Georgetown University

ANLY-580
Natural Language Processing
Transformer Models

Fall 2022
Lecture 11
Instructor: Chris Larson

Nov 02, 2022

The transformer architecture

Definition: A neural network architecture that uses self-attention (between layers) as a primary means of expressing relationships between the random variables in the sequence.

The first transformer (2017)

Original paper: *Attention is all you need* (2017)

30K citations

First neural language model to process sequence without use of recurrent connections or convolutions.

Reached SOTA BLEU / ROUGE scores

NMT w/encoder-decoder architecture

Uses position encodings

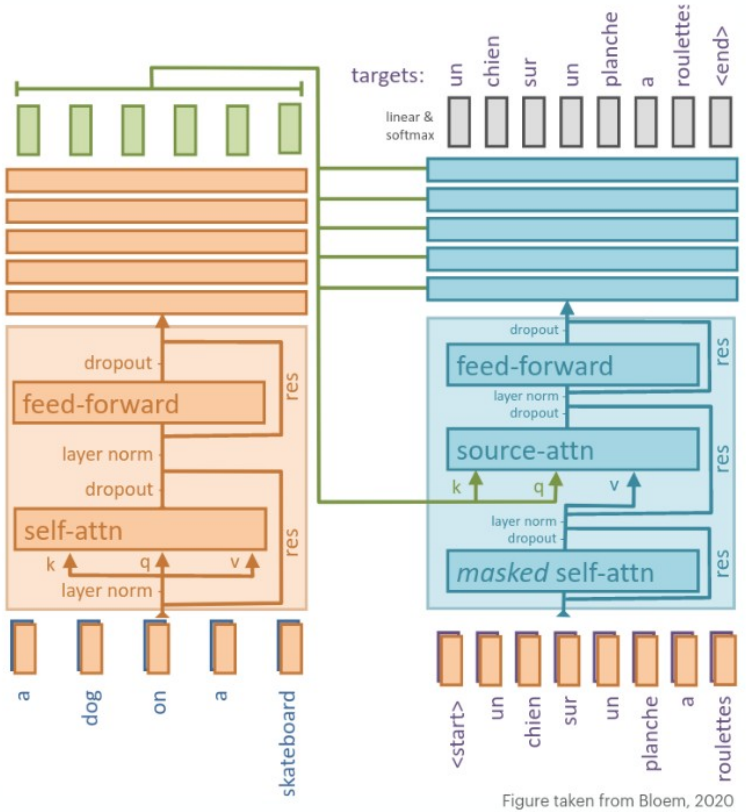


Figure taken from Bloem, 2020

Vaswani et al., *Attention is All You Need*, 2017

Attention and sequential structure

Position Embeddings

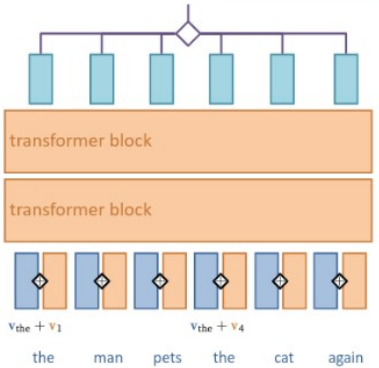
Conceptually simple

Easy to implement

Adds set of learnable parameters

Maximum context length at test time limited to max sequence length in training set

Embedding quality diminishes (in theory) with *t*

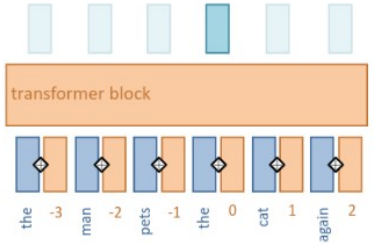


Relative position embeddings

Captures the property of absolute position invariance which is desirable

Like absolute position embeddings, limited in context length at inference

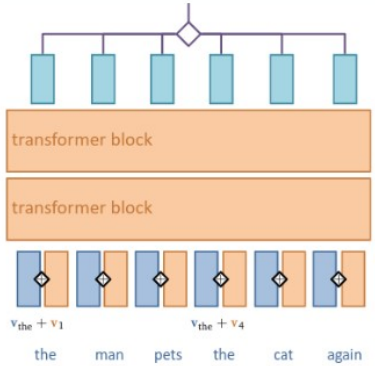
Conceptually and practically more complex; on the surface this results in (T-1)^2 different inputs because each input will have T-1 separate representations (there is a hack that turns this into 2T-1)



