


A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light greenish-blue. They are positioned diagonally, with the blue one partially covering the green one.

# Introduction to Transformers

Gholamreza Dar, Amirhassan Amirmahani

# YouTube Playlist



**Transformer and Language Models**

20 videos • 38 views • Updated today


Public ▾

✂ ↗ ⋮

Very good Introductory videos for learning about Language Models, Transformers, Attention and etc.  
Hand-picked and sorted by GHD and H08.  
Feb 2021


**GHD** by Gholamreza Dar and 1 other

≡ SORT




**AI Language Models & Transformers - Computerphile**  
Computerphile  
Added by [Gholamreza Dar](#) ⋮

WATCHED 20:39




**Illustrated Guide to Transformers Neural Network: A step by step explanation**  
The A.I. Hacker - Michael Phil  
Added by [Gholamreza Dar](#) ⋮

WATCHED 15:01




**GPT-2: Why Didn't They Release It? - Computerphile**  
Computerphile  
Added by [Gholamreza Dar](#) ⋮

WATCHED 9:11




**GPT3: An Even Bigger Language Model - Computerphile**  
Computerphile  
Added by [Gholamreza Dar](#) ⋮

WATCHED 25:57




**Attention in Neural Networks**  
CodeEmporium  
Added by [Gholamreza Dar](#) ⋮

WATCHED 11:19




**BERT Neural Network - EXPLAINED!**  
CodeEmporium  
Added by [Gholamreza Dar](#) ⋮

WATCHED 11:37




**LSTM is dead. Long Live Transformers!**  
Seattle Applied Deep Learning  
Added by [Gholamreza Dar](#) ⋮

WATCHED 28:48



**Transformer Neural Networks - EXPLAINED! (Attention is all you need)**  
CodeEmporium  
Added by [Gholamreza Dar](#) ⋮

WATCHED 13:05

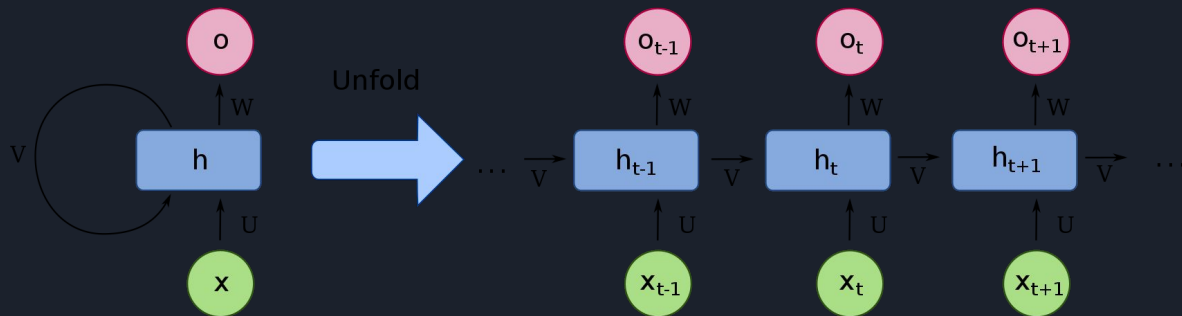


**NLP for Developers: Transformers | Rasa**  
Rasa  
Added by [Amir Hassan Amir Mahani](#) ⋮

WATCHED 5:34

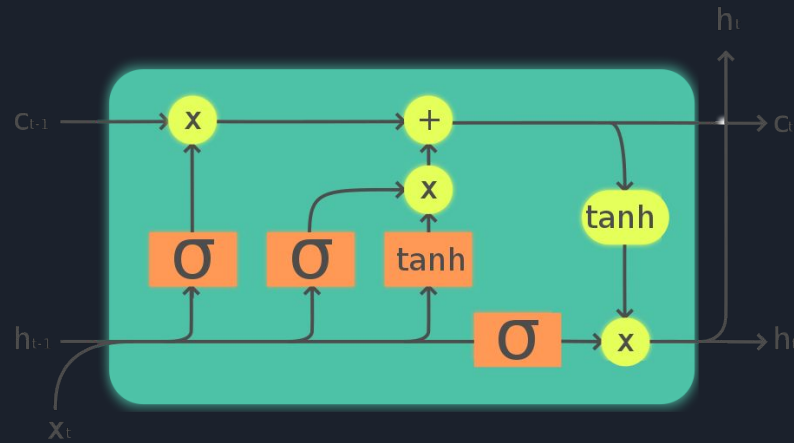
# Recurrent neural network (1986)

