

DOTE 6635: Artificial Intelligence for Business Research

# Pretraining

Renyu (Philip) Zhang

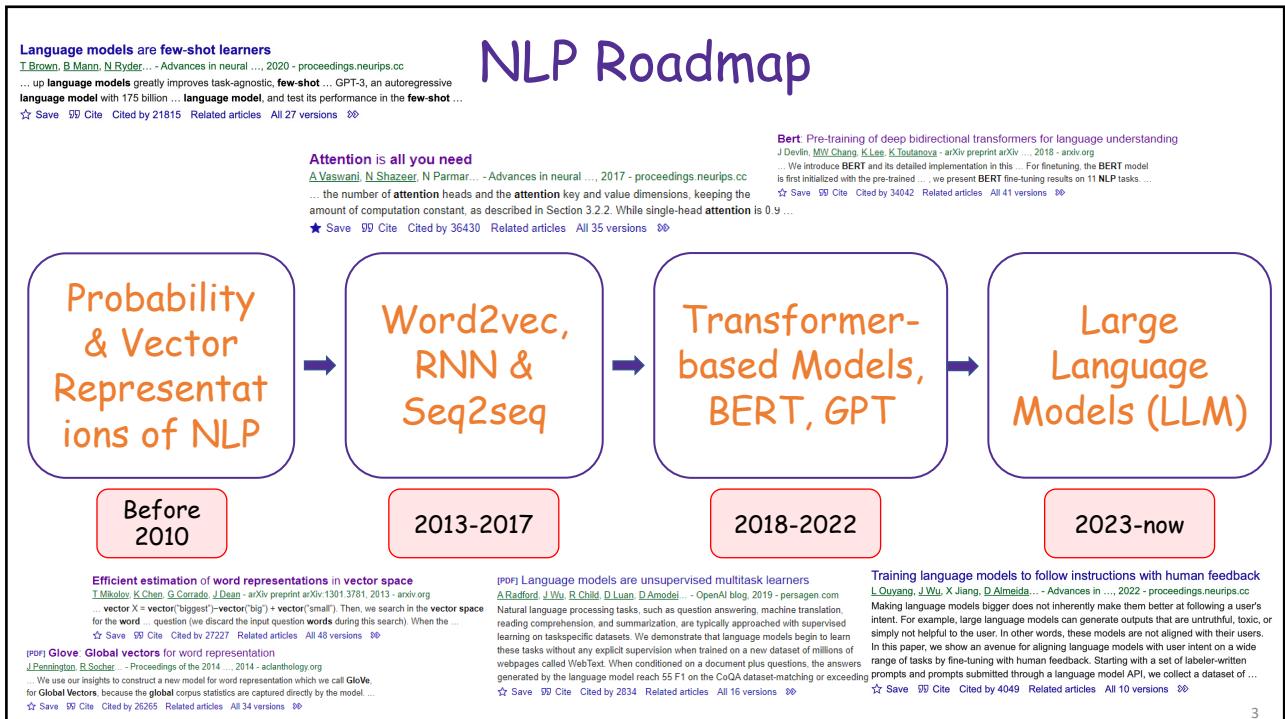
1

## Agenda

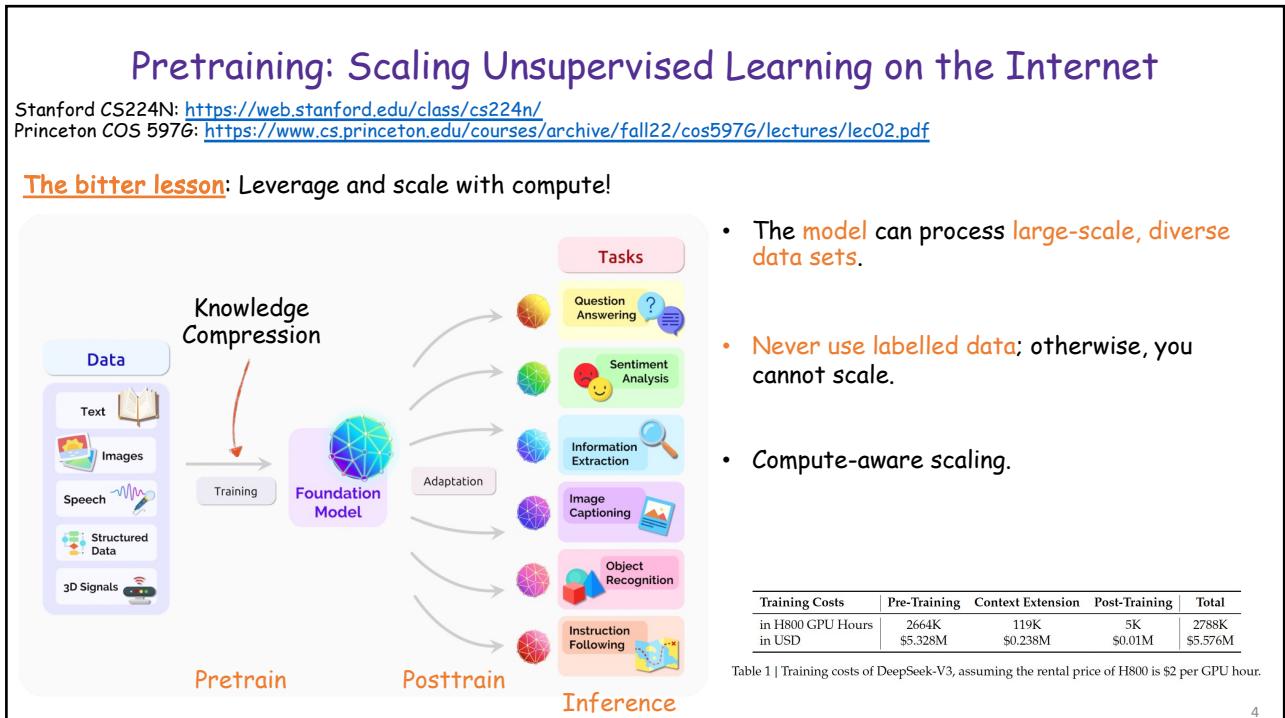
- BERT: Bidirectional Encoder Representations from Transformers
- GPT: Generative Pretrained Transformers
- Applications in Econ/Business Research

