Large Language Models

CSCI 601 471/671
NLP: Self-Supervised Models

https://self-supervised.cs.jhu.edu/sp2023/

JOHNS HOPKINS UNIVERSITY

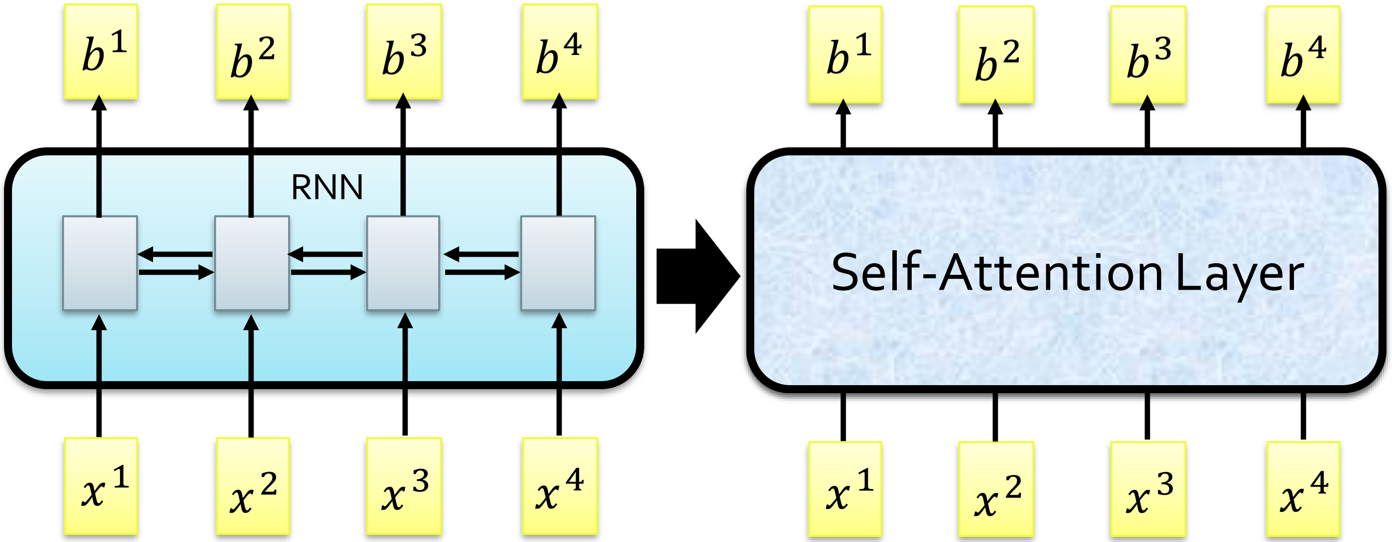
[Slide credit: Chris Tanner, Jacob Devlin and many others]

Logistics Update

- The midterm:
  - will be on March 7 during class time.
  - I will not be here; Adam (TA) will run the show.
  - it will be on paper
  - It will be based on the ideas you have seen in homework and lectures. If you understand them, you're set!
  - Scope HW 1-5 and lectures until last Thursday (Feb 23)

Recap: Self-Attention

- b^i is obtained based on the whole input sequence.
- can be parallelly computed.

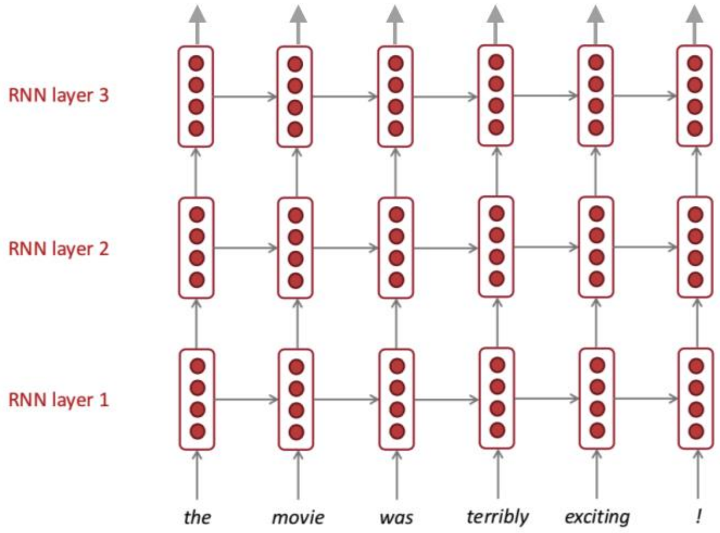


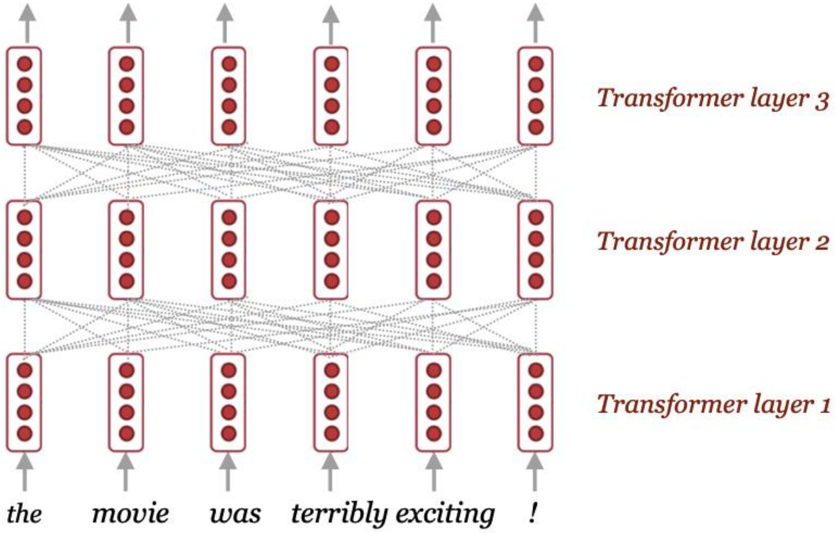
Idea: replace any thing done by RNN with self-attention.

