Fantastic Transformers and Where to Find Them

Current Trends and Future Perspectives in Neural Natural Language Processing

Gabriele Sarti

Research Scientist at Aindo S.r.l.

Deep Learning Course, University of Trieste

May 27th, 2021

Plan for today 📋

What I will cover:

- ✅ Intro to Neural NLP
- ✅ Transformers & Transfer Learning
- ✅ Current trends in NLP
- ✅ Limitations and open questions

What will be left out:

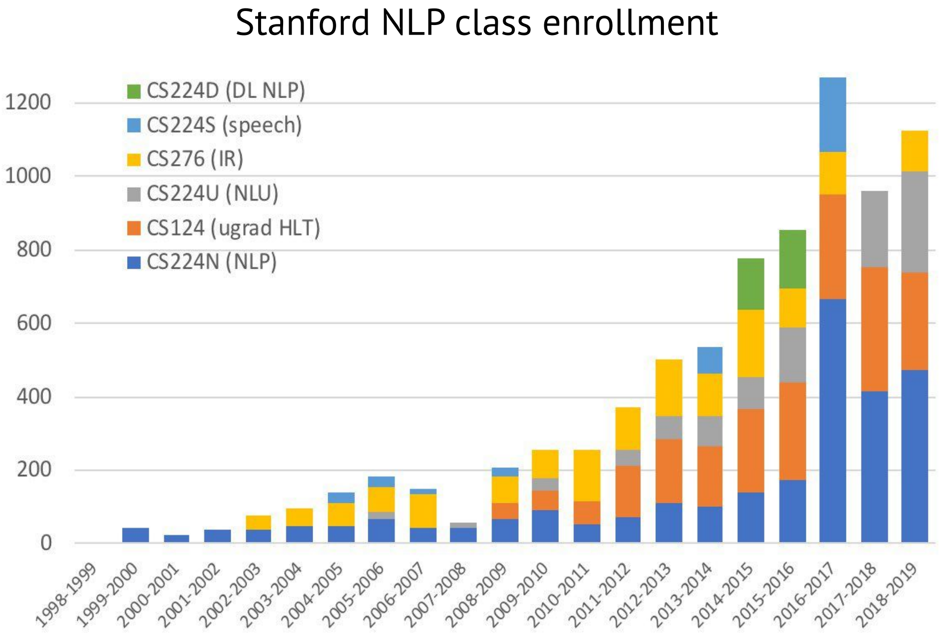
- ❌ Low-level architectural details
- ❌ Coding examples
- ❌ (Mostly) non-NLP subfields

Main goal: Provide distilled understanding to investigate further.

Introduction to Neural NLP

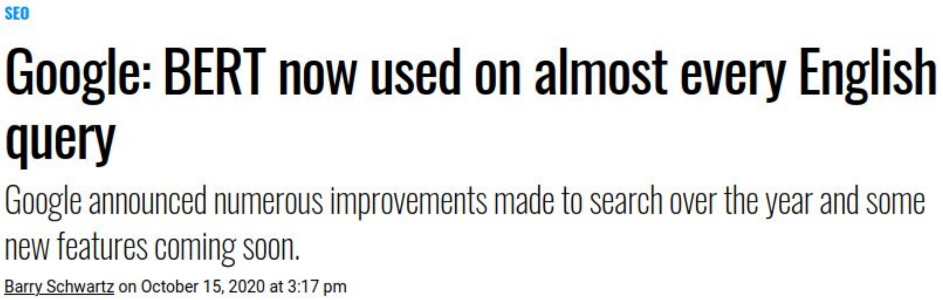
A Growing Interest for NLP 📈

- Stanford now teaches **10x students** in NLP w.r.t 1999–2004, and 2x w.r.t 2012–2014.
- NLP-first startups (HuggingFace, Rasa, Gong, etc.) raised **> 200M US$** in 2020.
- Open sourcing heavily adopted:
  - HuggingFace with 10k models, >50k model downloads per day, >400 contributors on Github.
  - 🌸 BigScience Workshop - European CERN for NLP



Source: stateof.ai 2020 report

The Turn Things Have Taken Since 2018 ↩️

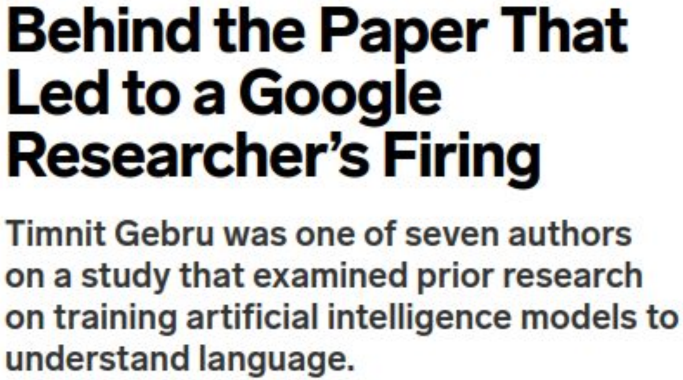


SEO

Google: BERT now used on almost every English query

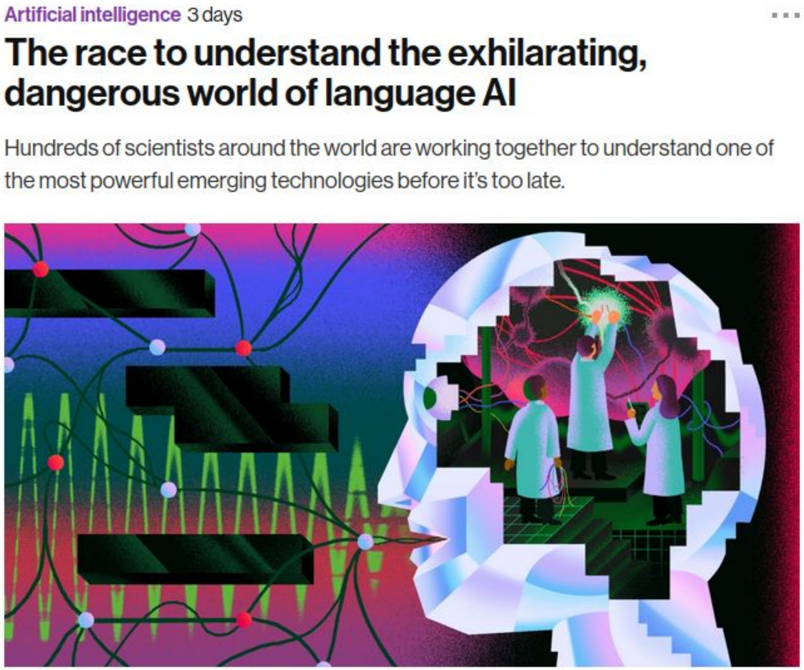
Google announced numerous improvements made to search over the year and some new features coming soon.

Barry Schwartz on October 15, 2020 at 3:17 pm



Behind the Paper That Led to a Google Researcher's Firing

Timnit Gebru was one of seven authors on a study that examined prior research on training artificial intelligence models to understand language.