2

2

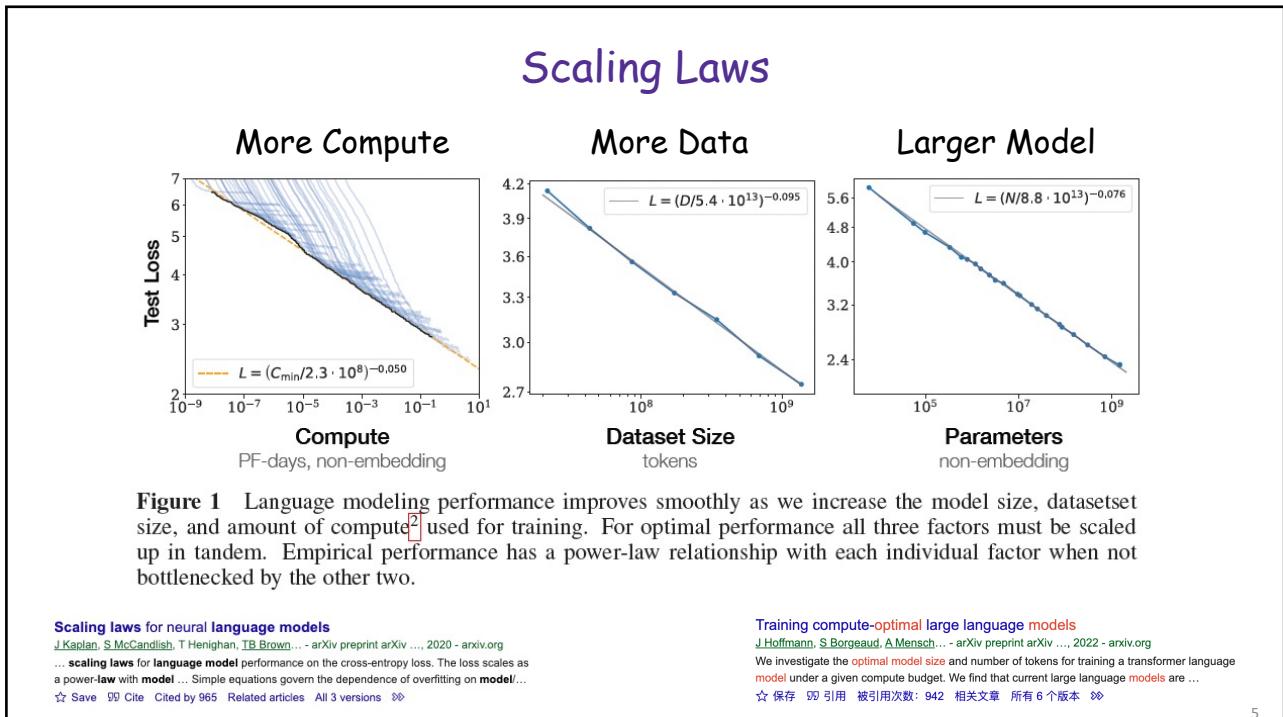


3



4

4



5

## Jason Wei's Typical Day

 **Jason Wei**   
@\_jasonwei

My typical day as a Member of Technical Staff at OpenAI:

[9:00am] Wake up

[9:30am] Commute to Mission SF via Waymo. Grab avocado toast from Tartine

[9:45 am] Recite OpenAI charter. Pray to optimization Gods. Learn the Bitter Lesson

[10:00am] Meetings (Google Meet). Discuss how to train larger models on more data

[11:00am] Write code to train larger models on more data. pair= @hwchung27

6

6

## The Bitter Lesson

- References: <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>  
<https://www.youtube.com/watch?v=vbVfAqPI8ng>
- The biggest lesson that can be read from 70 years of AI research is that **general methods that leverage computation** are ultimately the most effective, and by a large margin.
- Leveraging domain knowledge (short-term & specific) vs. Leveraging computation (long-term & general).
- Bitter lesson: Leveraging domain knowledge is **self-satisfying** and **intellectually inspiring**, but plateaus in the long-run or even inhibits further progress.



Prof. Richard Sutton

7

7

## Jason Wei's Typical Day (Cont'd)

[12:00pm] Lunch at the canteen (vegan, gluten-free)
[1:00pm] Actually train large models models on more data
[2:00pm] Debug infra issues (why the fck did I pull from master?)
[3:00pm] Babysit model training. Play with Sora
[4:00pm] Prompt engineer aforementioned large models trained on more data
[4:30pm] Short break, sit on avocado chair. Wonder how good Gemini Ultra actually is
[5:00pm] Brainstorm potential algorithmic improvements for models
[5:05pm] Conclude that algorithmic changes are too risky. Safer to just scale compute and data
[6:00pm] Dinner. Clam chowder with Roon
[7:00pm] Commute back home
[8:00pm] Have a wine and get back to coding. Ballmer's peak is coming
[9:00pm] Analyze experimental runs. I have a love/hate relationship with wandb
[10:00pm] Launch experiments to run overnight and get results by tomorrow morning
[1:00am] Experiments actually get launched
[1:15am] Bedtime. Satya and Jensen watch from above. Compression is all you need. Good night

