# The BERTs family

Longer, bigger, smaller, smarter

# Pre-trained transformer types (by training method)

Encoder only (autoencoder)

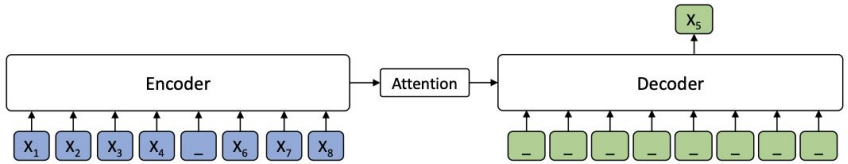
- BERT, ALBERT, RoBERTa
- Seq classification, token classification
- Masked words and predict

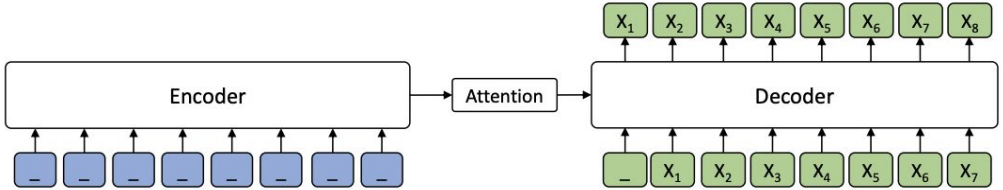
Encoder-decoder (seq2seq)

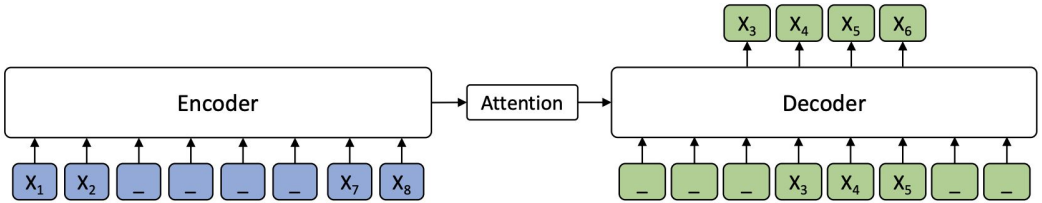
- MASS, BART, T5
- Machine translation, text summary
- Mask phrases

Decoder only (autoregressive)

- GPT, CTRL
- Text generation
- Predict next word




https://huggingface.co/transformers/summary.html https://arxiv.org/pdf/1905.02450.pdf

# CTRL

For language generation, with style control signal

Style (embedding)

Wikipedia *Anarchism is* a political philosophy that advocates the abolition of all forms of hierarchy and domination, including capitalism, patriarchy, racism, sexism, heterosexism and other oppressive social structures.\n\nThe term "anarchism" was coined by Pierre-Joseph Proudhon in his book "The

Books *Anarchism is* the \n only true and practical form of Socialism. It has been said that Socialism \n will never be established until it can be proved to work in practice. The \n same may be said, with equal truth, about Anarchism. But if we are not able \n to prove its practicability at present, then let us

CTRL: A CONDITIONAL TRANSFORMER LANGUAGE MODEL FOR CONTROLLABLE GENERATION

Horror *A knife* handle pulled through the open hole in the front. I jumped when the knife hit.\n\nEyes widened in horror. Her scream was the only sound I heard besides her sobs.\n\nThe spider touched her feet as it started to dig into the top of her arch. The creature's tears began to

Reviews *A knife* is a tool and this one does the job well.\n\nRating: 4.0\n\nI bought these for my husband who has been using them to cut up his own meat since he got them. He says they are very sharp so be careful when you use them, but that doesn't seem like much of an issue because he's used

# Can use url as control signal

Condition on topic, dates, etc.

Links https://www.cnn.com/2014/09/20/us-president-meets-british-pm
JUST WATCHED\n\nObama meets with British PM\n\nMUST WATCH\n\nStor
highlights\n\nPresident Barack Obama met with Britain's Prime Minister David Cameron

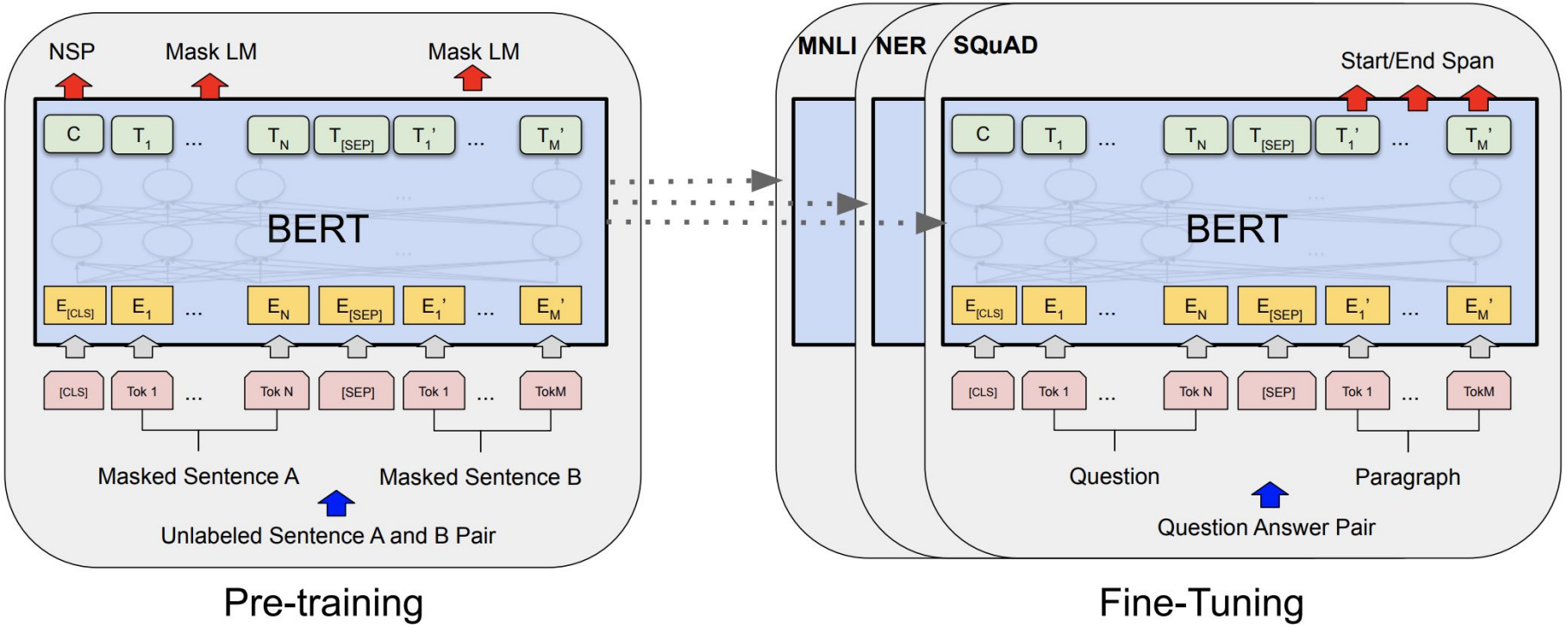
Links https://www.cnn.com/2018/09/20/us-president-meets-british-pm
JUST WATCHED\n\nTrump and May meet for first time\n\nMUST WATCH\n\nWashingto
(CNN) President Donald Trump, who has been criticized by some in the UK over his decision to leav
the European Union, met with British Prime Minister Theresa May, a White House official said o
Thursday.