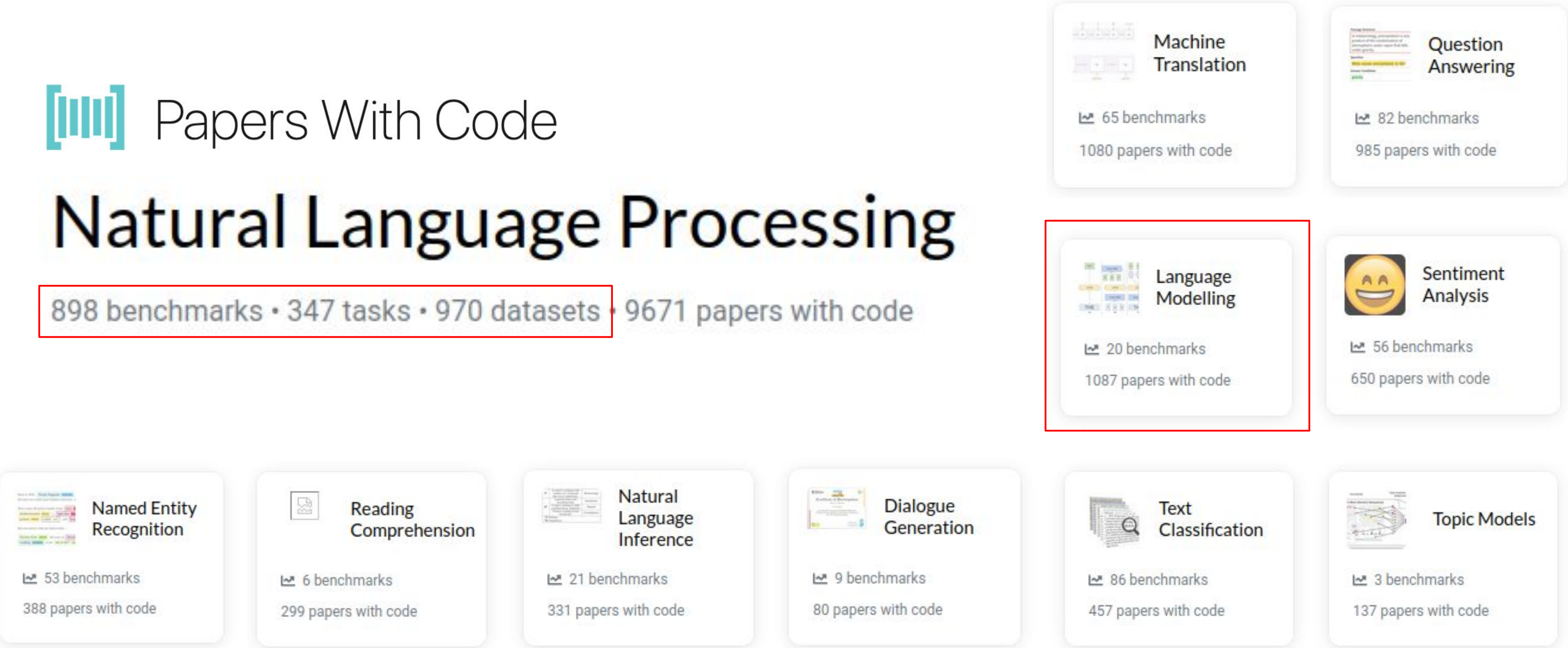


Artificial intelligence 3 days

The race to understand the exhilarating, dangerous world of language AI

Hundreds of scientists around the world are working together to understand one of the most powerful emerging technologies before it's too late.

A Wide World of NLP Tasks 🌍



Papers With Code

Natural Language Processing

898 benchmarks • 347 tasks • 970 datasets • 9671 papers with code

Machine Translation
65 benchmarks
1080 papers with code

Question Answering
82 benchmarks
985 papers with code

Language Modelling
20 benchmarks
1087 papers with code

Sentiment Analysis
56 benchmarks
650 papers with code

Named Entity Recognition
53 benchmarks
388 papers with code

Reading Comprehension
6 benchmarks
299 papers with code

Natural Language Inference
21 benchmarks
331 papers with code

Dialogue Generation
9 benchmarks
80 papers with code

Text Classification
86 benchmarks
457 papers with code

Topic Models
3 benchmarks
137 papers with code

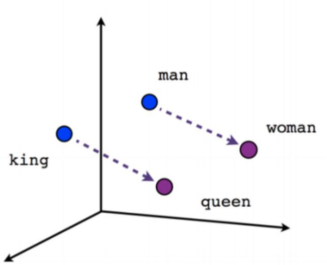
Non-Contextual Word Embeddings

Dense vectors used to represent text.

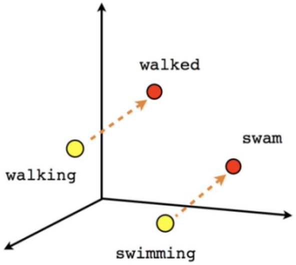
Good: Concise, semantic similarity.

Bad: Not useful for *polysemy*.