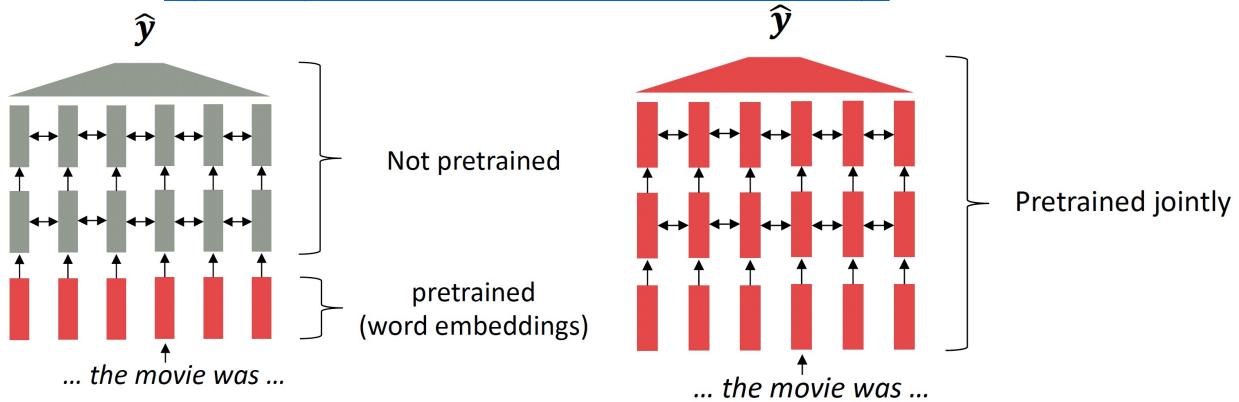
8

8

## From Pretrained Word Embeddings to Pretrained Models

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>

Princeton COS 597G: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



[Recall, *movie* gets the same word embedding,  
no matter what sentence it shows up in]

[This model has learned how to represent  
entire sentences through pretraining]

9

9

## From Pretrained Word Embeddings to Pretrained Models

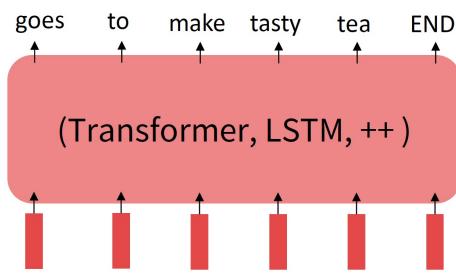
Stanford CS224N: <https://web.stanford.edu/class/cs224n/>

Princeton COS 597G: <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)

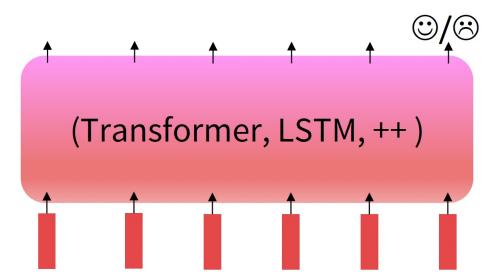
Lots of text; learn general things!



$\hat{\theta}$  by approximating  $\min_{\theta} \mathcal{L}_{\text{pretrain}}(\theta)$

### Step 2: Finetune (on your task)

Not many labels; adapt to the task!



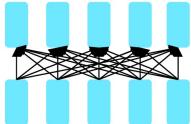
approximates  $\min_{\theta} \mathcal{L}_{\text{finetune}}(\theta)$ , starting at  $\hat{\theta}$

10

10

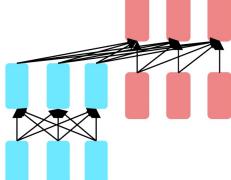
## Three Pretraining Architectures

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Princeton COS 597G: <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**

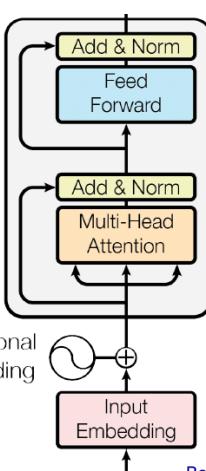
- Cannot condition on future.
- Example: GPT
- All (very) large language models are decoders.

11

11

## BERT: Bidirectional Encoder Representations from Transformers

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Princeton COS 597G: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



Nx  
N=12 or 24

- 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.



[Predict these!]