"Neural machine translation by jointly learning to align and translate" Bahdanau etl. 2014;
"Attention is All You Need" Vaswani et al. 2017

[adopted from Hung-yi Lee]

Recap: RNN vs Transformer



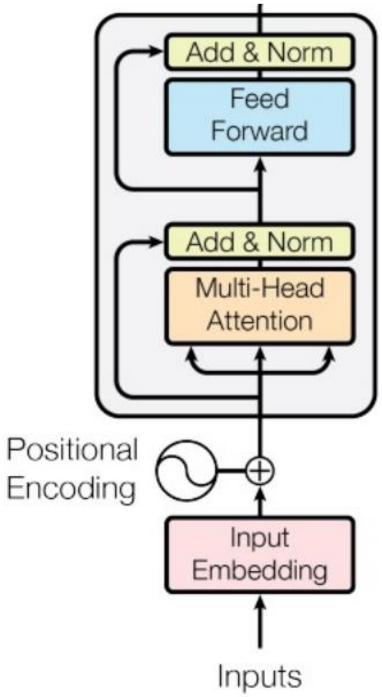


Recap: Attention Block

Given input $\mathbf{x}$:

$$Q = \mathbf{W}^q \mathbf{x}$$
$$K = \mathbf{W}^k \mathbf{x}$$
$$V = \mathbf{W}^v \mathbf{x}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{h}}\right)V$$



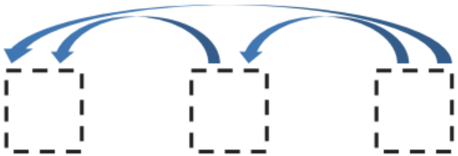
[Attention Is All You Need, Vaswani et al. 2017]

Recap: Transformer [Vaswani et al. 2017]

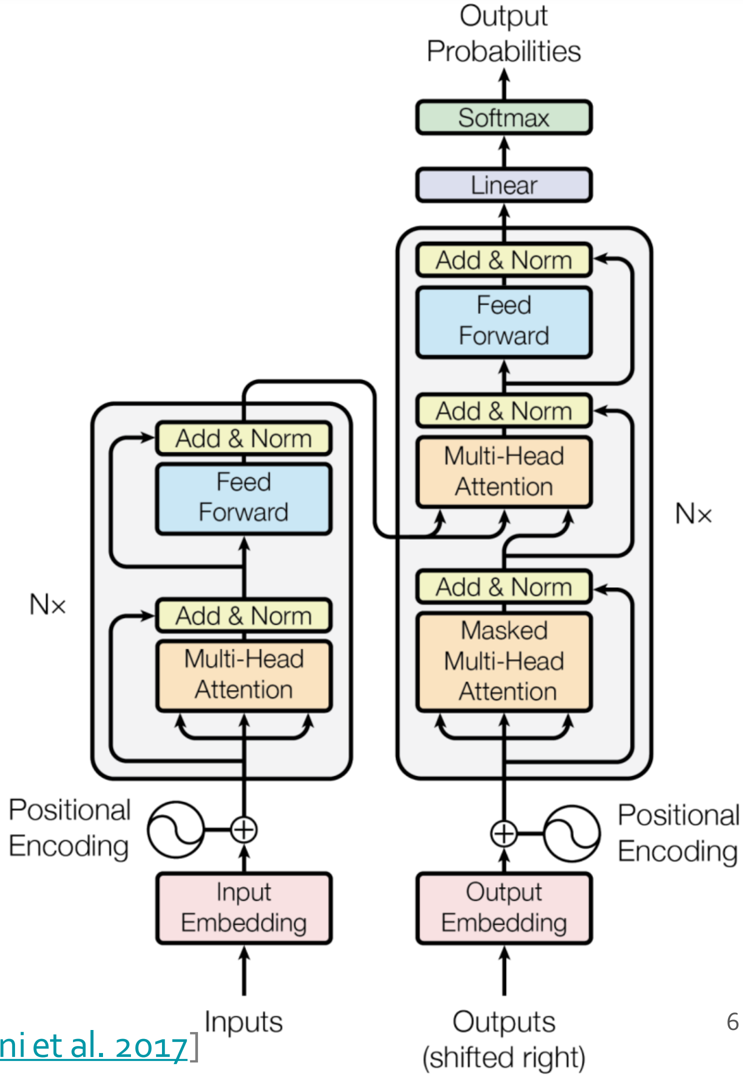
- An encoder-decoder architecture
- 3 forms of attention

Encoder-Decoder Attention

Encoder Self-Attention

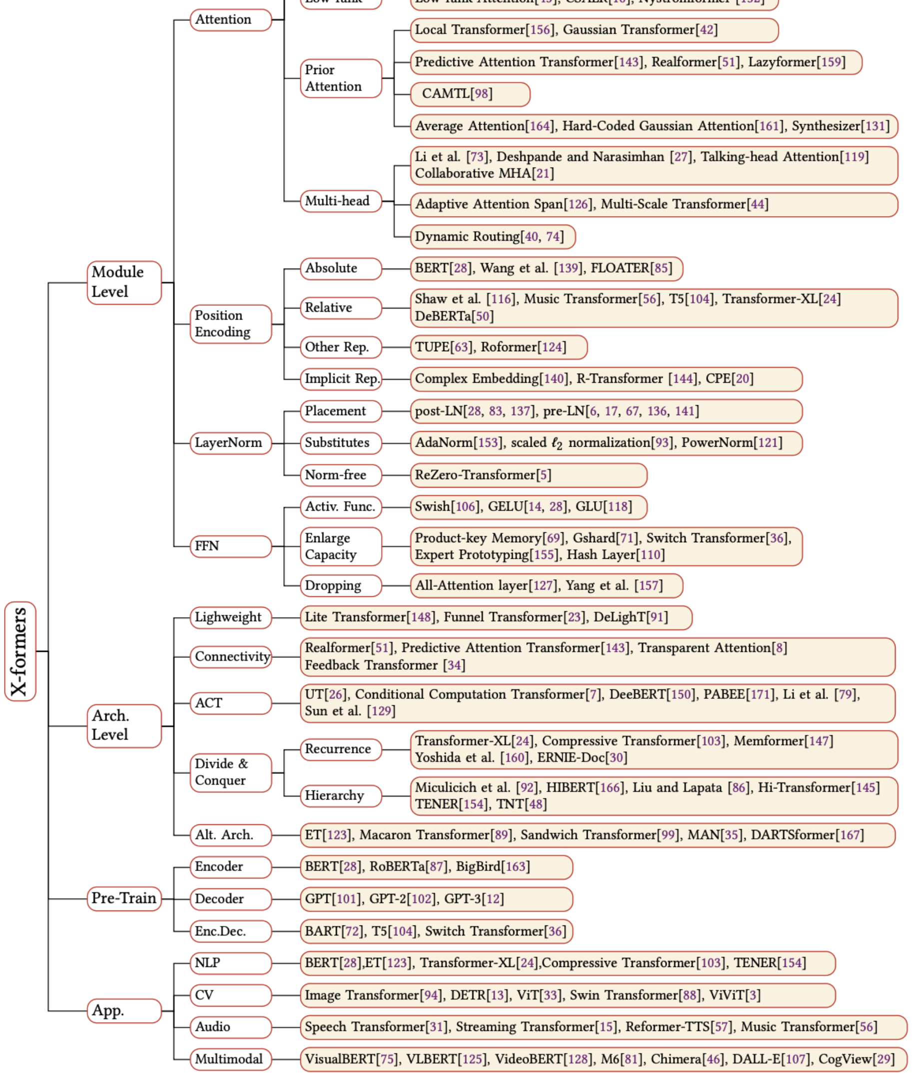


MaskedDecoder Self-Attention



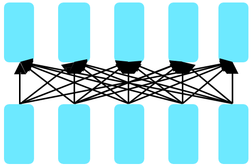
[Attention Is All You Need, Vaswani et al. 2017]

