DOTE 6635: Artificial Intelligence for Business Research

Pretraining

Renyu (Philip) Zhang

1

Agenda

- BERT: Bidirectional Encoder Representations from Transformers
- GPT: Generative Pretrained Transformers

2

2

NLP Roadmap

Language models are few-shot learners
T Brown, B Mann, N Ryder… - Advances in neural …, 2020 - proceedings.neurips.cc
… up **language models** greatly improves task-agnostic, **few-shot** … GPT-3, an autoregressive **language model** with 175 billion … **language model**, and test its performance in the **few-shot** …
☆ Save 99 Cite Cited by 21815 Related articles All 27 versions »

Attention is all you need
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… the number of **attention** heads and the **attention** key and value dimensions, keeping the amount of computation constant, as described in Section 3.2.2. While single-head **attention** is 0.9 …
★ Save 99 Cite Cited by 36430 Related articles All 35 versions »

Bert: Pre-training of deep bidirectional transformers for language understanding
J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv …, 2018 - arxiv.org
… We introduce **BERT** and its detailed implementation in this … For finetuning, the **BERT** model is first initialized with the pre-trained … , we present **BERT** fine-tuning results on 11 **NLP** tasks. …
☆ Save 99 Cite Cited by 34042 Related articles All 41 versions »

Probability & Vector Representations of NLP → Word2vec, RNN & Seq2seq → Transformer-based Models, BERT, GPT → Large Language Models (LLM)

Before 2010 | 2013-2017 | 2018-2022 | 2023-now

Efficient estimation of word representations in vector space
T Mikolov, K Chen, G Corrado, J Dean - arXiv preprint arXiv:1301.3781, 2013 - arxiv.org
… **vector** X = **vector**("biggest")−**vector**("big") + **vector**("small"). Then, we search in the **vector space** for the **word** … question (we discard the input question **words** during this search). When the …
☆ Save 99 Cite Cited by 27227 Related articles All 48 versions »

[PDF] **Glove: Global vectors** for word representation
J Pennington, R Socher … - Proceedings of the 2014 …, 2014 - aclanthology.org
… We use our insights to construct a new model for word representation which we call **GloVe**, for **Global Vectors**, because the **global** corpus statistics are captured directly by the model. …
☆ Save 99 Cite Cited by 26265 Related articles All 34 versions »

[PDF] Language models are unsupervised multitask learners
A Radford, J Wu, R Child, D Luan, D Amodei … - OpenAI blog, 2019 - persagen.com
Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on taskspecific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset-matching or exceeding
☆ Save 99 Cite Cited by 2834 Related articles All 16 versions »

Training language models to follow instructions with human feedback
L Ouyang, J Wu, X Jiang, D Almeida… - Advances in …, 2022 - proceedings.neurips.cc
Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through a language model API, we collect a dataset of …
☆ Save 99 Cite Cited by 4049 Related articles All 10 versions »

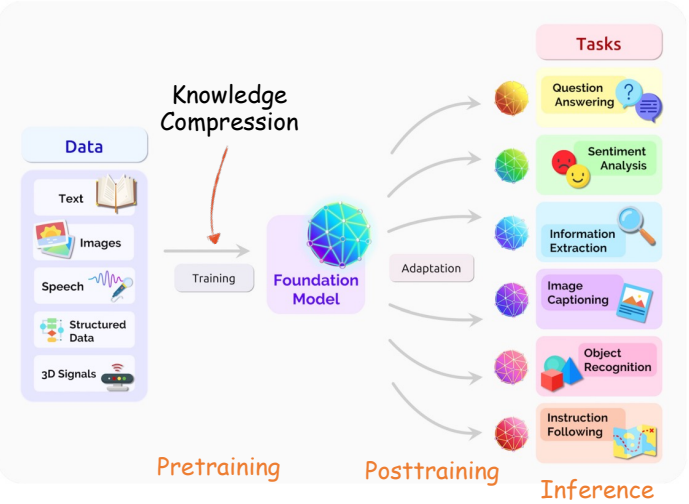
3

3

Pretraining: Scaling Unsupervised Learning on the Internet

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

The bitter lesson: Leverage and scale with compute!



- The model can process large-scale, diverse data sets.
- Never use labelled data; otherwise, you cannot scale.
- Compute-aware scaling.

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is $2 per GPU hour.