went to store

Transformer Encoder

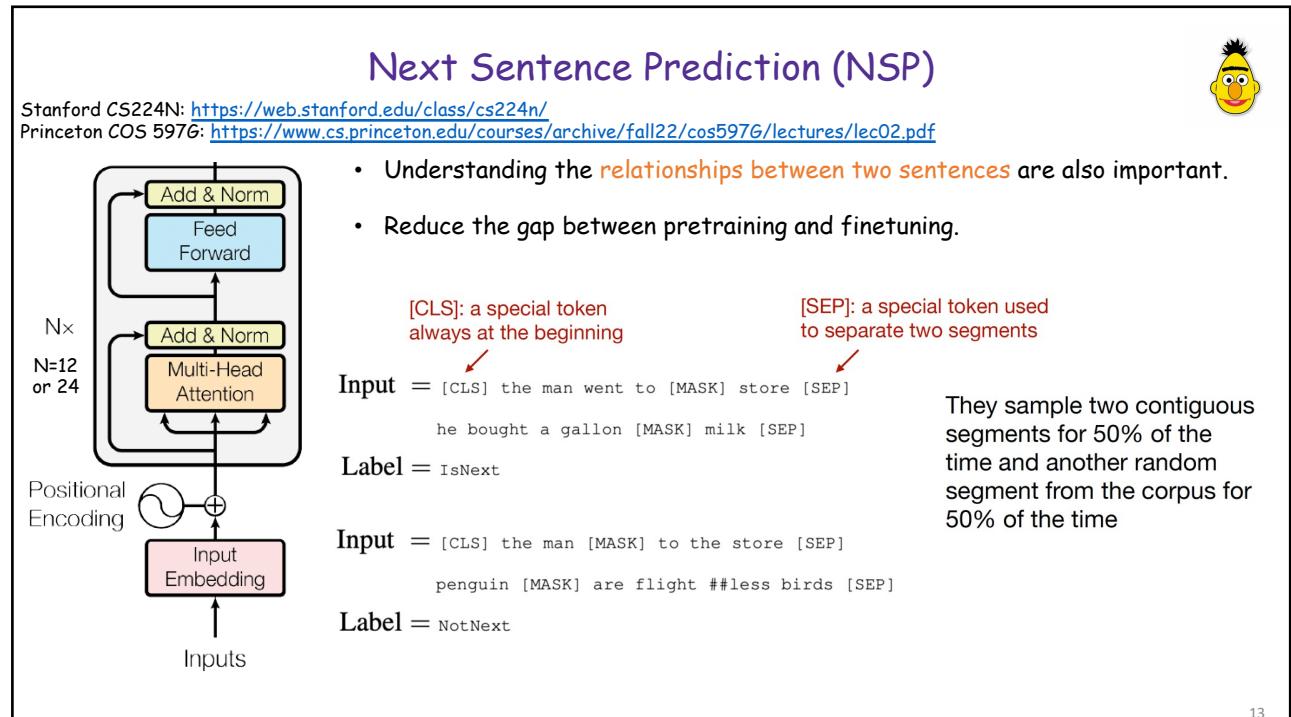
I pizza to the [M]

[Replaced] [Not replaced] [Masked]

Bert: Pre-training of deep **bidirectional** transformers for language understanding  
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 ☰