After Transformer …


X-formers
Module Level
Attention
Prior Attention
Local Transformer[156], Gaussian Transformer[42]
Predictive Attention Transformer[143], Realformer[51], Lazyformer[159]
CAMTL[98]
Average Attention[164], Hard-Coded Gaussian Attention[161], Synthesizer[131]
Multi-head
Li et al. [73], Deshpande and Narasimhan [27], Talking-head Attention[119]
Collaborative MHA[21]
Adaptive Attention Span[126], Multi-Scale Transformer[44]
Dynamic Routing[40, 74]
Position Encoding
Absolute
BERT[28], Wang et al. [139], FLOATER[85]
Relative
Shaw et al. [116], Music Transformer[56], T5[104], Transformer-XL[24]
DeBERTa[50]
Other Rep.
TUPE[63], Roformer[124]
Implicit Rep.
Complex Embedding[140], R-Transformer [144], CPE[20]
LayerNorm
Placement
post-LN[28, 83, 137], pre-LN[6, 17, 67, 136, 141]
Substitutes
AdaNorm[153], scaled ℓ_2 normalization[93], PowerNorm[121]
Norm-free
ReZero-Transformer[5]
FFN
Activ. Func.
Swish[106], GELU[14, 28], GLU[118]
Enlarge Capacity
Product-key Memory[69], Gshard[71], Switch Transformer[36],
Expert Prototyping[155], Hash Layer[110]
Dropping
All-Attention layer[127], Yang et al. [157]
Arch. Level
Lighweight
Lite Transformer[148], Funnel Transformer[23], DeLighT[91]
Connectivity
Realformer[51], Predictive Attention Transformer[143], Transparent Attention[8]
Feedback Transformer [34]
ACT
UT[26], Conditional Computation Transformer[7], DeeBERT[150], PABEE[171], Li et al. [79],
Sun et al. [129]
Divide & Conquer
Recurrence
Transformer-XL[24], Compressive Transformer[103], Memformer[147]
Yoshida et al. [160], ERNIE-Doc[30]
Hierarchy
Miculicich et al. [92], HIBERT[166], Liu and Lapata [86], Hi-Transformer[145]
TENER[154], TNT[48]
Alt. Arch.
ET[123], Macaron Transformer[89], Sandwich Transformer[99], MAN[35], DARTSformer[167]
Pre-Train
Encoder
BERT[28], RoBERTa[87], BigBird[163]
Decoder
GPT[101], GPT-2[102], GPT-3[12]
Enc.Dec.
BART[72], T5[104], Switch Transformer[36]
App.
NLP
BERT[28],ET[123], Transformer-XL[24],Compressive Transformer[103], TENER[154]
CV
Image Transformer[94], DETR[13], ViT[33], Swin Transformer[88], ViViT[3]
Audio
Speech Transformer[31], Streaming Transformer[15], Reformer-TTS[57], Music Transformer[56]
Multimodal
VisualBERT[75], VLBERT[125], VideoBERT[128], M6[81], Chimera[46], DALL-E[107], CogView[29]

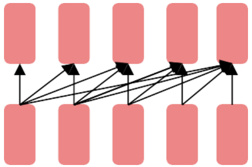
Impact of Transformers

- A building block for a variety of LMs



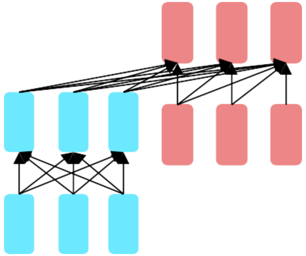
Encoders

- ❖ Examples: BERT, RoBERTa, SciBERT.
- ❖ Captures bidirectional context. Wait, how do we pretrain them?



Decoders

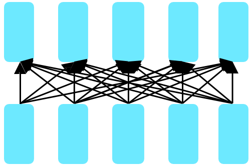
- ❖ Examples: GPT-2, GPT-3, LaMDA
- ❖ Other name: causal or auto-regressive language model
- ❖ Nice to generate from; can't condition on future words



Encoder-Decoders

- ❖ Examples: Transformer, T5, Meena
- ❖ What's the best way to pretrain them?

BERT



Encoders

BERT

Bidirectional Encoder Representations from Transformers

BERT

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

Like Bidirectional LSTMs (ELMo), let's look in **both** directions

BERT

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

Let's only use Transformer Encoders, no Decoders

BERT

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

It's a language model that builds rich representations
via self-supervised learning (pre-training)

BERT (2018)

- Transformer based network to learn representations of language

- Improvements
    - Bi-directional LSTM -> Self-attention
    - Massive data
    - Masked-LM objective

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

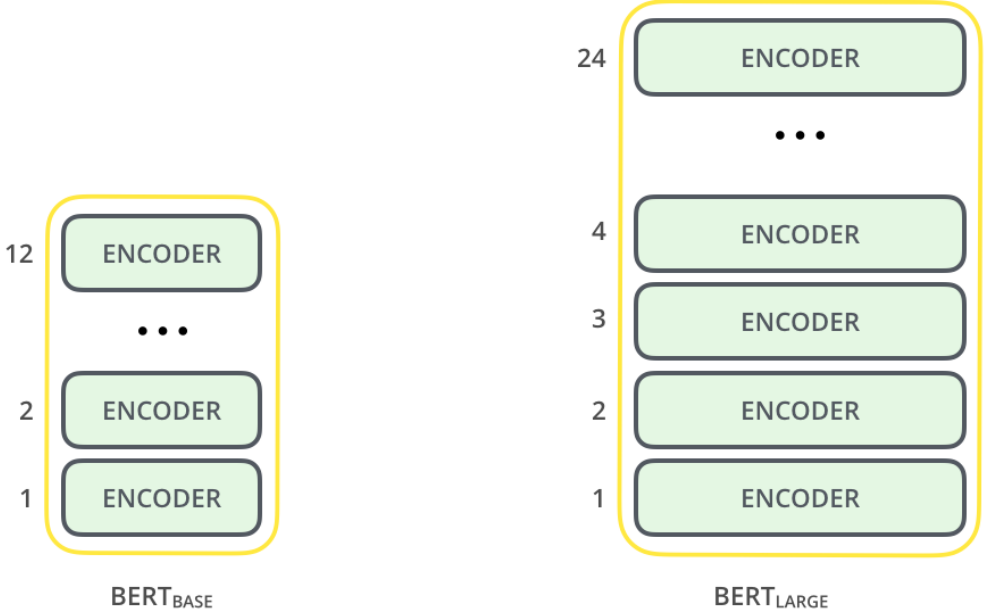
BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute im-

