BERT and GPT-1

Graeme Bates, Mingyang Cen
10/27/25

Review of Transformers

- Embeddings
  - Embedding Tokens into High Dimensional Euclidean Space
- Attention
  - Context plays a major role in Natural Language. Attention allows embedding to take into account the context of a given token.

Tokenization

BPE (byte pair encoding).

Start with taking each letter as a token, then merge together the tokens that appear the most together into one. Repeat until the desired vocabulary size is reached.

Ex: then that fox ate the rabbit. → (th)en (th)at fox ate (th)e rabbit.
→ (th)en (th)(at) fox (at)e (th)e rabbit. → (th)en(th)(at) fox (at)e(th)e rabbit.

Tokens: (th) (th) (at) [space] a b e f h i n o r t x

GPT-1(Radford et al) Motivation

- Labeled Data is hard to come by, but there is an abundance of unlabeled text.
- Natural Language has inherent structures that are common to all tasks.
- Pre-training on unlabeled text allows for a “regularization scheme” for more fine tuned tasks downstream.
- Task Specific Fine-Tuning of a Pre-Trained model requires less labeled data then traditional NLP schemes.
- Text Embeddings with attention allow for long-range linguistic structure.

GPT-1 in action

- Broadly speaking, GPT-1 is pre-training with transformers to train next token prediction $\min_\theta \mathbb{E}_{x\sim D}\left[-\log P_\theta(x_{n+1}|x_1, x_2, ..., x_n)\right]$ here theta represents the learned parameters of the NN
- Minimizing the loss between the actual next token y_{n+1} and predicted token $P_\theta(x_{n+1}|x_1, \ldots, x_n)$

GPT-1 Pre-training Data Sets

- Books Corpus Dataset containing 7,000 unpublished books
  - Long structure of books helps with long-range information training.
- Banord Benchmark containing shuffled sentences
  - This lowered token level perplexity

GPT-1 Pre-Training Architecture

- 12-Layer Decoder only Transformer with masked self attention heads
  - left to right nature of GPT-1, each token only vectorized with attention to previous tokens
- 768 Dimensional states (ie token embedding)
- Position-wise feedforward networks: 3072 dimensional inner states.
- Extensive use of Layer Normalization with weight initialization of N(0,0.2)

GPT-1 Pre-training Details

- Adam Optimizer
  - Momentum and Adaptive Learning Rate
- Max Learning Rate of 2.5e-4
- Cosine Scheduler for warm-up and cool down of Learning Rate.
  - Smoothly increase/decrease Learning rate
- Objective: $\min_{\theta} - \log P_{\theta}(x_{n+1} | x_{1:n})$

GPT-1 Supervised Fine Tuning

Pre-trained model ready for next token prediction ie. text generation

But this is of limited use, so the goal was to fine tune this pre-trained model to perform various more specialized tasks.

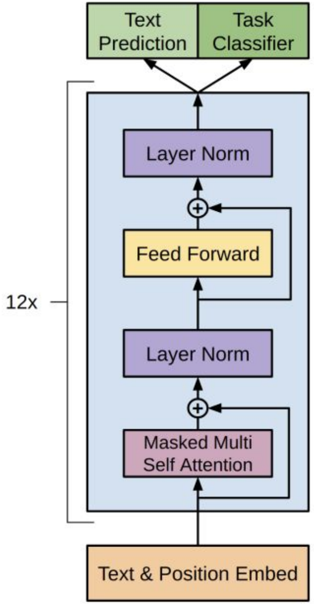
Final Linear Layer is added for fine-tuning and smaller learning rate is used to prevent canceling out pre-training.

Types of Tasks for Fine Tuning

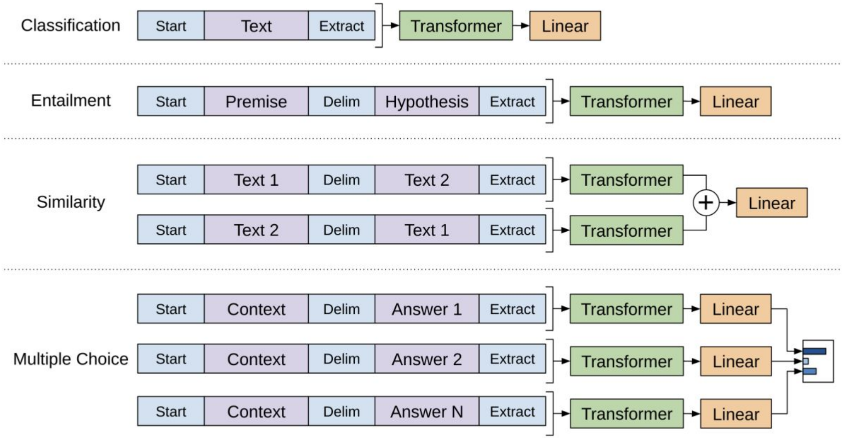
- Natural Language Inference
  - Recognizing the textual entailment
  - ie. judging relationship between two sentences; contradictory, supportive no relation etc…
- Question-Answering
  - Selecting best ending to multi-sentence stories
  - Completing middle/high school level exams.
- Classification
- Semantic Similarity