Links https://www.cnn.com/style/09/20/2018/george-clooney-interview
George Clooney on the future of his acting career\n\nBy\n\nUpdated 10:51 AM ET, Thu September 20, 2018\n\nChat with us in Facebook Messenger. Find out what's happening in the world as it unfolds.\n\nPhotos:George Clooney, 'Ocean's 8'\n\nActor George Clooney attends a photocall for "Ocean's 8" at Grauman's Chinese Theatre on August 31, 2018, in Los Angeles.\n\n...

Links https://www.cnn.com/politics/09/20/2018/george-clooney-interview
JUST WATCHED\n\nGeorge Clooney on the Trump administration\n\nMUST WATCH\n\n(CNN) Actor and activist George Clooney, who has been a vocal critic of President Donald Trump, said he is "ready to go back into the political arena" after his role in an anti-Trump documentary was cut from theaters this week.\n\n...

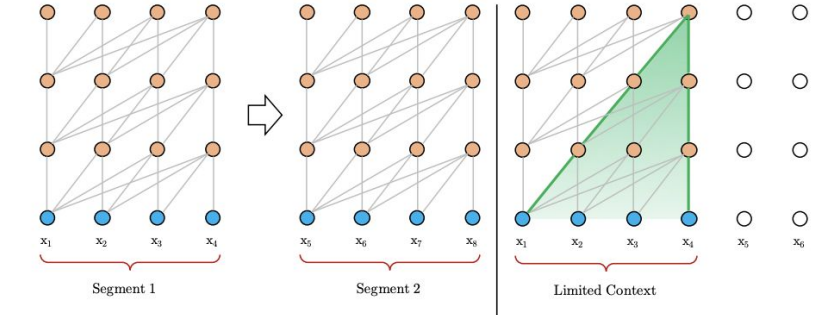
# BERT

Full transformer
Masked LM to learn bi-directional LM
Next sentence prediction to learn discourse

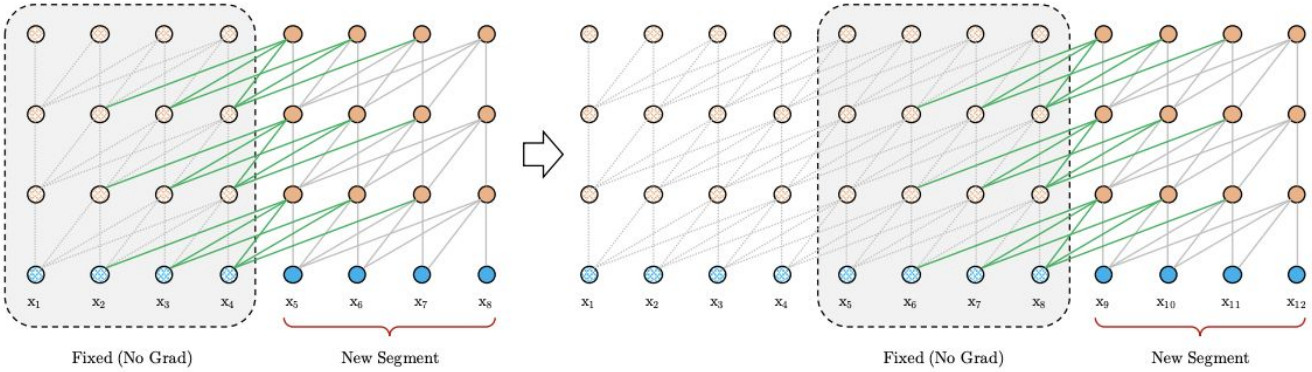



https://arxiv.org/abs/1810.04805

# XL-transformer




Chunking limits the length of context




Let next chunk attention has access to previous chunk

(a) Training phase.

Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

# XLNet

- Transformer-XL + permutation LM
- Mask LM creates mismatch between training and test.
  - Use different permutations to capture different context direction
- Add attention masks to simulate random permutation orders