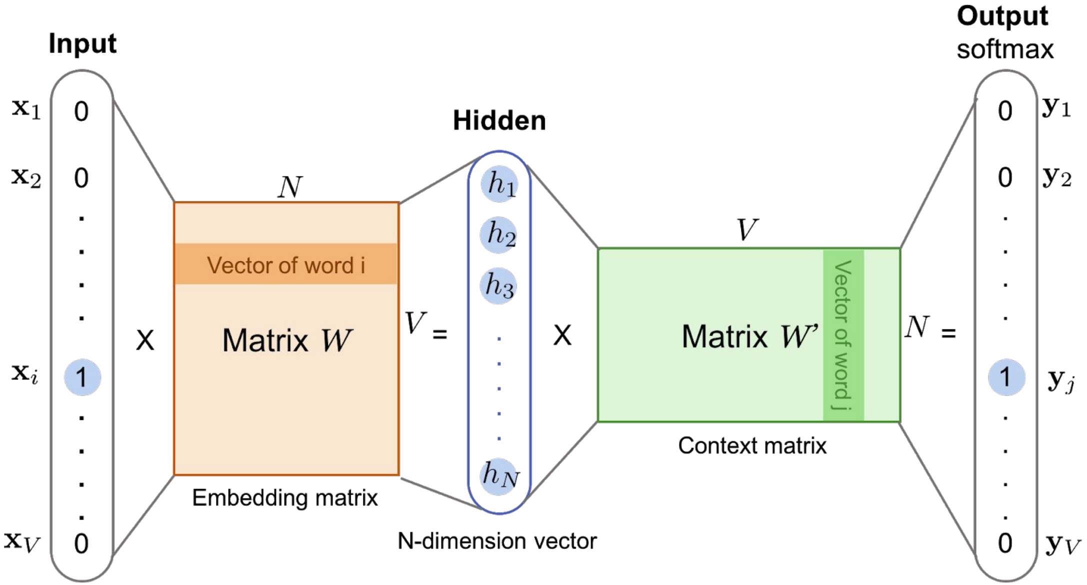
Popular: **Word2Vec, Glove, Fasttext**



Male-Female

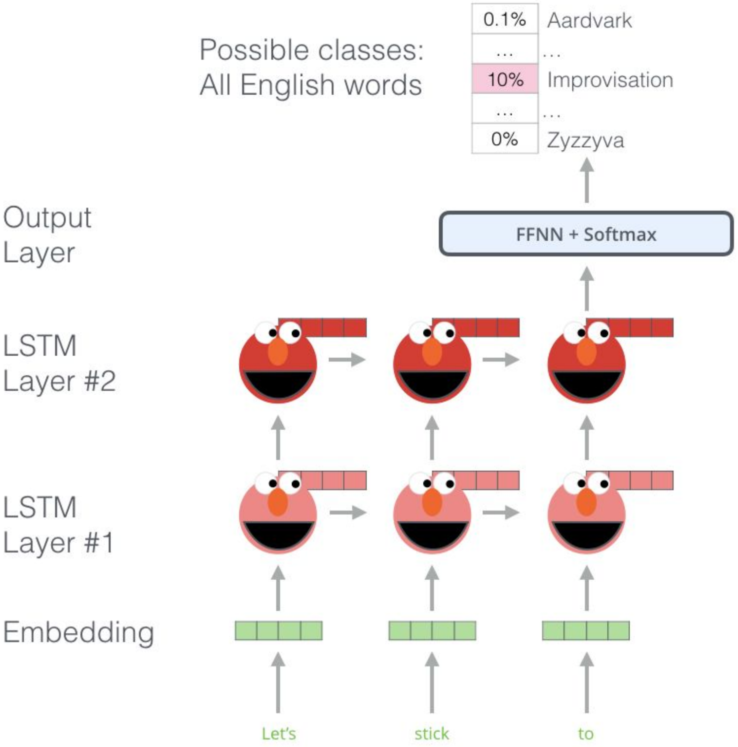


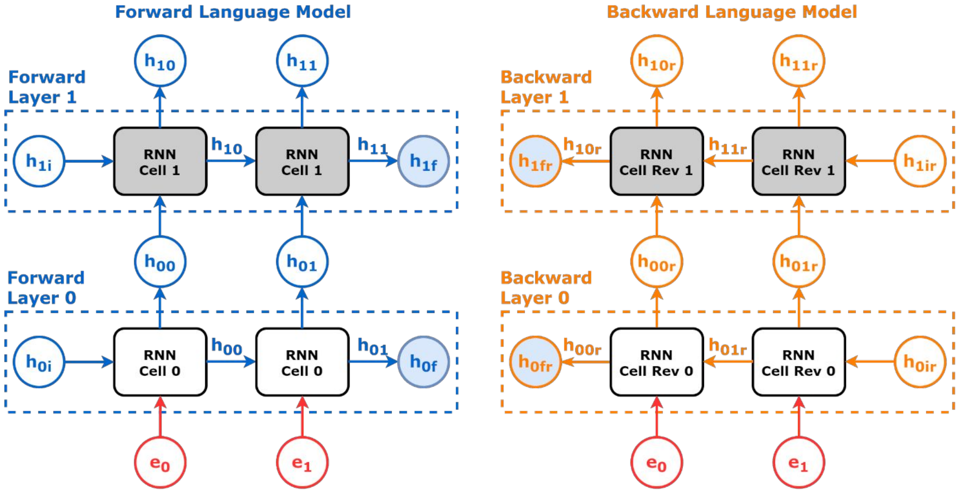
Verb tense

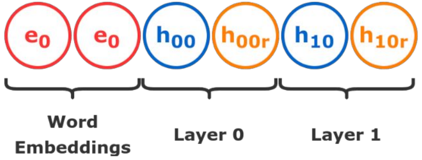


Mikolov et al. 2013, Pennington et al. 2014, Mikolov et al. 2017

ELMo: Contextualizing Word Embeddings







*Throw a **stick** to the dog*

vs

*Let's **stick** to the plan*

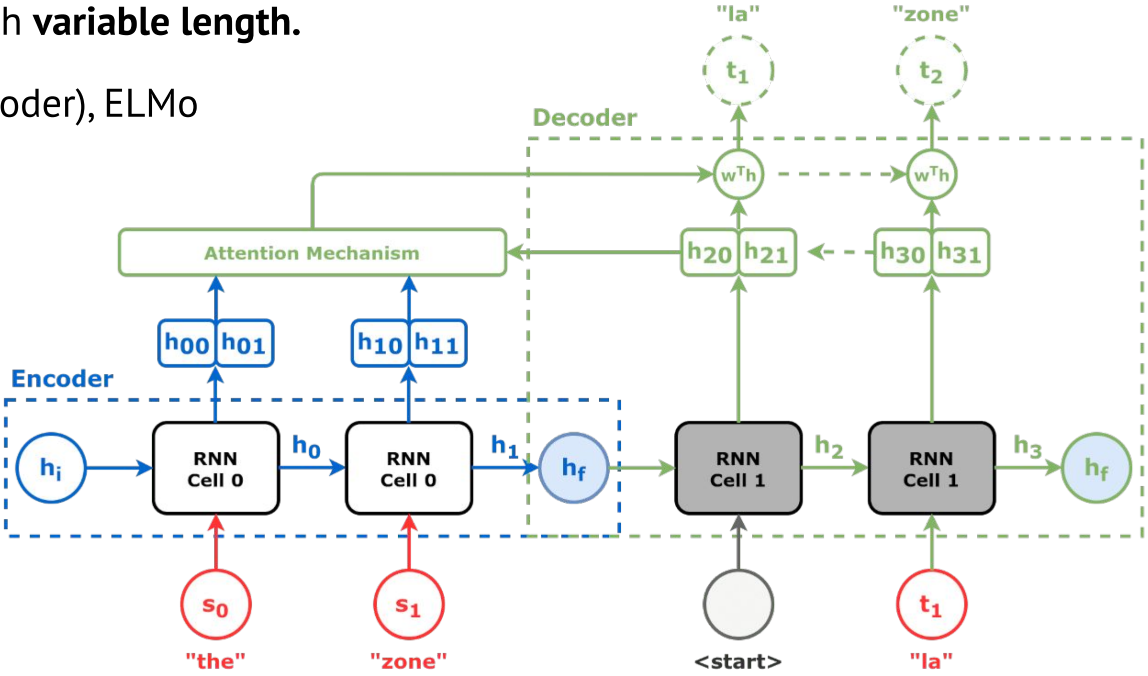
Peters et al. 2018

Before: Recurrent Neural Networks

- First model to successfully tackle seq2seq,
- Can be used to model inputs with **variable length.**
- Examples: Google Translate (decoder), ELMo

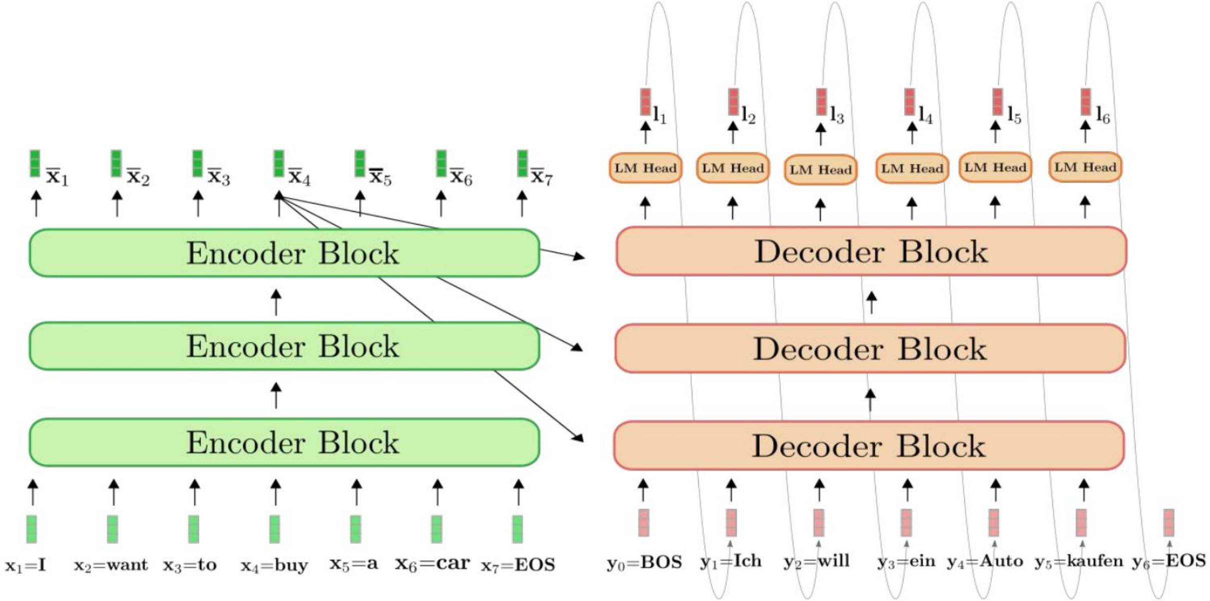
But:

- Difficult to parallelize.
- Ineffective for long-term dependencies.
- Use single state to encode all input information.



Now: Transformers

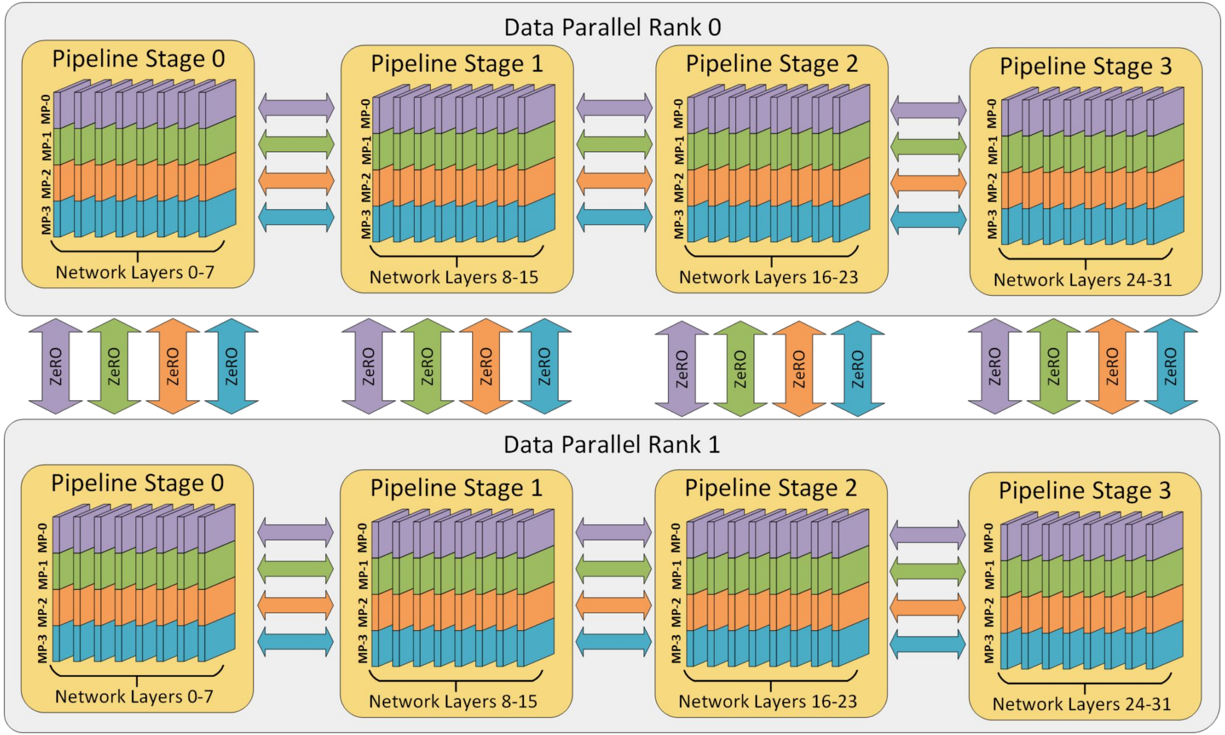
- Can also be used to model variable-length input.
- Highly parallelizable
- Effective at maintaining long distance relations.
- Less information loss by encoding inputs as sequences instead of using a single state.
- Examples: GPT-2, BERT, etc.