- Language Translation
- Text Summarization
- Next Sentence Prediction



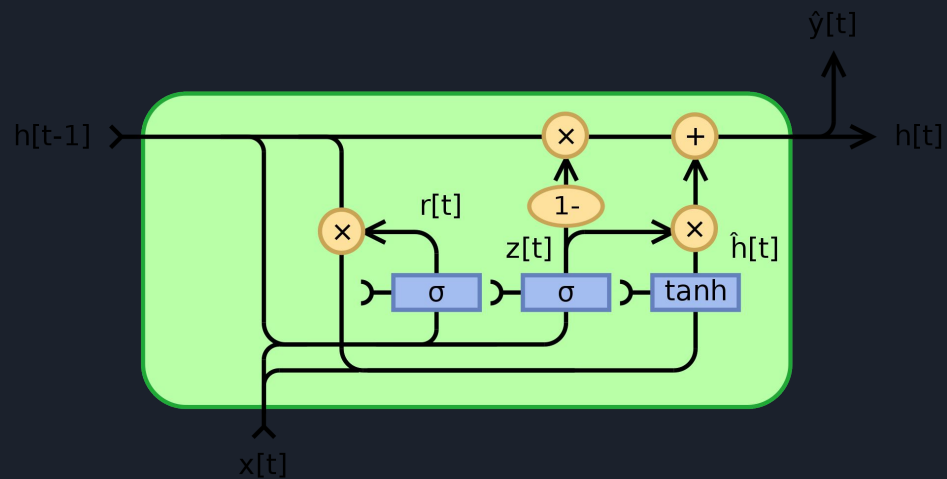
# LSTM (1997)

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) Cited by 18060

---

## 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

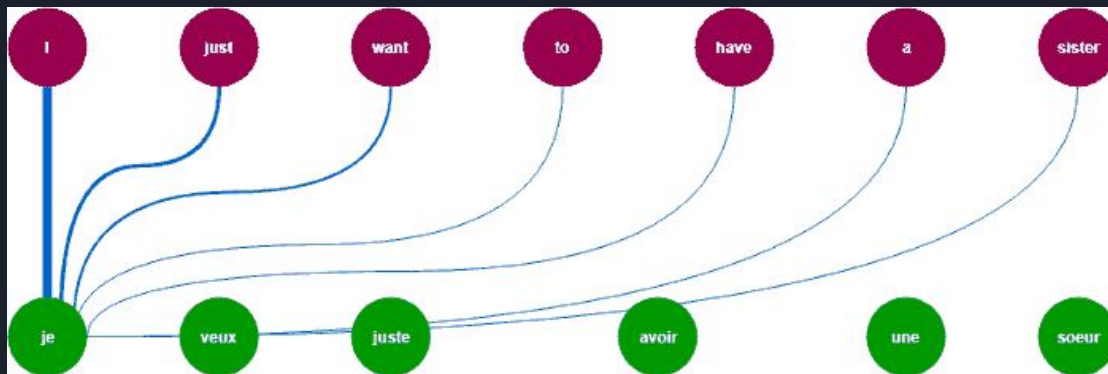
**Llion Jones\***  
Google Research  
llion@google.com

**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

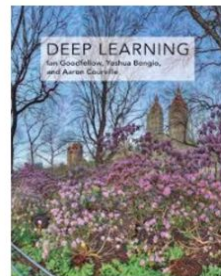


# Example : Searching for “Equivariance” in a book

ould require the same number of floating-point operations to  
still need to contain  $2 \times 319 \times 280 = 178,640$  entries. Con  
sufficient 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.