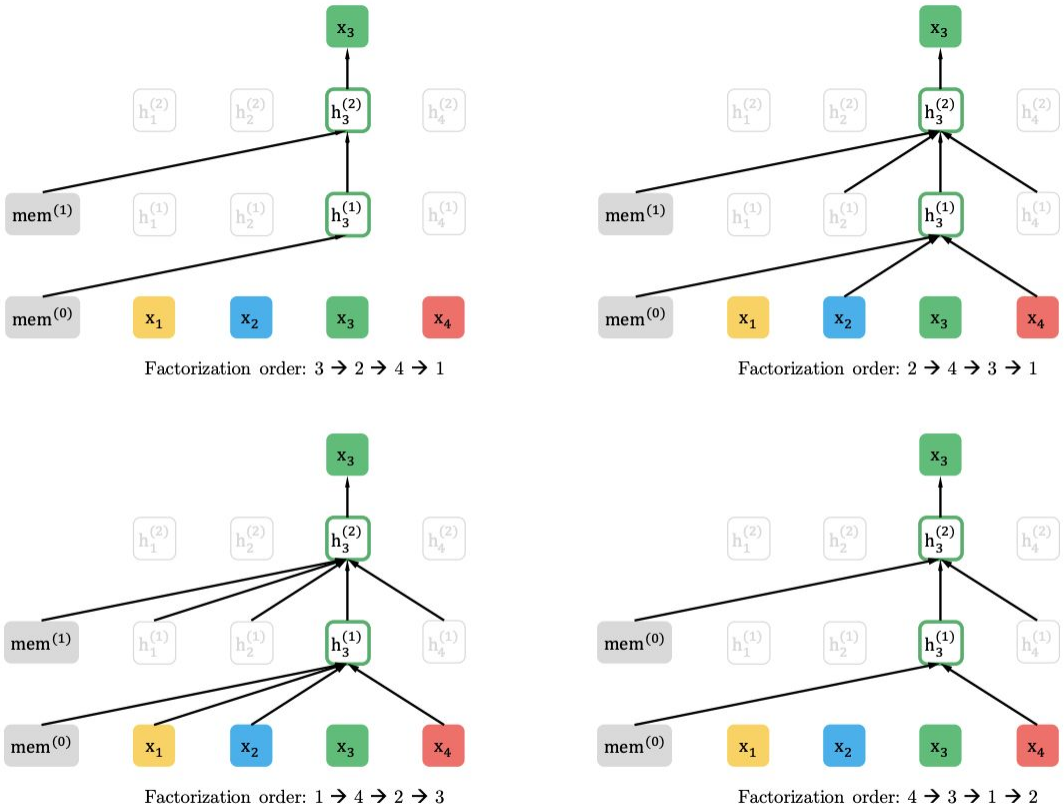



Figure 1: Illustration of the permutation language modeling objective for predicting $x_3$ given the same input sequence $\mathbf{x}$ but with different factorization orders.

Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context
XLNet: Generalized Autoregressive Pretraining for Language Understanding

# Roberta (Robustly optimized BERT approach)

A trick and tuning study

Dynamic masking > static

Next sentence prediction is not optimal

Larger batch + higher learning rate

| Masking | SQuAD 2.0 | MNLI-m | SST-2 |
|---|---|---|---|
| reference | 76.3 | 84.3 | 92.8 |
| *Our reimplementation:* | | | |
| static | 78.3 | 84.3 | 92.5 |
| dynamic | 78.7 | 84.0 | 92.9 |

| bsz | steps | lr | ppl | MNLI-m | SST-2 |
|---|---|---|---|---|---|
| 256 | 1M | 1e-4 | 3.99 | 84.7 | 92.7 |
| 2K | 125K | 7e-4 | **3.68** | **85.2** | **92.9** |
| 8K | 31K | 1e-3 | 3.77 | 84.6 | 92.8 |

RoBERTa: A Robustly Optimized BERT Pretraining Approach

# Side note on large batch size training

## Don't Decay the Learning Rate, Increase the Batch Size

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, Quoc V. Le

It is common practice to decay the learning rate. Here we show one can usually obtain the same learning curve on both training and test sets by instead increasing the batch size during training. This procedure is successful for stochastic gradient descent (SGD), SGD with momentum, Nesterov momentum, and Adam. It reaches equivalent test accuracies after the same number of training epochs, but with fewer parameter updates, leading to greater parallelism and shorter training times. We can further reduce the number of parameter updates by increasing the learning rate $\epsilon$ and scaling the batch size $B \propto \epsilon$. Finally, one can increase the momentum coefficient $m$ and scale $B \propto 1/(1-m)$, although this tends to slightly reduce the test accuracy. Crucially, our techniques allow us to repurpose existing training schedules for large batch training with no hyper-parameter tuning. We train ResNet-50 on ImageNet to $76.1\%$ validation accuracy in under 30 minutes.

Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour
Don't Decay the Learning Rate, Increase the Batch Size

# Megatron-LM

Distributed training

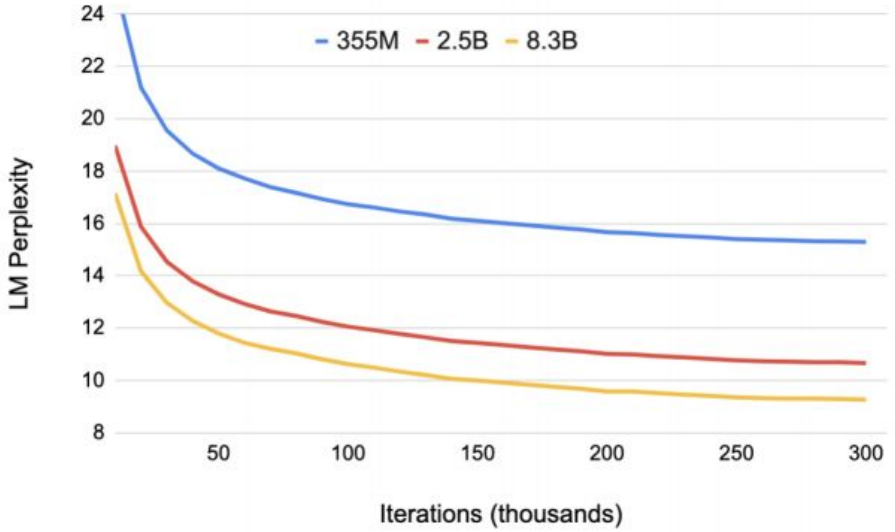
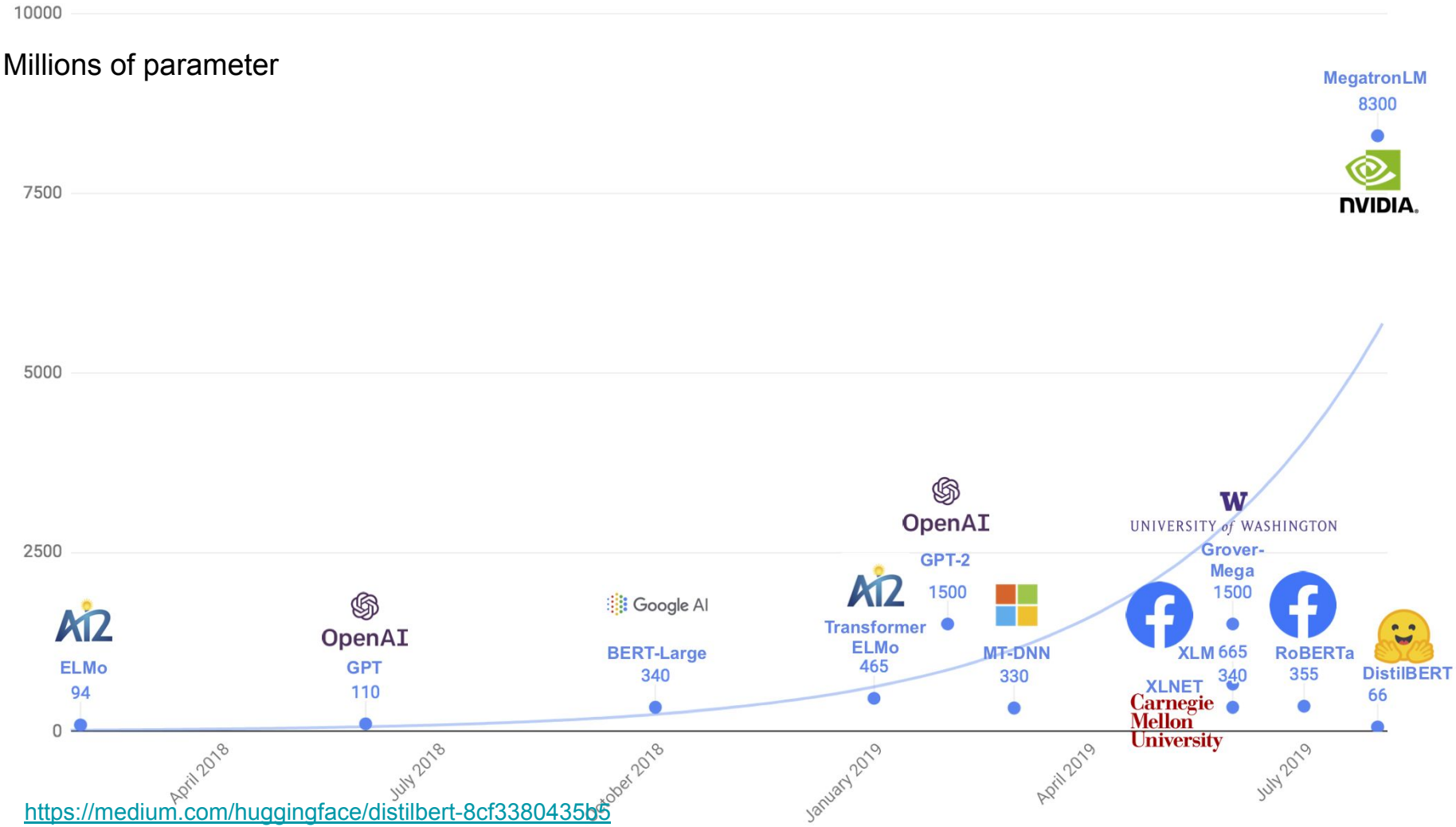