The Hardware Lottery 🎲

“A model is just as good as the hardware it runs on”

GPUs & TPUs → Transformers

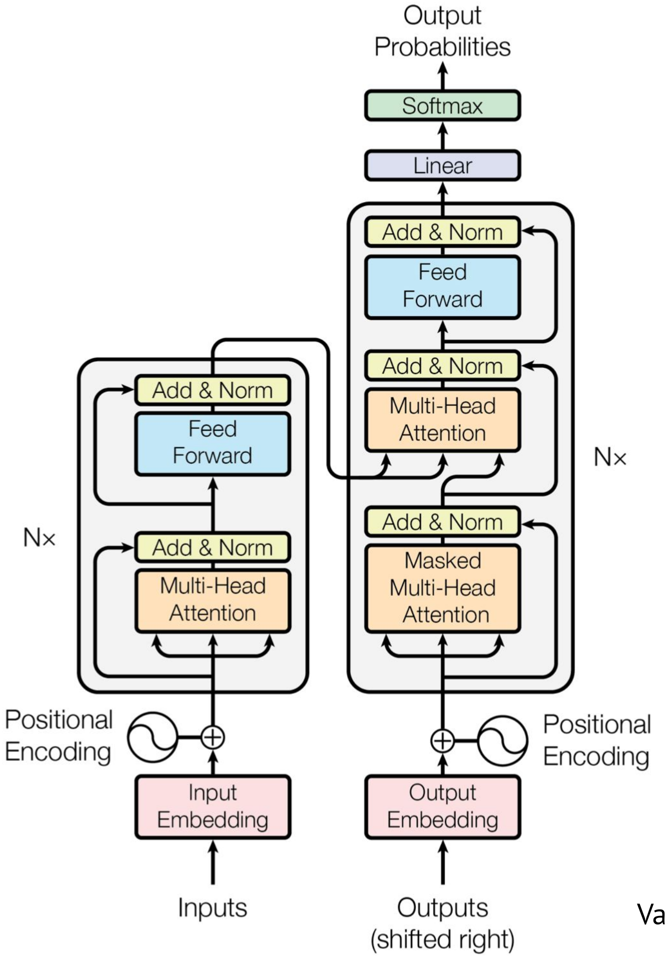


Hooker 2020

The Transformer Architecture

Attention is All You Need

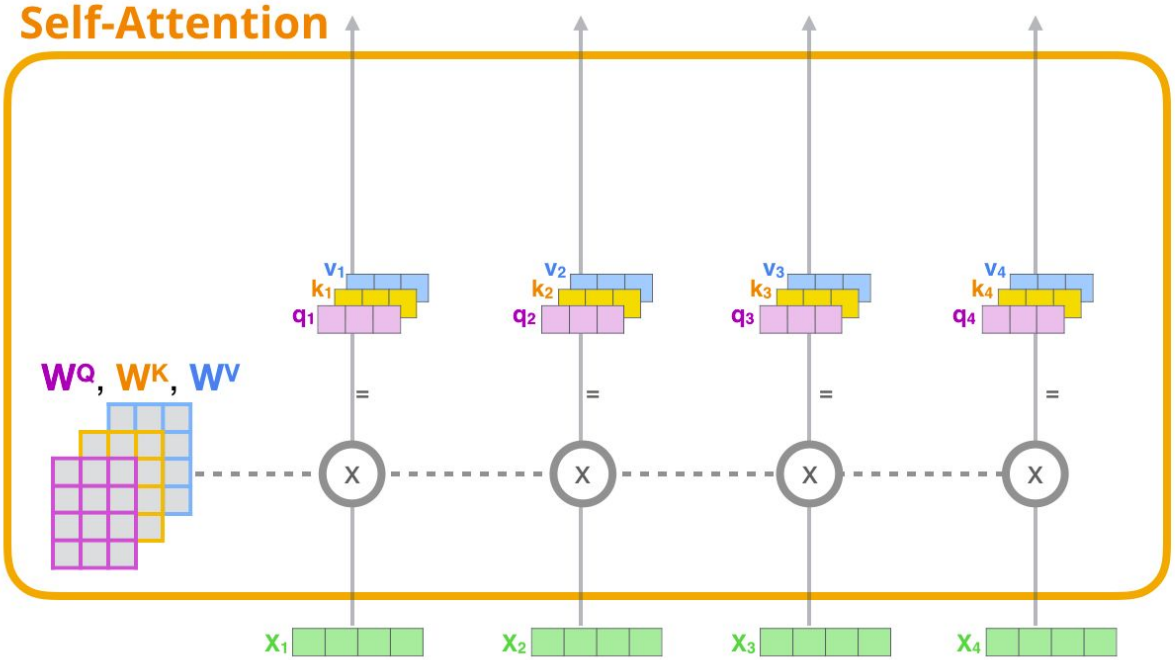
- The model is the first using only attention, without any recurrent operation.
- Encoder-decoder architecture, later dropped by many notable examples



Vaswani et al. 2017

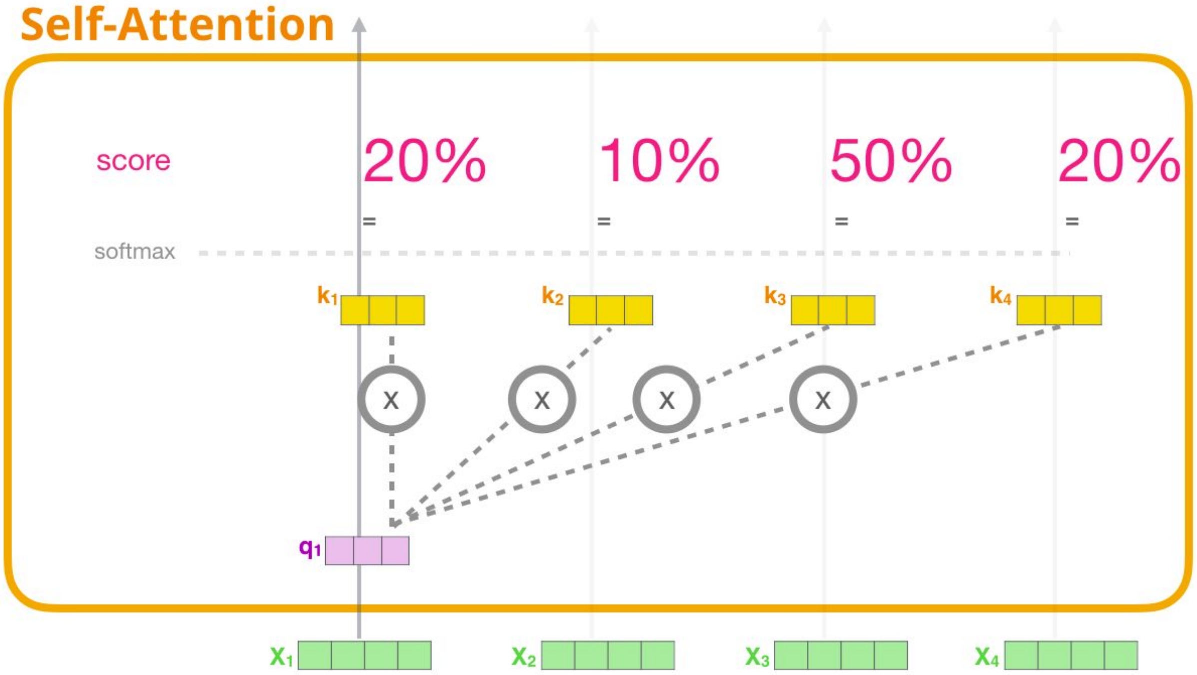
Self-attention

1) For each input token, create a query vector, a key vector, and a value vector by multiplying by weight Matrices W^Q, W^K, W^V