GPT-1 Structure

The transformer structure is as follows:



These are the transformations applied to turn a labeled dataset into a sequence of natural language tokens for fine-tuning. Special blue tokens indicate delimiters.



GPT-1 Was it Worth It?

- GPT-1 set state of the art benchmarks on 9 of the 12 studied tasks.

GPT-1 Was it Worth It?

- Natural Language Inference

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

GPT-1 Outperforms on 4 of 5 baselines
RTE dataset happens to be the smallest

GPT-1 Was it Worth It?

- Question-Answering

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

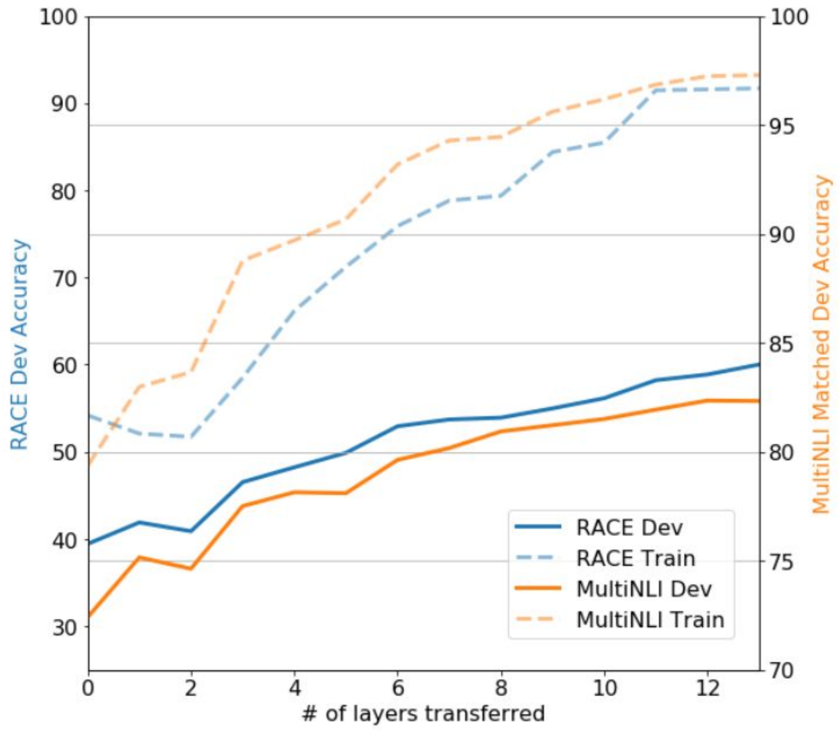
GPT-1 Outperforms on all datasets

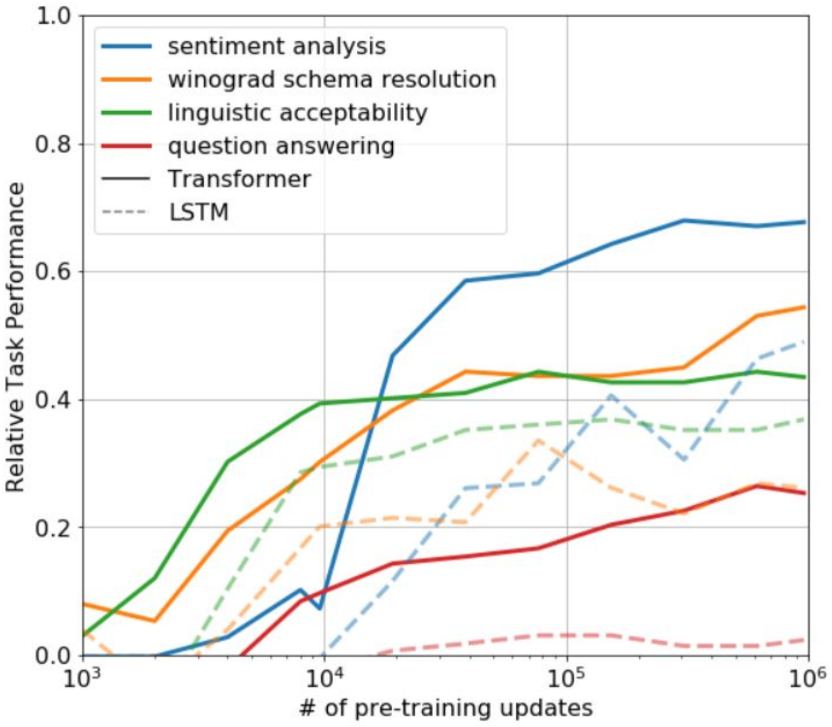
GPT-1 Was it Worth It?