Position encodings

Captures the property of absolute position invariance which is desirable

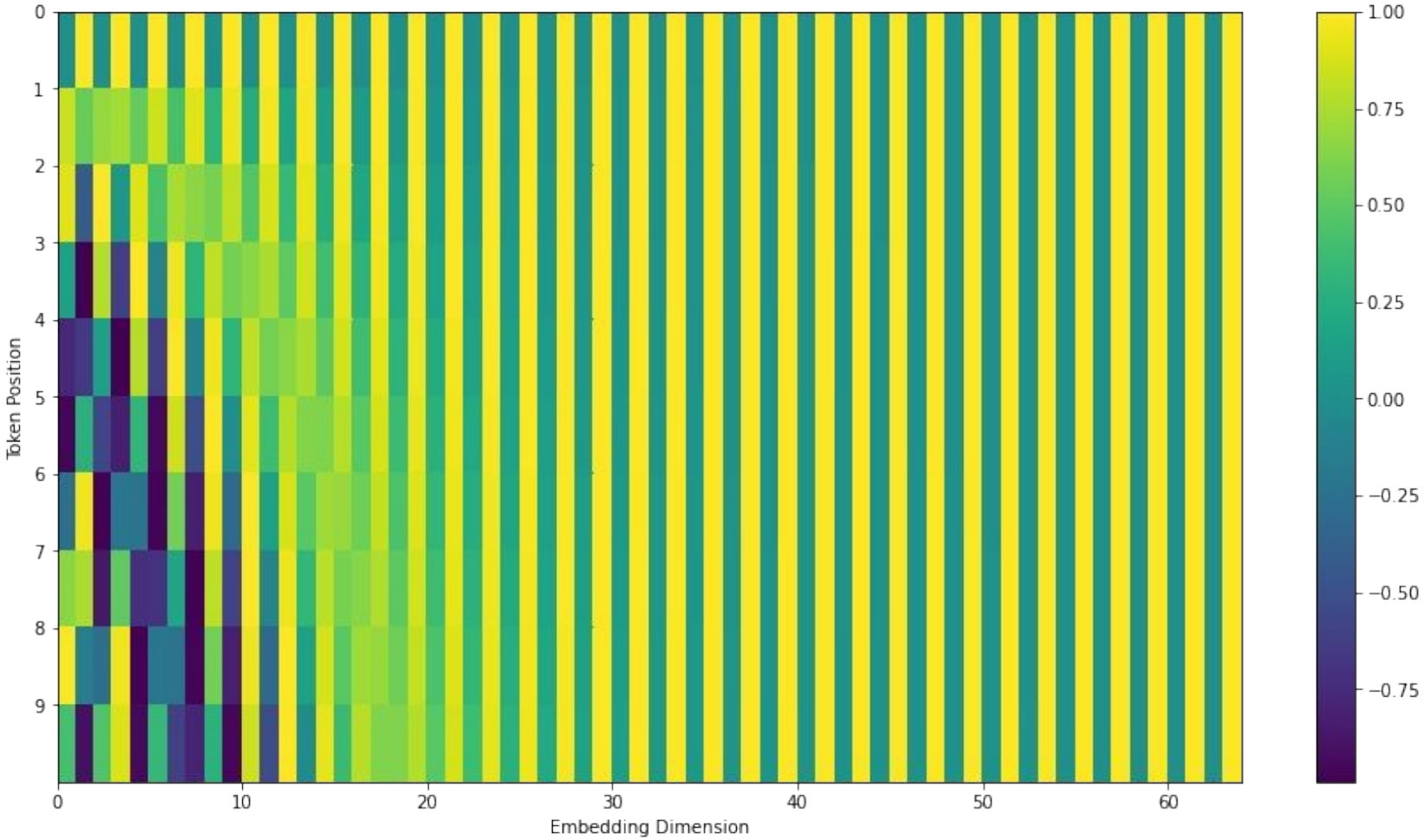
Conceptually and practically more complex; on the surface this results in (T-1)^2 different inputs because each input will have T-1 separate representations (there is a hack that turns this into 2T-1)