*Figure 7.* Validation set perplexity. All language models are trained for 300k iterations. Larger language models converge noticeably faster and converge to lower validation perplexities than their smaller counterparts.




https://medium.com/huggingface/distilbert-8cf3380435b5

# Distill bert

Knowledge distillation to get smaller models

Reduce the # of transformer layers by half. Use tricks in Roberta.

Use KL-divergence between teacher and student model

“Cheaper training”
eight 16GB V100 GPUs for approximately three and a half days

| | Nb of parameters (millions) | Inference Time (s) |
|---|---|---|
| GLUE BASELINE (ELMo + BiLSTMs) | 180 | 895 |
| BERT base | 110 | 668 |
| DistilBERT | 66 | 410 |

| | Macro Score | CoLA | MNLI | MNLI-MM | MRPC | | QNLI | QQP | | RTE | SST-2 | STS-B | | WNLI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mcc | acc | acc | acc | f1 | acc | acc | f1 | acc | acc | pearson | spearmanr | acc |
| GLUE BASELINE (ELMo + BiLSTMs) | 68.7 | 44.1 | 68.6 (avg) | | 70.8 | 82.3 | 71.1 | 88.0 | 84.3 | 53.4 | 91.5 | 70.3 | 70.5 | 56.3 |
| BERT base | 78.0 | 55.8 | 83.7 | 84.1 | 86.3 | 90.5 | 91.1 | 90.9 | 87.7 | 68.6 | 92.1 | 89.0 | 88.6 | 43.7 |
| DistilBERT | 75.2 | 42.5 | 81.6 | 81.1 | 82.4 | 88.3 | 85.5 | 90.6 | 87.7 | 60.0 | 92.7 | 84.5 | 85.0 | 55.6 |

https://github.com/huggingface/transformers https://medium.com/huggingface/distilbert-8cf3380435b5

# ALBERT

Want higher hidden units without growing the model. Factorized embedding matrix

VxE -> VxE+ExH

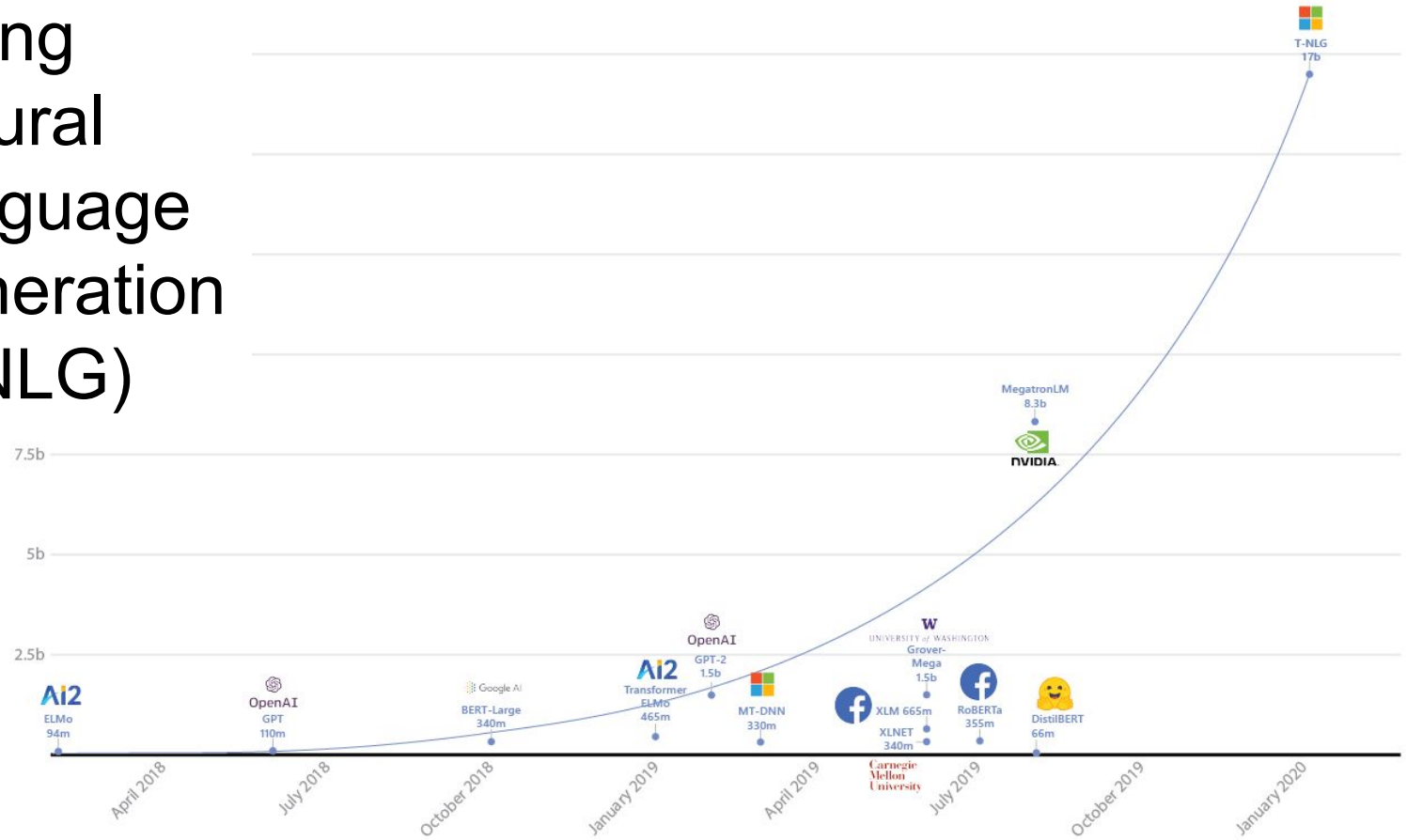
Share attention layer parameters across layers. More stable training as a side effect.