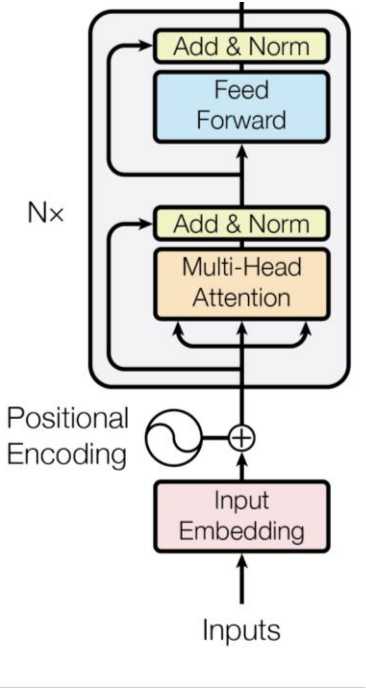
There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-

BERT: Architecture

- Stacks of Transformer encoders”

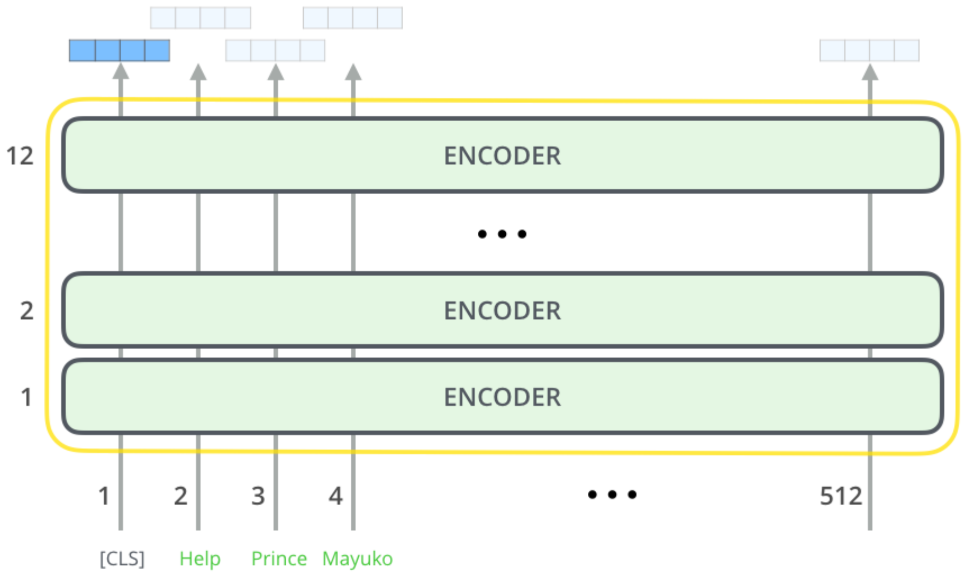


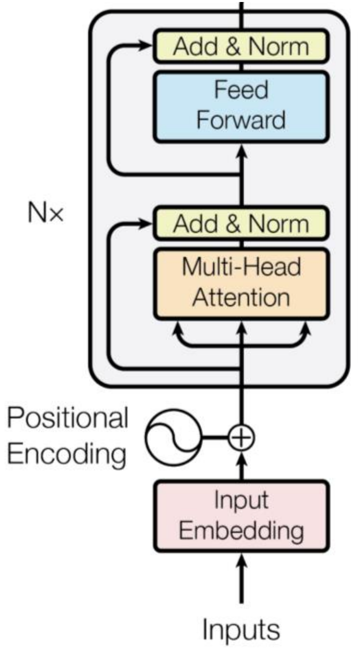


[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

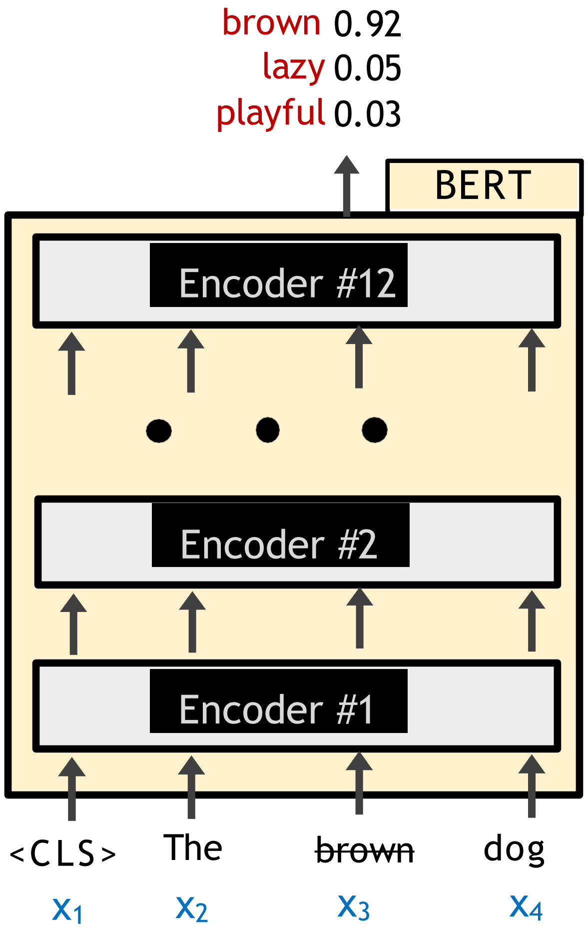
BERT: Architecture

- Model output dimension: 512





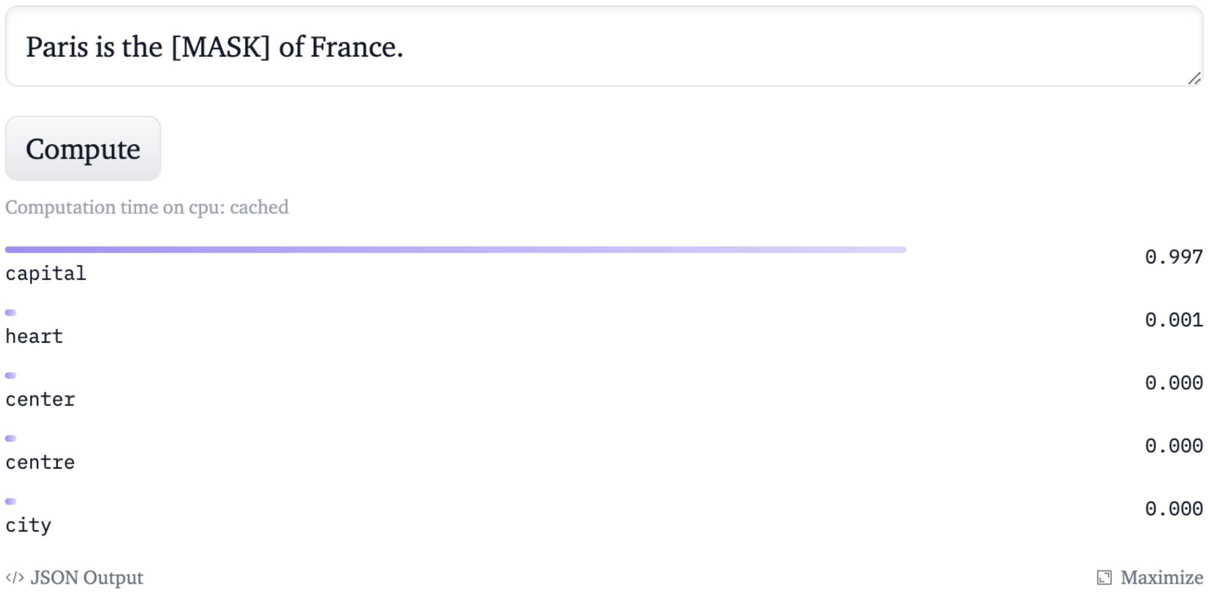
[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]



BERT is trained to uncover masked tokens.