Taken from The Annotated Transformer

Embedding illustrations taken from Bloem (2020)

Attention and sequential structure



BERT (2018)

BERT

The most popular/used/studied Transformer model to date

30K citations [1]

Trained on Wikipedia and Book Corpus (~10K books)

Wordpiece embedding [2]

Non-causal, uses masked LM procedure for pretraining

340 M parameters in total

LM prediction head: 4 ReLU (D=4096)

LM training took 4 days on 64 TPU cores

$$D = 1024$$

$$N_H = 16$$

$$N_B = 24$$

$$T_{max} = 1024$$

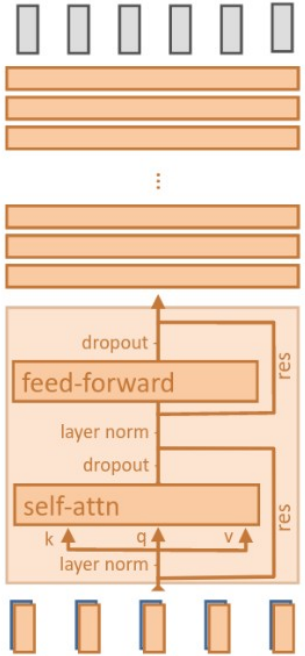
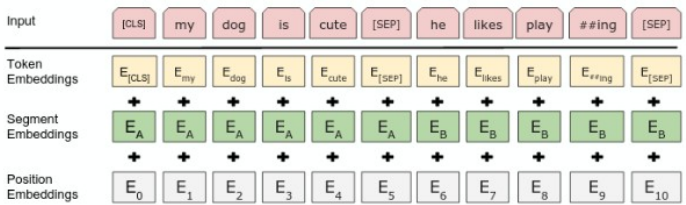


Figure taken from Bloem, 2020



BERT Input Features [1]

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

Ablation Study [1]

[1] Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018)*, [2] Wu et al., *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (2016)*