4

4

Scaling Laws

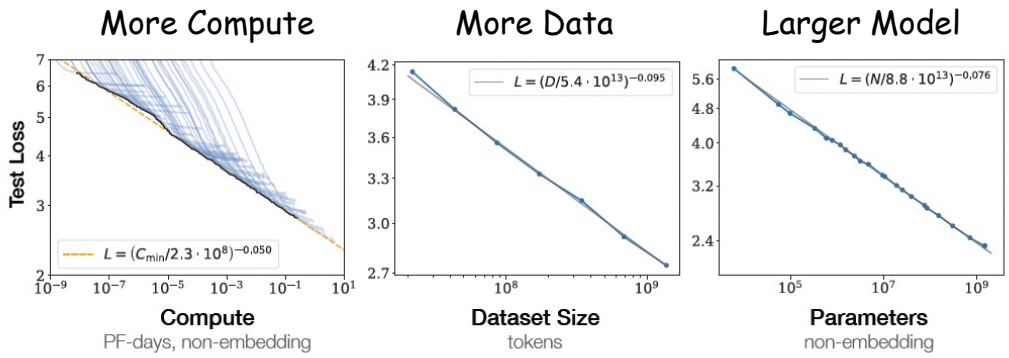


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

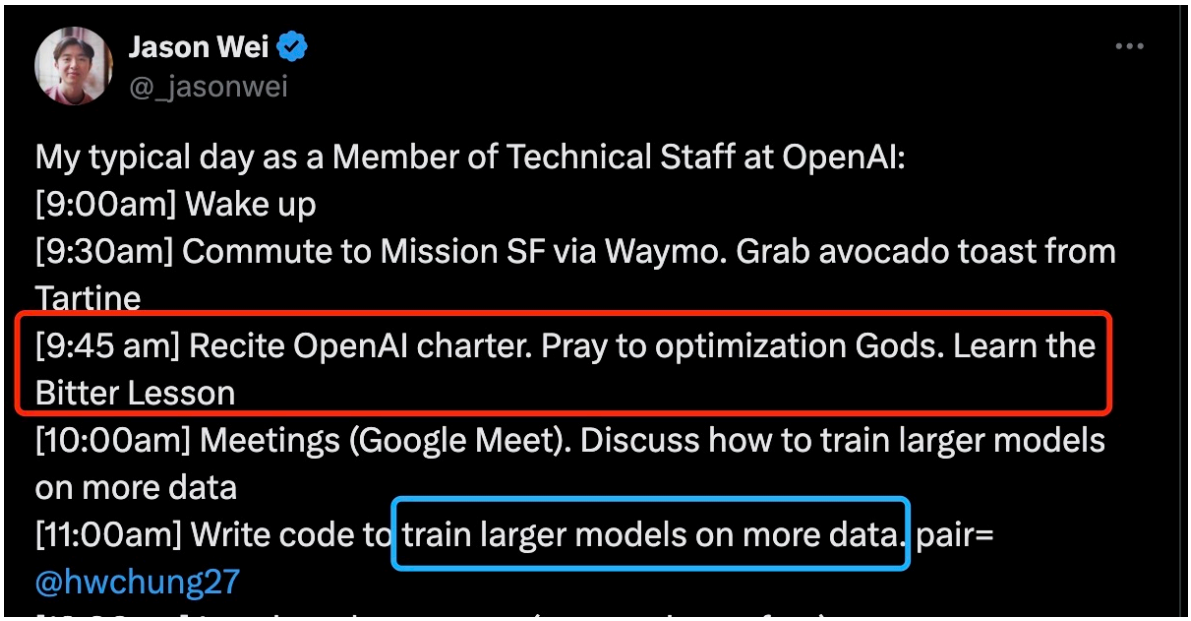
Scaling laws for neural **language models**
J Kaplan, S McCandlish, T Henighan, TB Brown… - arXiv preprint arXiv …, 2020 - arxiv.org
… **scaling laws** for **language model** performance on the cross-entropy loss. The loss scales as a power-**law** with **model** … Simple equations govern the dependence of overfitting on **model**/…
☆ Save 99 Cite Cited by 965 Related articles All 3 versions »

Training compute-optimal large language models
J Hoffmann, S Borgeaud, A Mensch… - arXiv preprint arXiv …, 2022 - arxiv.org
We investigate the optimal model size and number of tokens for training a transformer language model under a given compute budget. We find that current large language models are …
☆ 保存 99 引用 被引用次数：942 相关文章 所有 6 个版本 »

5

5

Jason Wei's Typical Day



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)

