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

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 »

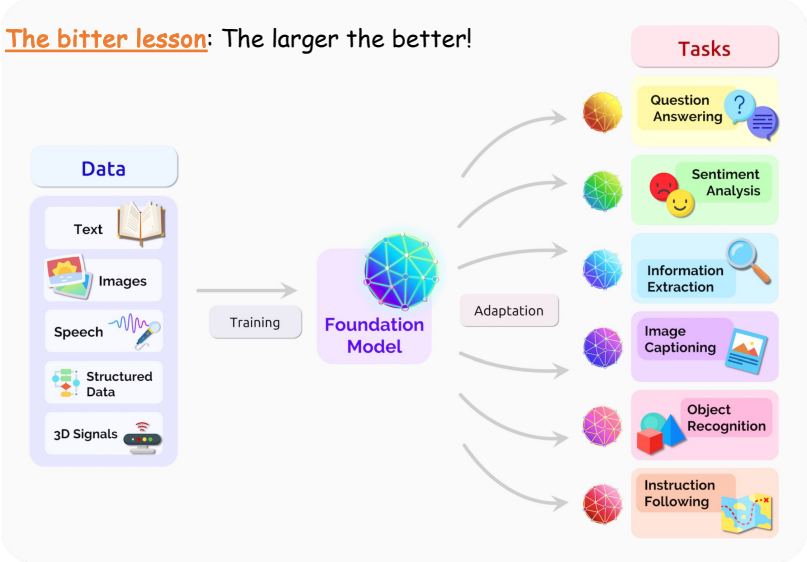
Training language models to follow instructions with human feedback
L Ouyang, J Wu, X Jiang, D Almeida… - Advances in …, 2022 - proceedings.neurips.cc
Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs 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

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



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

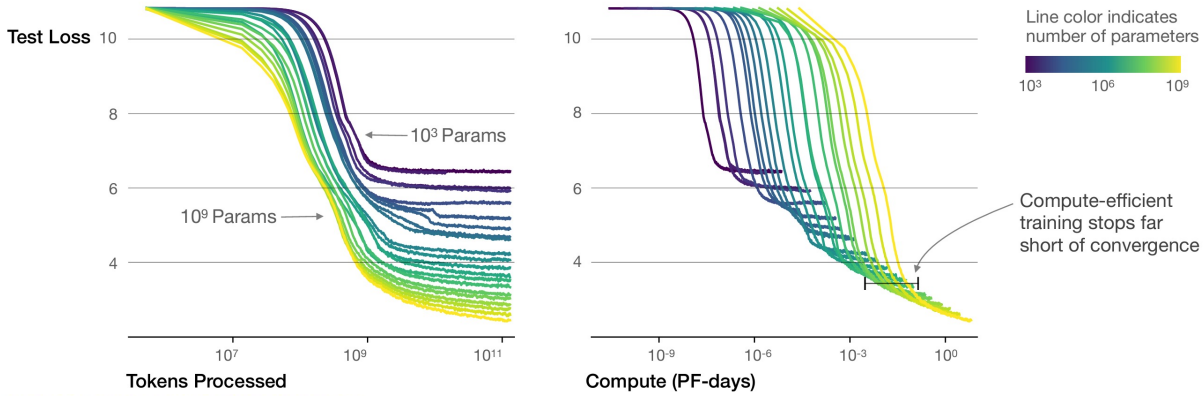
4

4

Scaling Laws

Larger models require **fewer samples** to reach the same performance

The optimal model size grows smoothly with the loss target and compute budget



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**/…

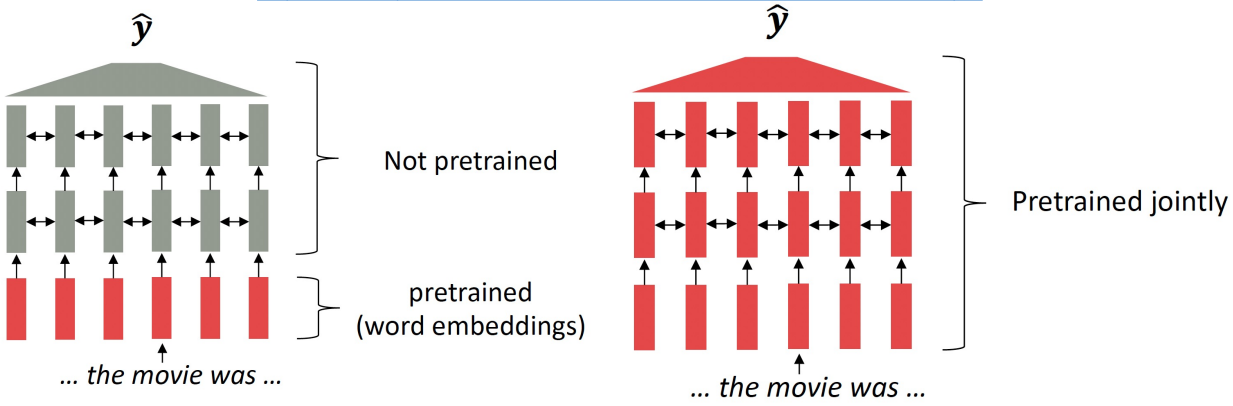
☆ Save 99 Cite Cited by 965 Related articles All 3 versions »

