



Transformers (no not that kind...)

Areeb Gani, Michael Ilie, Vijay Shanmugam

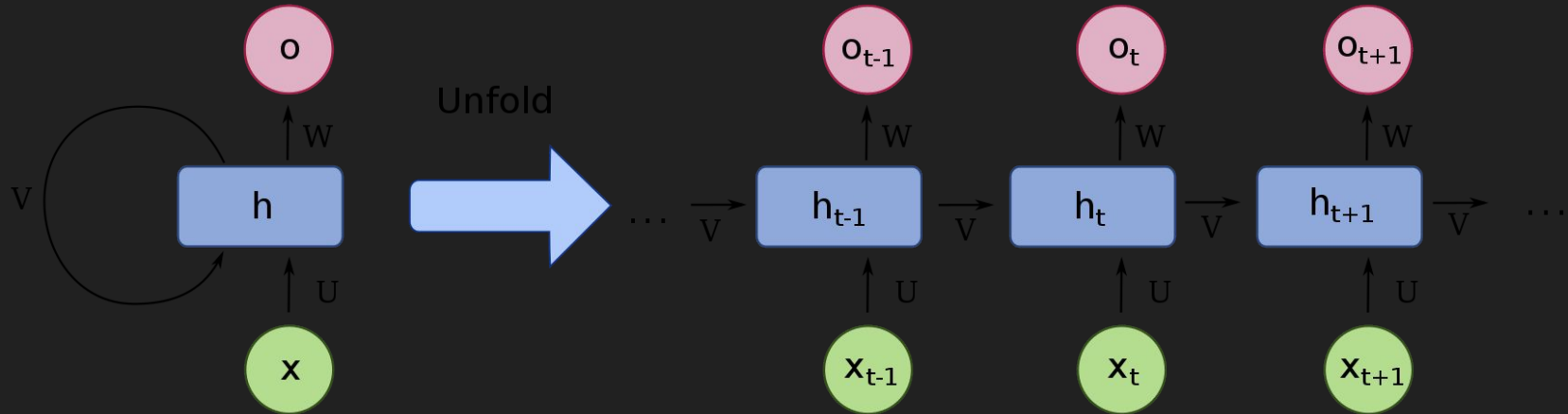
Welcome!



ml.mbhs.edu

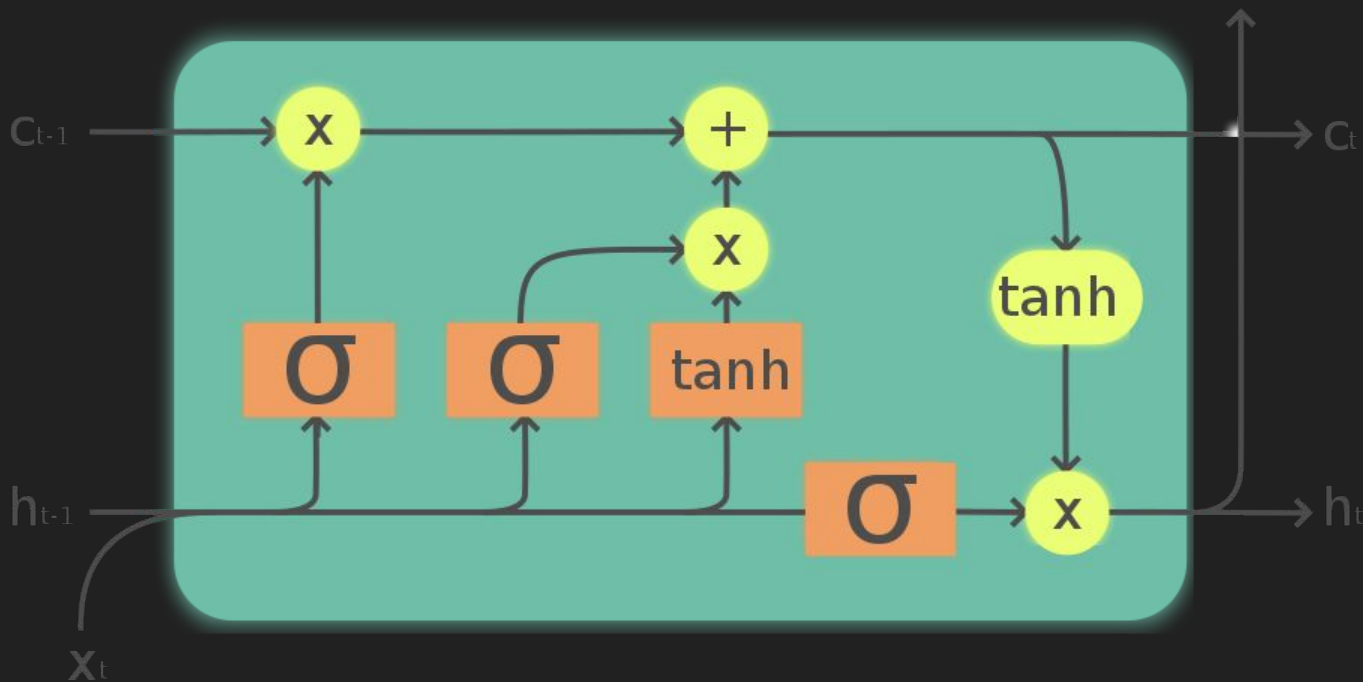
Recurrent neural network (1986)