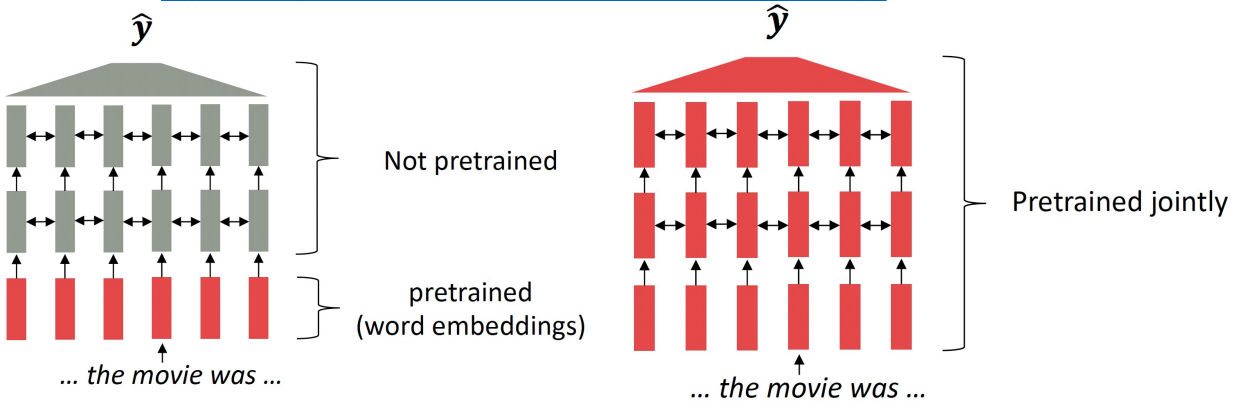
[12:00pm] Lunch at the canteen (vegan, gluten-free)
[1:00pm] Actually train large models models on more data
[2:00pm] Debug infra issues (why the fck did I pull from master?)
[3:00pm] Babysit model training. Play with Sora
[4:00pm] Prompt engineer aforementioned large models trained on more data
[4:30pm] 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

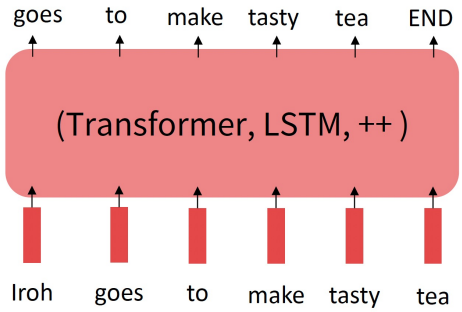
From Pretrained Word Embeddings to Pretrained Models

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)

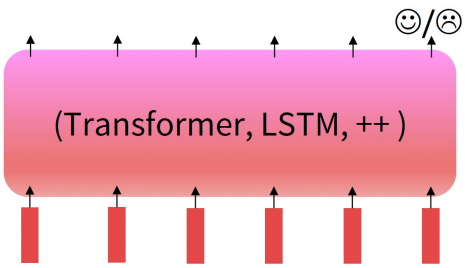
Lots of text; learn general things!



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

Step 2: Finetune (on your task)

Not many labels; adapt to the task!



… the movie was …

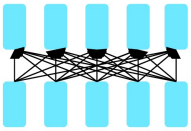
approximates $\min_{\theta} \mathcal{L}_{\text{finetune}}(\theta)$, starting at $\hat{\theta}$

10

10

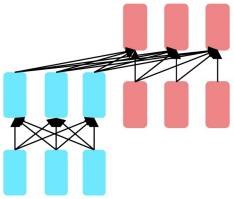
Three Pretraining Architectures

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



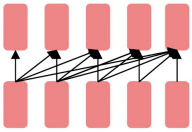
Encoders

- Can condition on future.
- Example: BERT.



Encoder-Decoders

- Combining encoder and decoder.
- Example: T5



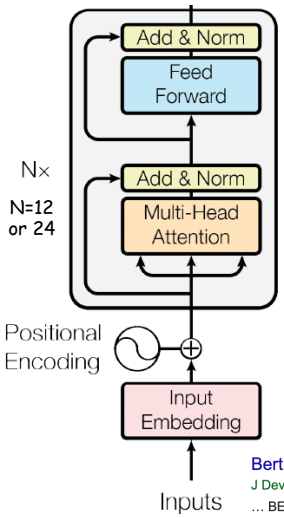
Decoders

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

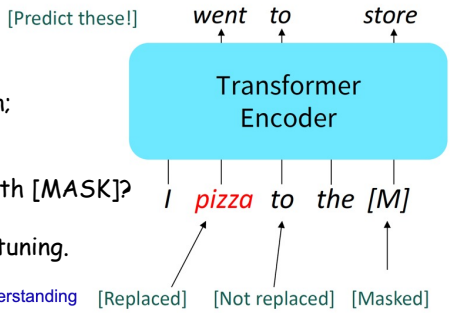
11

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



- 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.



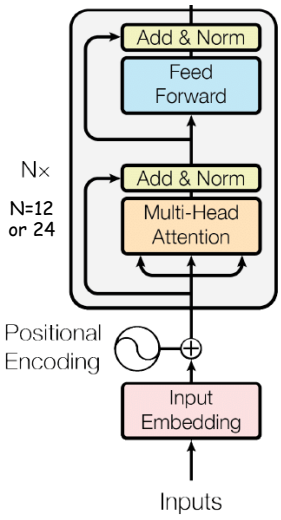
Bert: Pre-training of deep **bidirectional** transformers for language understanding
J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv …, 2018 - arxiv.org
… BERT, which stands for **Bidirectional Encoder Representations** from Transformers. Unlike … 2018), BERT is designed to pretrain deep **bidirectional representations** from unlabeled text by …
☆ Save 99 Cite Cited by 93230 Related articles All 46 versions »

RoBERTa: Dynamic Masking

12

Next Sentence Prediction (NSP)

Stanford CS224N: https://web.stanford.edu/class/cs224n/
Princeton COS 597G: 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.
- RoBERTa: Training without NSP.