5

5

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



[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]

6

6

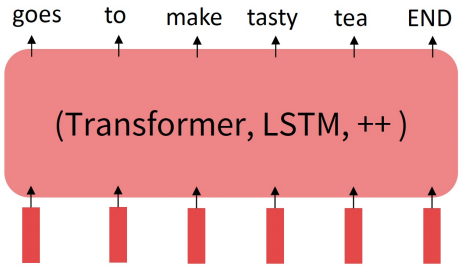
From Pretrained Word Embeddings to Pretrained Models

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)

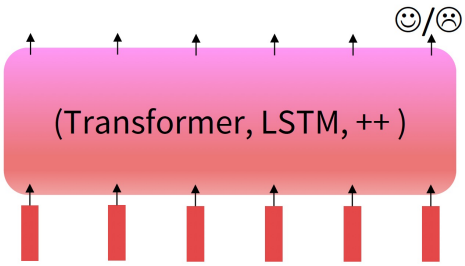
Lots of text; learn general things!



$\hat{\theta}$ 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 …

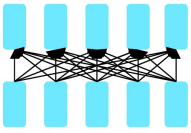
approximates $\min_{\theta} \mathcal{L}_{\text{finetune}}(\theta)$, starting at $\hat{\theta}$

7

7

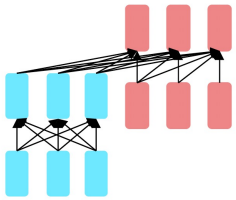
Three Pretraining Architectures

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



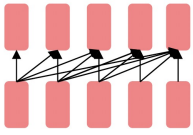
Encoders

- Can condition on future.
- Example: BERT.



Encoder-Decoders

- Combining encoder and decoder.
- Example: T5



Decoders

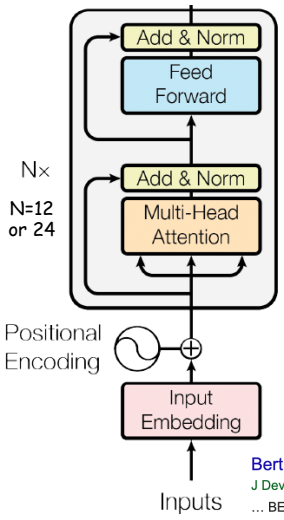
- Cannot condition on future.
- Example: GPT
- All (very) large language models are decoders.

8

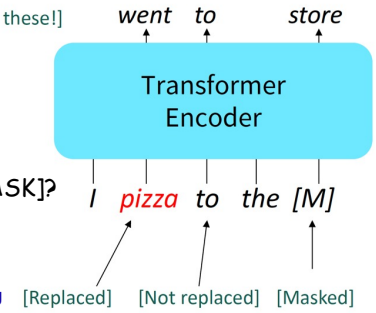
8

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.



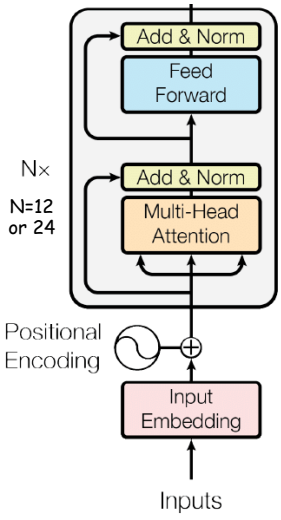
Bert: Pre-training of deep **bidirectional** transformers for language understanding
J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv …, 2018 - arxiv.org
… BERT, which stands for **Bidirectional Encoder Representations** from Transformers. Unlike … 2018), BERT is designed to pretrain deep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