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



Chapter 9

Convolutional Networks

Convolutional networks (LeCun, 1988), 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 handwritten digits, which are typically of size  $10 \times 10$  grid of pixels, and image data, which are typically of size  $224 \times 224$  grid of pixels. Convolutional networks have been successfully employed in a variety of applications. The name “convolutional neural network” indicates that the network employs a mathematical operation called convolution. Convolution is a specialized kind of matrix multiplication. 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 describe the operations behind using convolution in a neural network. We then describe an operation called pooling, which allows convolutional networks to simplify. Usually, the operations used in a convolutional neural network have not been used previously in the definition of neural networks as used in other fields, such as engineering or pure mathematics. We describe neural networks on the architecture level, that are widely used in practice for neural networks. We also show how convolution can be applied to many kinds of data, with different numbers of dimensions. We then discuss issues of making convolution more efficient. Convolutional networks played an important role in recent success stories in image recognition. We discuss these recent success stories, then conclude with comments about the role of convolutional networks in the history of deep learning. The topic of this chapter does not exhaust a topic in itself. In the architecture of your convolutional network, the goal of this chapter is to describe the kind of units that convolutional networks provide, while chapter 11 describes general problems



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 ...

YT-Michael Phi



# 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 ...

YT-Michael Phi

# RNN vs LSTM vs **Attention**

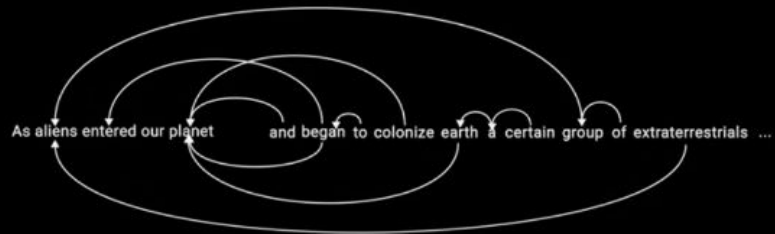
Attention Mechanism has an infinite reference window

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



YT-Michael Phi

# RNN vs LSTM vs **Attention**



YT-Michael Phi

# Attention example in images

A **bodybuilder** holding a **dumbbell**



YT - Computerphile

# Microsoft Attention GANs



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



# Self Attention

Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence.

Self-attention has been used successfully in a variety of tasks including reading comprehension, abstractive summarization, textual entailment and learning task-independent sentence representations

# Transformers

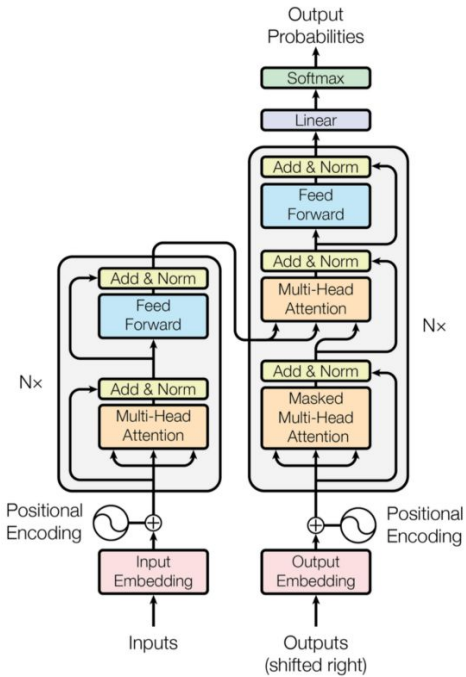
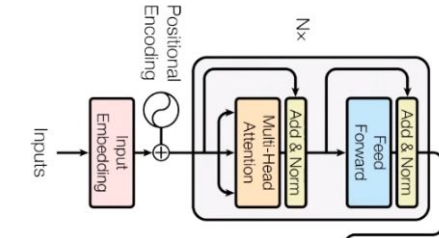


Figure 1: The Transformer - model architecture.

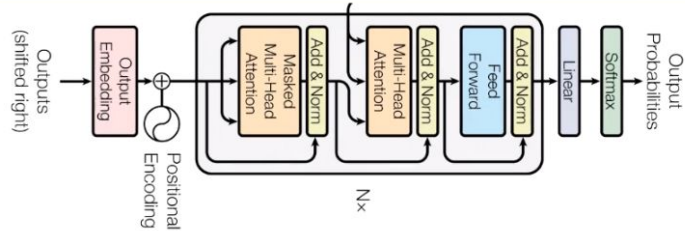


# Encoder and Decoder

## Transformer Flow



What is English? What is context?  
What is language!