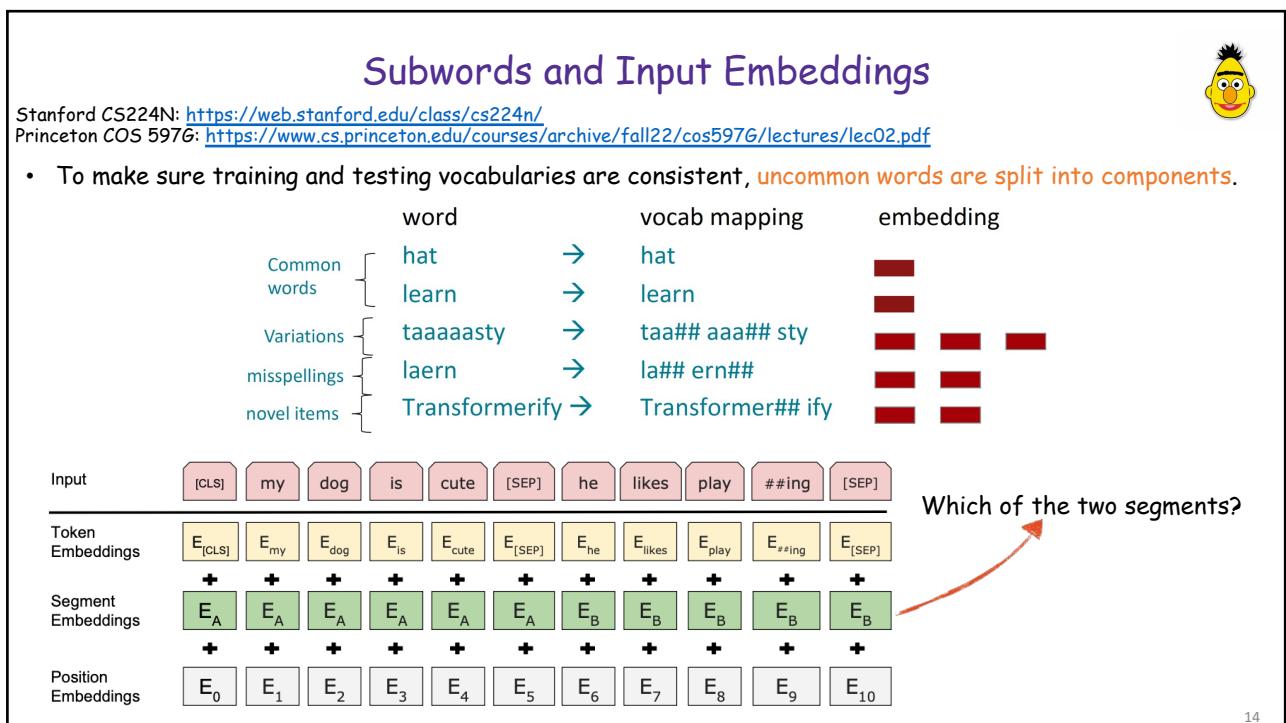
12

12



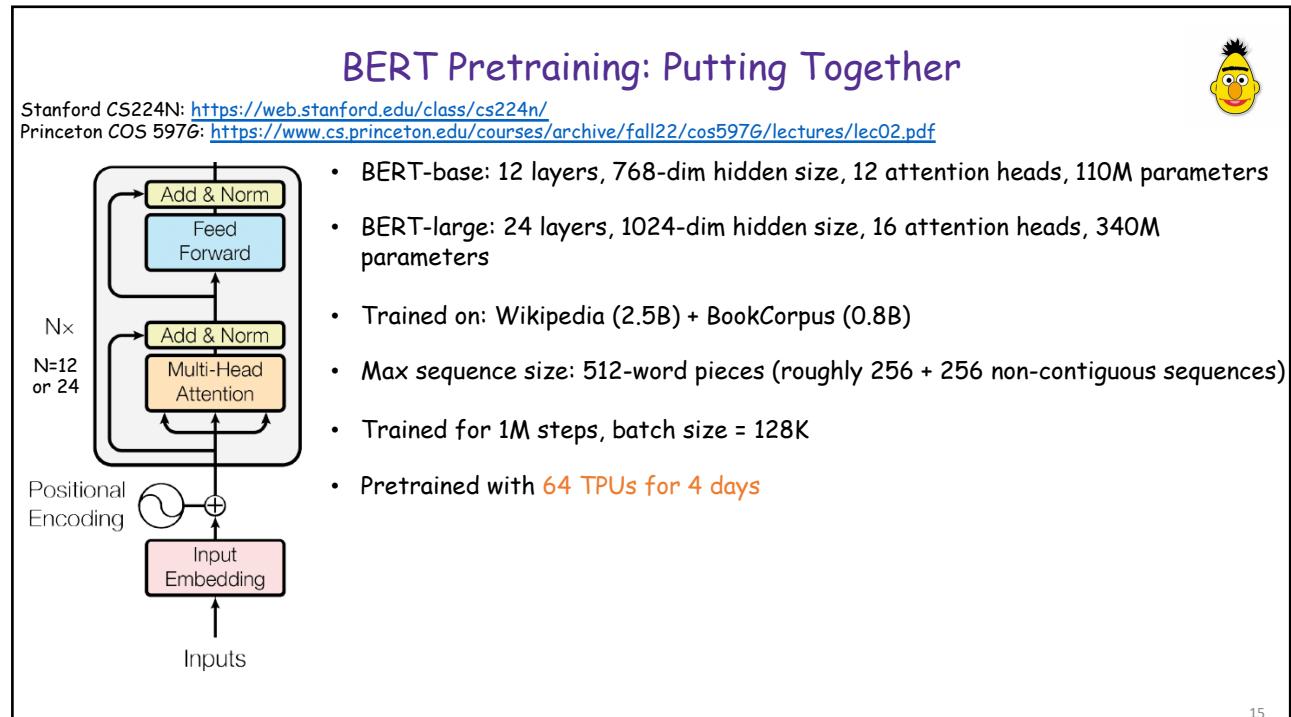
13

13



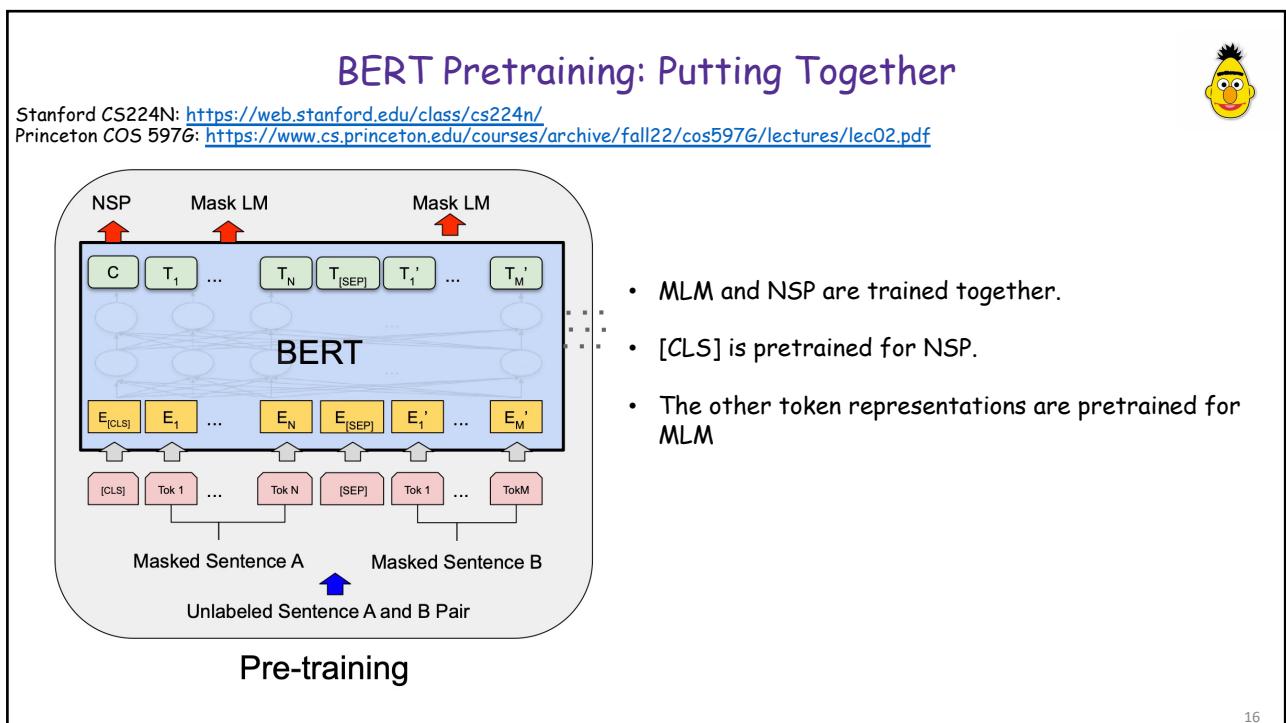
14

14



15

15



16

16

## Pretrain Once, Finetune Many Times



Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Princeton COS 597G: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>

**Sentence-Level Task**

- Sentence pair classification tasks:

**MNLI**      Premise: A soccer game with multiple males playing.  
Hypothesis: Some men are playing a sport.      {entailment, contradiction, neutral}

**Multi-genre Natural Language Inference**: Predict the relationship between two sentences.

**QQP**      Q1: Where can I learn to invest in stocks?  
Q2: How can I learn more about stocks?      {duplicate, not duplicate}