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`

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

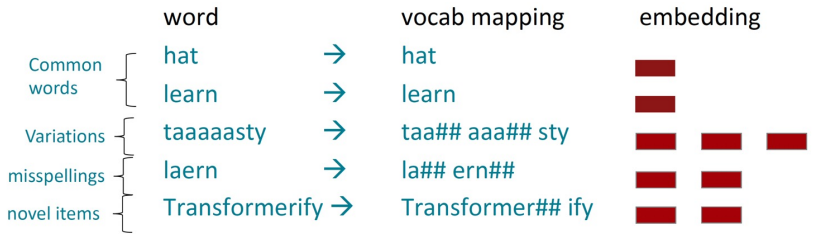
10

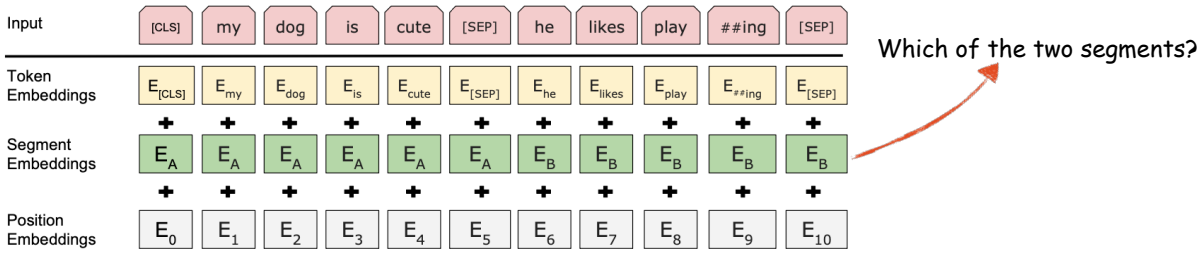
10

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.



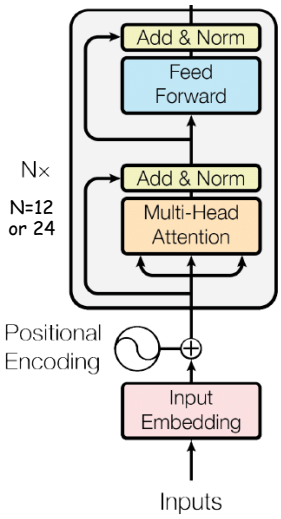


11

11

BERT Pretraining: Putting Together

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



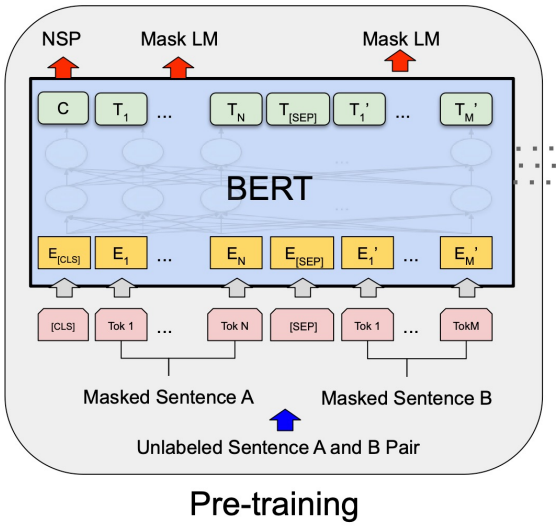
- BERT-base: 12 layers, 768-dim hidden size, 12 attention heads, 110M parameters
- BERT-large: 24 layers, 1024-dim hidden size, 16 attention heads, 340M parameters
- Trained on: Wikipedia (2.5B) + BookCorpus (0.8B)
- Max sequence size: 512 word pieces (roughly 256 + 256 non-contiguous sequences)
- Trained for 1M steps, batch size = 128K
- Pretrained with 64 TPUs for 4 days