Probing BERT Masked LM

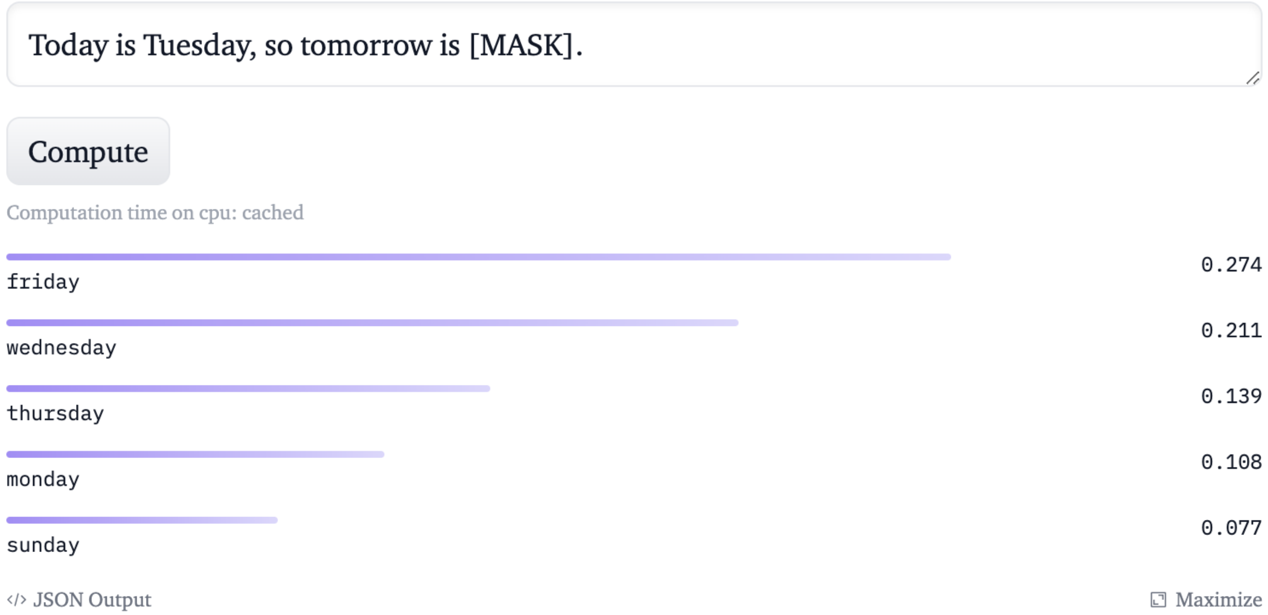
- Making words forces BERT to use context in both directions to predict the masked word.



https://huggingface.co/bert-base-uncased

Probing BERT Masked LM

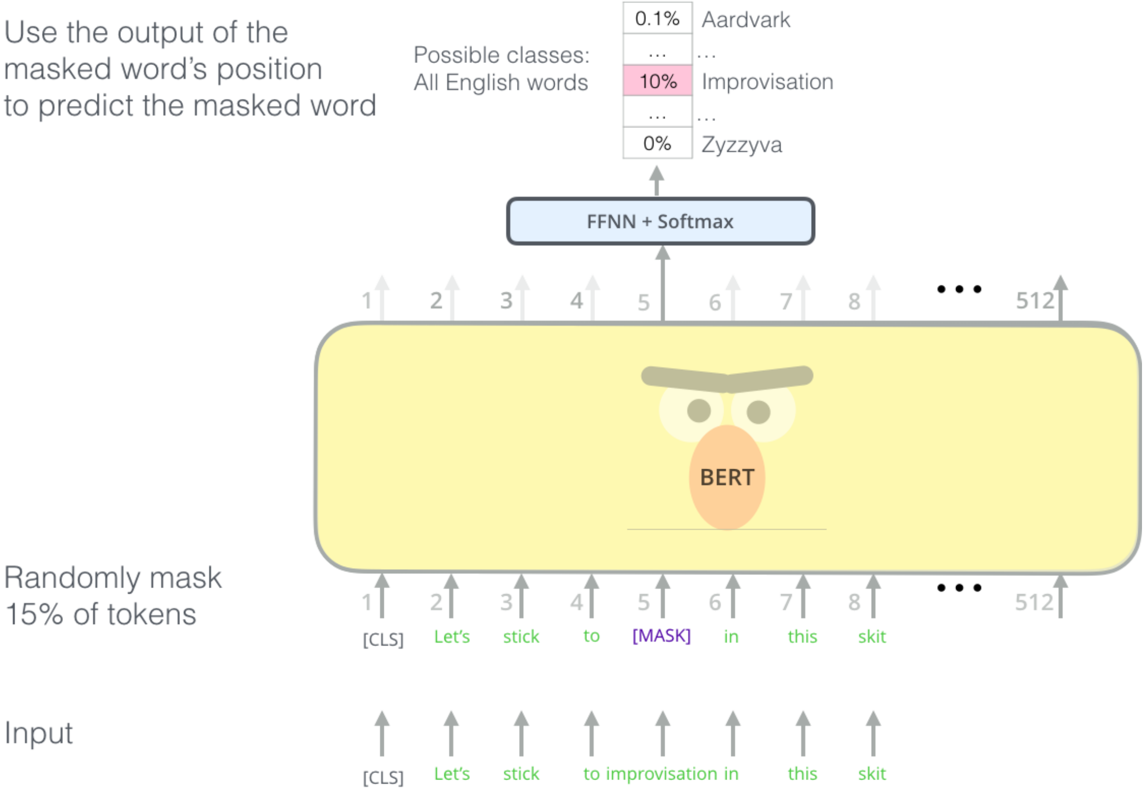
- Making words forces BERT to use context in both directions to predict the masked word.



https://huggingface.co/bert-base-uncased

BERT: Pre-training Objective (1): Masked Tokens

- Randomly mask 15% of the tokens and train the model to predict them.