[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`

Input = `[CLS] the man [MASK] to the store [SEP]`
`penguin [MASK] are flight ##less birds [SEP]`
Label = `NotNext`

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

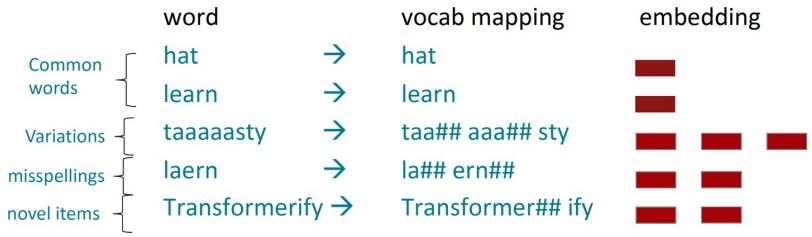
13

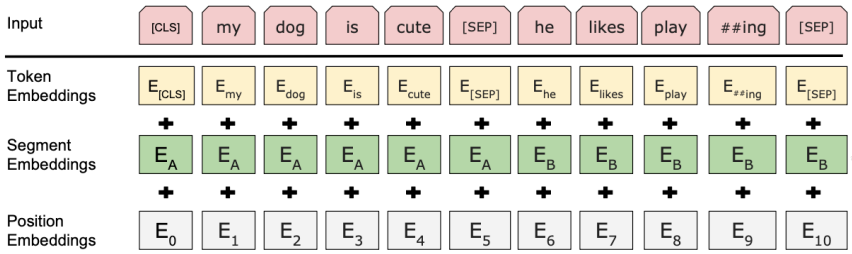
13

Subwords and Input Embeddings

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



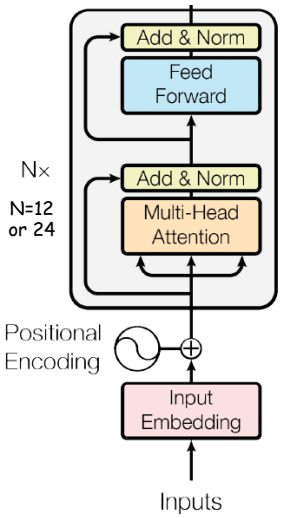


14

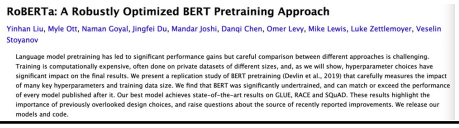
14

BERT Pretraining: Putting Together

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
- Trained on: Wikipedia (2.5B Tokens) + BookCorpus (0.8B Tokens)
  - RoBERTa: BookCorpus (16GB), CC-News (76GB), OpenWebText (38GB), Stories (31GB)
- Max sequence size: 512-word pieces (roughly 256 + 256 non-contiguous sequences)
- Trained for 1M steps, batch size = 256
  - Batch size increased to 8K for RoBERTa
- Pretrained with 64 TPUs for 4 days


RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov

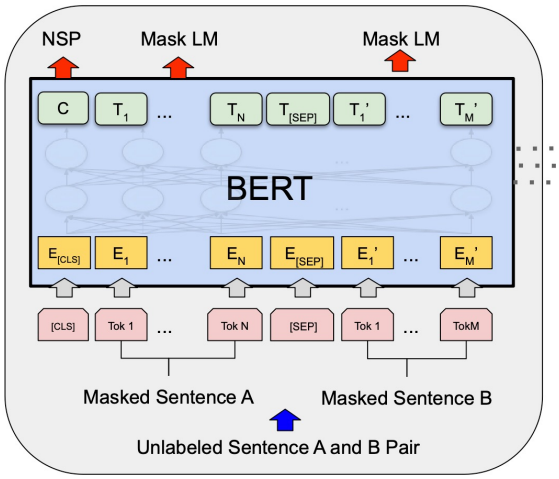
Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining (Devlin et al., 2019) that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD. These results highlight the importance of previously overlooked design choices, and raise questions about the source of recently reported improvements. We release our models and code.