NSP is easy compared to LM tasks (multi-task imbalance)
Next sentence prediction (random sentence) -> Sentence order prediction (swapped sentence or not)

| Model | | Parameters | Layers | Hidden | Embedding | Parameter-sharing | Avg | Speedup |
|---|---|---|---|---|---|---|---|---|
| BERT | base | 108M | 12 | 768 | 768 | False | 82.1 | 17.7x |
| | large | 334M | 24 | 1024 | 1024 | False | 85.1 | 3.8x |
| | xlarge | 1270M | 24 | 2048 | 2048 | False | 76.7 | 1.0 |
| ALBERT | base | 12M | 12 | 768 | 128 | True | 80.1 | 21.1x |
| | large | 18M | 24 | 1024 | 128 | True | 82.4 | 6.5x |
| | xlarge | 59M | 24 | 2048 | 128 | True | 85.5 | 2.4x |
| | xxlarge | 233M | 12 | 4096 | 128 | True | **88.7** | 1.2x |

Some experiments show dropout hurt performance

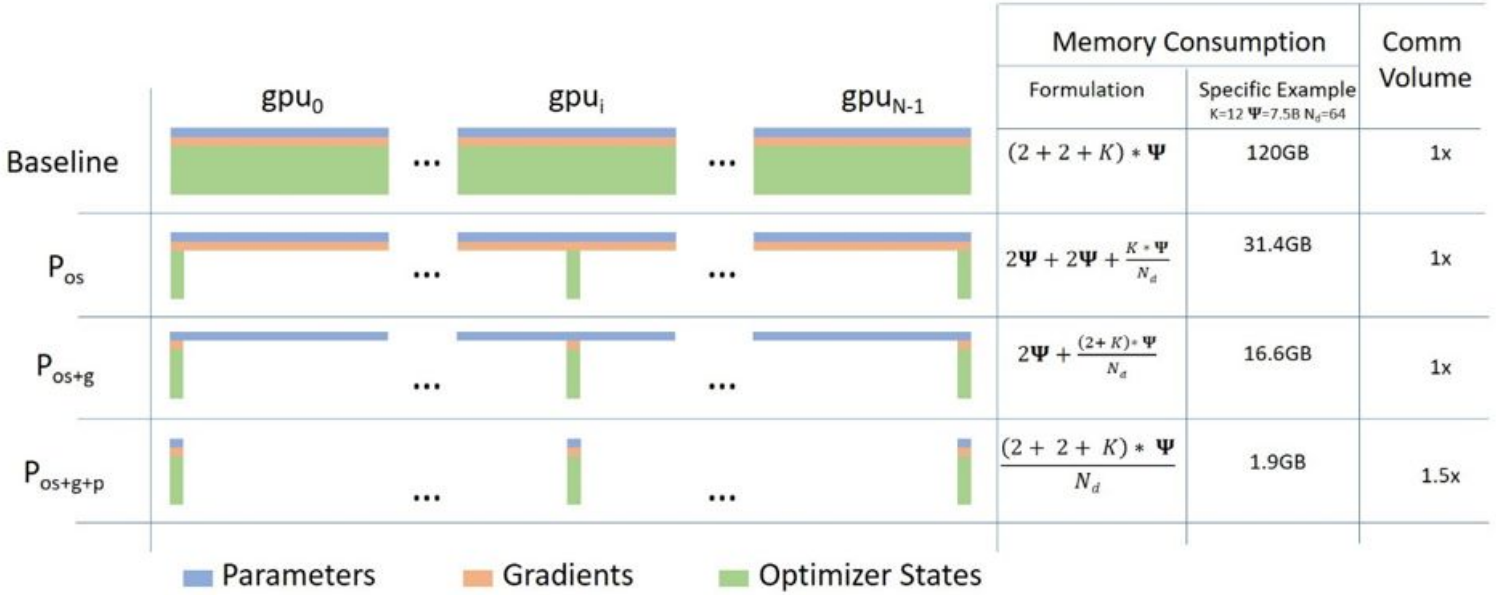
# Turing Natural Language Generation (T-NLG)




https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

# Zero & DeepSpeed

Clever management of training parameters across GPUs and machines




| | Memory Consumption | | Comm Volume |
|---|---|---|---|
| | Formulation | Specific Example K=12 Ψ=7.5B $N_d$=64 | |
| Baseline | $(2+2+K)*\Psi$ | 120GB | 1x |
| $P_{os}$ | $2\Psi+2\Psi+\frac{K*\Psi}{N_d}$ | 31.4GB | 1x |
| $P_{os+g}$ | $2\Psi+\frac{(2+K)*\Psi}{N_d}$ | 16.6GB | 1x |
| $P_{os+g+p}$ | $\frac{(2+2+K)*\Psi}{N_d}$ | 1.9GB | 1.5x |

https://www.microsoft.com/en-us/research/blog/zero-deepspeed-new-system-optimizations-enable-training-models-with-over-100-billion-parameters/