- Semantic Similarity and Classification

Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

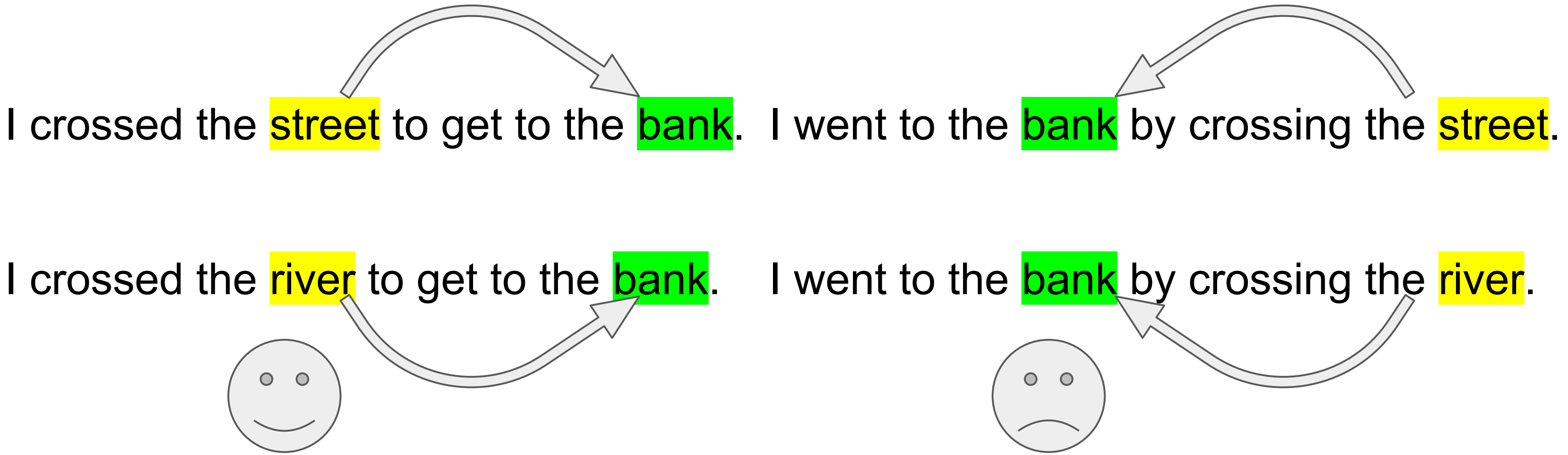
GPT-1 Analysis



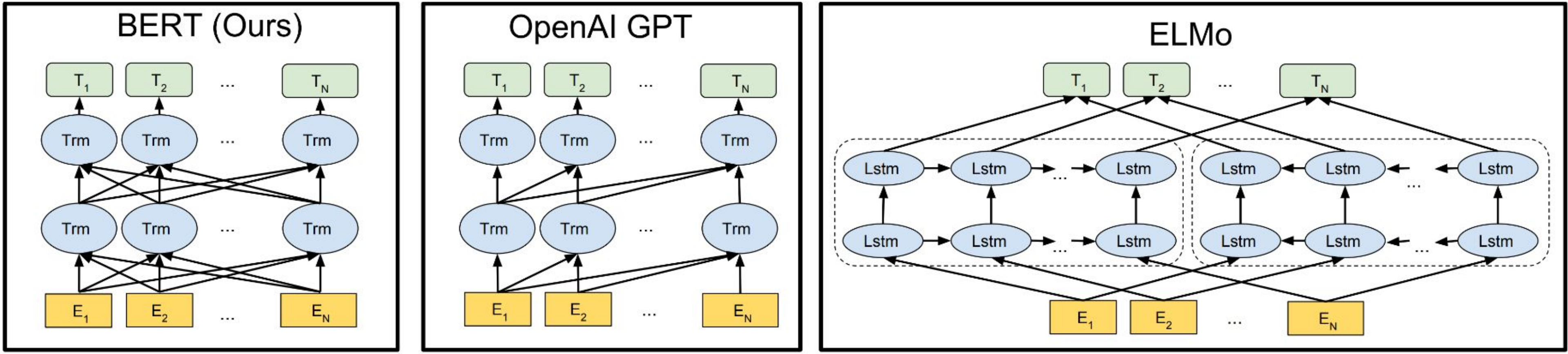


BERT Motivation (Devlin et al.)