15

15

BERT Pretraining: Putting Together

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



Pre-training

- MLM and NSP are trained together.
- [CLS] is pretrained for NSP.
- The other token representations are pretrained for MLM

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

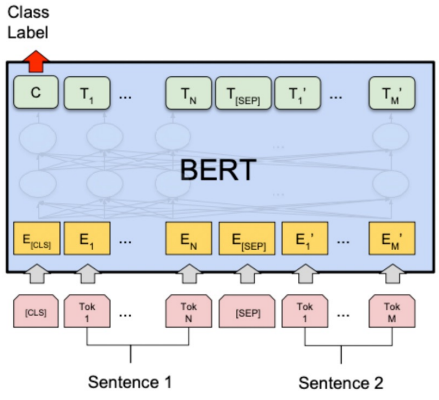
17

17

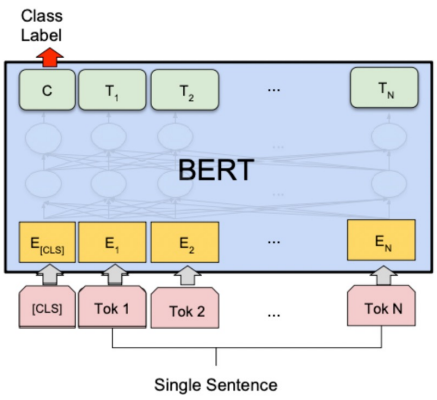
Pretrain Once, Finetune Many Times

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

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

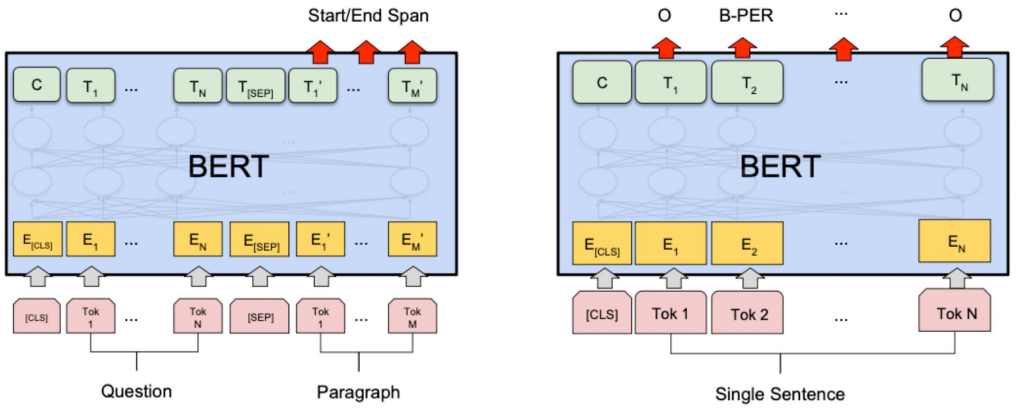
19

19

Pretrain Once, Finetune Many Times

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



(c) Question Answering Tasks: SQuAD v1.1

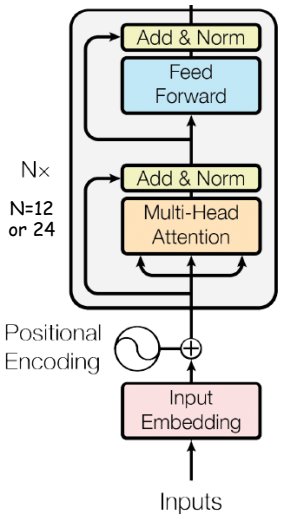
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

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

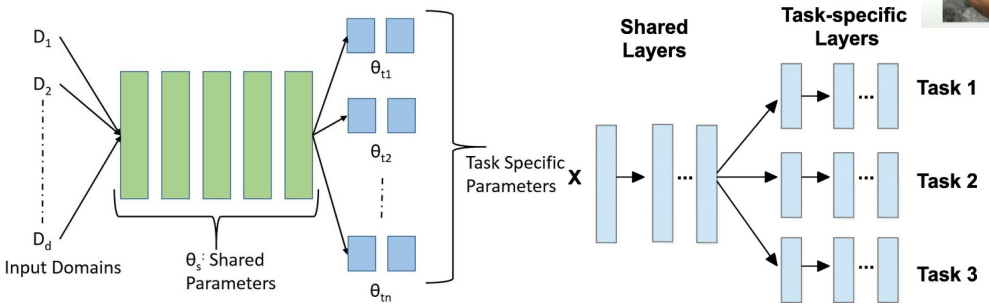
System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$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

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

21

21

Multi-Task Learning and Transfer Learning



- Transfer learning: Use the knowledge learned from solving one problem to solve another problem.
- Before BERT, transfer learning in NLP is used at feature representation level: e.g., directly use the prior word embeddings from word2vec as inputs of subsequent tasks.
- After BERT, transfer learning in NLP is mostly used with fine-tuning: e.g., take a pre-trained model and add trainable layers at the final stage to get certain outputs.