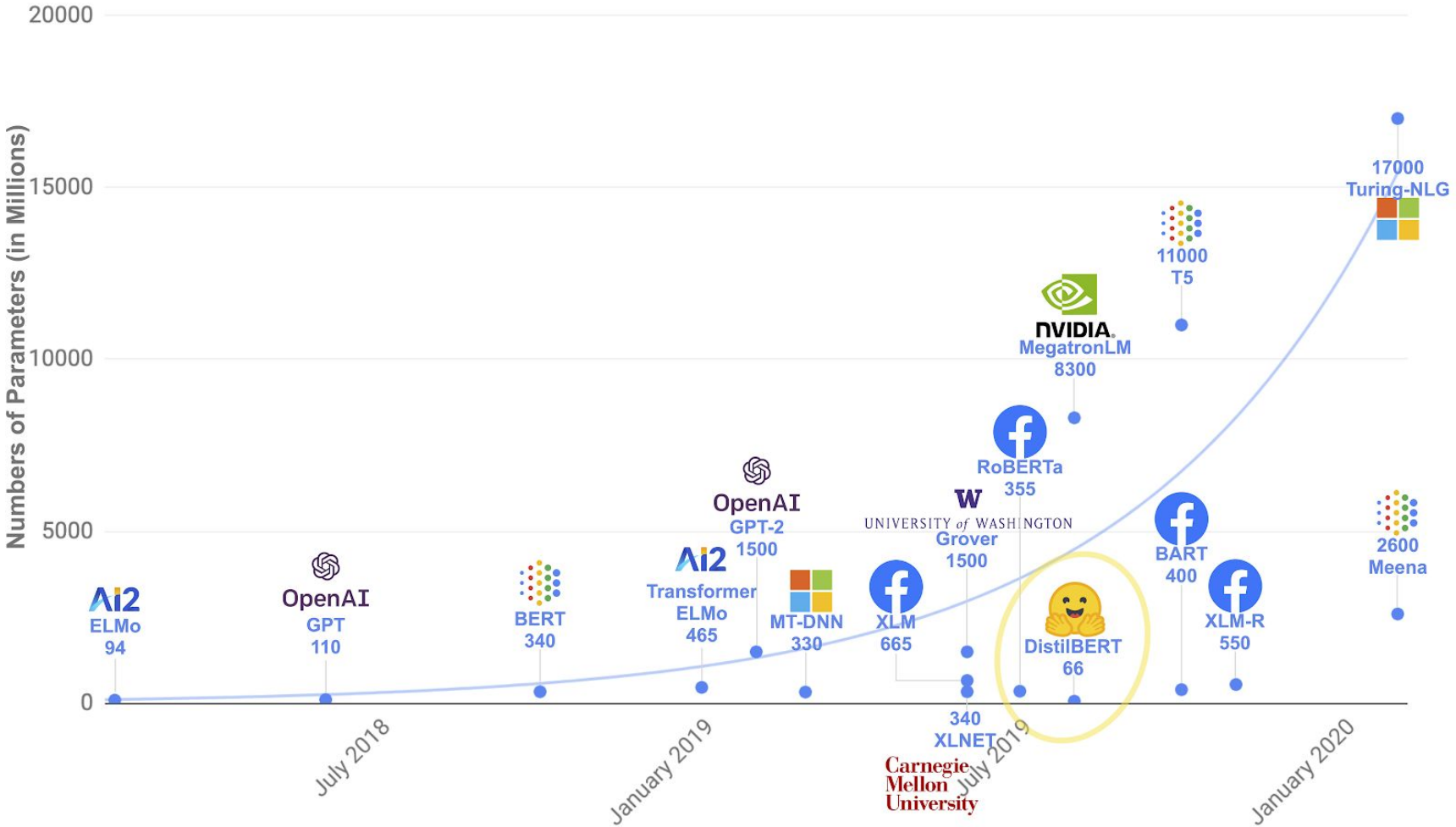
# T-NLG

| | LAMBADA (acc) strict | WikiText-103 (test adj. ppl) |
|---|---|---|
| Open AI GPT-2 1.5B | 52.66 (63.24)* | 17.48 |
| Megatron-LM 8.3B | 66.51 | 10.81 |
| **T-NLG 17B** | 67.98 | 10.21 |

*Open AI used additional processing (stopword filtering) to achieve higher numbers than the model achieved alone. Neither Megatron nor T-NLG use this stopword filtering technique.

Can do Q/A by just LOTs of internet text

| | |
|---|---|
| **When did WW2 end?** | WW2 ended in 1945. |
| **How many people live in the US?** | There are over 300 million people living in the US. |


20000
15000
10000
5000
0
Numbers of Parameters (in Millions)
ELMo
94
OpenAI
GPT
110
BERT
340
Transformer
ELMo
465
OpenAI
GPT-2
1500
MT-DNN
330
XLM
665
340
XLNET
Grover
1500
RoBERTa
355
MegatronLM
8300
DistilBERT
66
11000
T5
BART
400
XLM-R
550
17000
Turing-NLG
2600
Meena
July 2018
January 2019
July 2019
January 2020
Carnegie Mellon University
UNIVERSITY of WASHINGTON