GPT-1 was trained on **constrained** self-attention, where every token can only have its attention computed with ones to its left. BERT fixes this and allows every token to attend to every other.



BERT Structure



ELMo/BiLSTM was a previous “feature-based” model: essentially a left-to-right RNN-style (long short-term memory) model combined with a right-to-left model. As you can see, BERT is bidirectional in every layer, making it superior.

BERT Details

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

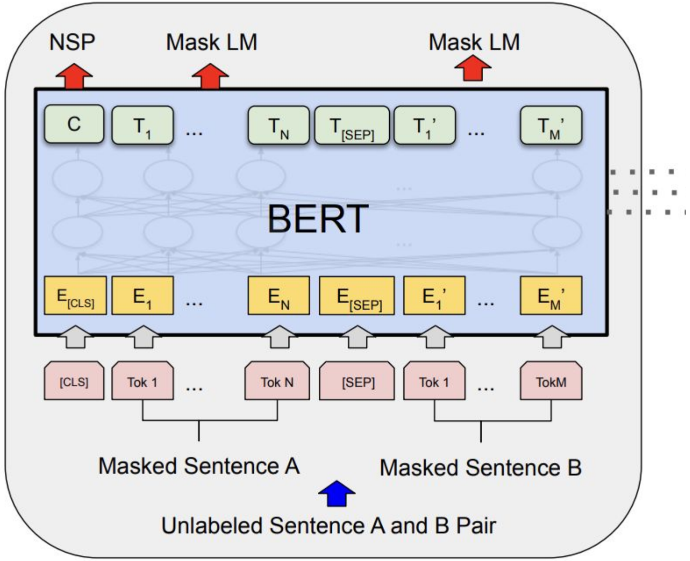
Effectively an encoder-only transformer.

Instead of next-token prediction, the model is pre-trained on two tasks:

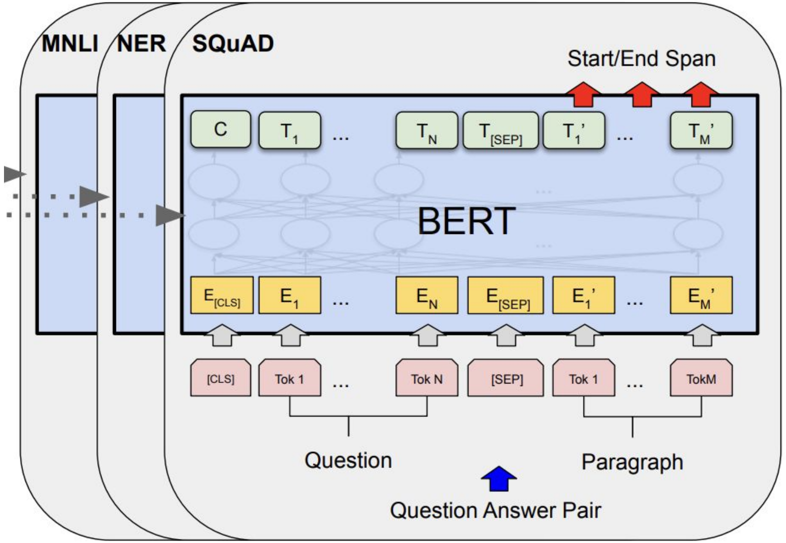
- Masked LM/Cloze: Give the model a sequence of tokens, but choose some randomly to hide or replace with random. The model is challenged to find the correct words.
- A sequence of two sentences (one after the other) is chosen from the training data. The second is replaced with a random unrelated sentence with a 50% probability.

Same fine-tuning process as GPT.