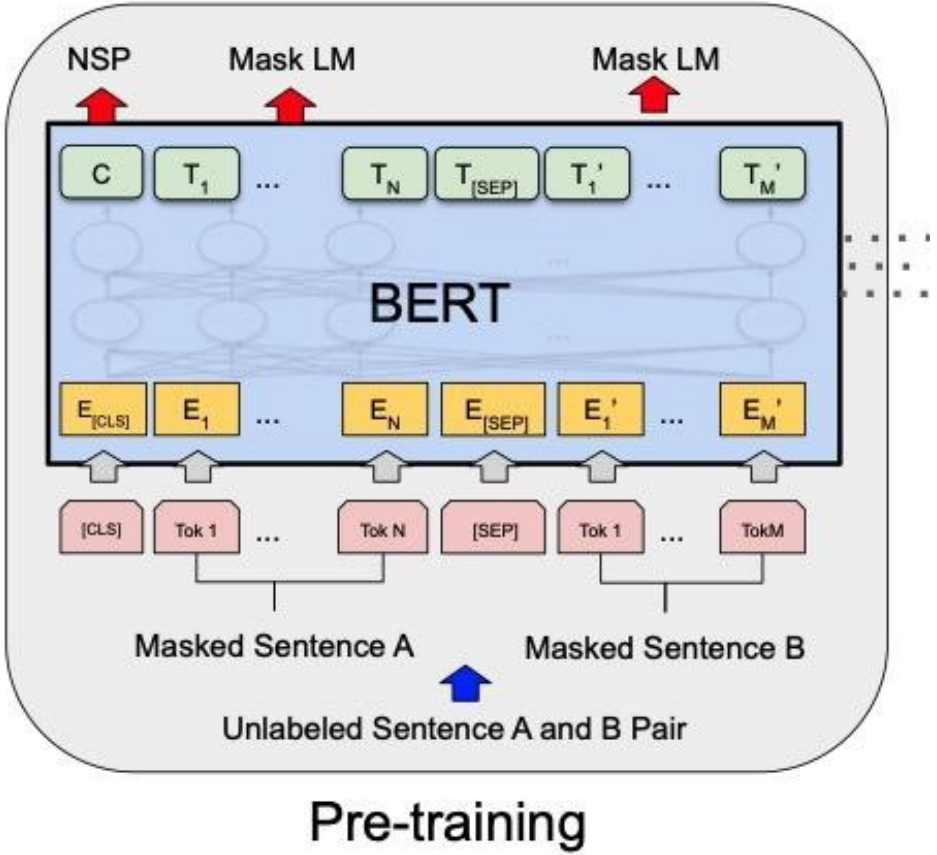
BERT

Key innovation:
self-supervised pre-training
… plus previous innovations:

- self-attention/transformers
- ELMo language model
- lots of GPUs
- …

Why is self-supervised training a game-changer?