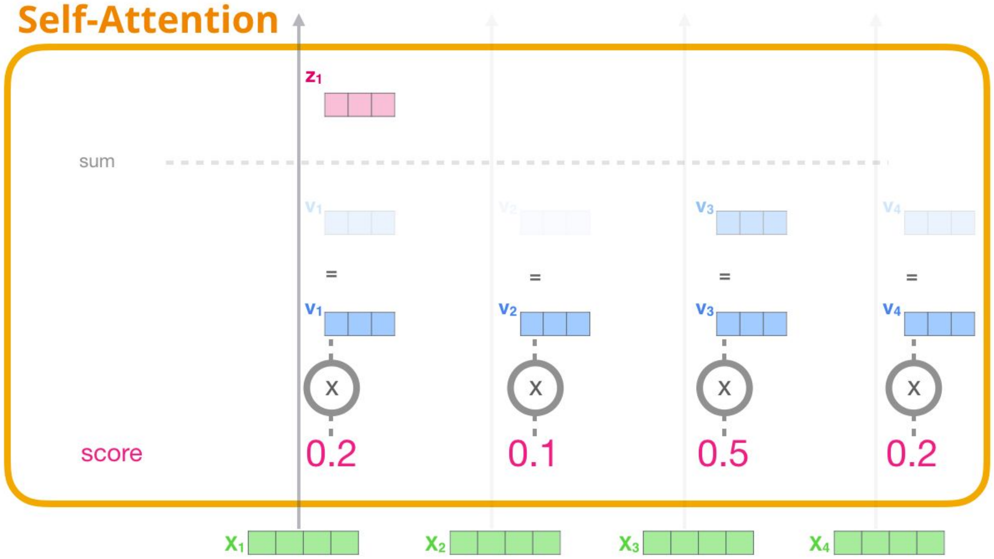
Self-attention

2) Multiply (dot product) the current query vector, by all the key vectors, to get a score of how well they match



(Multi-Head) Self-attention

3) Multiply the value vectors by the scores, then sum up

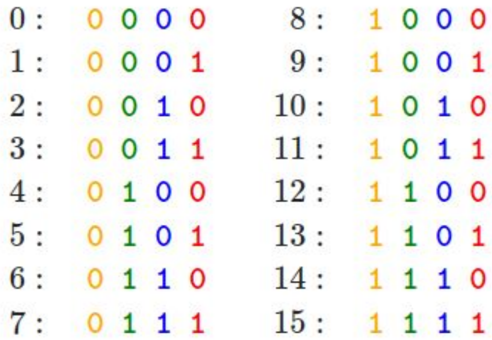


$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

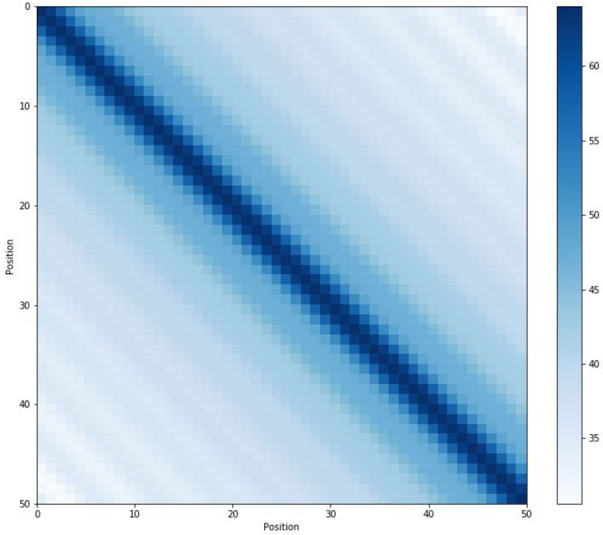
$$\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}(\mathrm{head_1},...,\mathrm{head_h})W^O$$
$$\text{where}\ \mathrm{head_i}=\mathrm{Attention}(QW_i^Q,KW_i^K,VW_i^V)$$

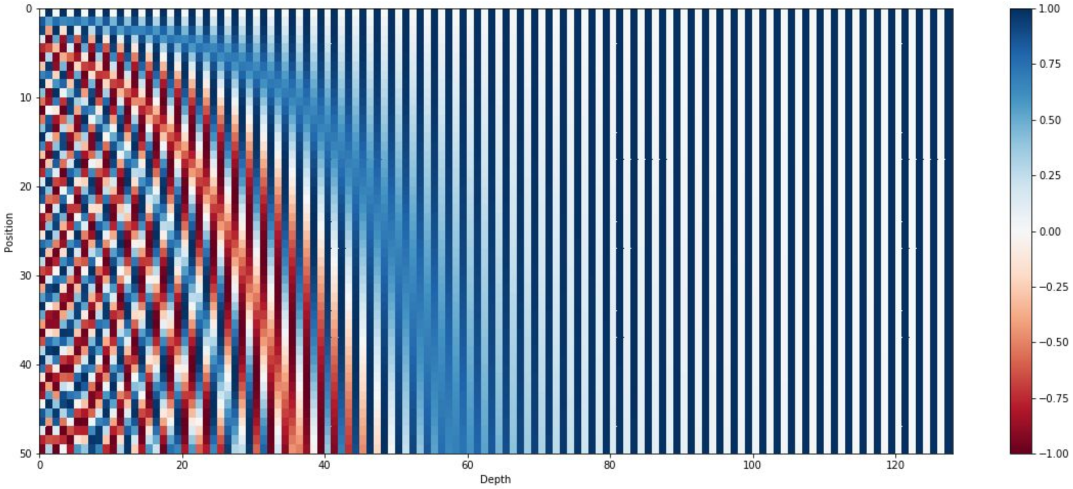
Positional Encodings

1. Identify a position deterministically and univocally
2. Provide distance information between positions

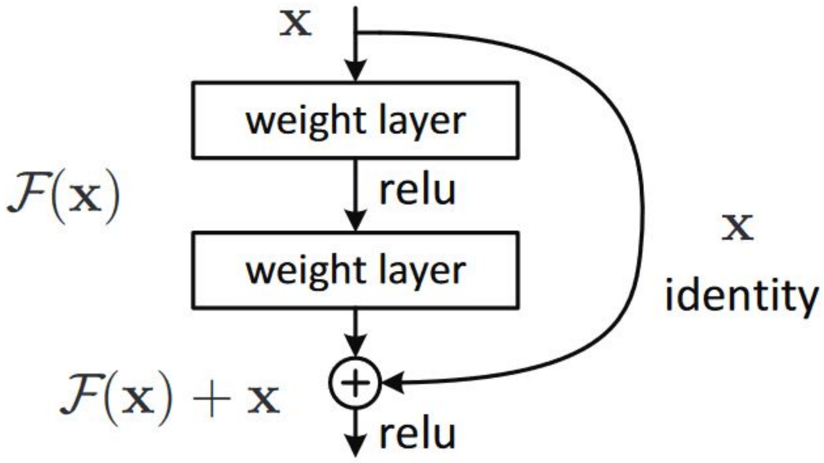


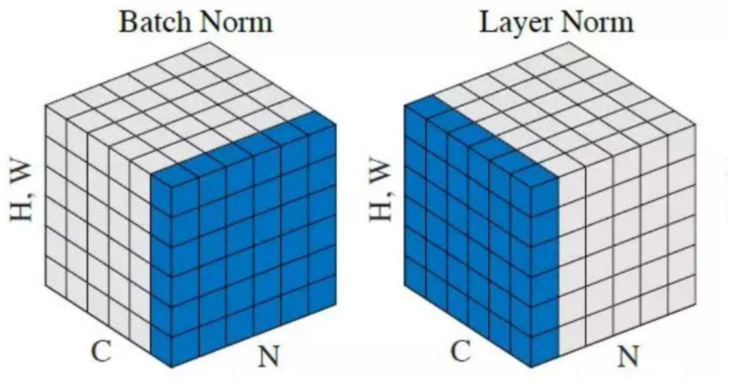
0 :	0	0	0	0	8 :	1	0	0	0
1 :	0	0	0	1	9 :	1	0	0	1
2 :	0	0	1	0	10 :	1	0	1	0
3 :	0	0	1	1	11 :	1	0	1	1
4 :	0	1	0	0	12 :	1	1	0	0
5 :	0	1	0	1	13 :	1	1	0	1
6 :	0	1	1	0	14 :	1	1	1	0
7 :	0	1	1	1	15 :	1	1	1	1