BERT Details

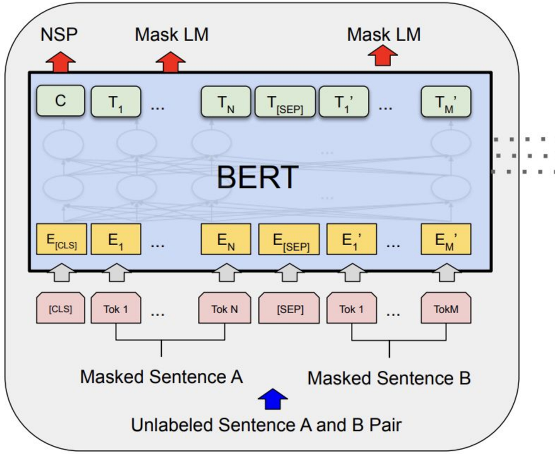


Pre-training



Fine-Tuning

BERT Pre-training: Cloze task



Pre-training

My dog is **hairy**.

→ My dog is **[MASK]**. (80%)

→ My dog is **hairy**. (10%)

→ My dog is **apple**. (10%)

The model's task is to output the probability distribution of the actual, non-masked tokens as the T_i.

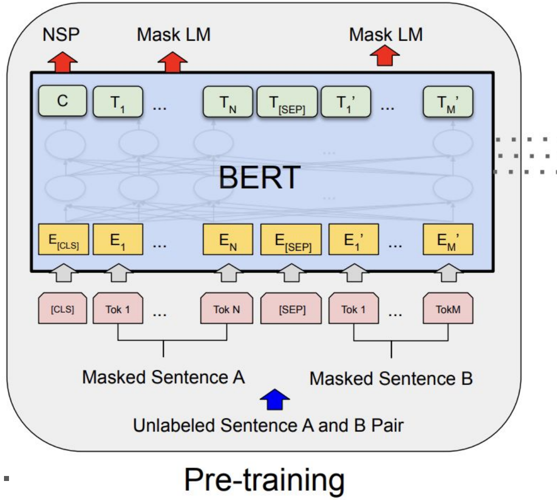
BERT Pre-training: Next Sentence Prediction

The man went to the store. [SEP] He bought a gallon of milk.

→ **IsNext**

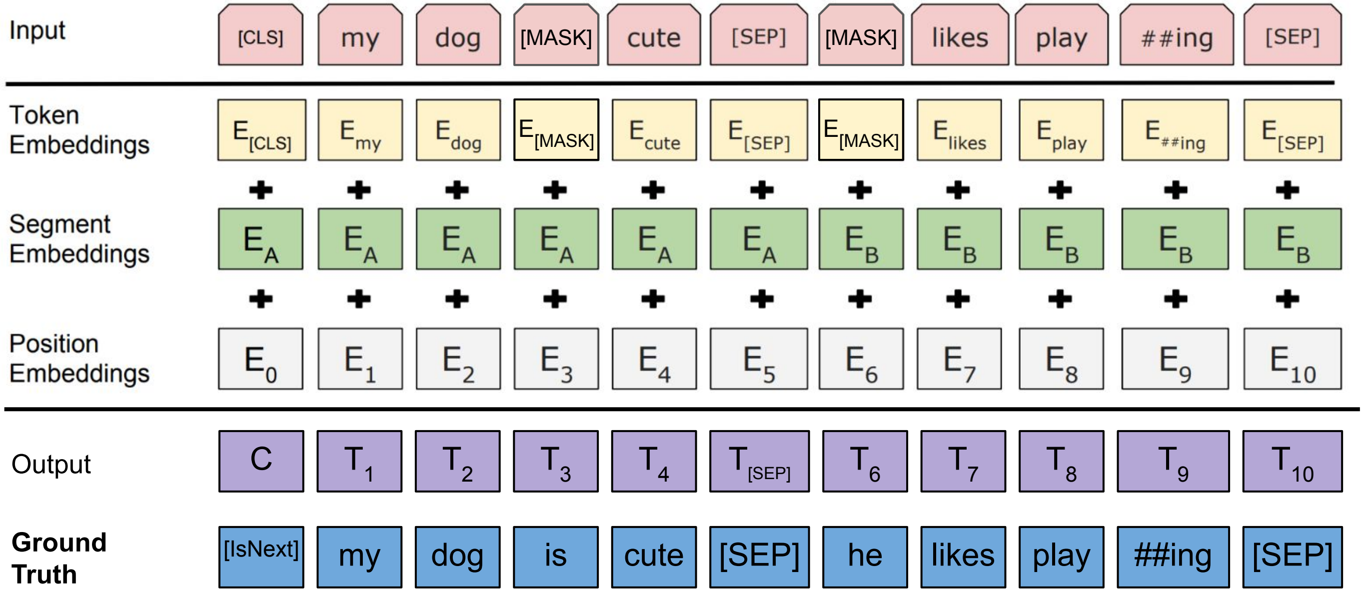
The man went to the store. [SEP] Penguins are flightless birds.