1. Labels are provided by text
2. Generic language representations amenable to fine-tuning

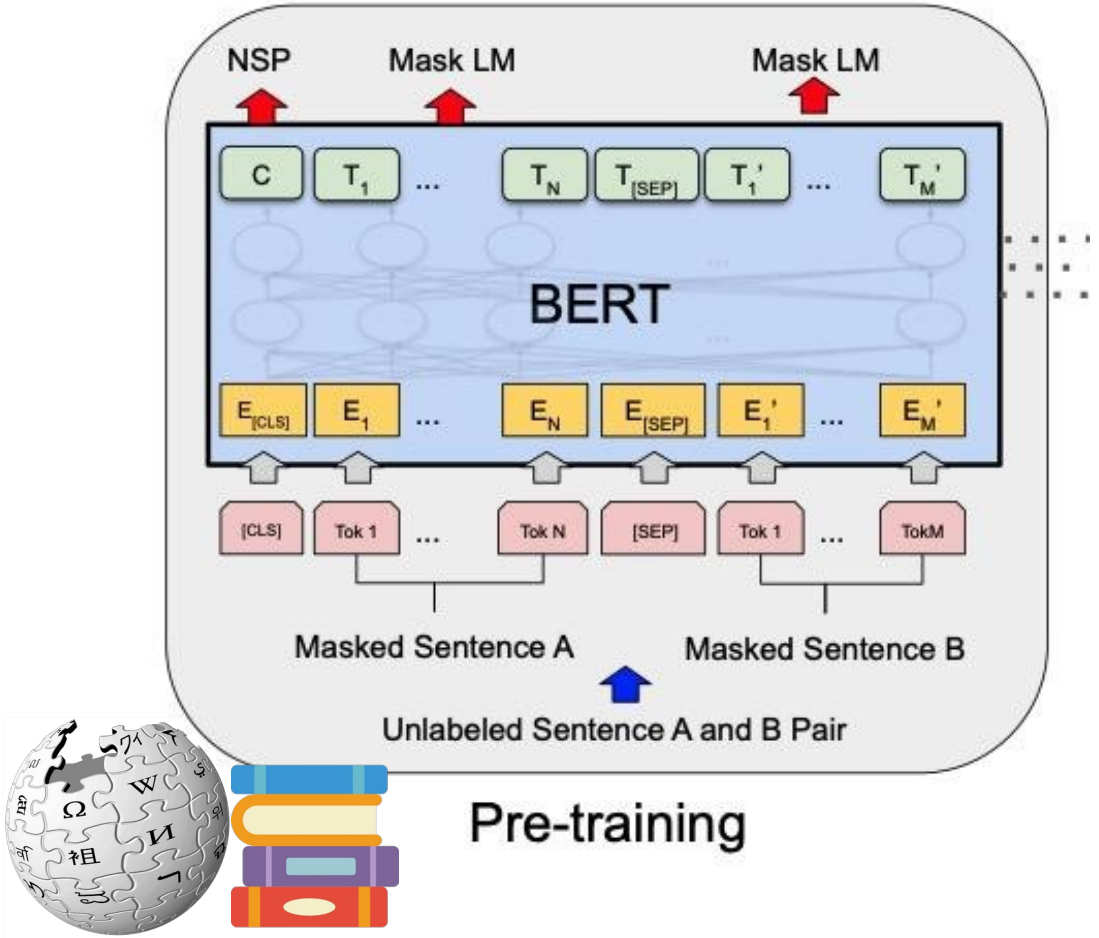


Pre-training

Training BERT

Two language modeling tasks:

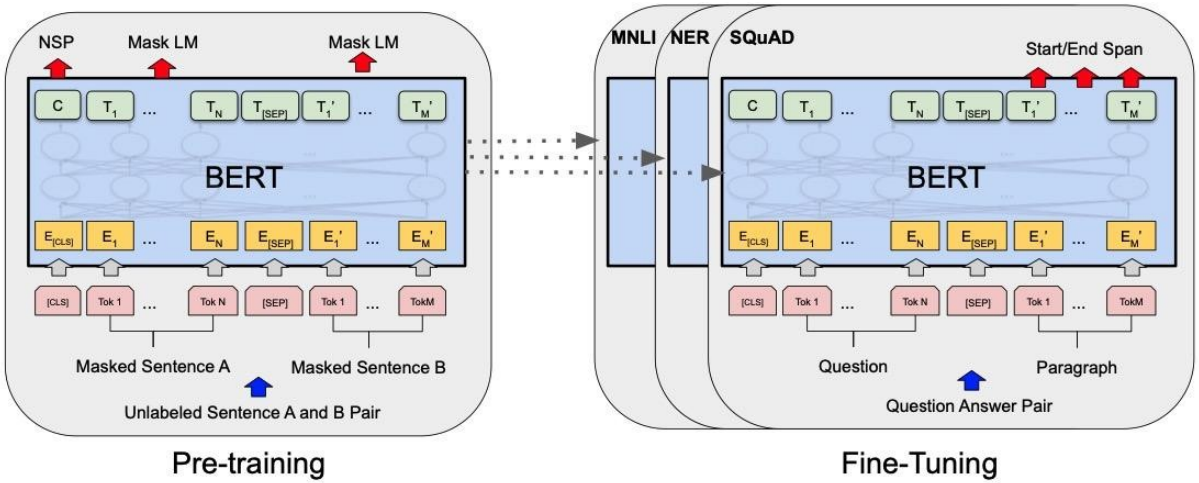
1. Predict token replaced by [MASK] (15% of inputs randomly masked)
2. Predict the next sentence as you feed in two sentences at a time (Randomly flip sentence order 50% of the time)



BERT Applications

Modern Recipe

1. Start with a pre-trained model
2. Fine-tune the model to your particular task



GPT (2018-present)

GPT 2

Autoregressive model (i.e., causal, only backwards connections)

Uses BytePair tokens

Trained on WebText dataset (~8M high quality web pages)

Positional embeddings

1.5B parameters

$D = 768$

$N_H = 12$

$N_B = 48$

$T_{max} = 2048$

Brown et al., *Language Models are Unsupervised Multitask Learners* (2018)

GPT 3

" " except 175B parameters

LM pretraining on Common Crawl dataset (~1T words)

$D = 12288$

$N_H = 96$

$N_B = 96$

$T_{max} = 2048$

Brown et al., *Language Models are Few-Shot Learners* (2020)