- Language Translation
- Text Summarization
- Next Sentence Prediction



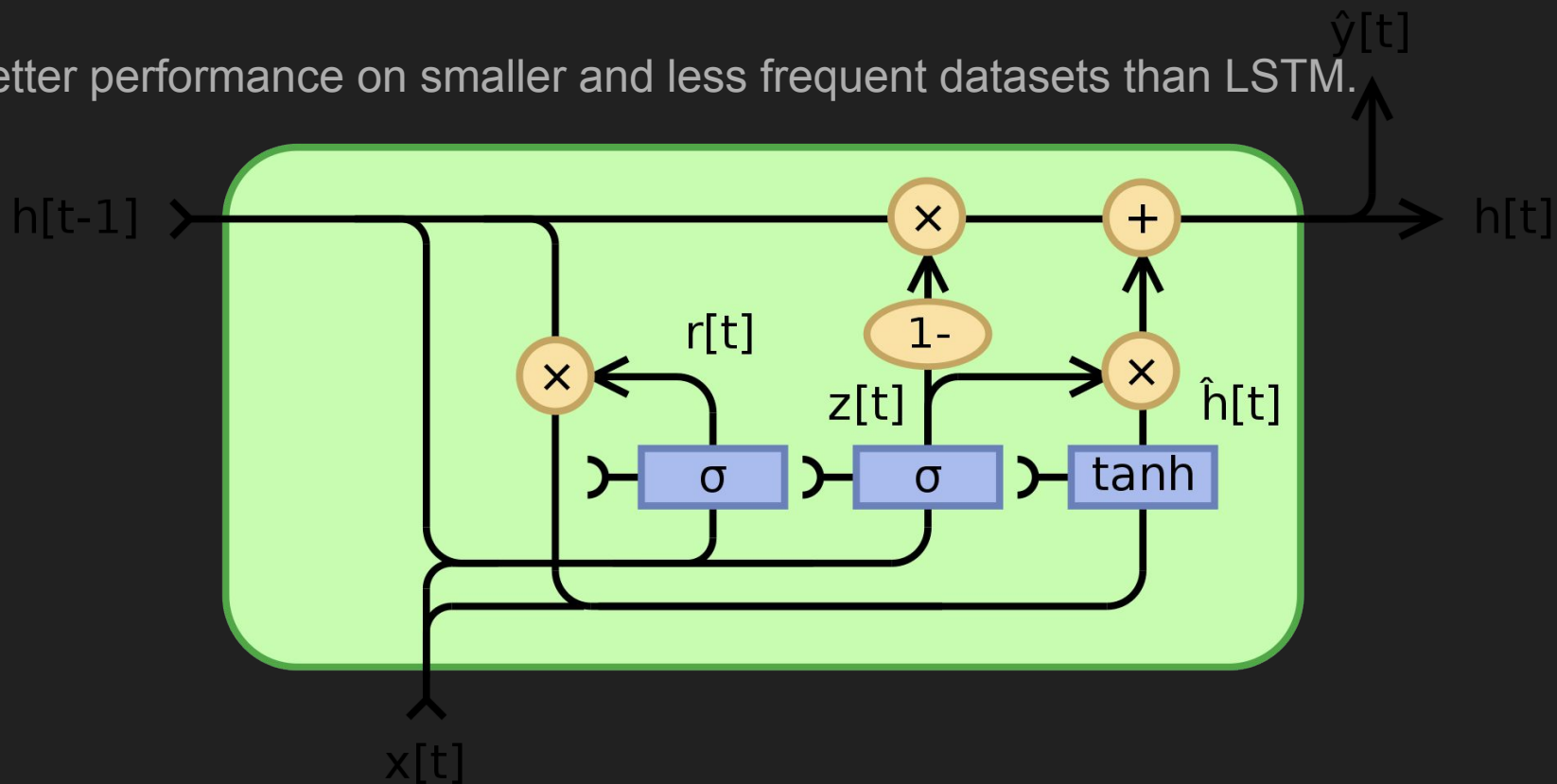
LSTM (1997)

RNNs using LSTM units partially solve the vanishing gradient problem



GRU (2014)

Better performance on smaller and less frequent datasets than LSTM.



Attention is all you need (2017)

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

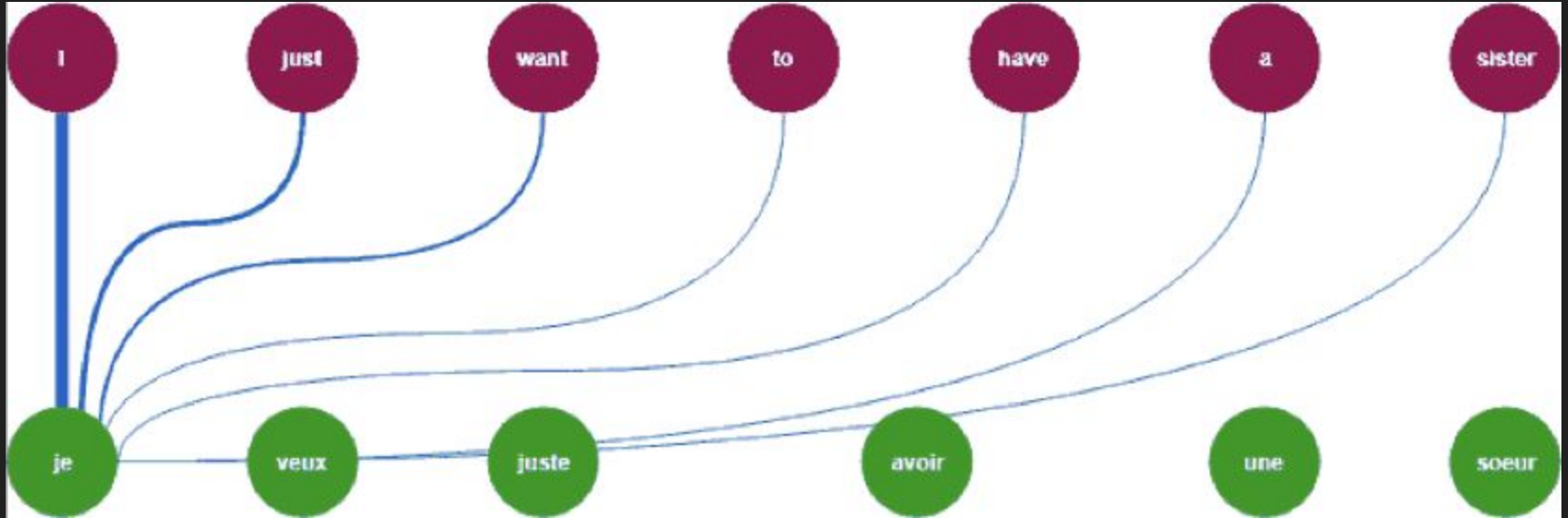
Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Attention in Neural networks