12

12

BERT Pretraining: Putting Together

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



Pre-training

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

13

13

Pretrain Once, Finetune Many Times

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

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

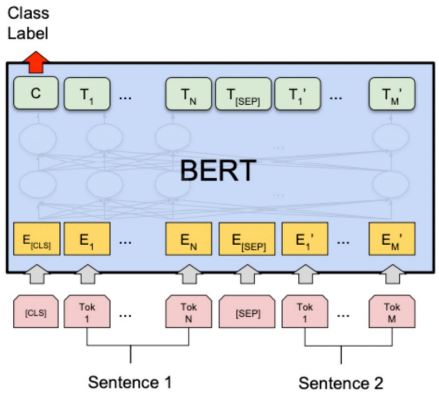
14

14

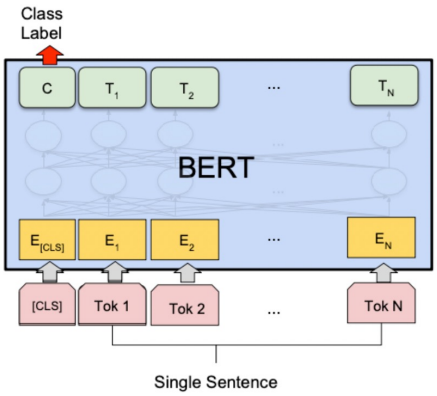
Pretrain Once, Finetune Many Times

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

sentence-level tasks



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



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

15

15

Pretrain Once, Finetune Many Times

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

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

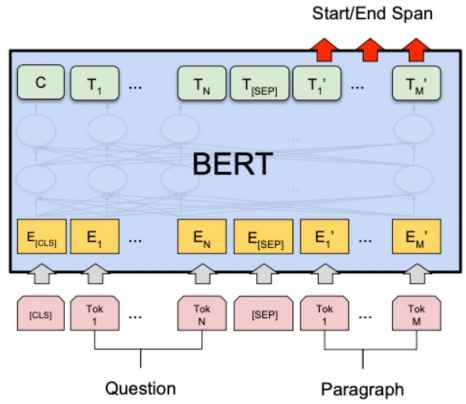
16

16

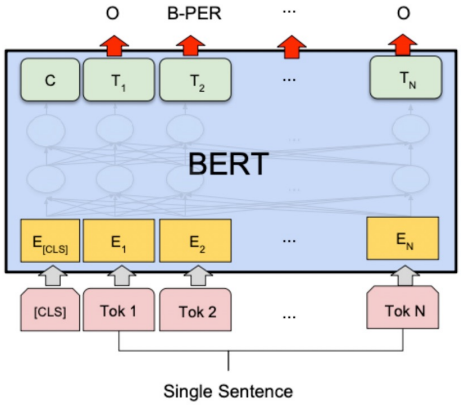
Pretrain Once, Finetune Many Times

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

token-level tasks



(c) Question Answering Tasks: SQuAD v1.1



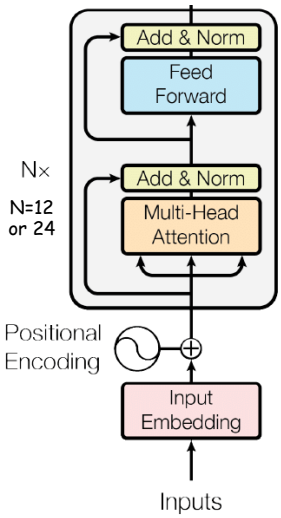
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

17

17

BERT was the State-of-The-Art

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



- 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
$BERT_{BASE}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**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.

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 Sekking Alpha 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?

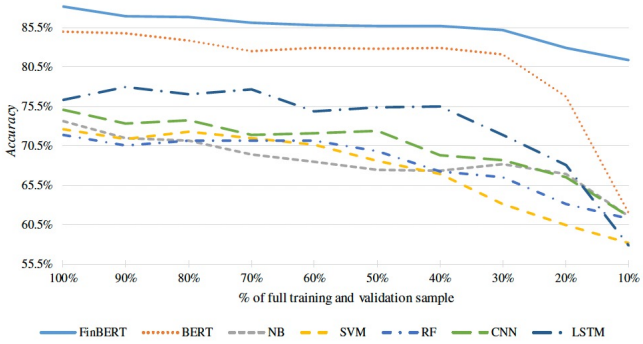
FinBERT: A large language model for **extracting information** from **financial text**

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 »

Figure 1 Sentiment classification accuracy across sample sizes



19

19

Agenda

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

20

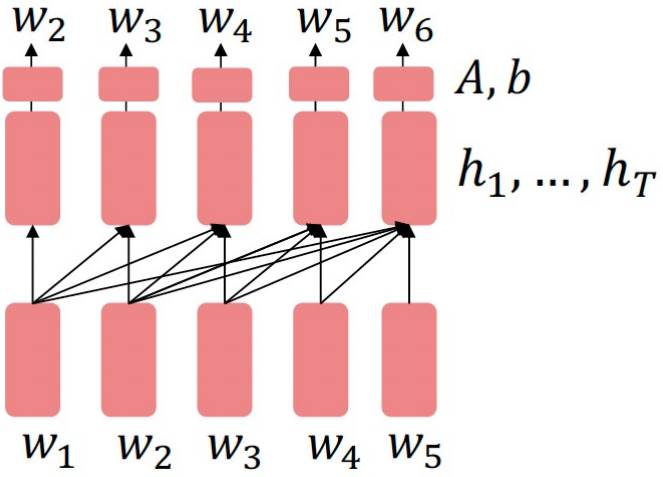
20

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 $Pr(W_n|W_1,W_2,\ldots W_{n-1})$ 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 »