→ **NotNext**



BERT Embedding and Output Format

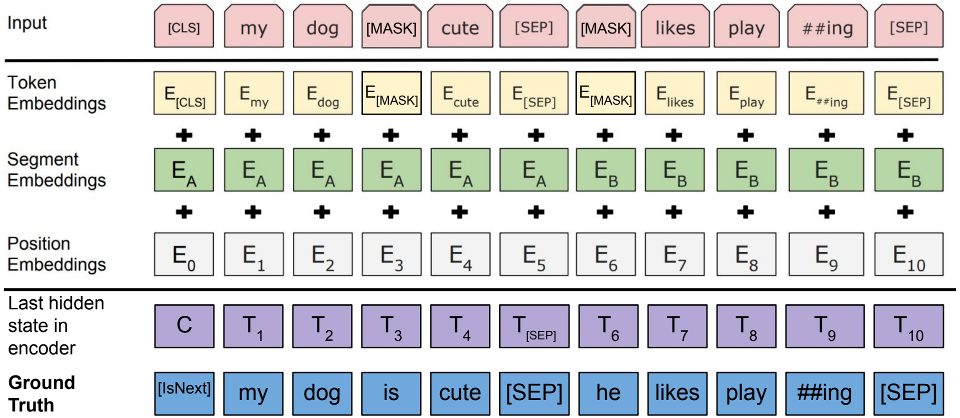
The two tasks, Cloze and NSP, are combined.
BERT uses a combination of three embeddings: a token embedding, a segment embedding, and a position embedding.



BERT Embedding and Output Format

This embedding allows fine-tuning on several natural language processing tasks:

Is this text grammatical? Is this movie review positive? Do these two sentences mean the same thing? Which part of this article answers this question? Does it even answer the question? Does this statement imply the other?



BERT Ablation Study Results

Every part of BERT is important!
The number of attention heads, embedding dimension, parameters, etc. cause smooth improvement.

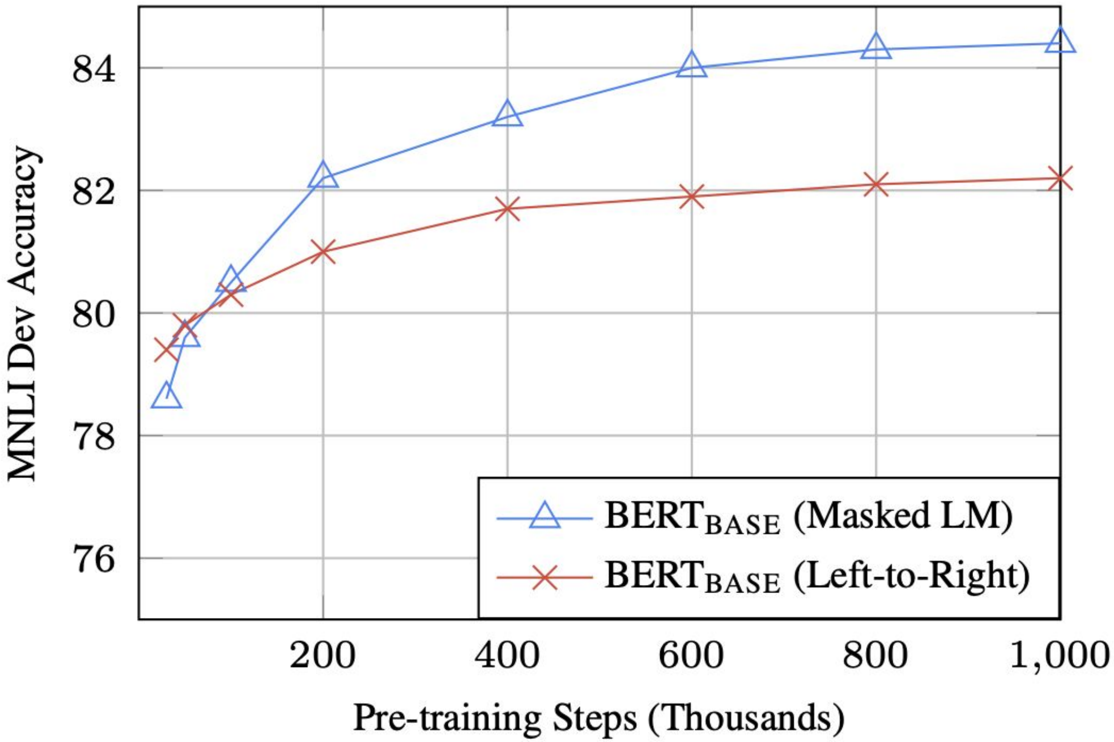
Tasks	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
	Dev Set				
$\mathrm{BERT}_{\mathrm{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Cloze task only

GPT-1-style Cloze

ELMo +
GPT-1-style Cloze

BERT Ablation Study Results



Scaling Laws

Kaplan et al.

