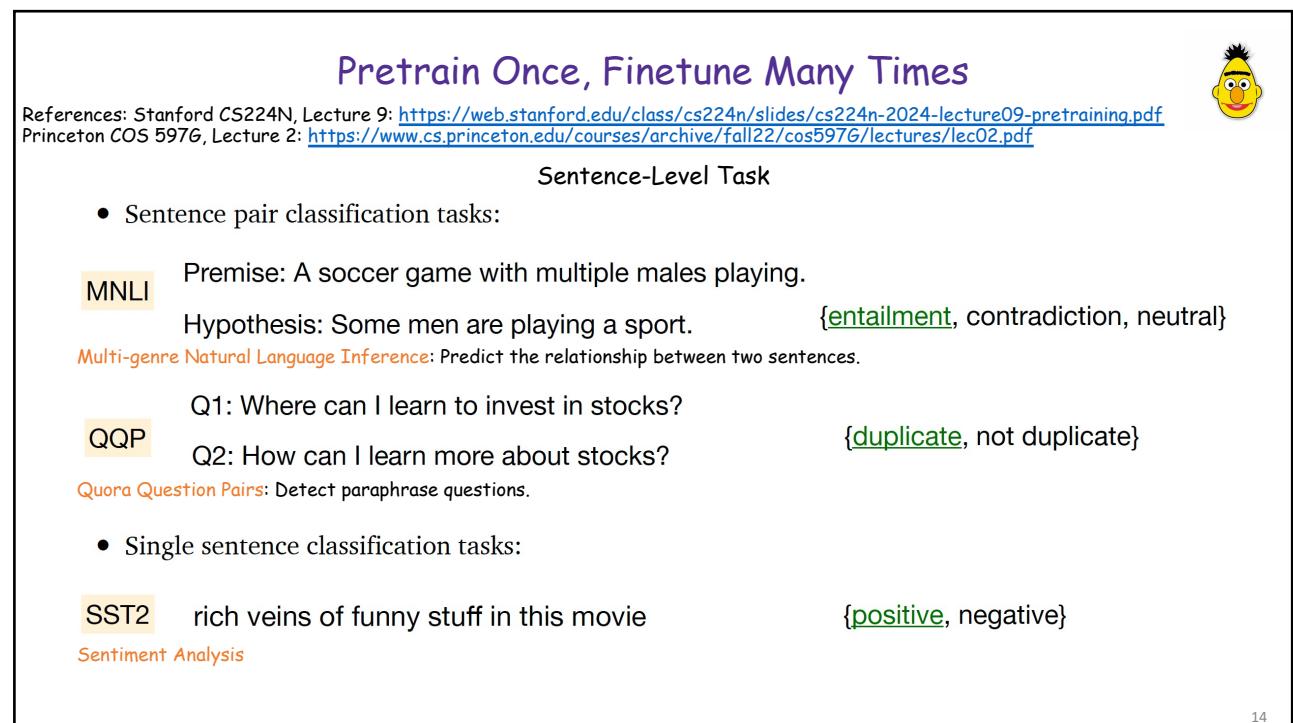
11



12

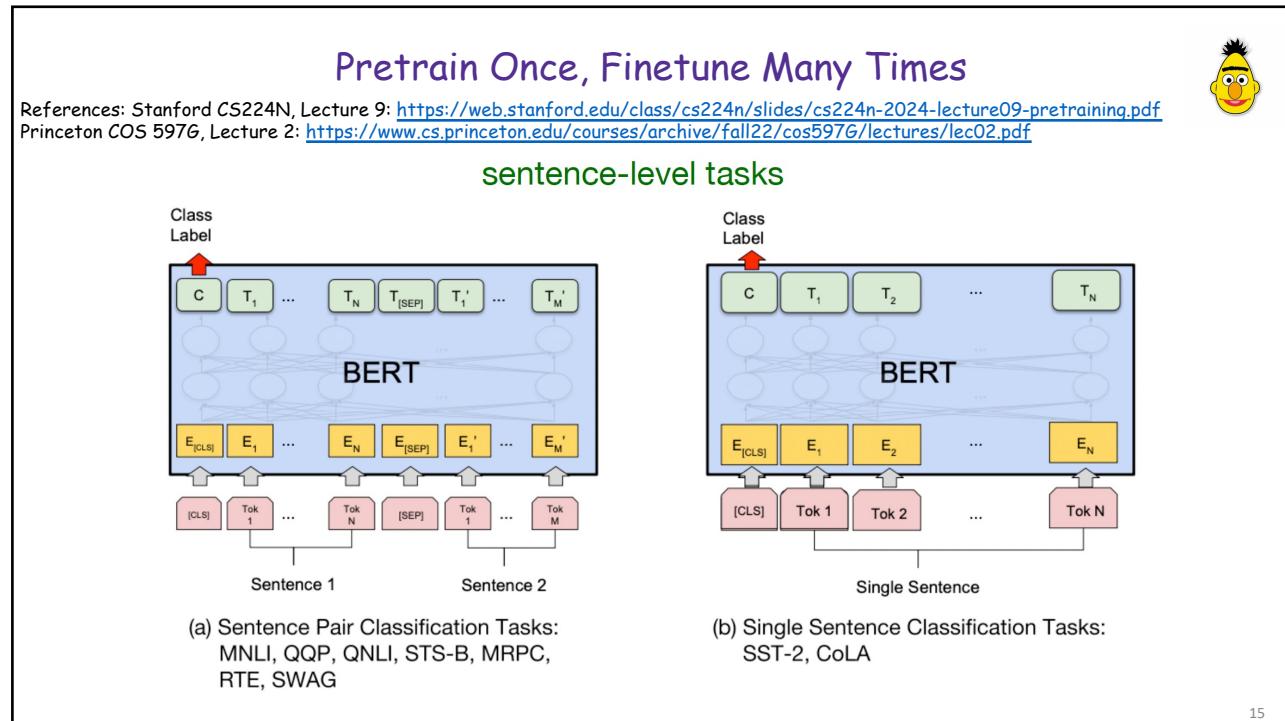


13



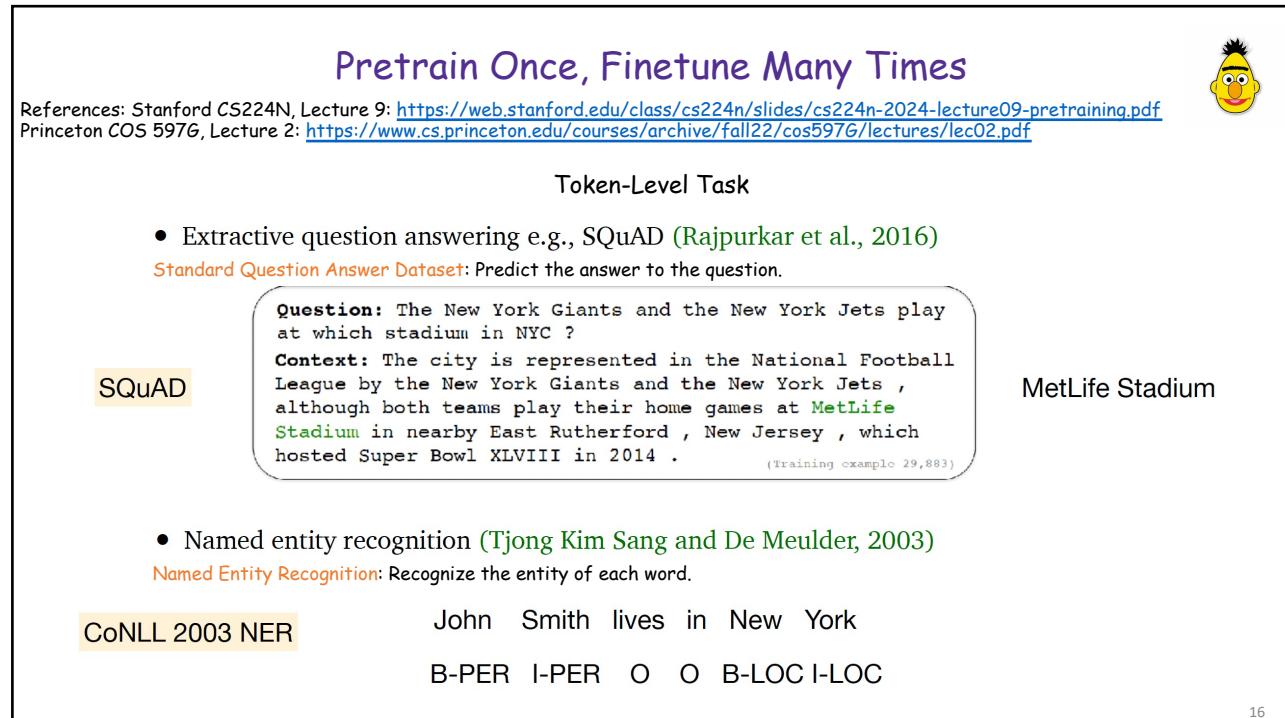
14

14



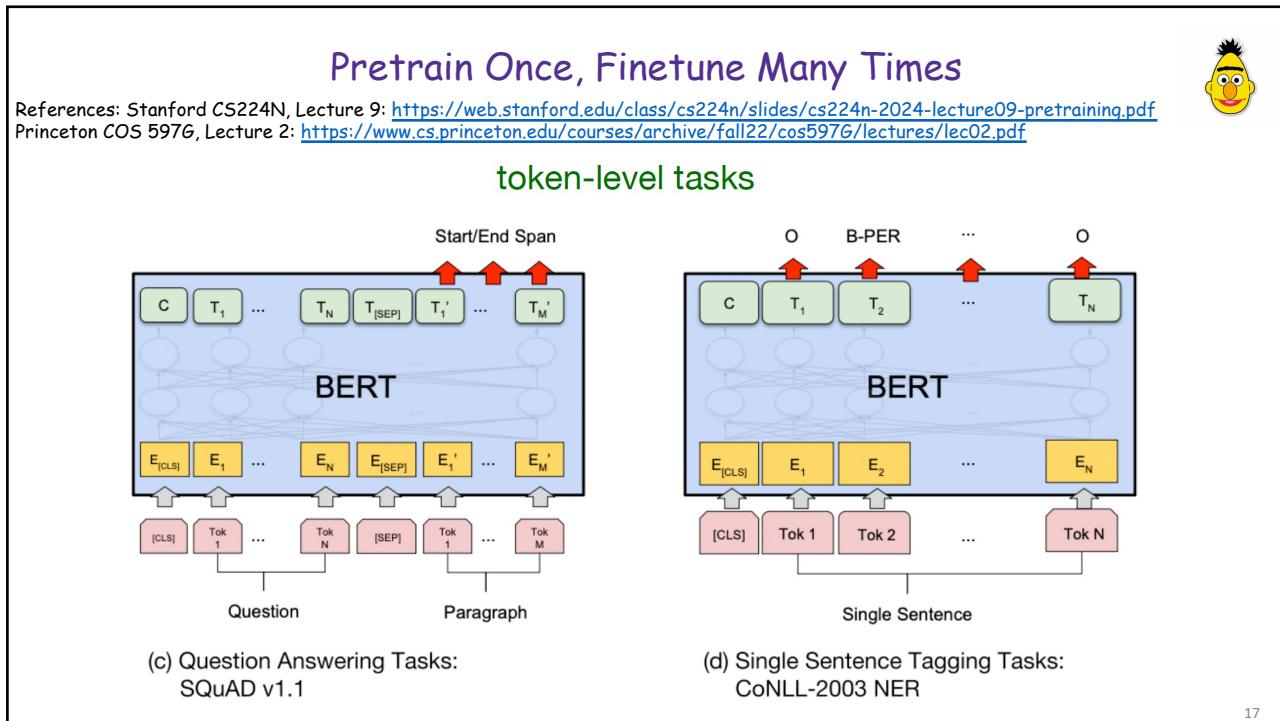
15

15

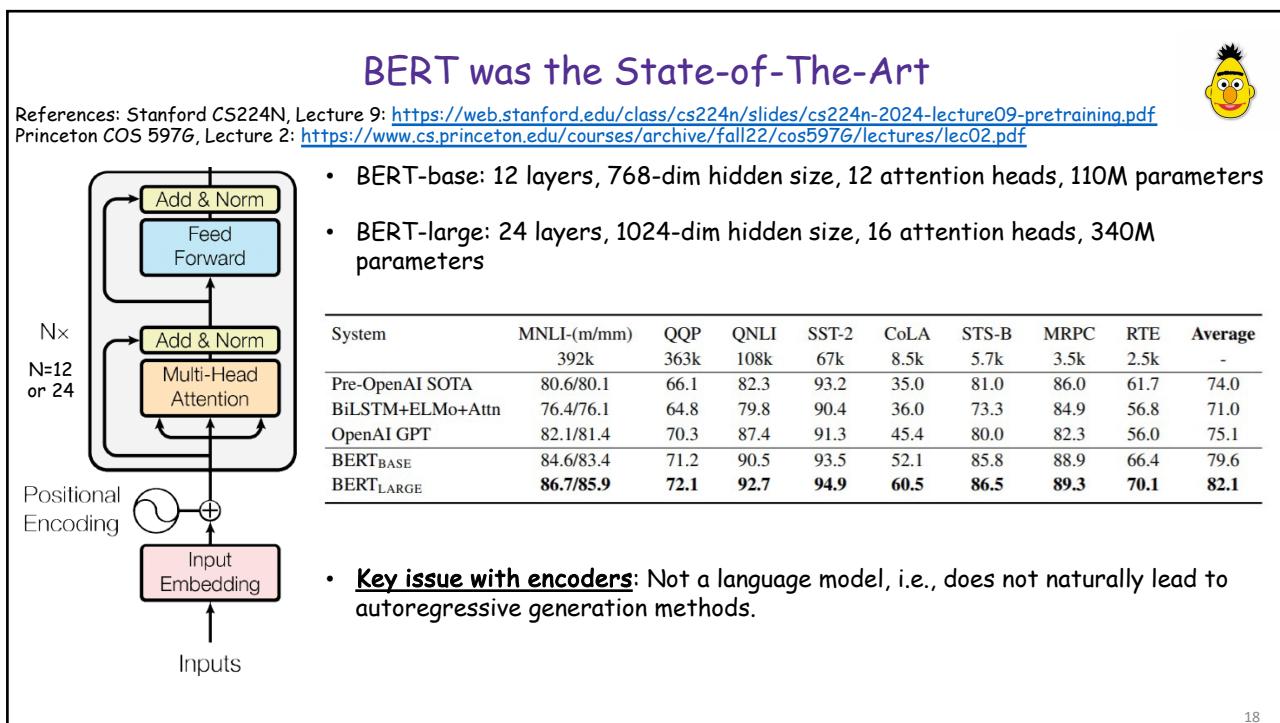


16

16



17



18

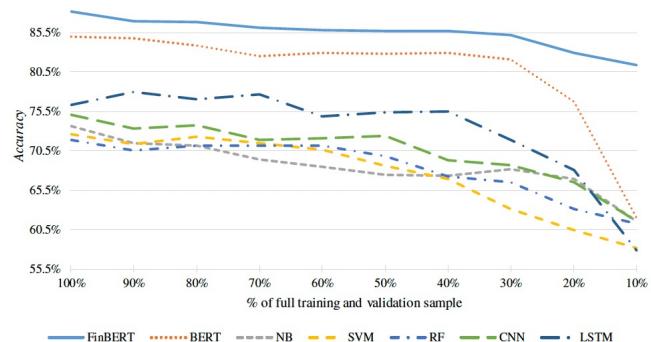
18

Revisit FinBERT



- Pretrain BERT-base using financial datasets (4.9B tokens in total) with 4 P100 GPUs (100G memory):
 - Corporate annual and quarterly filings from SEC's EDGAR website (1994-2019).
 - Financial analyst reports from Thomson Intestext database (2003-2012).
 - Earnings conference call transcripts from the SeekingAlpha website (2004-2019).
- Finetuning and evaluation:**
 - Sentiment analysis 10,000 sentences
 - 36% positive
 - 46% neutral
 - 18% negative
- Can FinBERT beat GPT-4 or Claude-3 in tasks related to financial texts?
 • How can we make **fair comparisons?**

Figure 1 Sentiment classification accuracy across sample sizes



FinBERT: A large language model for extracting information from financial text

AH Huang, H Wang, Y Yang - Contemporary Accounting ..., 2023 - Wiley Online Library

... model that adapts to the **finance** domain. We show that **FinBERT** incorporates **finance** knowledge and can better summarize contextual **information** in **financial texts**. Using a sample of ...