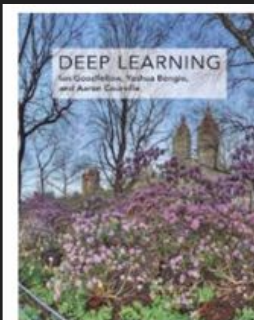


Example : Searching for “Equivariance” in a book

could require the same number of floating-point operations to
still need to contain $2 \times 319 \times 280 = 178,640$ entries. Con
efficient way of describing transformations that apply the sam
a small local region across the entire input. Photo credit

ear function for detecting edges in an image.

of convolution, the particular form of parameter sharing ca
property called **equivariance** to translation. To say a fun
as that if the input changes, the output changes in the sa
action $f(x)$ is equivariant to a function g if $f(g(x)) = g(f$
lution, if we let g be any function that translates the inp
the convolution function is equivariant to g . For exampl
ing image brightness at integer coordinates. Let g be a f
ge function to another image function, such that $I' = g(I$
with $P(x, y) = I(x - 1, y)$. This shifts every pixel of I one
apply this transformation to I , then apply convolution, th
as if we applied convolution to I' , then applied the transfo



Chapter 9
Convolutional Networks

Convolutional networks [Lec15a, 1985] also known as convolutional neural networks, or CNNs, are a specialized kind of neural network for processing data that has a known grid-like topology. Examples include time-series data, which can be thought of as a 1-D grid taking samples at regular time intervals, and image data, which can be thought of as a 2-D grid of pixels. Convolutional networks have been tremendously successful in practical applications. The name “convolutional neural network” indicates that the network employs a mathematical operation called convolution. Convolution is a specialized kind of inner-product. Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.

In this chapter, we first describe what convolution is. Next, we explain the motivation behind using convolution in a neural network. We then describe an operation called pooling, which almost all convolutional networks require. Finally, the operations used in a convolutional neural network show us an important property in the definition of convolution as used in other fields, such as engineering or time-series analysis. We describe neural networks in the standard function that are widely used in practice for neural networks. We also show how convolution can be applied to some kinds of data with different notions of dimension. We then discuss means of making convolution more efficient. Convolutional networks stand out as an example of neuroscience principles informing deep learning. We discuss three neuroscientific principles that motivate this perspective about the role convolutional networks have played in the history of deep learning. One goal of this chapter is to show the architecture of a typical convolutional network. The goal of this chapter is to describe the kinds of tools that convolutional networks provide, with chapter 10 describing general guidelines



1 Week



1 minute

RNN vs LSTM vs Attention

Recurrent Neural Networks has a short reference window

As aliens entered our planet

and began to colonize earth a certain group of extraterrestrials ...

RNN vs LSTM vs Attention

GRU's and LSTM's have a longer reference window than RNN's

As aliens entered our planet

and began to colonize earth a certain group of extraterrestrials ...

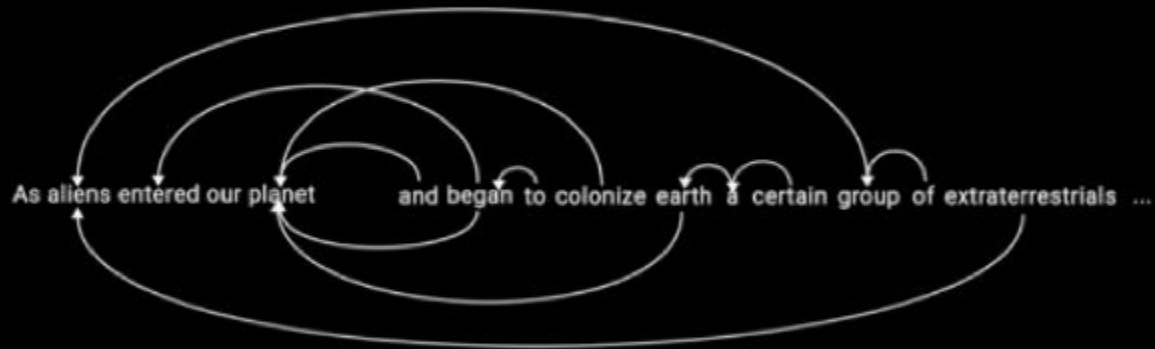
RNN vs LSTM vs Attention

Attention Mechanism has an infinite reference window

As aliens entered our planet and began to colonize earth a certain group of extraterrestrials ...