Skip Connections and Layer Normalization



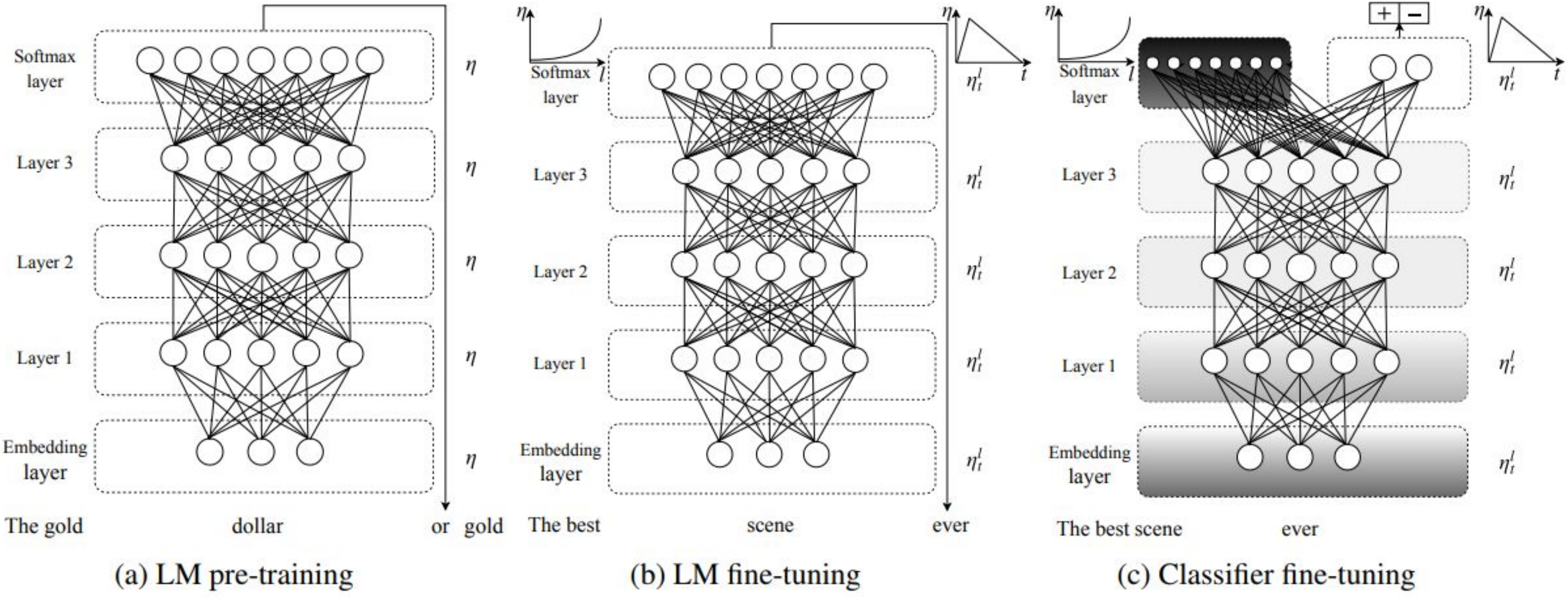


- Help with converge and vanishing gradients
- Preserve positional information

- Improve generalization, reduce covariate shift
- Less movement → Speed up training