☆ Save ⌂ Cite Cited by 144 Related articles Web of Science: 22 ☰

19

19

Agenda

- BERT: Bidirectional Encoder Representations from Transformers
- GPT: Generative Pretrained Transformers
- Applications in Econ/Business Research

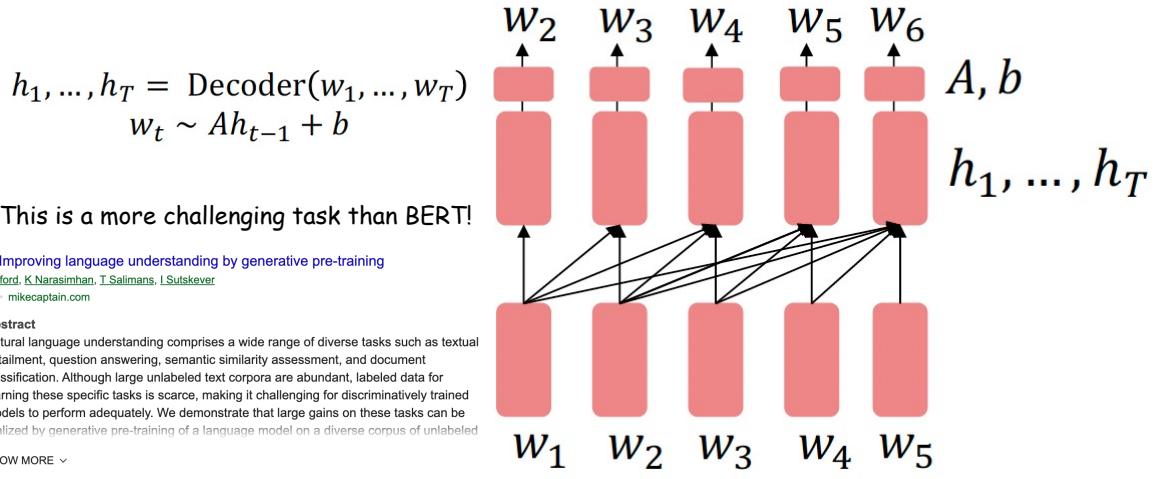
20

20

Pretraining Decoders

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>

- Key idea: Pretrain decoders as language models $\Pr(W_n | W_1, W_2, \dots, W_{n-1})$ via autoregression.



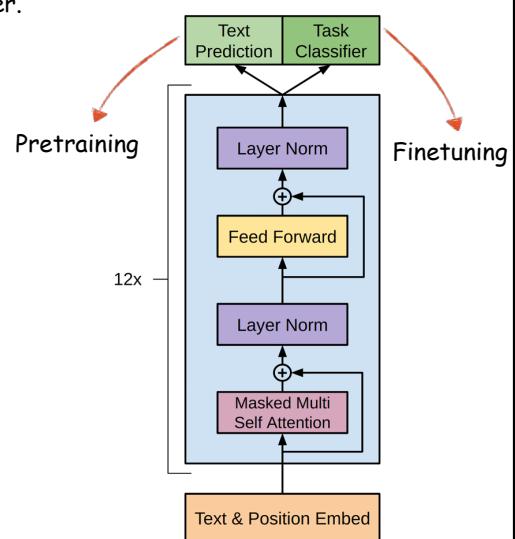
21

21

GPT-1

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>

- Architecture: Only masked self-attention, but deeper and larger.
- 12 layers of transformer decoders, 117M parameters.
- 768-dim hidden states, 3072-dim MLP hidden layers.
- Trained on BooksCorpus of over 7,000 unique books.