**Quora Question Pairs**: Detect paraphrase questions.

- Single sentence classification tasks:

**SST2**      rich veins of funny stuff in this movie      {positive, negative}  
**Sentiment Analysis**

17

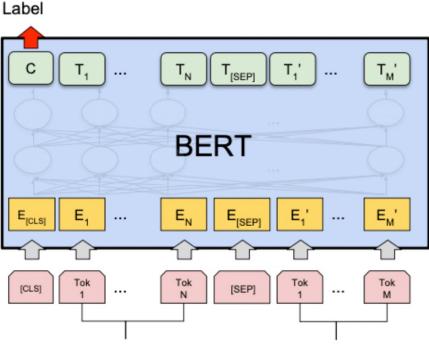
17

## Pretrain Once, Finetune Many Times

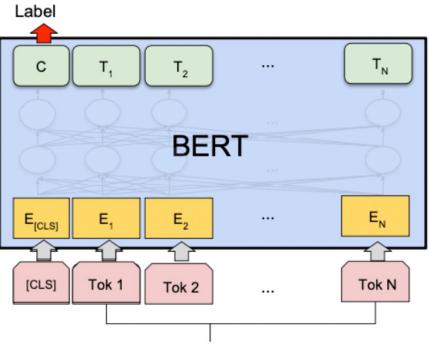


Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Princeton COS 597G: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>

**sentence-level tasks**



**(a) Sentence Pair Classification Tasks:**  
MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



**(b) Single Sentence Classification Tasks:**  
SST-2, CoLA

18

18

## Pretrain Once, Finetune Many Times

Bert

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Princeton COS 597G: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>

**Token-Level Task**

- Extractive question answering e.g., SQuAD (Rajpurkar et al., 2016)  
**Standard Question Answer Dataset:** Predict the answer to the question.

SQuAD

**Question:** The New York Giants and the New York Jets play at which stadium in NYC ?

**Context:** The city is represented in the National Football League by the New York Giants and the New York Jets , although both teams play their home games at **MetLife Stadium** in nearby East Rutherford , New Jersey , which hosted Super Bowl XLVIII in 2014 .

(Training example 29,883)

MetLife Stadium

- Named entity recognition (Tjong Kim Sang and De Meulder, 2003)  
**Named Entity Recognition:** Recognize the entity of each word.

CoNLL 2003 NER

John Smith lives in New York

B-PER I-PER O O B-LOC I-LOC

19

19

## Pretrain Once, Finetune Many Times

Bert

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Princeton COS 597G: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>

**token-level tasks**

**Start/End Span**

BERT

Question      Paragraph

**Single Sentence**

BERT

Single Sentence

(c) Question Answering Tasks:  
SQuAD v1.1

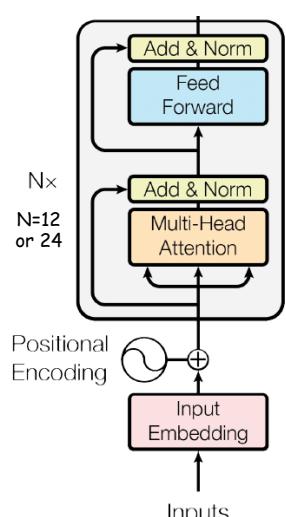
(d) Single Sentence Tagging Tasks:  
CoNLL-2003 NER

20

20

## BERT was the State-of-The-Art

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Princeton COS 597G: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec02.pdf>



- BERT-base: 12 layers, 768-dim hidden size, 12 attention heads, 110M parameters
- BERT-large: 24 layers, 1024-dim hidden size, 16 attention heads, 340M parameters

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT <sub>BASE</sub>	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT <sub>LARGE</sub>	<b>86.7/85.9</b>	<b>72.1</b>	<b>92.7</b>	<b>94.9</b>	<b>60.5</b>	<b>86.5</b>	<b>89.3</b>	<b>70.1</b>	<b>82.1</b>

- **Key issue with encoders:** Not a language model, i.e., does not naturally lead to autoregressive generation methods.

21

21

## FinBERT

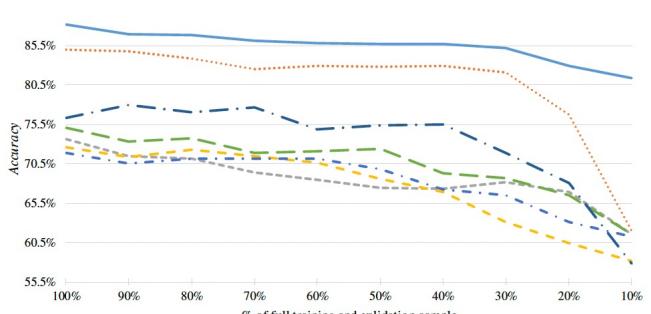


- Pretrain BERT-base using financial datasets (4.9B tokens in total) with 4 P100 GPUs (100G memory):
  - Corporate annual and quarterly filings from SEC's EDGAR website (1994-2019).
  - Financial analyst reports from Thomson Intestext database (2003-2012).
  - Earnings conference call transcripts from the SeekingAlpha website (2004-2019).
- Finetuning and evaluation:
  - Sentiment analysis 10,000 sentences
  - 36% positive
  - 46% neutral
  - 18% negative
- Can FinBERT beat GPT-4, Claude-3.5 or DeepSeek-V3 in tasks related to financial texts?
  - How can we make **fair comparisons?**