RNN vs LSTM vs **Attention**



Attention example in images

A **bodybuilder** holding a dumbbell



Microsoft Attention GANs



Figure 1. Example results of the proposed AttnGAN. The first row

Transformers

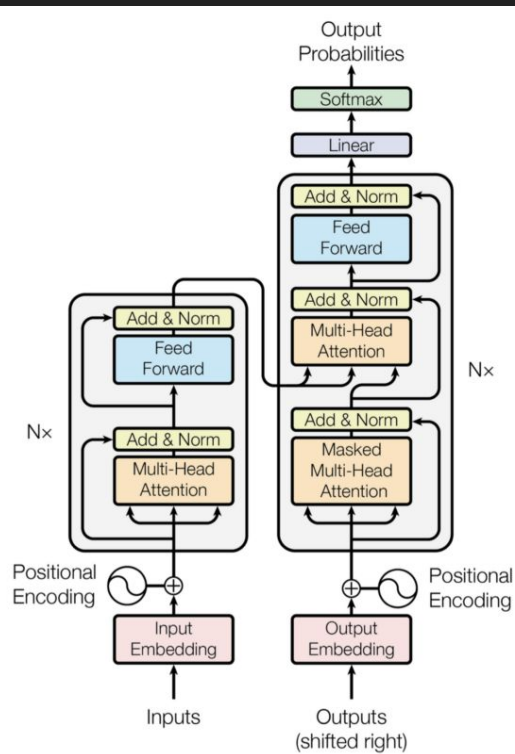
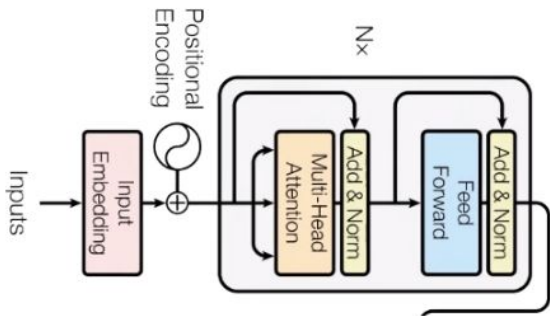


Figure 1: The Transformer - model architecture.

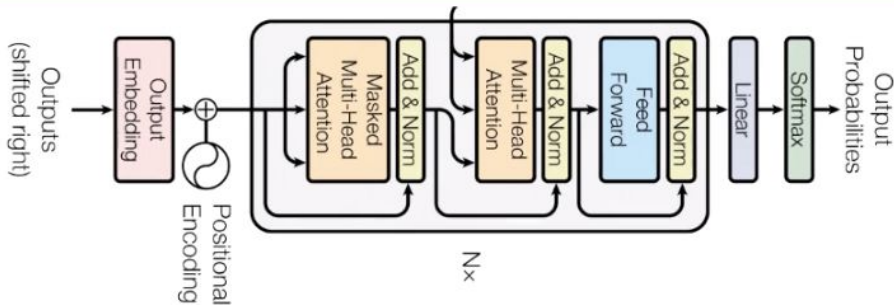
Encoder and Decoder

Transformer Flow



What is English? What is context?

What is language!



How to map English words to French words?

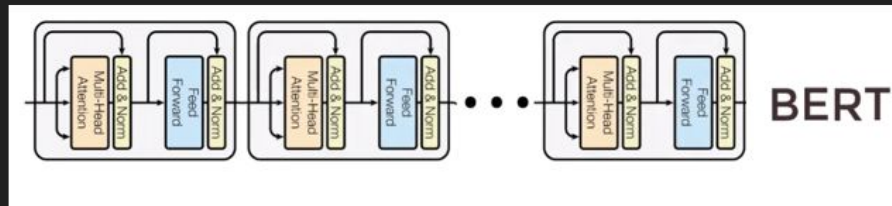
What is language!