[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

BERT: Pre-training Objective (1): Masked Tokens

store

Galon

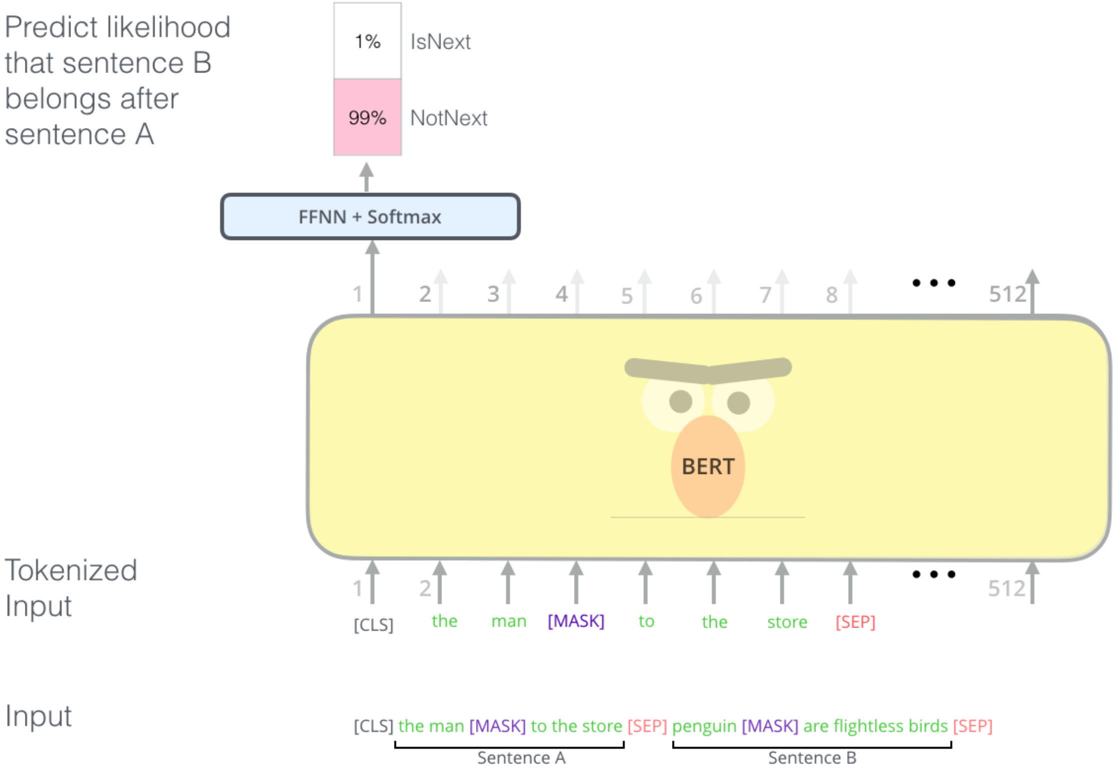
`the man went to the [MASK] to buy a [MASK] of milk`

- Too little masking: Too expensive to train
- Too much masking: Underdefined (not enough context)

[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

BERT: Pre-training Objective (2): Sentence Ordering

- Predict sentence ordering
- 50% correct ordering, and 50% random incorrect ones



[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

BERT: Pre-training Objective (2): Sentence Ordering

- Learn relationships between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

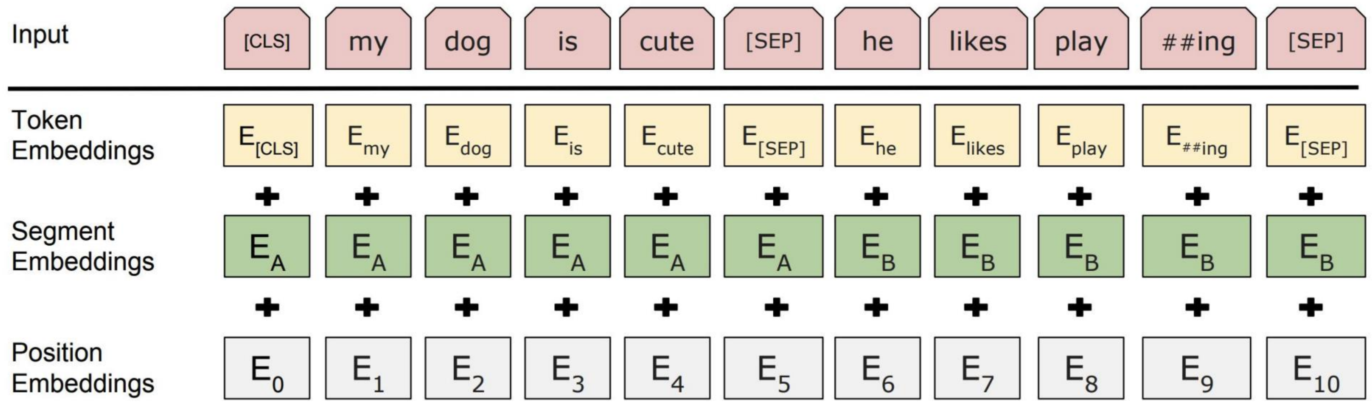
Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

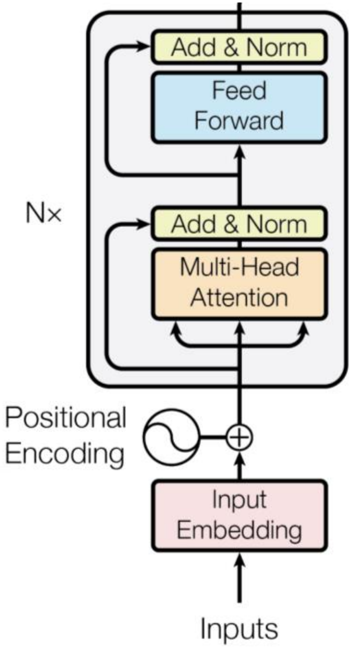
Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

BERT: Input Representation

- Use 30,000 WordPiece vocabulary on input.
- Each token is sum of three embeddings
  - Addition to transformer encoder: sentence embedding





[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

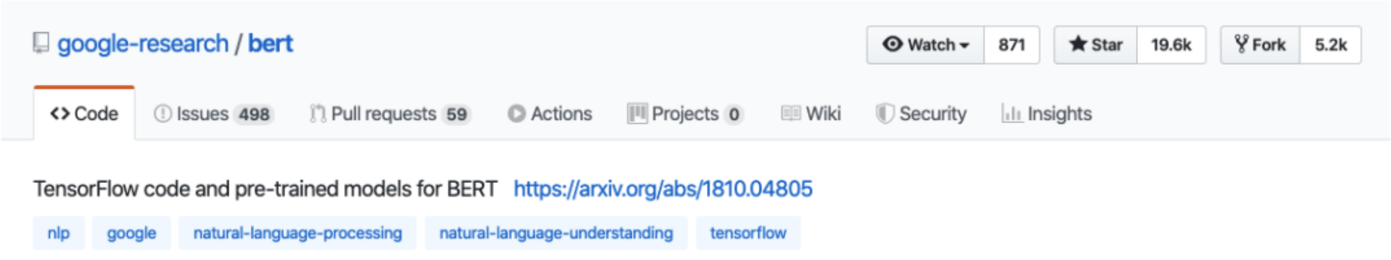
Training

- Trains model on unlabeled data over different pre-training tasks (self-supervised learning)
- **Data:** Wikipedia (2.5B words) + BookCorpus (800M words)
- **Training Time:** 1M steps (~40 epochs)
- **Optimizer:** AdamW, 1e-4 learning rate, linear decay
- **BERT-Base:** 12-layer, 768-hidden, 12-head
- **BERT-Large:** 24-layer, 1024-hidden, 16-head
- Trained on 4x4 or 8x8 TPUs for 4 days