22

22

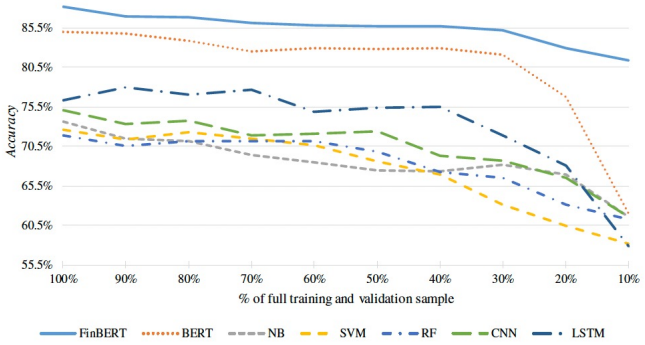
FinBERT

- 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?

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 99 Cite Cited by 144 Related articles Web of Science: 22 »

23

23

Voice of Monetary Policy

The Voice of Monetary Policy[†]

By Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera[*]

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)[1]

Monetary policy is 98 percent talk and 2 percent action, and communication is a big part.
—Ben Bernanke (2022)[2]

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

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

- Use BERT to predict the sentiment of FOMC texts; RoBERTa as a robustness check; or directly use FinBERT for sentiment analysis.

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

- A positive tone of FR chairs leads to significant increases in share prices: How to say is as important as what to say.

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 99 Cite Cited by 118 Related articles All 30 versions Web of Science: 9 »

24

24

Agenda

- BERT: Bidirectional Encoder Representations from Transformers
- GPT: Generative Pretrained Transformers

25

Pretraining Decoders

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

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

$$h_1, \ldots, h_T = \text{Decoder}(w_1, \ldots, w_T)$$
$$w_t \sim Ah_{t-1} + b$$

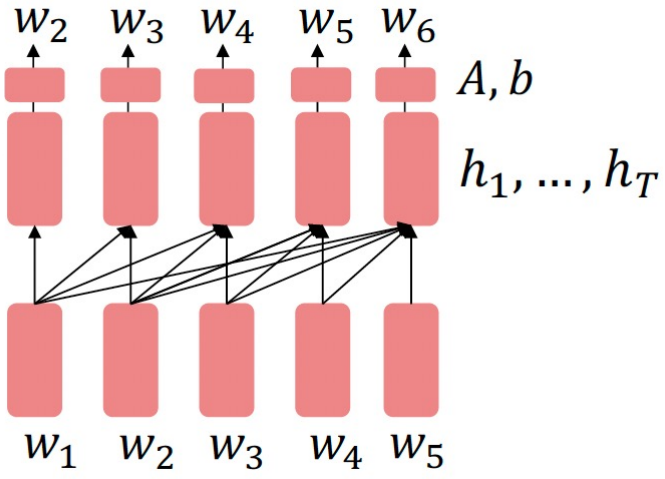
This is a more challenging task than BERT!

[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 99 Cite Cited by 8363 Related articles All 15 versions »

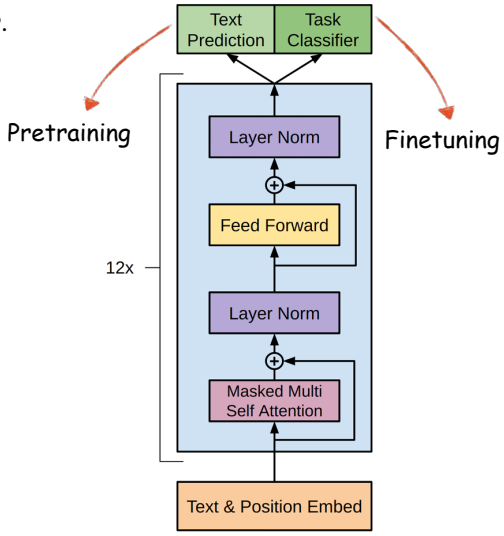


26

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

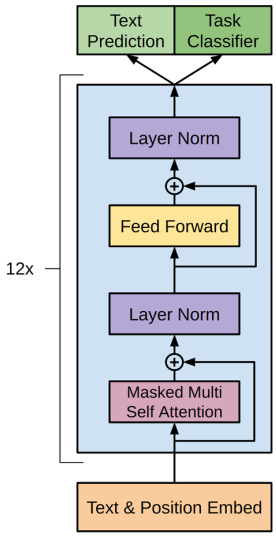
☆ Save 99 Cite Cited by 8363 Related articles All 15 versions »