BERT and GPT

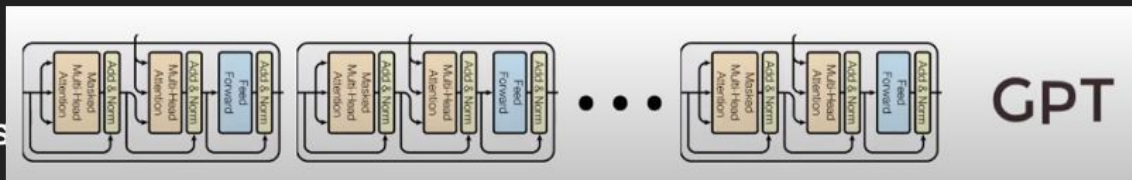
BERT (2018) by Google

A stack of Encoders



GPT (2018) by OpenAI

A stack of Decoders



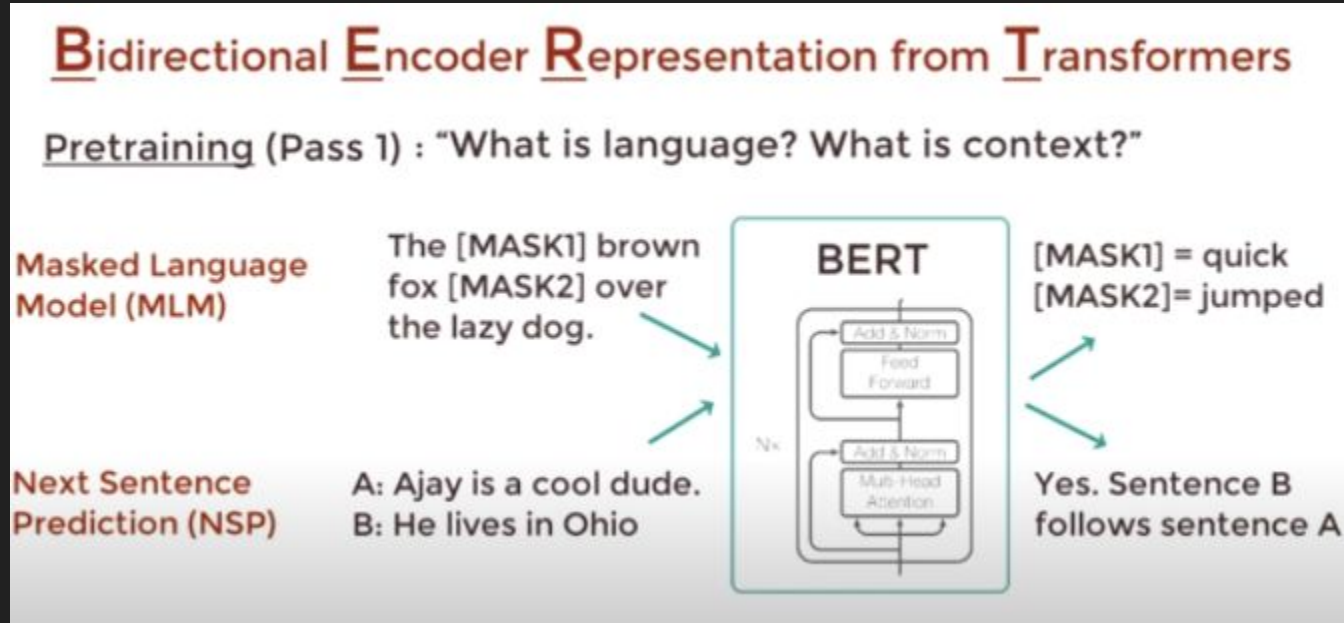
BERT(Bidirectional Encoder Representations from Transformers)

state-of-the-art performance on a number of NLU tasks

- GLUE (General Language Understanding Evaluation)
- SQuAD (Stanford Question Answering Dataset)
- SWAG (Situations With Adversarial Generations)

BERT pre-training procedure

1. Masked Language Model
2. Next Sentence Prediction



GPT-1(Generative Pre-trained Transformer)

- GPT is a "transformer" model, which uses "attention" in place of previous recurrence- and convolution-based architectures.
- It showed how a generative model of language is able to acquire world knowledge and process long-range dependencies by pre-training on a diverse corpus with long stretches of contiguous text.

GPT-2 (1.5 Billion Parameters)

Dataset for pretraining : 40 GB of text
required tens of petaflop/s-days*

translates text, answers questions, summarizes passages, and generates text
output on a level that, while sometimes indistinguishable from that of humans

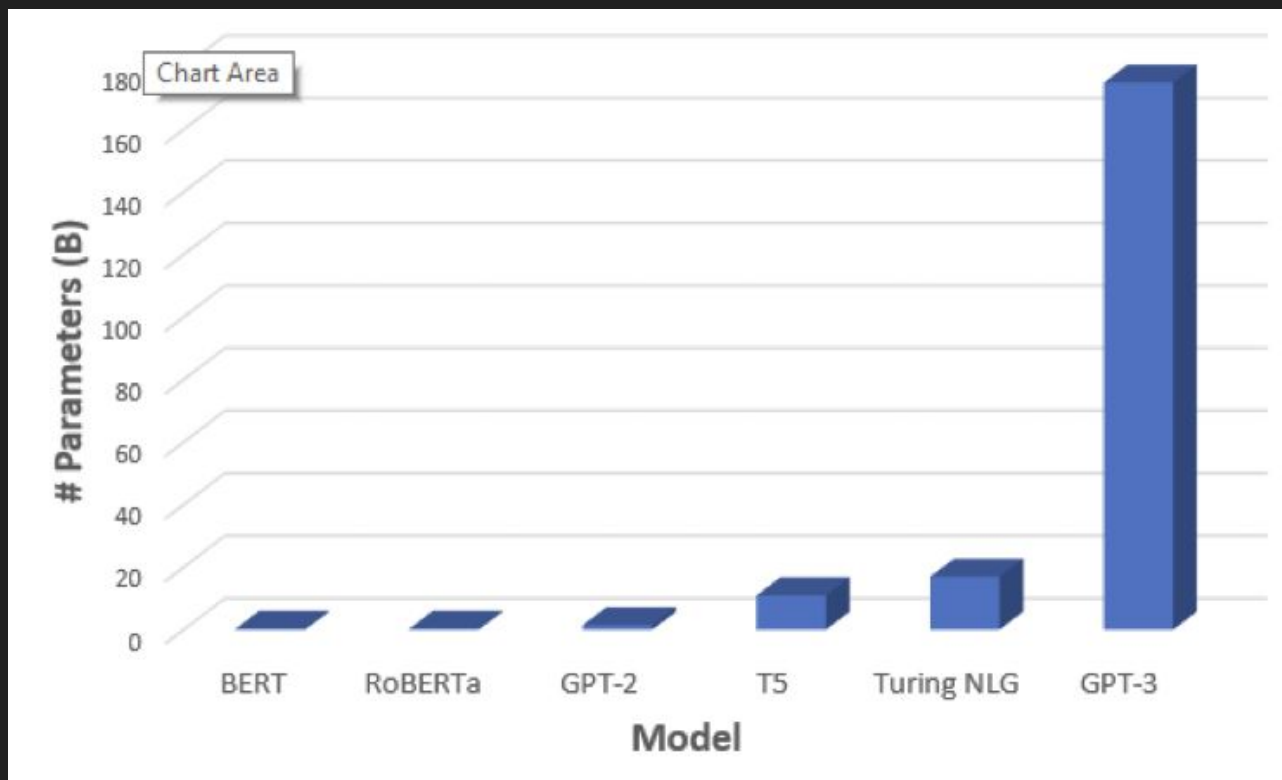
*One petaflop/s-day is approximately equal to 10^{20} neural net operations