[PDF] Improving language understanding by generative pre-training

A Radford, K Narasimhan, T Salimans, I Sutskever

2018 - mikedcaption.com

Abstract

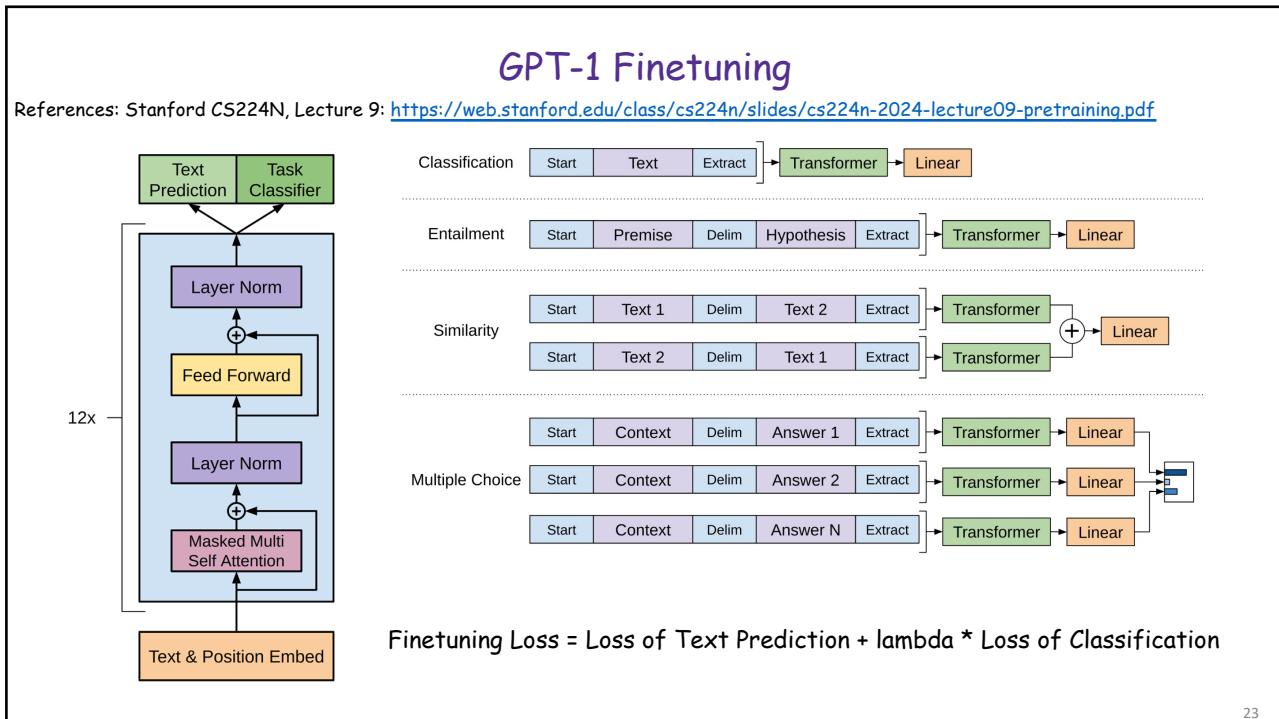
Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by generative pre-training of a language model on a diverse corpus of unlabeled