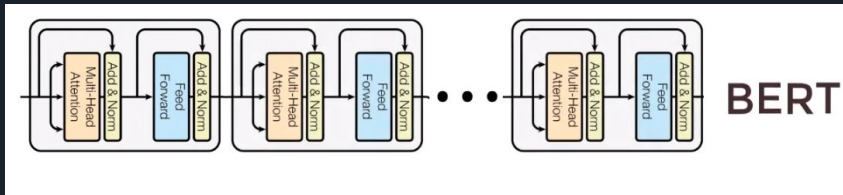
How to map English words to French words?  
What is language!



# BERT and GPT

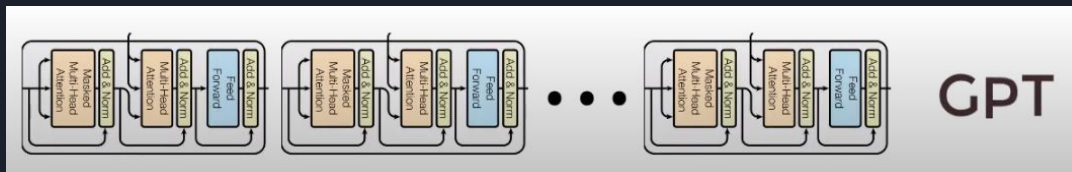
BERT (2018) by Google

A stack of Encoders



GPT (2018) by OpenAI

A stack of Decoders





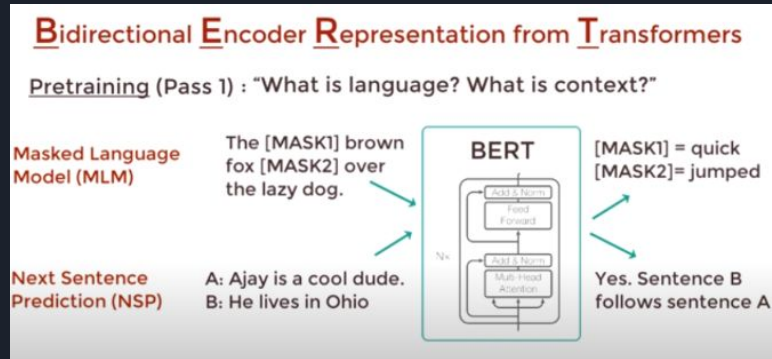
# BERT (Bidirectional Encoder Representations from Transformers)