GPT-3 (175 Billion Parameters)

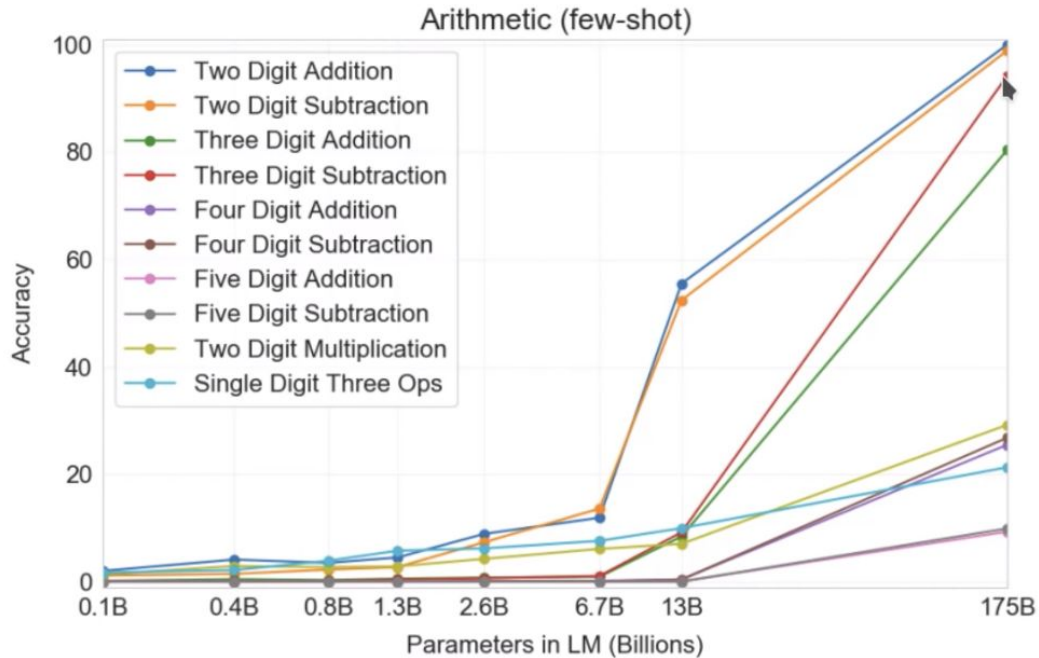
Dataset for pretraining : 570 GB of text
required several thousand petaflop/s-days*

*One petaflop/s-day is approximately equal to 1020 neural net operations

Comparison of sizes



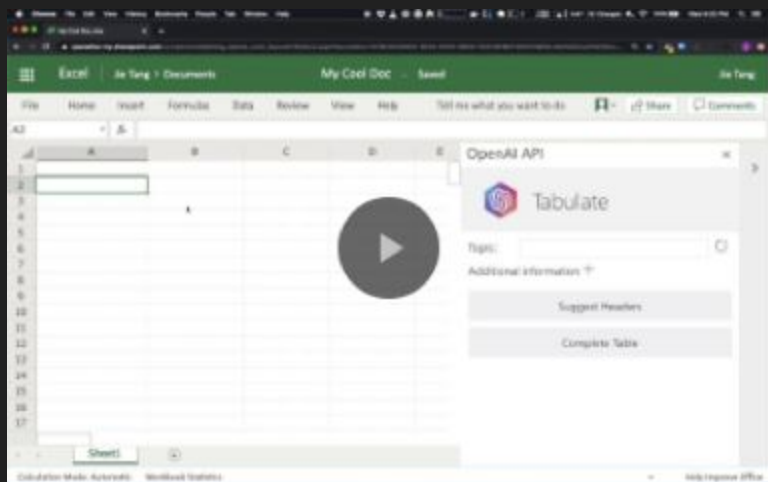
Size matters!