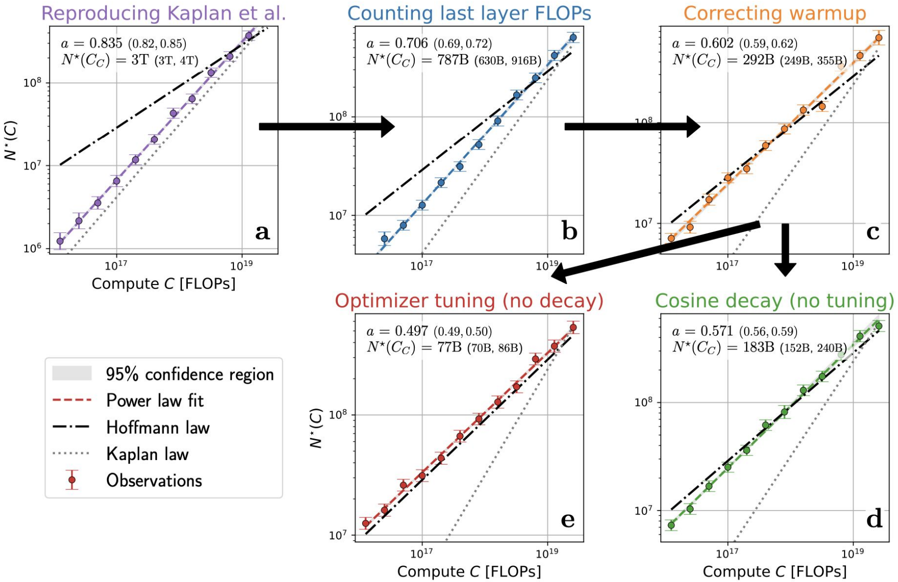
$N^{0.74} \sim D$ where N = params, D = dataset size.

10^3–10^9 non-embedding parameters in tests

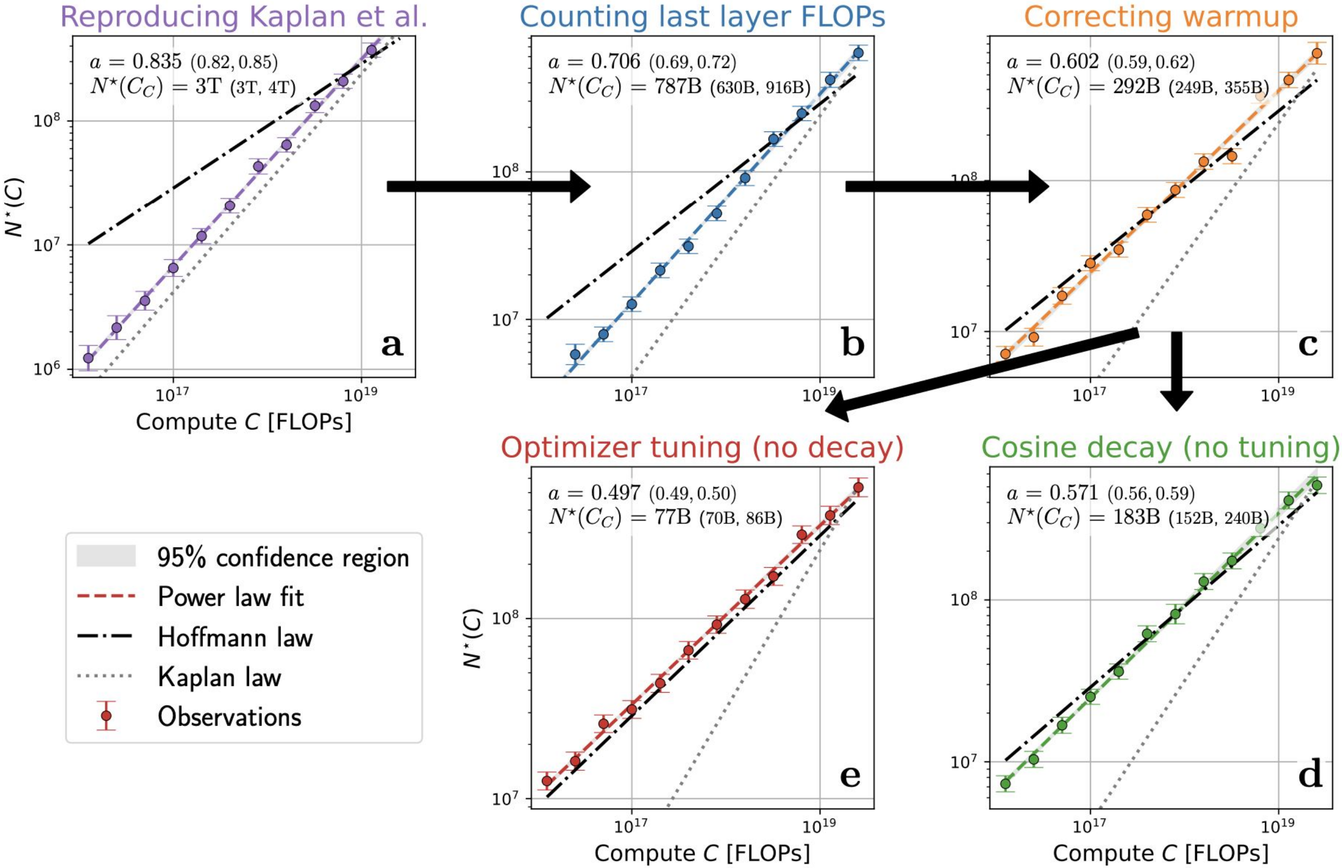
Hoffman et al.