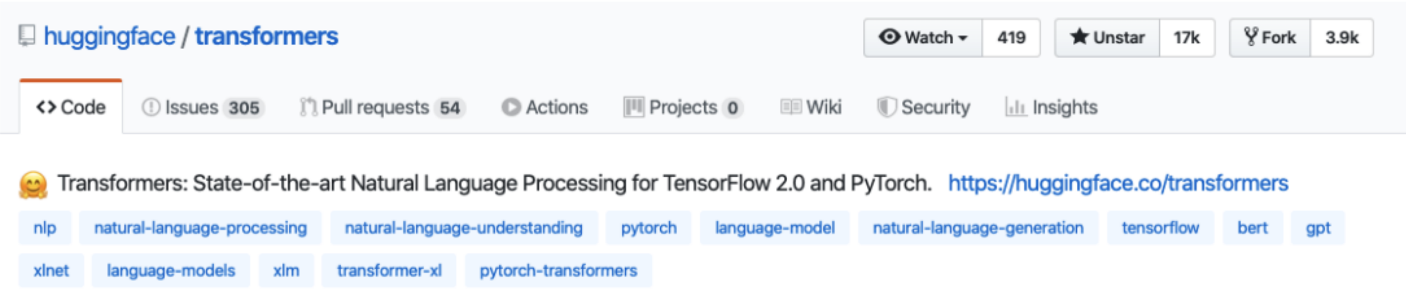
[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

BERT in Practice

TensorFlow: https://github.com/google-research/bert



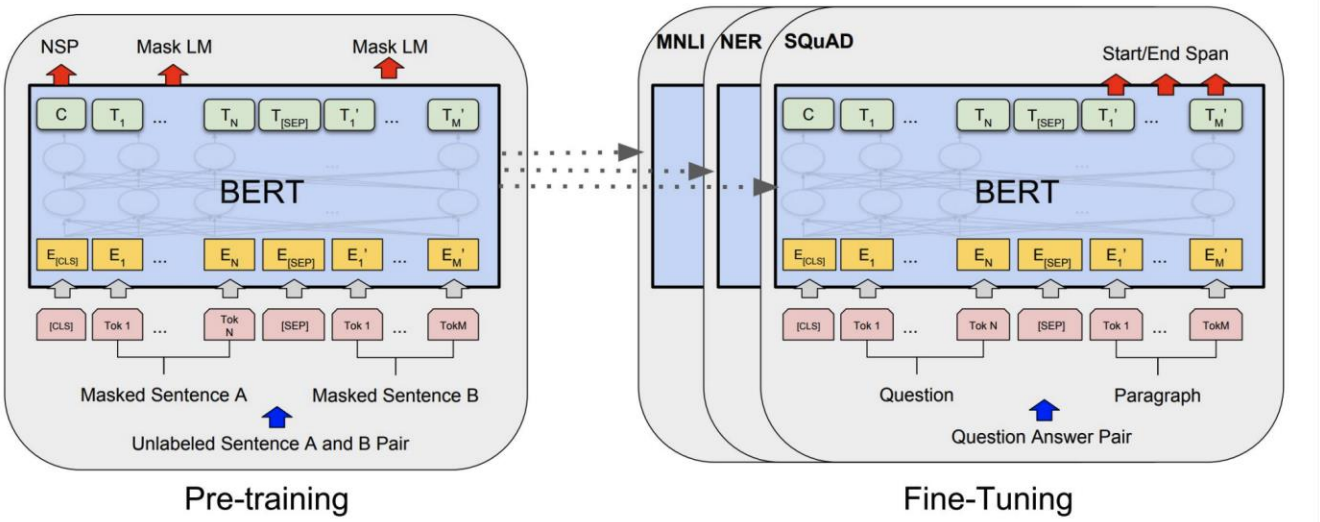
PyTorch: https://github.com/huggingface/transformers



Fine-tuning BERT

"Pretrain once, finetune many times."

- **Idea:** Make pre-trained model **usable** in **downstream tasks**
- Initialized with pre-trained model parameters
- Fine-tune model parameters using labeled data from downstream tasks



[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

An Example Result: SWAG

```
A girl is going across a set of monkey bars.  She
(i)  jumps up across the monkey bars.
(ii)  struggles onto the bars to grab her head.
(iii)  gets to the end and stands on a wooden plank.
(iv)  jumps up and does a back flip.
```

Leaderboard

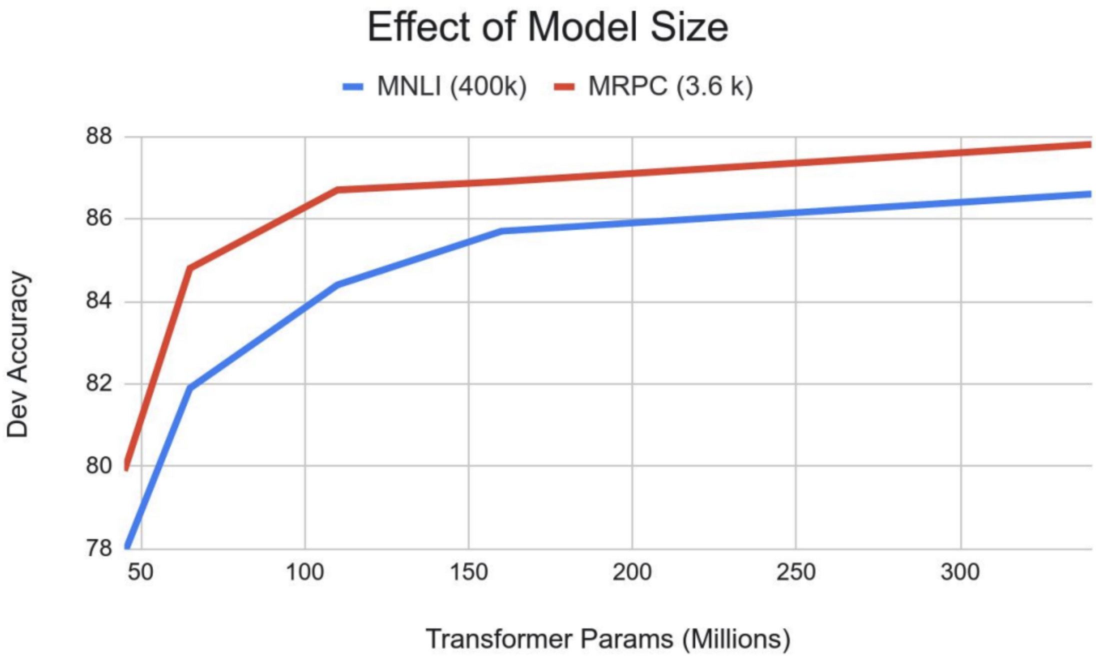
Human Performance (88.00%)
Running Best
Submissions