Some Practice Tutorials

- https://www.tensorflow.org/text/tutorials/text_generation (Tensorflow)
- https://www.tensorflow.org/text/tutorials/text_classification_rnn (Tensorflow)
- https://pytorch.org/tutorials/intermediate/char_rnn_classification_tutorial.html (Pytorch)
- https://pytorch.org/tutorials/beginner/nlp/sequence_models_tutorial.html (LSTM in Pytorch)

GPT-3 Demos



Transformers Drawbacks

- Very large models.
 - Memory and compute intensive to train
- Relatively young class of models
 - so we know less about them
- Might be worse for hierarchical data (Tran et al, ACL 2018)

Challenges

1. Transformer complexity
2. Longer sequences

Shrinking Transformer

Transformers are becoming both more accurate and larger (t5 has 11 billion parameters)

But there are ways to make them smaller without hurt performance:

1. Quantization
2. Distillation
3. Pruning
4. More specialized models

Quantization

Reduced number of bit needed to store the trained parameters in model

Convert 32 bit floating point to 8 bit integer

Problem: usually hardware dependant



Distillation

A new model is trained to predict the weights of one or more layers of the larger model

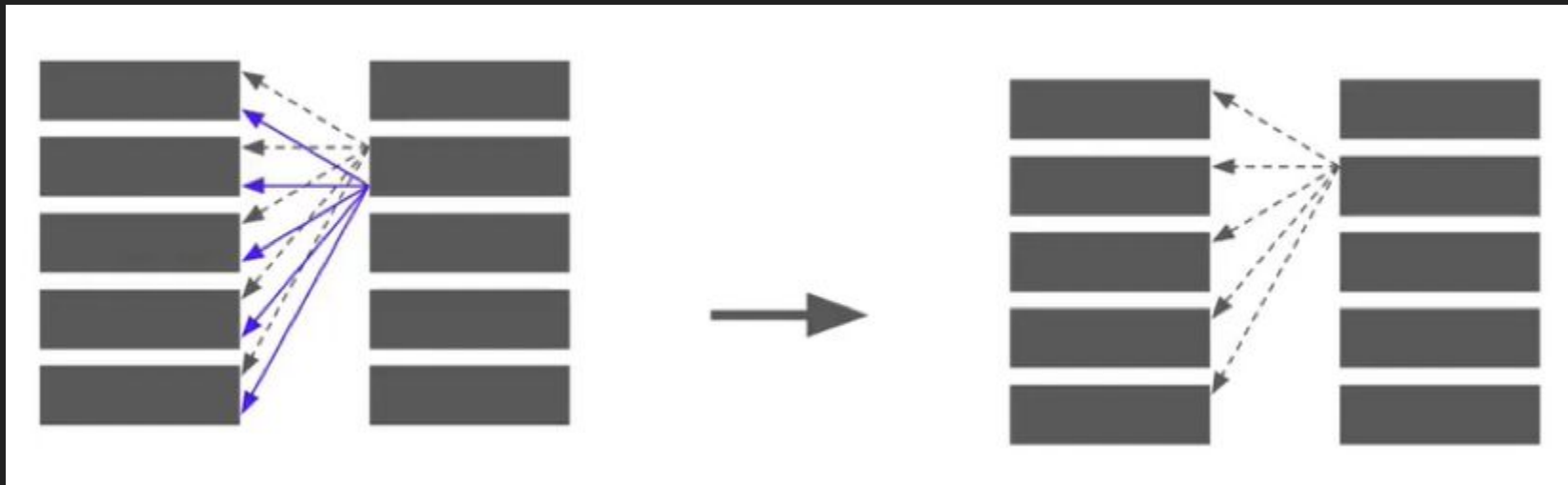
Up to 100x smaller and 15x faster

Problem: need more setup



Pruning

Remove attention heads based on how useful they are for a specific task
Up to 80% the heads of trained transformer heads can removed without significantly reducing accuracy



More specialized models

Train a special smaller model

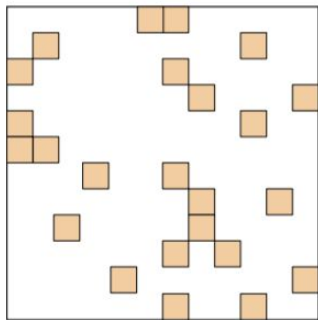
Really large nlp models (like bert & gpt) tend to be open domain



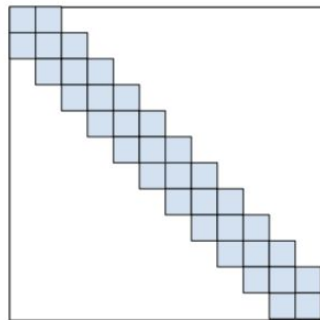
Why don't these methods hurt performance?

1. Really large transformers are bigger than they need to be for some tasks
2. There is a lot of redundancy in these models