# Longer, bigger, smaller, smarter

XLNet

Megatron

Roberta

Distill Bert

Albert

T-NLG

# Tricks to make better transformers

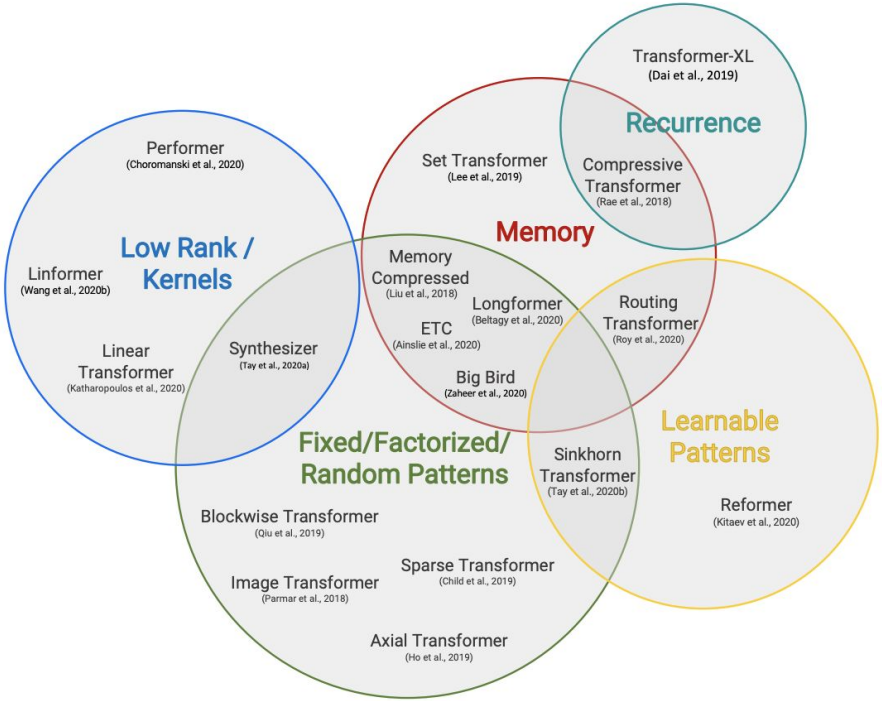
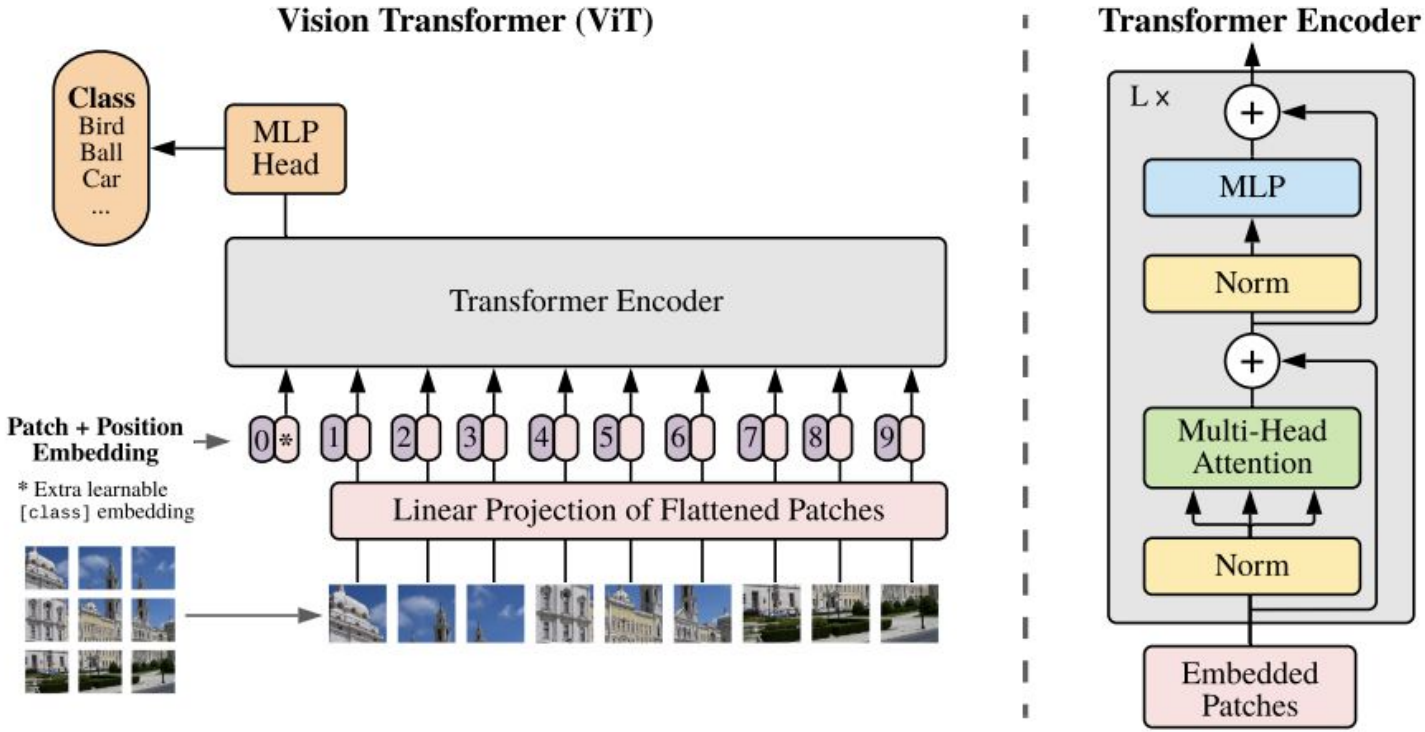



Figure 2: Taxonomy of Efficient Transformer Architectures.

Efficient Transformers: A Survey, 2020 https://arxiv.org/pdf/2009.06732.pdf

# Vision transformer




AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE
https://arxiv.org/pdf/2010.11929v1.pdf

# Dall E

GPT3 for images

Cluster image patches into tokens and combined with text to train an autoregressive model


(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

(c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign

(d) the exact same cat on the top as a sketch on the bottom

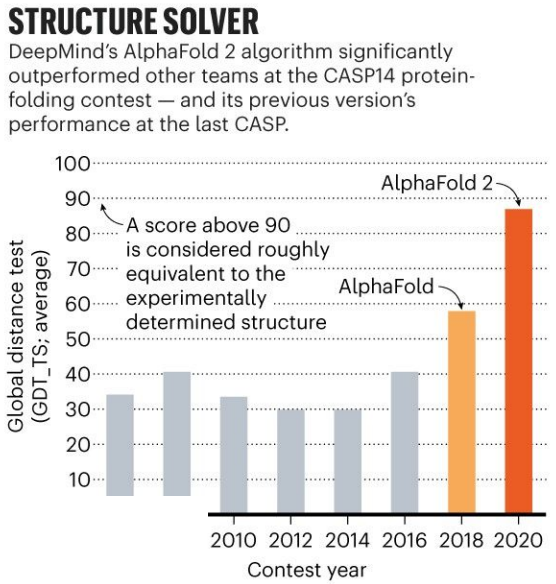
*Figure 2.* With varying degrees of reliability, our model appears to be able to combine distinct concepts in plausible ways, create anthropomorphized versions of animals, render text, and perform some types of image-to-image translation.