SHOW MORE ▾

☆ Save 59 Cite Cited by 8363 Related articles All 15 versions ☰

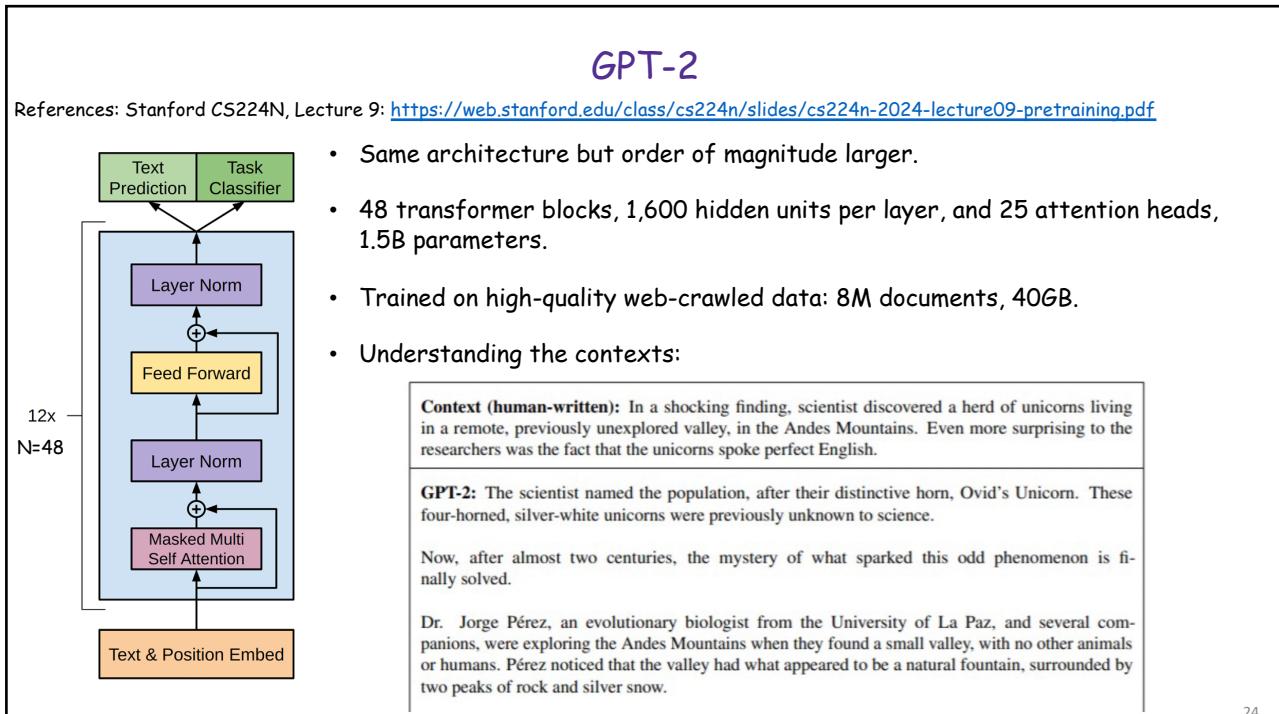
22

22



23

23

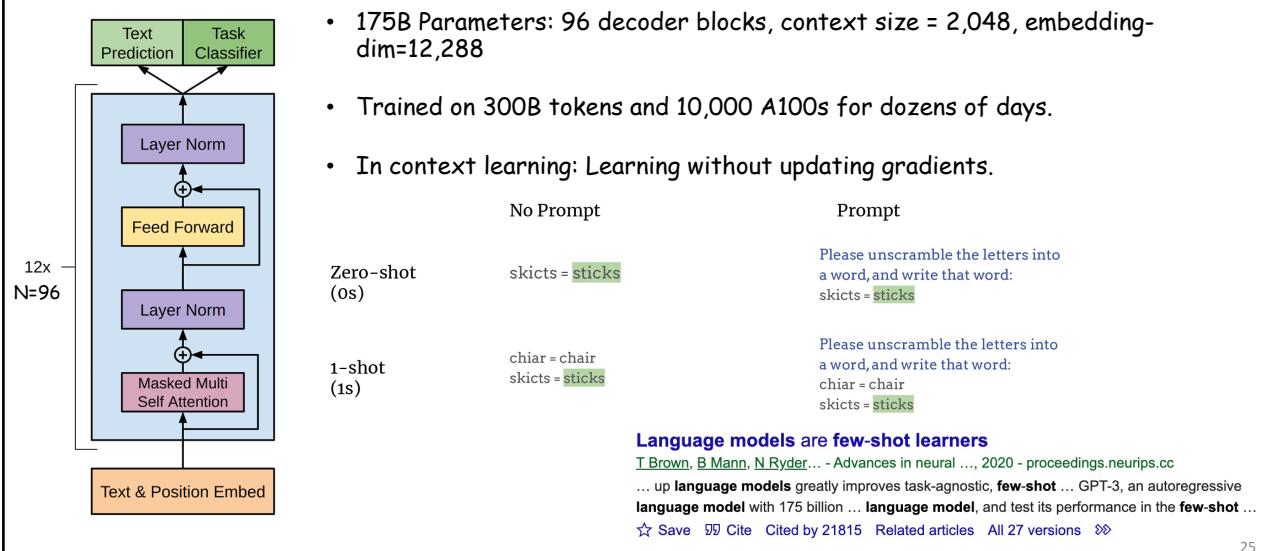


24

24

GPT-3

References: Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture09-pretraining.pdf>
 Princeton COS 597G, Lecture 4: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec04.pdf>



25

25

Agenda

- BERT: Bidirectional Encoder Representations from Transformers
- GPT: Generative Pretrained Transformers
- Applications in Econ/Business Research

26

26

Voice of Monetary Policy

The Voice of Monetary Policy[†]

By YURIY GORODNICHENKO, THO PHAM, AND OLEKSANDR TALAVERA^[2]

We develop a deep learning model to detect emotions embedded in press conferences after the Federal Open Market Committee meetings and examine the influence of the detected emotions on financial markets. We find that, after controlling for the Federal Reserve's actions and the sentiment in policy texts, a positive tone in the voices of Federal Reserve chairs leads to significant increases in share prices. Other financial variables also respond to vocal cues from the chairs. Hence, how policy messages are communicated can move the financial market. Our results provide implications for improving the effectiveness of central bank communications. (JEL D83, E31, E44, E52, E58, F31, G14)

How can a president not be an actor?