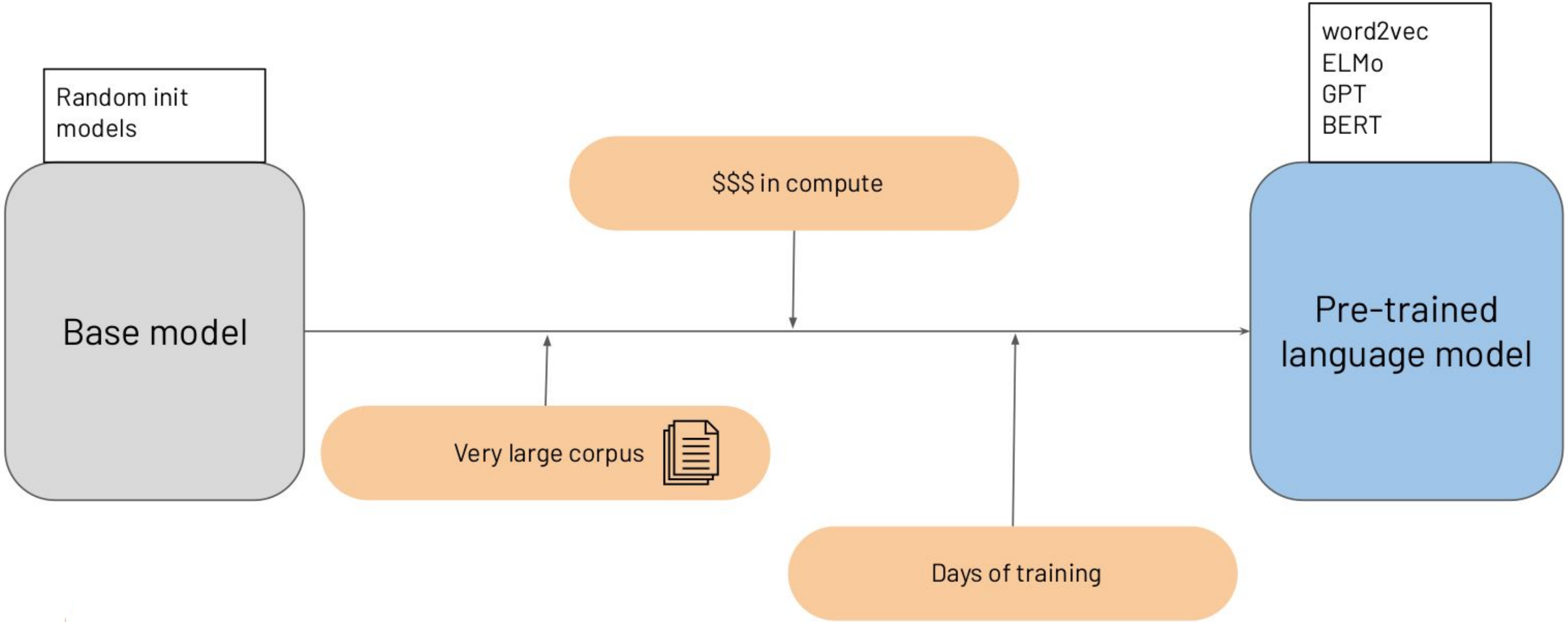
Transfer Learning in NLP

ULMFiT: Universal LM Pre-training & Fine-tuning

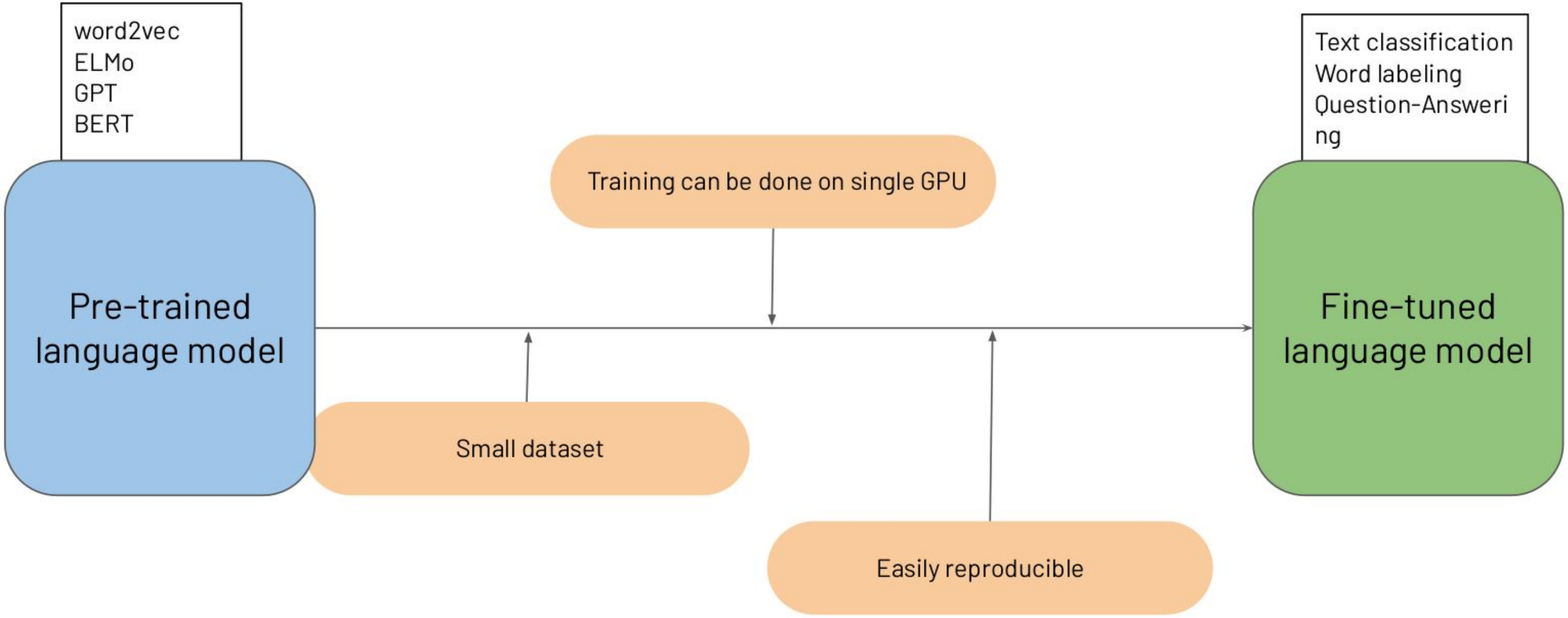


Howard and Ruder 2018

The pre-training procedure



The fine-tuning procedure

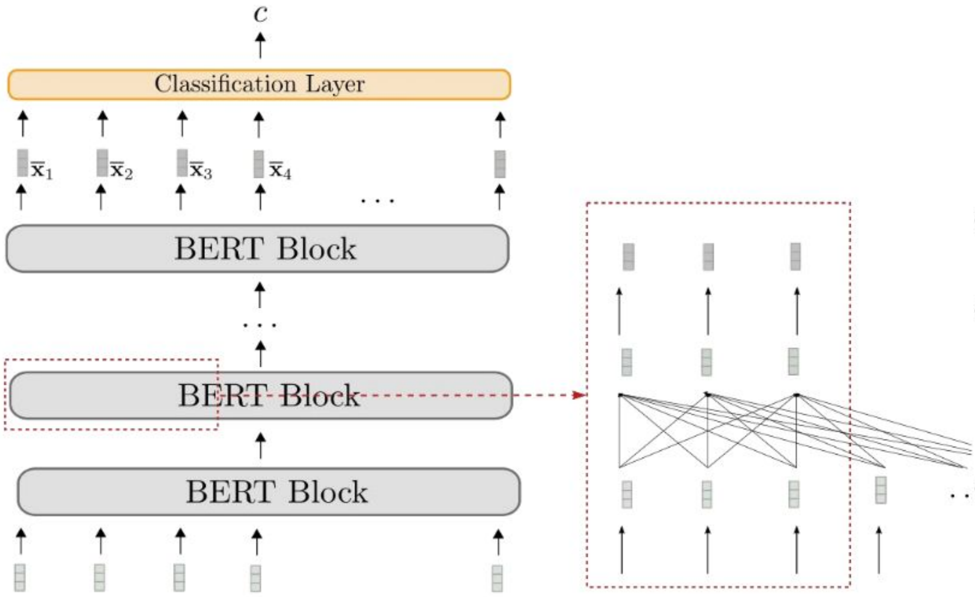


word2vec
ELMo
GPT
BERT

Text classification
Word labeling
Question-Answeri
ng

Pre-trained language model
Training can be done on single GPU
Small dataset
Easily reproducible
Fine-tuned language model

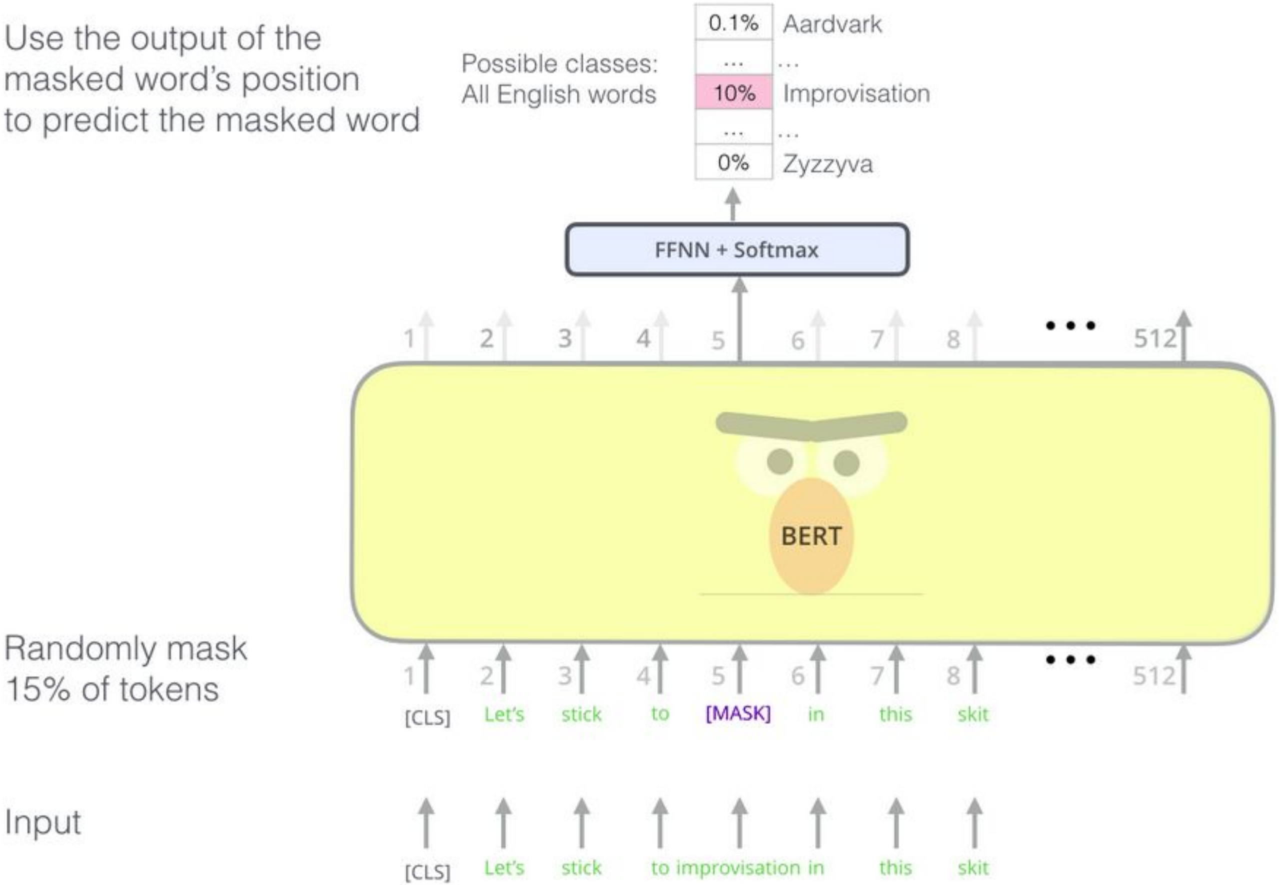
Autoencoding (Masked) Language Models


c
Classification Layer
$\overline{\mathbf{x}}_1$
$\overline{\mathbf{x}}_2$
$\overline{\mathbf{x}}_3$
$\overline{\mathbf{x}}_4$
...
BERT Block
...
BERT Block
BERT Block