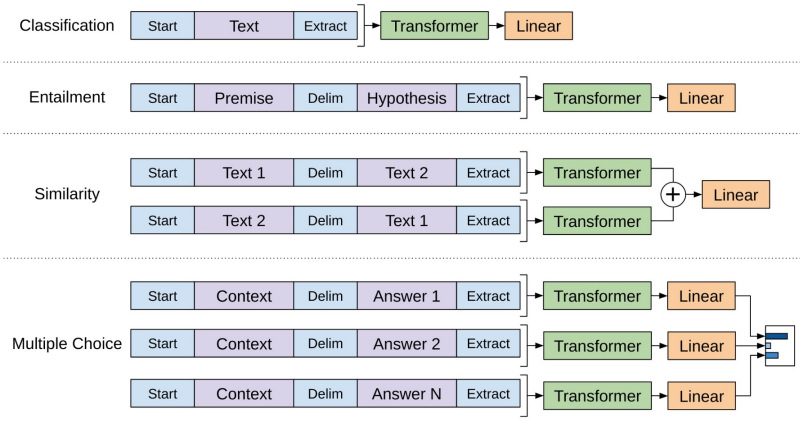
27

27

GPT-1 Finetuning

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





Finetuning Loss = Loss of Text Prediction + lambda * Loss of Classification

28

28

GPT-2

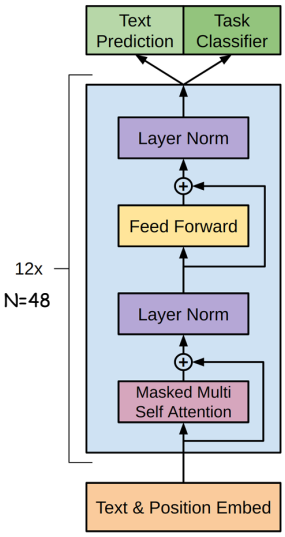
Stanford CS224N, Lecture 9: https://web.stanford.edu/class/cs224n/
Andrej Karpathy's video reproducing GPT-2: https://www.youtube.com/watch?v=l8pRSuU81PU

[PDF] Language models are unsupervised multitask learners
A Radford, J Wu, R Child, D Luan, D Amodei, I Sutskever
OpenAI blog, 2019 · insightcivic.s3.us-east-1.amazonaws ...

Abstract
Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on taskspecific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset-matching or

展开 ⌄

☆ 保存 99 引用 被引用次数：15792 相关文章 所有 31 个版本 »



- Same architecture but order of magnitude larger.
- 48 transformer blocks, 1,600 hidden units per layer, and 25 attention heads, 1.5B parameters.
- Trained on high-quality web-crawled data: 8M documents, 40GB.
- Understanding the contexts:

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

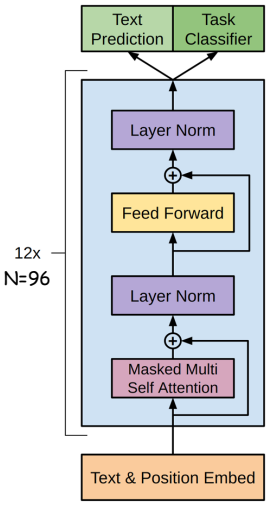
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

29

29

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



- 175B Parameters: 96 decoder blocks, context size = 2,048, embedding-dim=12,288
- Trained on 300B tokens and 10,000 A100s for dozens of days.
- In context learning: Learning without updating gradients.

	No Prompt	Prompt
Zero-shot (0s)	skicts = sticks	Please unscramble the letters into a word, and write that word: skicts = sticks
1-shot (1s)	chiar = chair skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair skicts = sticks

A 2B model in 2025 has a similar level of intelligence as the 175B GPT-3.

Language models are few-shot learners
T Brown, B Mann, N Ryder... - Advances in neural ..., 2020 - proceedings.neurips.cc
... up **language models** greatly improves task-agnostic, **few-shot** ... GPT-3, an autoregressive **language model** with 175 billion ... **language model**, and test its performance in the **few-shot** ...
☆ Save 99 Cite Cited by 21815 Related articles All 27 versions »

30

30

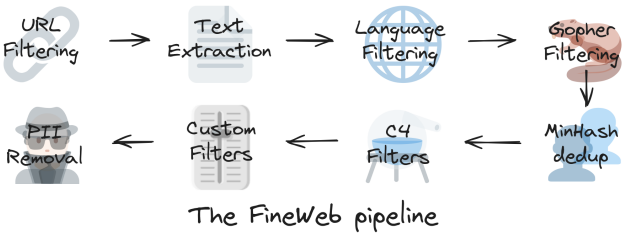
Pretraining Data for LLMs

Stanford CS224N: https://web.stanford.edu/class/cs224n/; FineWeb: https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1

- Orders of magnitude difference in sizes between labeled and unlabeled data:

Dataset	Tokens (~0.75 words)
SQuAD 2.0 [Rajpukar+ 2018]	< 50 Million
DCLM-pool [Li+ 2024]	240 Trillion
Estimated 'internet text' [Villalobos 2024]	510T (indexed), 3100T (total)

A *10 million* times gap in QA to indexed internet



The FineWeb pipeline

Source	Doc Type	UTF-8 bytes (GB)	Documents (millions)	Unicode words (billions)	Llama tokens (billions)
Common Crawl	web pages	9,812	3,734	1,928	2,479
GitHub	code	1,043	210	260	411
Reddit	social media	339	377	72	89
Semantic Scholar	papers	268	38.8	50	70
Project Gutenberg	books	20.4	0.056	4.0	6.0
Wikipedia, Wikibooks	encyclopedic	16.2	6.2	3.7	4.3
Total		**11,519**	**4,367**	**2,318**	**3,059**

https://arxiv.org/abs/2402.00159

- Common Crawl: https://commoncrawl.org/ (750T)
- SOTA models do not release their pretraining data (10~15T).
- Data ablation: Train language models on a dataset following a specific data curation decision.
- Evaluation of curated datasets: zero-shot in-context prompting.