N ~ D where N = params and D = dataset size

7×10^7–10^{10} parameters in tests



Scaling Laws



Data Sourcing

WebText (GPT-2): Scraped content of websites receiving outward links from trusted Reddit posts.

Common Crawl: Scraped websites

Various books, trusted websites, Wikipedia

Data Processing

- Deduplication using various hashing algorithms
- Heuristics to remove bad data
- Training small classifier models to tell apart good data and trash.
- Weighting different sources: better sources like books and Wikipedia can get 2-3 passes while less filtered sources like Common Crawl get less than 1.
- Separating different languages, removing formatting

Zero-shot and Few-shot Tasks

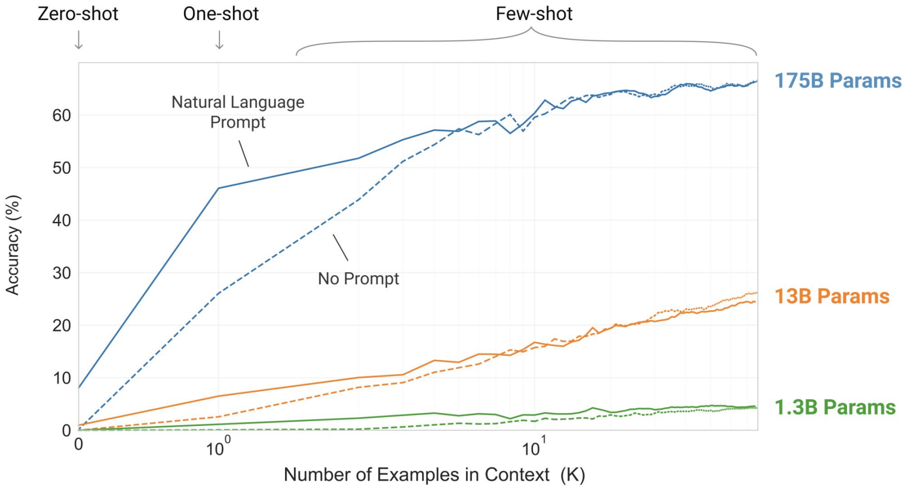
Very large models experience a sort of emergent behavior: if the task to be performed can be expressed in language, then the text (task description, examples of acceptable output, input) can be fed directly into a language model to produce good output. This gets even better after fine-tuning.

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Various benchmarks for GPT-2 before fine-tuning. Bold is better than state of the art at the time.

Zero-shot and Few-shot Tasks

(Task description, examples of task, input) → language model → Output



GPT-3: accuracy plotted against number of examples given.

References

GPT-1,2,3:
https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

https://arxiv.org/pdf/2005.14165

BERT:

https://arxiv.org/pdf/1810.04805

WordPiece tokenization:

https://huggingface.co/learn/llm-course/chapter6/6

References

Kaplan scaling laws:

https://arxiv.org/pdf/2001.08361

Chinchilla scaling laws:

https://arxiv.org/pdf/2203.15556

Reconciliation:

https://arxiv.org/pdf/2406.12907

https://arxiv.org/abs/2406.19146