**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 ☰

**Figure 1** Sentiment classification accuracy across sample sizes



% of full training and validation sample	FinBERT	BERT	NB	SVM	RF	CNN	LSTM
100%	~85.5%	~85.5%	~75.5%	~75.5%	~75.5%	~75.5%	~75.5%
90%	~85.0%	~85.0%	~75.0%	~75.0%	~75.0%	~75.0%	~75.0%
80%	~84.5%	~84.5%	~74.5%	~74.5%	~74.5%	~74.5%	~74.5%
70%	~84.0%	~84.0%	~74.0%	~74.0%	~74.0%	~74.0%	~74.0%
60%	~83.5%	~83.5%	~73.5%	~73.5%	~73.5%	~73.5%	~73.5%
50%	~83.0%	~83.0%	~73.0%	~73.0%	~73.0%	~73.0%	~73.0%
40%	~82.5%	~82.5%	~72.5%	~72.5%	~72.5%	~72.5%	~72.5%
30%	~82.0%	~82.0%	~72.0%	~72.0%	~72.0%	~72.0%	~72.0%
20%	~81.5%	~81.5%	~71.5%	~71.5%	~71.5%	~71.5%	~71.5%
10%	~81.0%	~81.0%	~71.0%	~71.0%	~71.0%	~71.0%	~71.0%

22

22

## Agenda

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

23

23

## Pretraining Decoders

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>

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

$$h_1, \dots, h_T = \text{Decoder}(w_1, \dots, w_T)$$

$$w_t \sim A h_{t-1} + b$$

This is a more challenging task than BERT!

[PPL: Improving language understanding by generative pre-training](#)

A Radford, K Narasimhan, T Salimans, I Sutskever

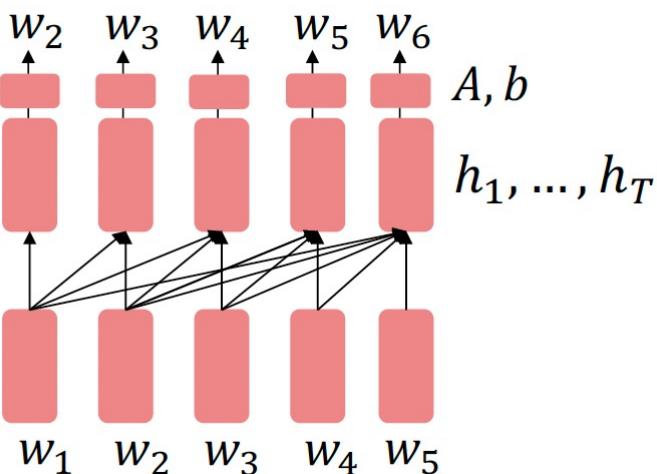
2018 - [mikecaptain.com](#)

### 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](#) 99 [Cite](#) Cited by 8363 [Related articles](#) All 15 versions [⊗⊗](#)



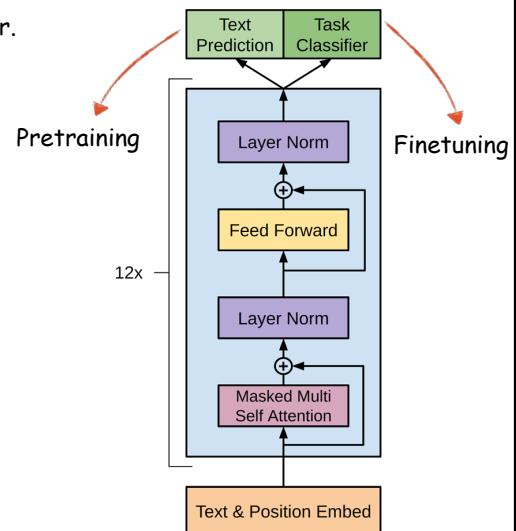
24

24

# GPT-1

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>

- 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 • mikecaptain.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 ▾](#)

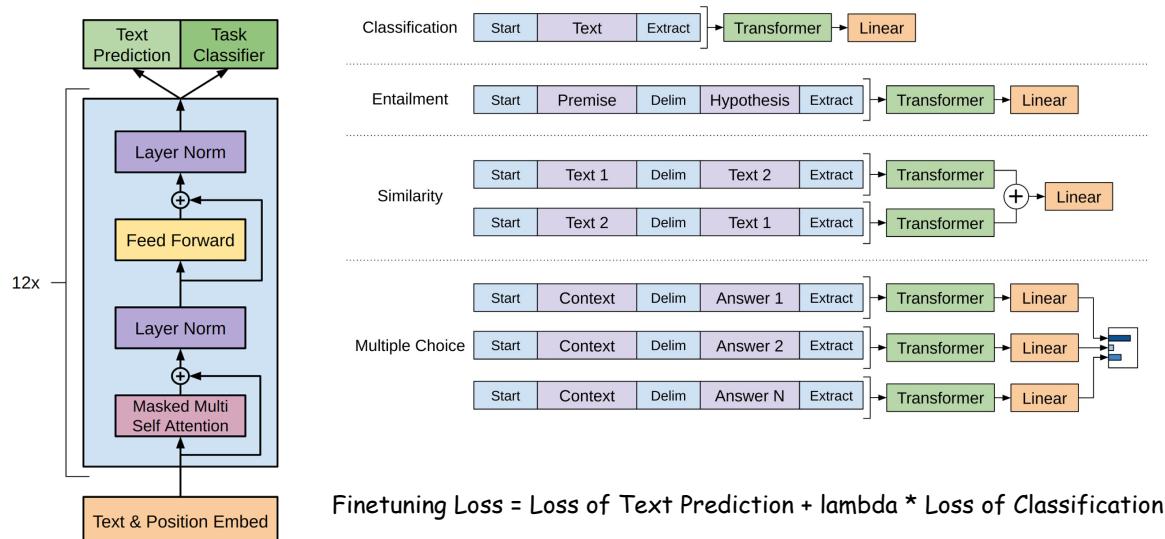
Save Cite Cited by 8363 Related articles All 15 versions