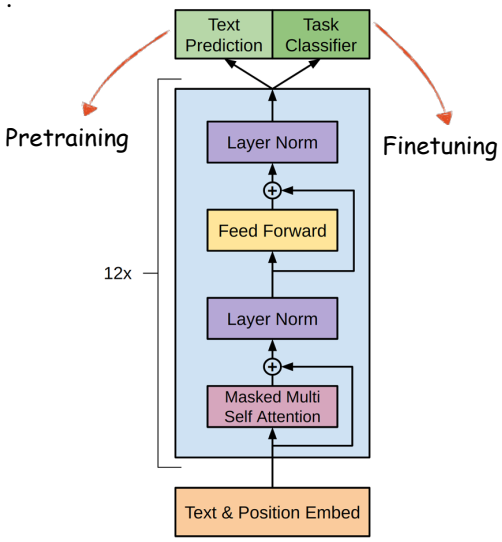
21

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.
- Byte-pair encoding with 40,000 merges.
- 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

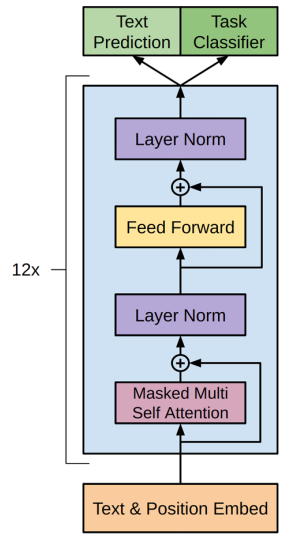
☆ Save 99 Cite Cited by 8363 Related articles All 15 versions »

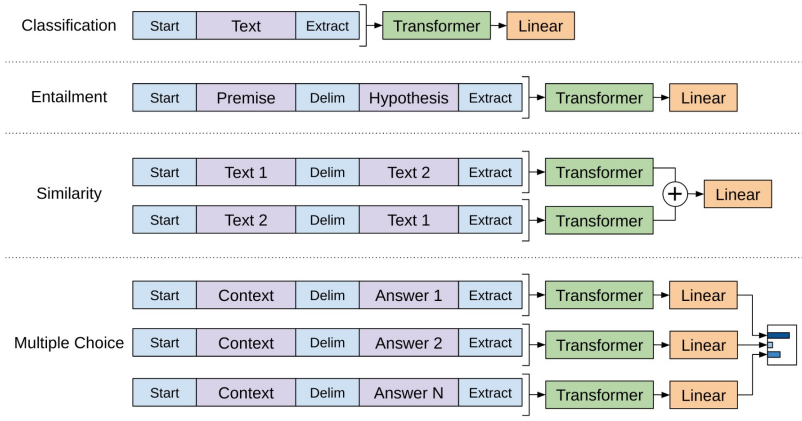
22

22

GPT-1 Finetuning

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





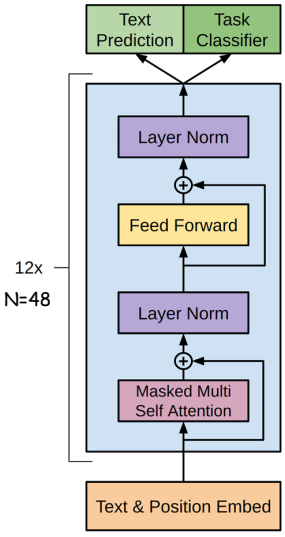
Finetuning Loss = Loss of Text Prediction + lambda * Loss of Classification

23

23

GPT-2

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



- 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.

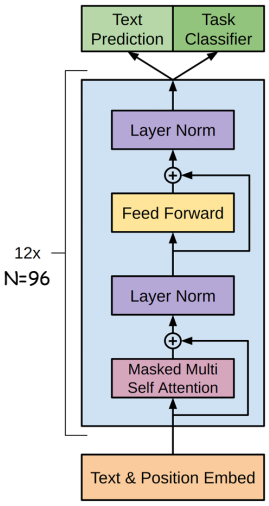
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock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

24

24

GPT-3

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

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 »

25