Rank	Model	Test Score
1	**BERT (Bidirectional Encoder Representations from Transfo...** *Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova* 10/11/2018	**86.28%**
2	**OpenAI Transformer Language Model** *Original work by Alec Radford, Karthik Narasimhan, Tim Salimans, ...* 10/11/2018	**77.97%**
3	**ESIM with ELMo** *Zellers, Rowan and Bisk, Yonatan and Schwartz, Roy and Choi, Yejin* 08/30/2018	**59.06%**
4	**ESIM with Glove** *Zellers, Rowan and Bisk, Yonatan and Schwartz, Roy and Choi, Yejin* 08/29/2018	**52.45%**

- Run each Premise + Ending through BERT.
- Produce logit for each pair on token 0 ([CLS])

[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

Effect of Model Size



- Big models help a lot
- Going from 110M -> 340M params helps even on datasets with 3,600 labeled examples
- Improvements have not *asymptoted*

[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

Why did no one think of this before?

- Concretely, why wasn’t contextual pre-training popular before 2018 with ELMo?
- Good results on pre-training is >1,000x to 100,000 more expensive than supervised training.

What Happened After BERT?

- RoBERTa (Liu et al., 2019)
  - Drops the next sentence prediction loss!
  - Trained on 10x data (the original BERT was actually under-trained)
  - Much stronger performance than BERT (e.g., 94.6 vs 90.9 on SQuAD)
  - Still one of the most popular models to date

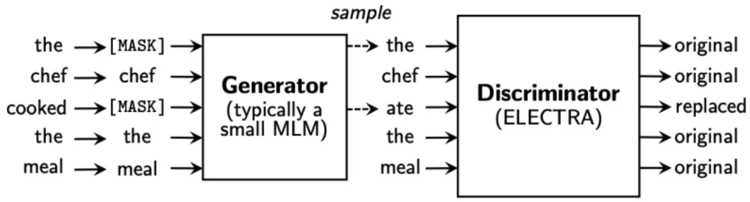
Model	**data**	**bsz**	**steps**	**SQuAD** (v1.1/2.0)	**MNLI-m**	**SST-2**
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7

What Happened After BERT?

- RoBERTa (Liu et al., 2019)
  - Drops the next sentence prediction loss!
  - Trained on 10x data (the original BERT was actually under-trained)
  - Much stronger performance than BERT (e.g., 94.6 vs 90.9 on SQuAD)
  - Still one of the most popular models to date

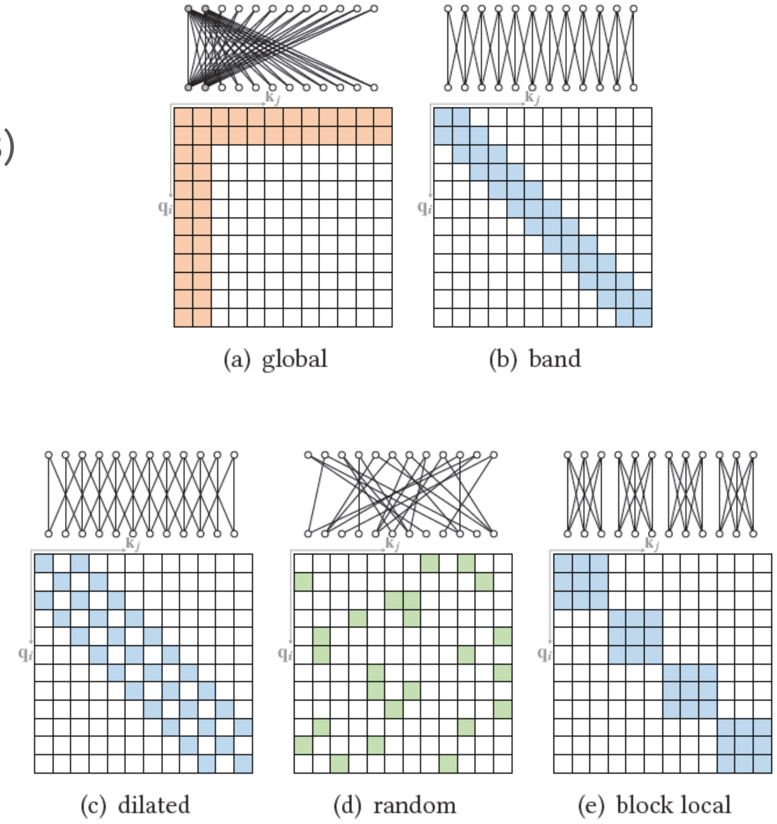
- ALBERT (Lan et al., 2020)
  - Increasing model sizes by sharing model parameters across layers
  - Less storage, much stronger performance but runs slower..

- ELECTRA (Clark et al., 2020)
  - Two models generator and discriminator
  - It provides a more efficient training method



What Happened After BERT?

- Models that handle long contexts (512 tokens)
  - Longformer, Big Bird, ...
- Multilingual BERT
  - Trained single model on 104 languages from Wikipedia. Shared 110k WordPiece vocabulary
- BERT extended to different domains
  - SciBERT, BioBERT, FinBERT, ClinicalBERT, ...
- Making BERT smaller to use
  - DistillBERT, TinyBERT, ...



Text generation using BERT

BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model

Alex Wang
New York University
alexwang@nyu.edu

Kyunghyun Cho
New York University
Facebook AI Research
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Mask-Predict: Parallel Decoding of Conditional Masked Language Models

Marjan Ghazvininejad* Omer Levy* Yinhan Liu* Luke Zettlemoyer
Facebook AI Research
Seattle, WA

Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis--Hastings

Kartik Goyal, Chris Dyer, Taylor Berg-Kirkpatrick

src	Der Abzug der franzsischen Kampftruppen wurde am 20. November abgeschlossen .
$t = 0$	The departure of the French combat completed completed on 20 November .
$t = 1$	The departure of French combat troops was completed on 20 November .
$t = 2$	The withdrawal of French combat troops was completed on November 20th .

Leveraging Pre-trained Checkpoints for Sequence Generation Tasks

Sascha Rothe, Shashi Narayan, Aliaksei Severyn

Summary Thus Far

- BERT and the family
- An encoder; Transformer-based networks trained on massive piles of data.
- Incredible for learning contextualized embeddings of words
- It's very useful to pre-train a large unsupervised/self-supervised LM then fine-tune on your particular task (replace the top layer, so that it can work)
- However, they were not designed to generate text.