25

25

## GPT-1 Finetuning

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>

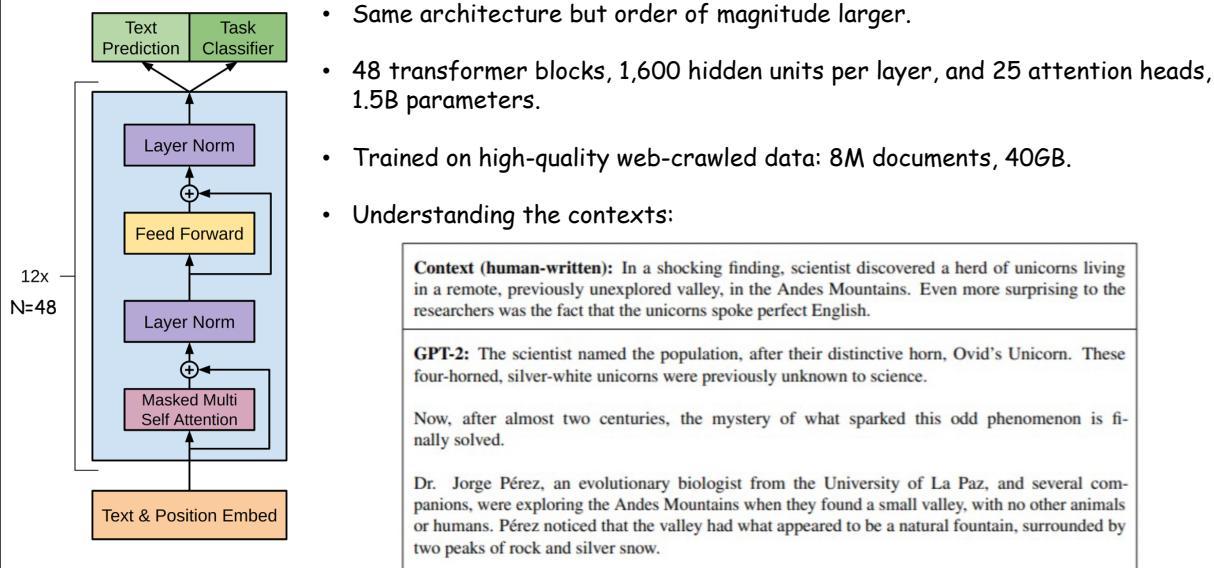


26

26

## GPT-2

Stanford CS224N, Lecture 9: <https://web.stanford.edu/class/cs224n/>

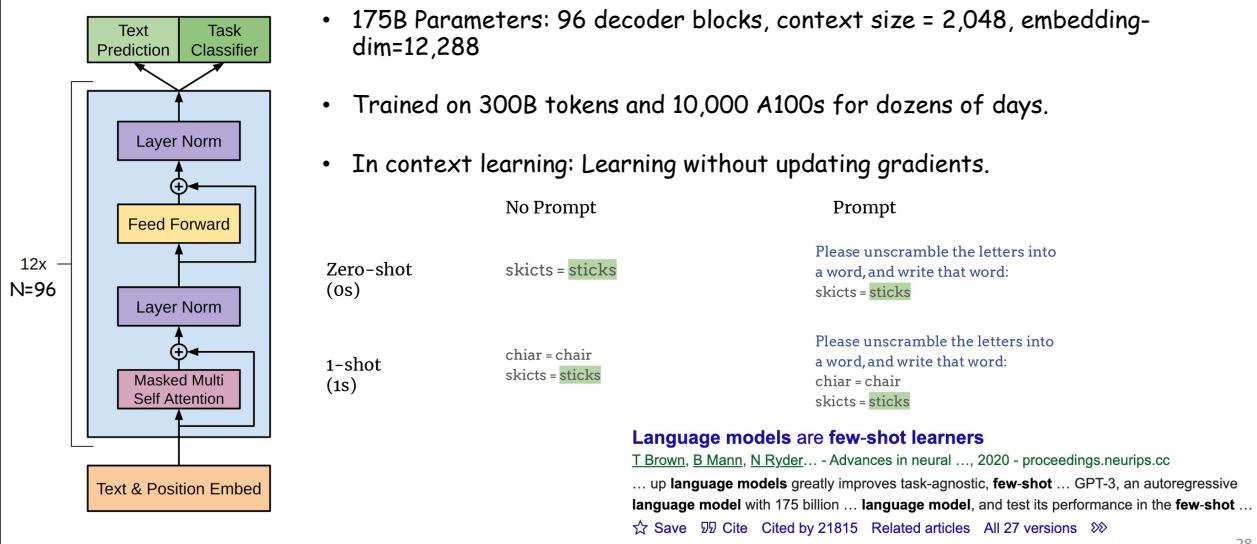


27

27

## GPT-3

Stanford CS224N: <https://web.stanford.edu/class/cs224n/>  
Princeton COS 597G: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec04.pdf>



28

28

# Agenda

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

29

29

# Voice of Monetary Policy

## The Voice of Monetary Policy<sup>†</sup>

By YURIY GORODNICHENKO, THO PHAM, AND OLEKSANDR TALAVERA<sup>‡</sup>

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)<sup>§</sup>

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

—Ben Bernanke (2022)<sup>¶</sup>

- 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 30

30

15

# 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 new employees can work remotely one or more days per week rose more than three-fold in the U.S and by a factor of five or more in Australia, Canada, New Zealand and the U.K. These movements 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](#) [PDF Cite](#) [Cited by 36](#) [Related articles](#) [All 20 versions](#) [»»](#)

31