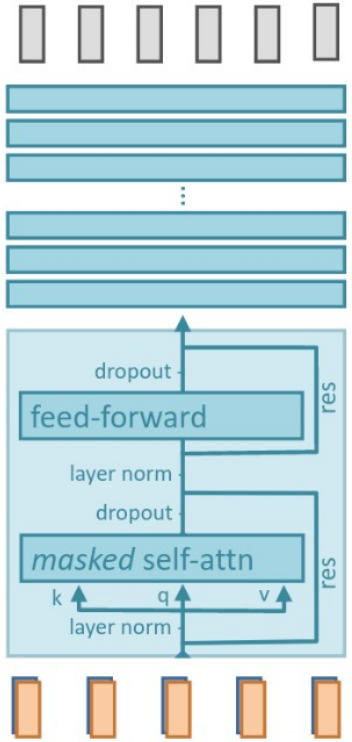


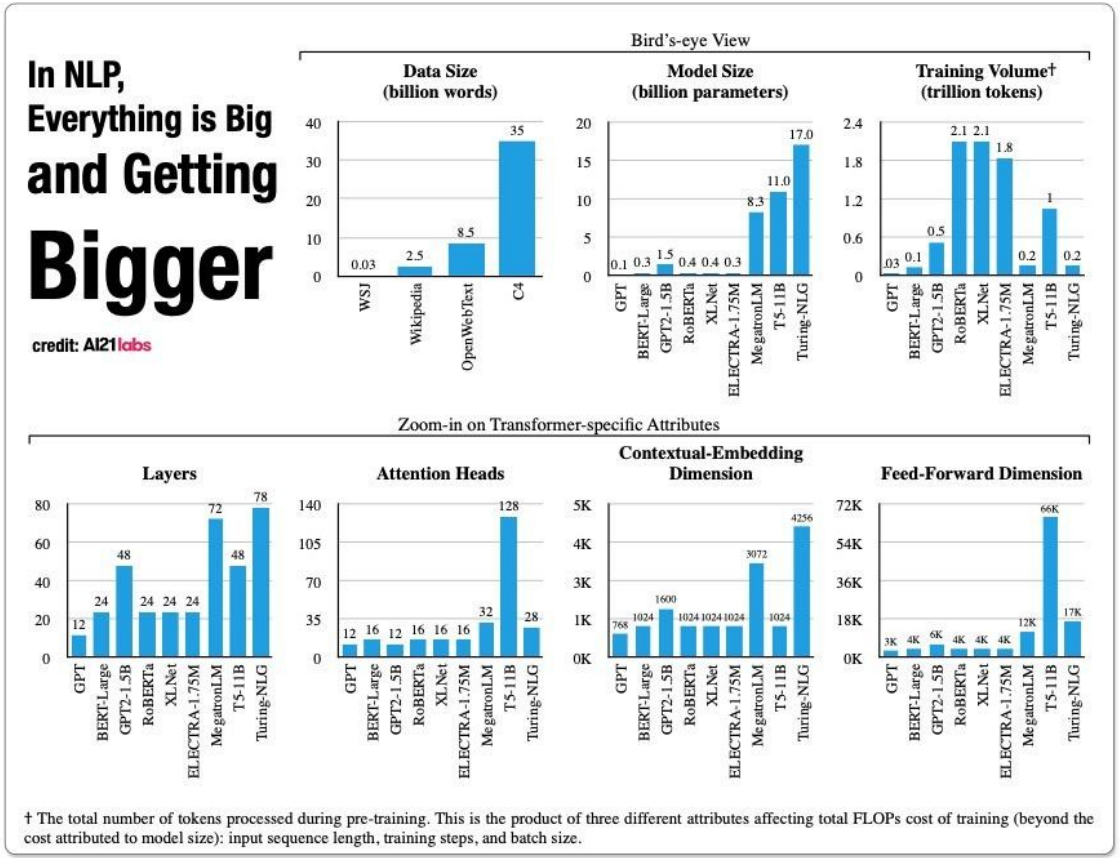
Figure taken from Bloem, 2020

Text generation with GPT-3

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Brown et al., *Language Models are Few-Shot Learners* (2020)

Attention-based models are optimized for parallel computation, This generally means they can be made larger



Garbage-in garbage-out still applies!