31

31

Tokenization

Stanford CS224N: https://web.stanford.edu/class/cs224n/; Ticktokenizer: https://tiktokenizer.vercel.app/

- Most used tokenization (except Google): Byte Pair Encoding (BPE)
- Efficiency vs. Effectiveness

Algorithm 1 Byte-pair encoding (Sennrich et al., 2016; Gage, 1994)

```
Input: set of strings D, target vocab size k
procedure BPE(D, k)
    V ← all unique characters in D
          (about 4,000 in English Wikipedia)
    while |V| < k do                ▷ Merge tokens
        t_L, t_R ← Most frequent bigram in D
        t_NEW ← t_L + t_R           ▷ Make new token
        V ← V + [t_NEW]
        Replace each occurrence of t_L, t_R in
            D with t_NEW
    end while
    return V
end procedure
```

Iteration	Corpus	Vocabulary
0	AACGCACTATATA	{A,T,C,G}
1	A A C G C A C T A T A T A	{A,T,C,G,TA}
2	A A C G C A C TA TA TA	{A,T,C,G,TA, AC}
3	A AC G C AC TA TA TA	.....

Figure 2: Illustration of the BPE vocabulary constructions.

Monolingual models – 30-50k vocab

Model	Token count
Original transformer	37000
GPT	40257
GPT2/3	50257
T5/T5v1.1	32128
LLaMA	32000

Multilingual / production systems 100-250k

Model	Token count
mT5	250000
PaLM	256000
GPT4	100276
BLOOM	250680
DeepSeek	100000
Qwen 15B	152064
Yi	64000

Multilingual vocabularies are much larger.

Stanford CS336: https://stanford-cs336.github.io/spring2024/

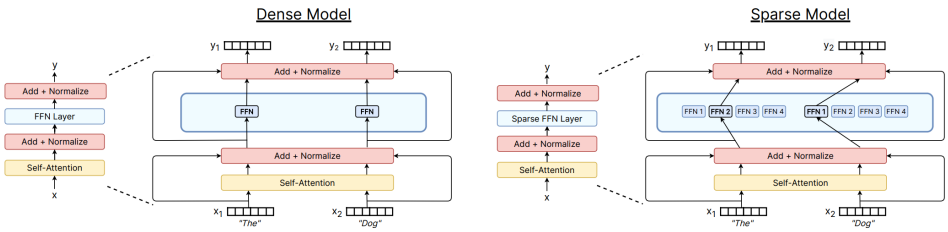
32

32

Mixture of Experts (MoE)

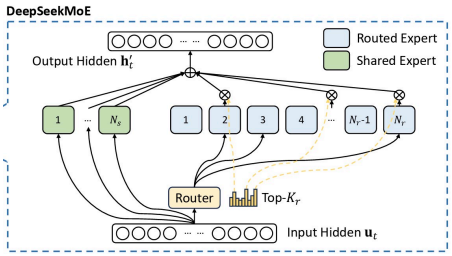
Stanford CS336: https://stanford-cs336.github.io/spring2024/; DeepSeek-V3: https://arxiv.org/pdf/2412.19437v1
Princeton COS 597G: https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec04.pdf

- MoE: Leverage sparsity for more efficient training
  - DeepSeek-V3 has 671B parameters, but each token only activates 37B.



Open-sourced MoEs: Qwen, DeepSeek, Grok, Gemma, Mistral, etc.

GPT-4 is believed to use the MoE architecture as well.



https://www.youtube.com/watch?v=P7txFafuUOE

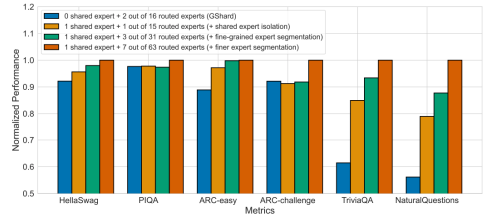


Figure 3 | Ablation studies for DeepSeekMoE. The performance is normalized by the best performance for clarity in presentation. All compared models have the same number of parameters and activated parameters. We can find that fine-grained expert segmentation and shared expert isolation both contribute to stronger overall performance.

Load-balancing: Experts have balanced loads during training.

33

33

A pretrained LLM is called a base model, which is essentially a next-token simulator that approximates ALL the texts on the Internet.

Hallucination is inevitable!!!

To make the base model aligned with human behaviors in specific contexts, we need posttraining.

34

34