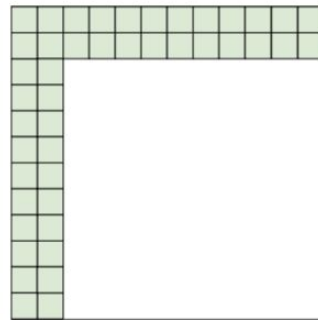
Bigbird



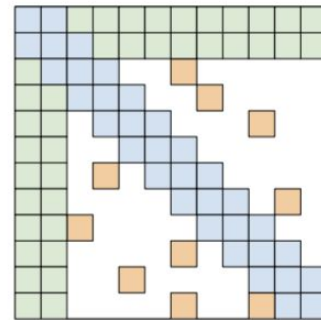
(a) Random attention



(b) Window attention

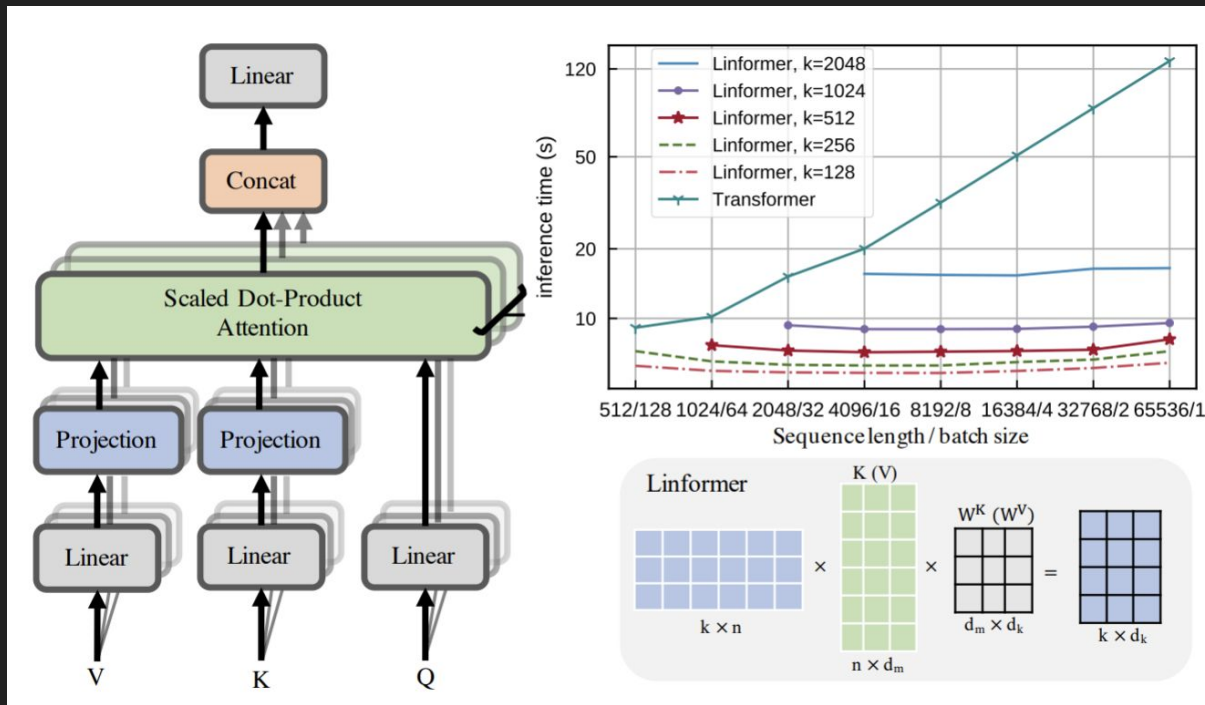


(c) Global Attention

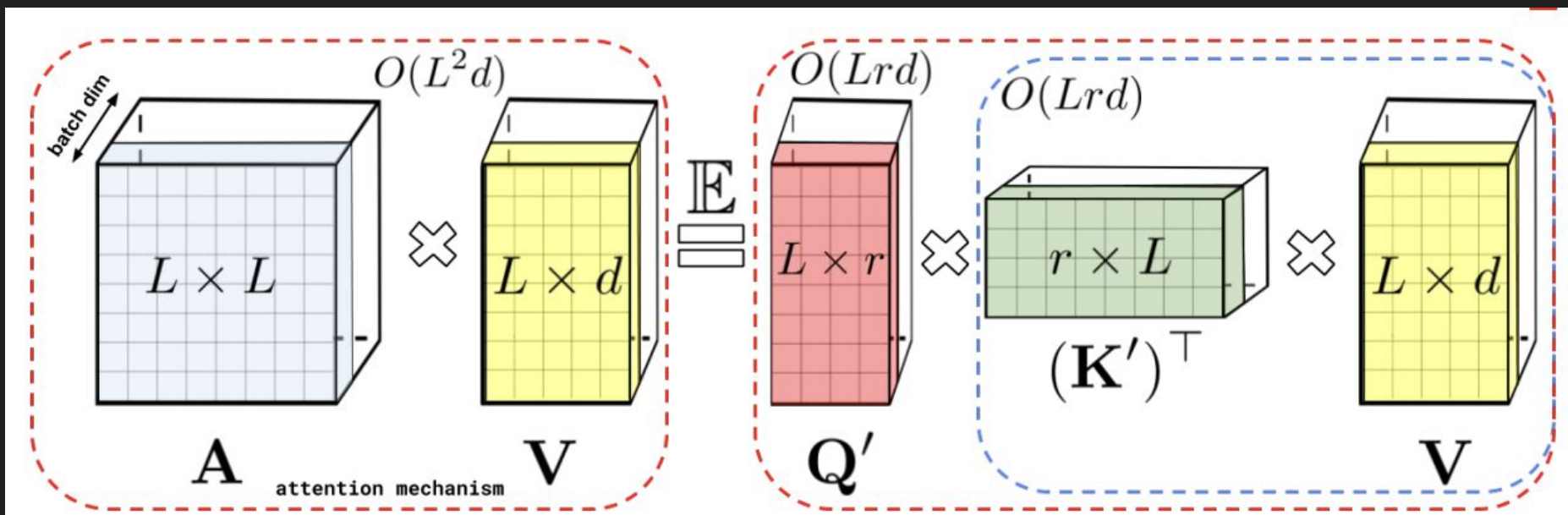


(d) BIGBIRD

Linformer



Performer



Join Our Groups

- Sign up for Discord (<https://discord.gg/3Z5YuPqt>)
- Join Deepnote (<https://deepnote.com/join-team?token=af3af0284bc8497>)
- Fill out our form (<https://forms.gle/Fr31aFLWx8cHdtTY8>)
 - Join mailing list + Github organization
- Next week: Text Analysis/NLP (Natural Language Processing)