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

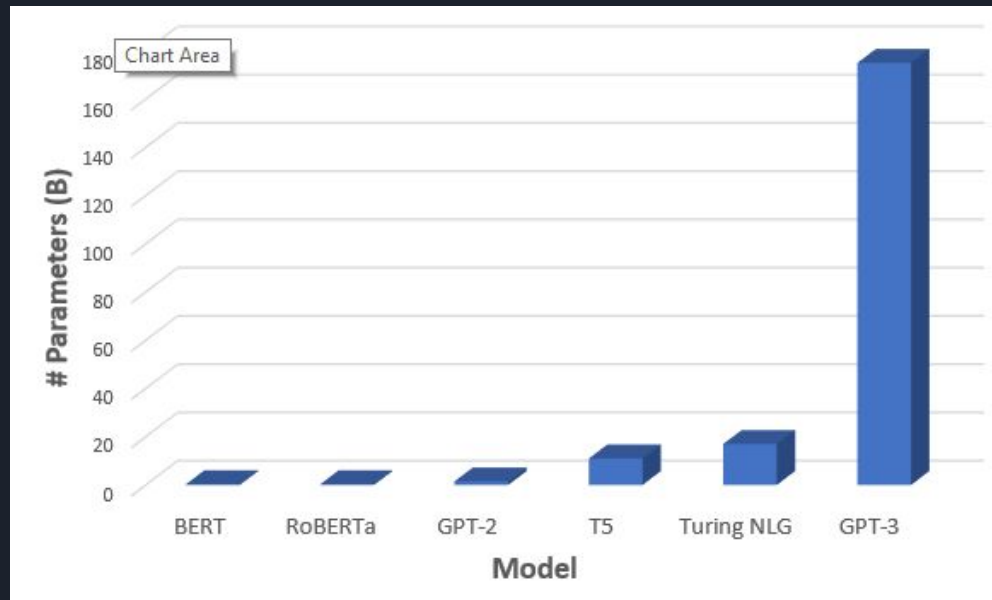


# GPT-3 (175 Billion Parameters)

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

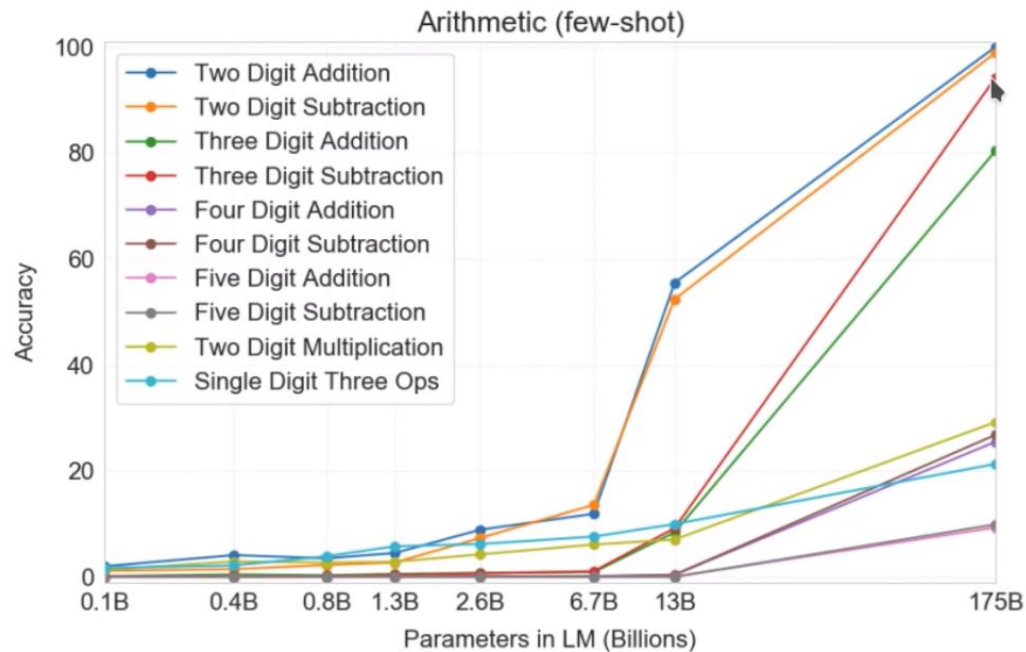
\*One petaflop/s-day is approximately equal to  $10^{20}$  neural net operations

# Comparison of sizes

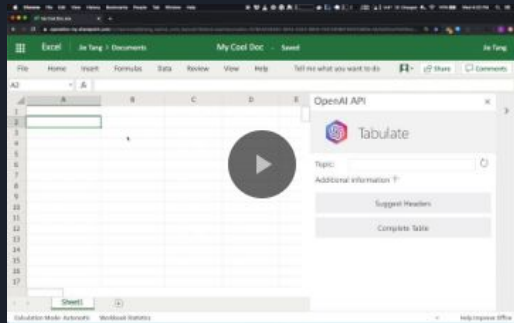




# Size Matters!



# 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:

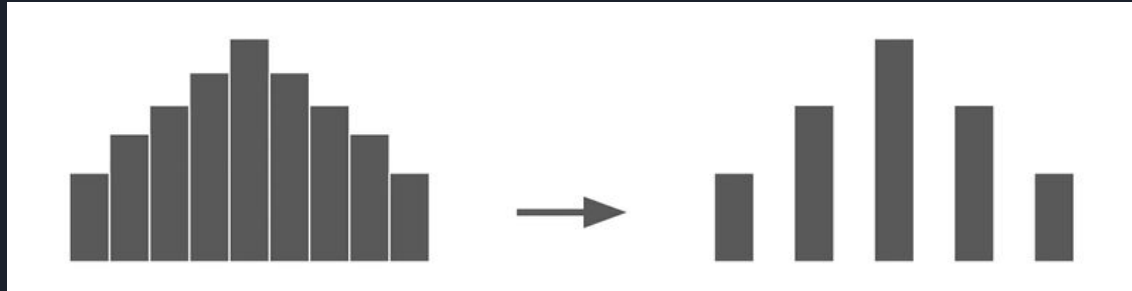
1. Quantization
2. Distillation
3. Pruning
4. More specialized models