—Ronald Reagan (1980)

As Chairman, I hope to foster a public conversation about what the Fed is doing to support a strong and resilient economy. And one practical step in doing so is to have a press conference like this after every one of our scheduled FOMC meetings. ... [This] is only about improving communications.

—Jerome Powell (2018)^[3]

Monetary policy is 98 percent talk and 2 percent action, and communication is a big part.

—Ben Bernanke (2022)^[4]

- Use an MLP of 3 hidden layers to predict the voice tone of FOMC press conferences.

$$\text{VoiceTone} = \frac{\text{Positive answers} - \text{Negative answers}}{\text{Positive answers} + \text{Negative answers}},$$

- Use BERT to predict the sentiment of FOMC texts.

$$\text{TextSentiment} = \frac{\text{Dovish text} - \text{Hawkish text}}{\text{Dovish text} + \text{Hawkish text}},$$

- A positive tone of FR chairs leads to significant increases in share prices: **How to say is as important as what to say.**
- ?Seemed to suggest that using FinBERT saves the finetuning in sentiment analysis?

The voice of monetary policy

[Y.Gorodnichenko, T.Pham, O.Talavera](#) - American Economic Review, 2023 - aeaweb.org

... on recent advances in **voice** recognition technology and classify the **voice** tone of the Fed chairs into a spectrum of emotions. We, then, study how variations in **voice** tone (emotions) can ...

[☆ Save](#) [⤒ Cite](#) [Cited by 118](#) [Related articles](#) [All 30 versions](#) [Web of Science: 9](#) [⤓](#) [27](#)

27

Remote Work

Remote Work across Jobs, Companies, and Space

Stephen Hansen, Peter John Lambert, Nicholas Bloom,
Steven J. Davis, Raffaella Sadun & Bledi Taska

WORKING PAPER 31007 DOI 10.3386/w31007 ISSUE DATE March 2023

The pandemic catalyzed an enduring shift to remote work. To measure and characterize this shift, we examine more than 250 million job vacancy postings across five English-speaking countries. Our measurements rely on a state-of-the-art language-processing framework that we fit, test, and refine using 30,000 human classifications. We achieve 99% accuracy in flagging job postings that advertise hybrid or fully remote work, greatly outperforming dictionary methods and also outperforming other machine-learning methods. From 2019 to early 2023, the share of postings that say neither workers can work remotely nor require employees to come in three-fold in the U.S. and by a factor of five or more in Australia, Canada, New Zealand and the U.K. These developments are highly non-uniform across and within cities, industries, occupations, and companies. Even when zooming in on employers in the same industry competing for talent in the same occupations, we find large differences in the share of job postings that explicitly offer remote work.

- Pre-trained transformers are used for some **downstream tasks** (similarity measurement, concept detection, conception relationship characterization, text-metadata association, etc.).
- Use **DistilBERT** pre-trained on 1M text chunks of job vacancy postings to measure the **Work-from-homeness** of the 250 M jobs (**Work from Home Algorithmic Measure**), achieving 99% accuracy that outperforms dictionary-based methods.
- The number of WFM jobs has risen significantly since 2019 and it differs w.r.t. different industries.

Remote work across jobs, companies, and space

[S.Hansen, PJ.Lambert, N.Bloom, SJ.Davis, R.Sadun...](#) - 2023 - nber.org

The pandemic catalyzed an enduring shift to remote work. To measure and characterize this shift, we examine more than 250 million job vacancy postings across five English-speaking countries. Our measurements rely on a state-of-the-art language-processing framework that we fit, test, and refine using 30,000 human classifications. We achieve 99% accuracy in flagging job postings that advertise hybrid or fully remote work, greatly outperforming dictionary methods and also outperforming other machine-learning methods. From 2019 to ...

[☆ Save](#) [⤒ Cite](#) [Cited by 36](#) [Related articles](#) [All 20 versions](#) [⤓](#)

28

28