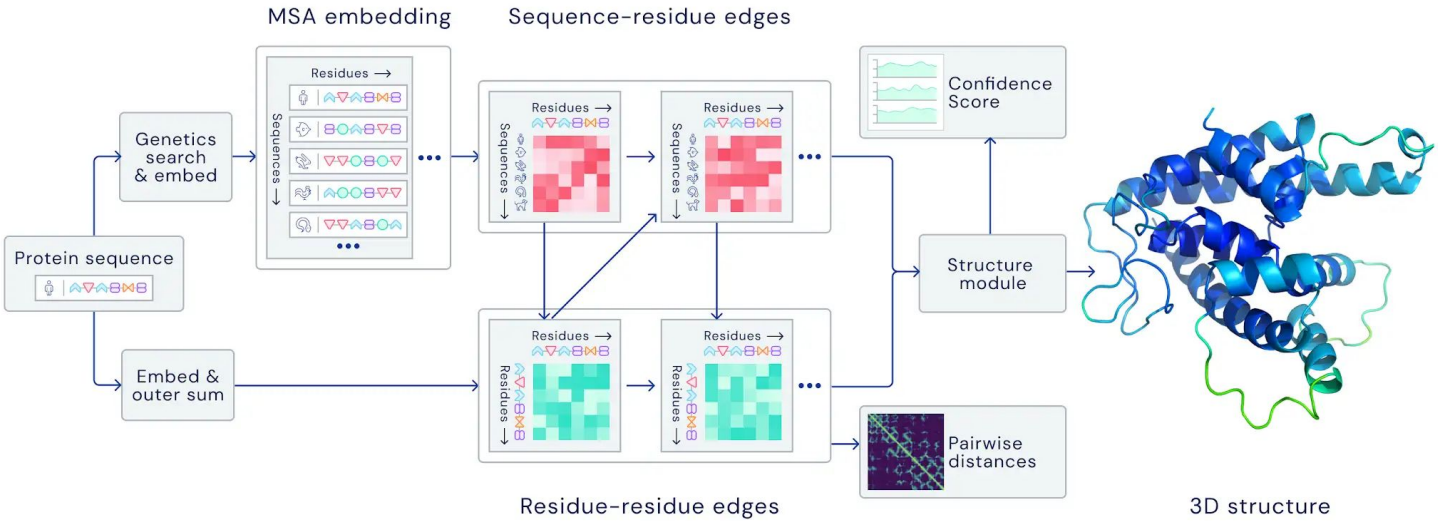
https://openai.com/blog/dall-e/
https://arxiv.org/pdf/2102.12092.pdf

# Alphafold2

https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

https://www.nature.com/articles/d41586-020-03348-4



# Homework

Working with WangchanBERTa

- A Roberta model
- Different pretrained models with different domains/tokenizations

# Schedule

| | | | |
|---|---|---|---|
| 29-Mar-2021 | 3-Apr-2021 | 10 | Recent Research in NLP<br>Project Announcement + Paper Announcement |
| 5-Apr-2021 | 10-Apr-2021 | 11 | NLP Application 1 (Guest); 5-Apr-2021 |
| 12-Apr-2021 | 17-Apr-2021 | | Songkran Holiday |
| 19-Apr-2021 | 24-Apr-2021 | 12 | Paper Presentation & Progress Report |
| 26-Apr-2021 | 1-May-2021 | 13 | NLP Application 2 (Guest); 26-Apr-2021 |
| 3-May-2021 | 8-May-2021 | 14 | Project Presentation |

# HOW TO READ A SCIENTIFIC ARTICLE

# 2 Paper types

- Review article/tutorial
  - Give insights about the field
  - Useful for learning about a new field
  - Read multiple to avoid the author’s bias
  - Title usually has “review” or “tutorial”
- Primary research article
  - More details on the experiments and results

# Parts of an article

- Abstract
- Introduction
- Methods
- Results and discussion
- Conclusion
- Reference

# Things to look for before reading an article

- Publication date
- Author names
  - Previous and newer publications
- Keywords
- Acknowledgements and funding sources

REFORMER: THE EFFICIENT TRANSFORMER

**Nikita Kitaev***
U.C. Berkeley & Google Research
kitaev@cs.berkeley.edu

**Łukasz Kaiser***
Google Research

**Anselm Levskaya**
Google Research
{lukaszkaiser,levskaya}@google.com

https://arxiv.org/pdf/2001.04451.pdf

# Getting the big picture

- Read the abstract
- Read the introduction
  - What is the research question?
  - What is the method?
  - What had been done? How is it different from other work?
- Look at figures and results

Tip: keep track of terms you don’t understand

# First reading

- Reread the introduction
- Skim methods
- Read results and discussion
  - Does the figures make sense now?
- Write on the article!

# Understanding the article

- Reread the article (until you get what you want)
- Check references for parts you don’t understand
- Reread the abstract
  - Does your understanding match the abstract?
- Note down important points. This might come in handy when you write you paper/thesis!

# Evaluating the article

- Does the method make sense?
  - What are the limitations that the authors mention?
  - Are there other limitations?
  - Can it be used in other situations?
- Are the experiments legitimate?
  - The sample size is big enough?
  - What kind of dataset is used? How big?
  - The evaluation criterion is sound?
- Have these results been reproduced?
  - Look for articles that cite this paper

# ML paper checklist

- What is being done?
- How is it being done?
  - How is it different from previous work
- What is the dataset?
  - Nature of dataset
  - How many training/testing samples? How many classes/vocab size?
- Evaluation metric
  - What are the baselines?
- Practicality
  - Prone to parameter tuning?
  - Computing resource / Runtime (training and testing)

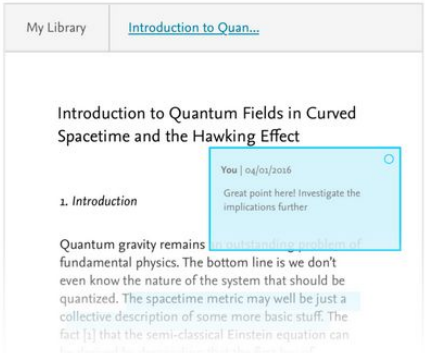
# Useful tools

- https://scholar.google.com
  - For finding other articles by the same authors or paper that cites the article

- https://www.mendeley.com/
  - Reference manager

**Annotate as you read**

Easily add your thoughts on documents in your own library, even from mobile devices. For ease of collaboration, you can also share documents with groups of colleagues and annotate them together.



# Useful resources

Most famous NLP papers have blogs explaining them nowadays

Huggingface usually implements a paper within a couple weeks

https://lilianweng.github.io/lil-log/

Yannic's youtube channel
https://www.youtube.com/channel/UCZHmQk67mSJgfCCTn7xBfew

# Project

Group of 3-5 people

Anything text/NLP related

Must has some application component (cannot be purely basic NLP task)

| | | | |
|---|---|---|---|
| | | | Recent Research in NLP |
| 29-Mar-2021 | 3-Apr-2021 | 10 | Project Announcement + Paper Announcement |
| 5-Apr-2021 | 10-Apr-2021 | 11 | NLP Application 1 (Guest); 5-Apr-2021 |
| 12-Apr-2021 | 17-Apr-2021 | | Songkran Holiday |
| 19-Apr-2021 | 24-Apr-2021 | 12 | Paper Presentation & Progress Report |
| 26-Apr-2021 | 1-May-2021 | 13 | NLP Application 2 (Guest); 26-Apr-2021 |
| 3-May-2021 | 8-May-2021 | 14 | Project Presentation |

# Summary

Different ways to pretrained models

Different ideas of optimizing transformers

Transformer for everything