# 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



YT-Rasa

# 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

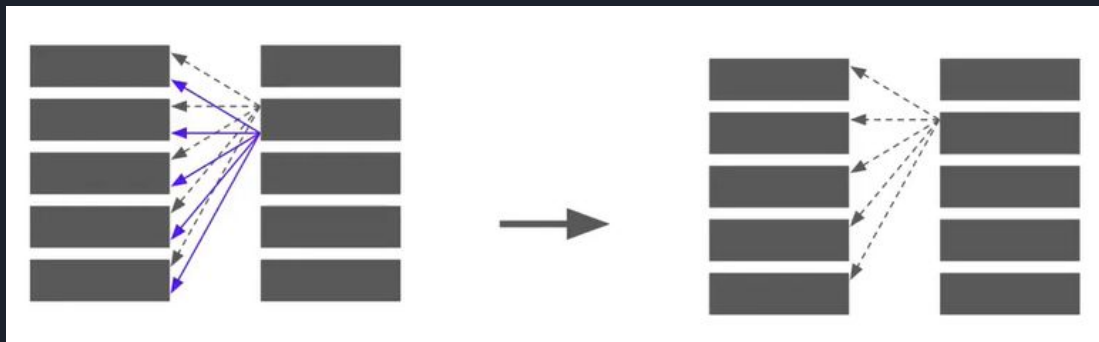


YT-Rasa

# 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



YT-Rasa

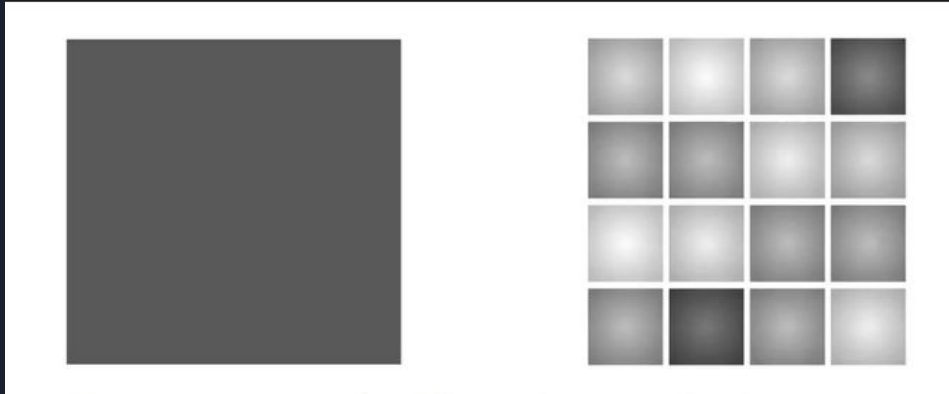




# More specialized models

Train a special smaller model

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



YT-Rasa

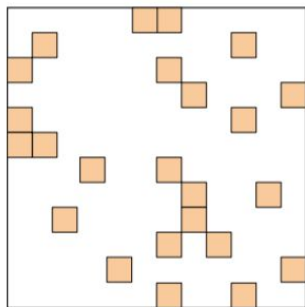


# Why don't these methods hurt performance?

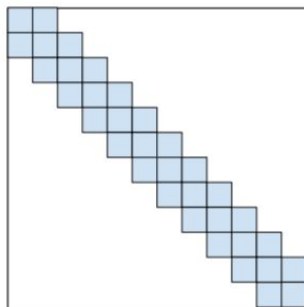
1. Really large transformers are bigger than they need to be for some tasks
2. There is a lot of redundancy in these models

# Bigbird

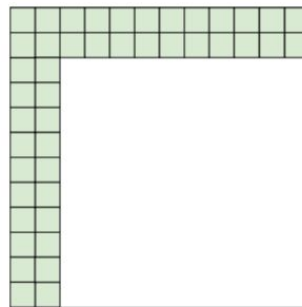
last revised 8 Jan 2021



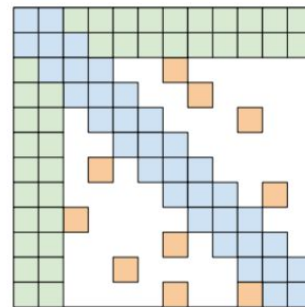
(a) Random attention



(b) Window attention



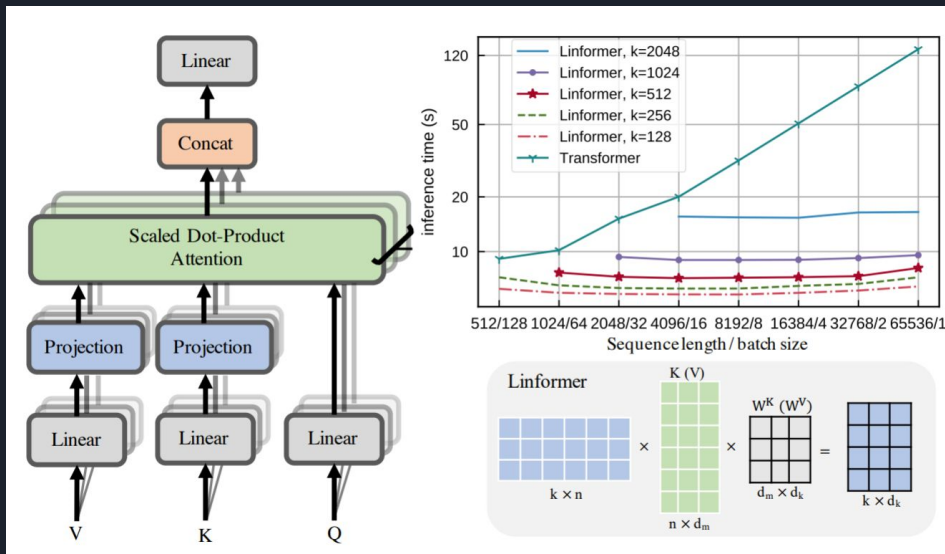
(c) Global Attention



(d) BIGBIRD

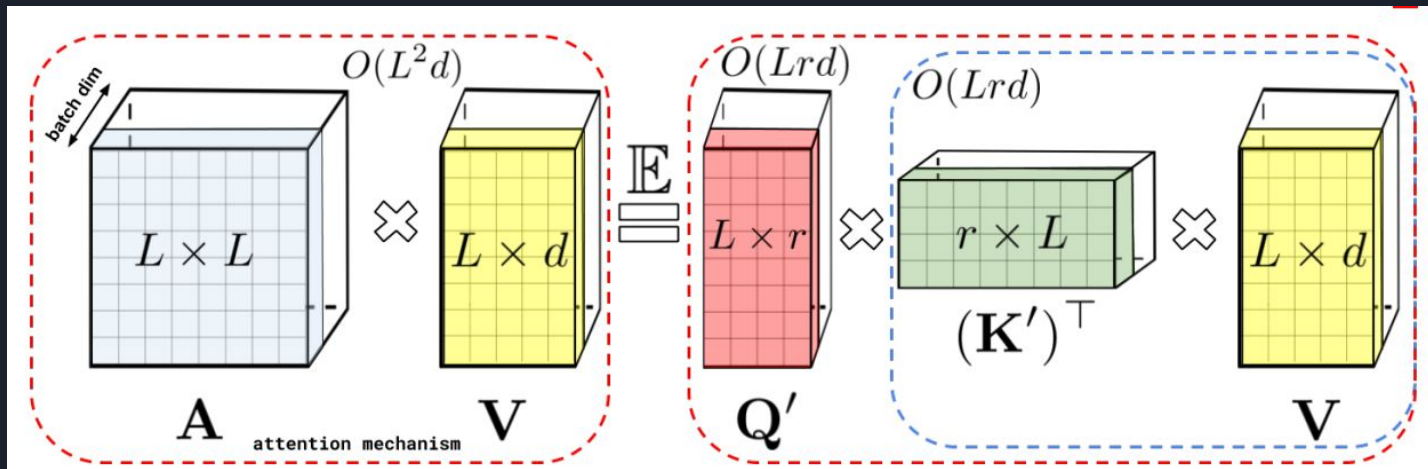
# Linformer

last revised 14 Jun 2020



# Performer

last revised 16 Feb 2021





# References

[Recurrent neural network - Wikipedia](#)

[Long short-term memory - Wikipedia](#)

[Gated recurrent unit - Wikipedia](#)

[Transformer and language models - YouTube](#) (playlist)

[BigBird - arxiv](#)

[Linformer - arxiv](#)

[Performer - arxiv](#)

# Thanks

[hasanmahani08@gmail.com](mailto:hasanmahani08@gmail.com)

[rezadar1378@gmail.com](mailto:rezadar1378@gmail.com)