- Mathematical Model: P(class | “input seq”)
- Tasks: Natural Language **Understanding** e.g. sentiment classification, named entity recognition, ...
- Prominent Models: BERT, ALBERT, DistilBERT

Devlin et al. 2019, Lan et al. 2019
Sanh et al. 2019

Masked Language Modeling



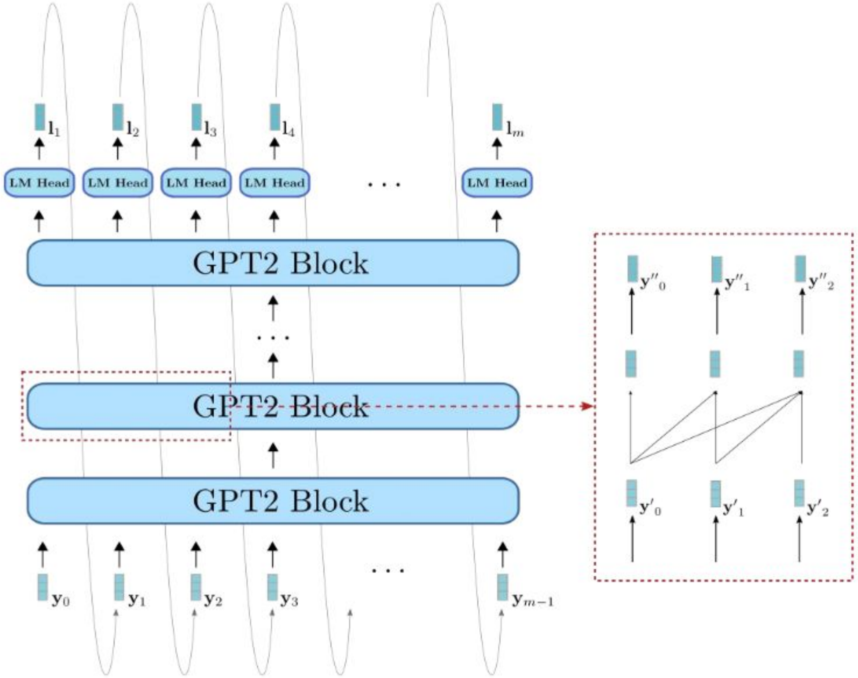
Use the output of the masked word's position to predict the masked word

Possible classes: All English words

Randomly mask 15% of tokens

Input

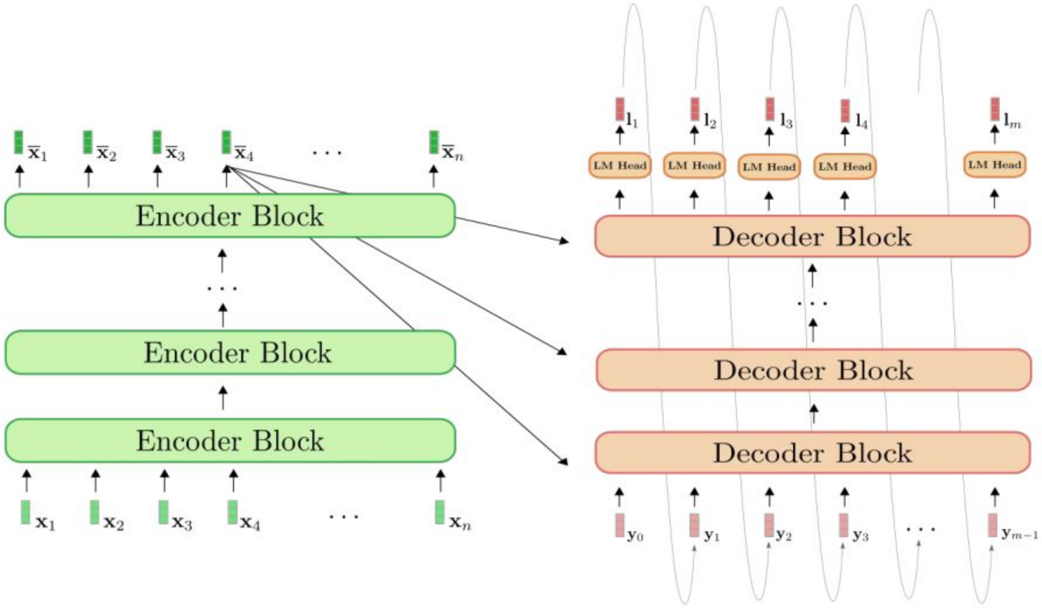
Autoregressive (Causal) Language Models



- Mathematical Model: P(out_seq_i | out_seq_0:i-1)
- Tasks: Natural Language **Generation**, especially open-domain generation
- Prominent Models: GPT1, GPT2, GPT3

Radford et al. 2018, Radford et al. 2019
Brown et al. 2020

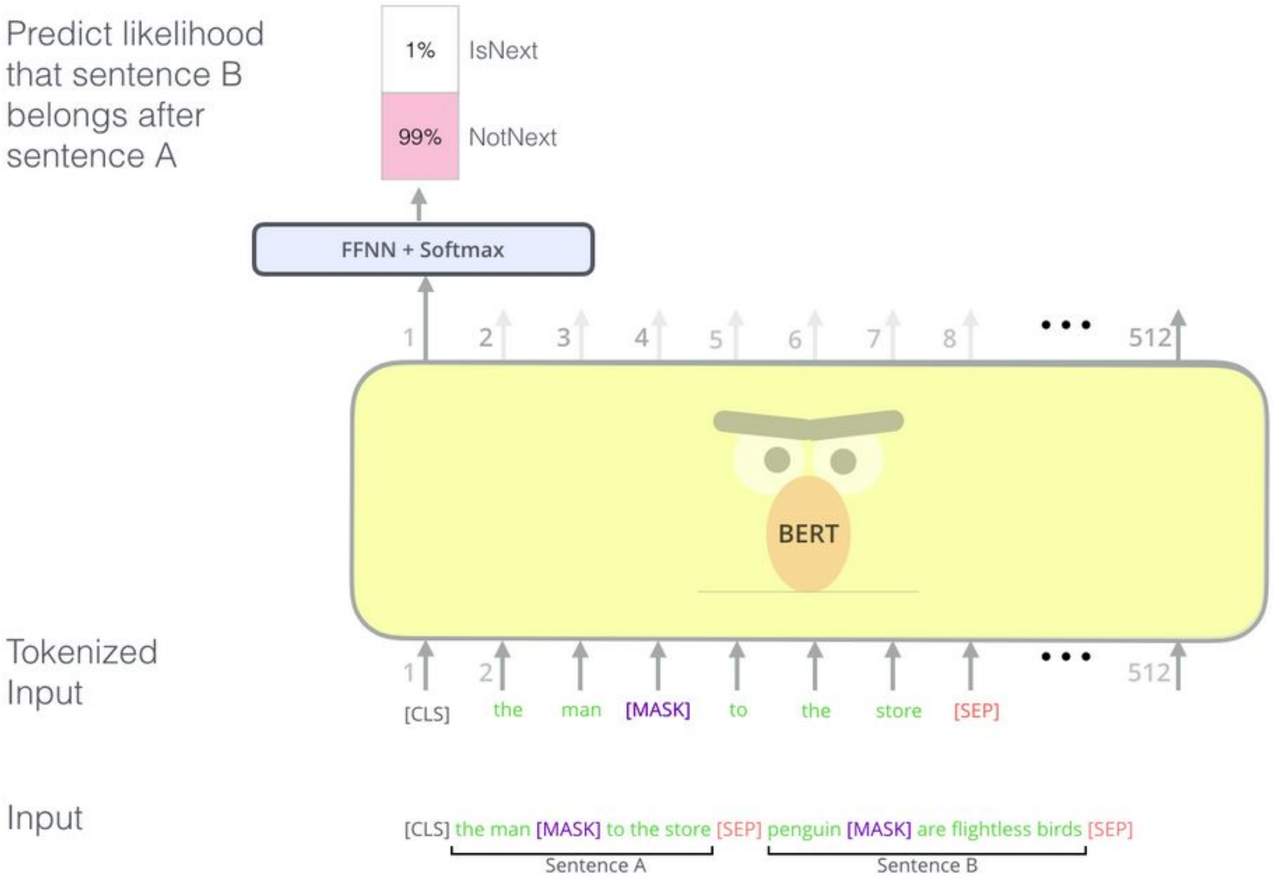
Sequence-to-sequence Language Models



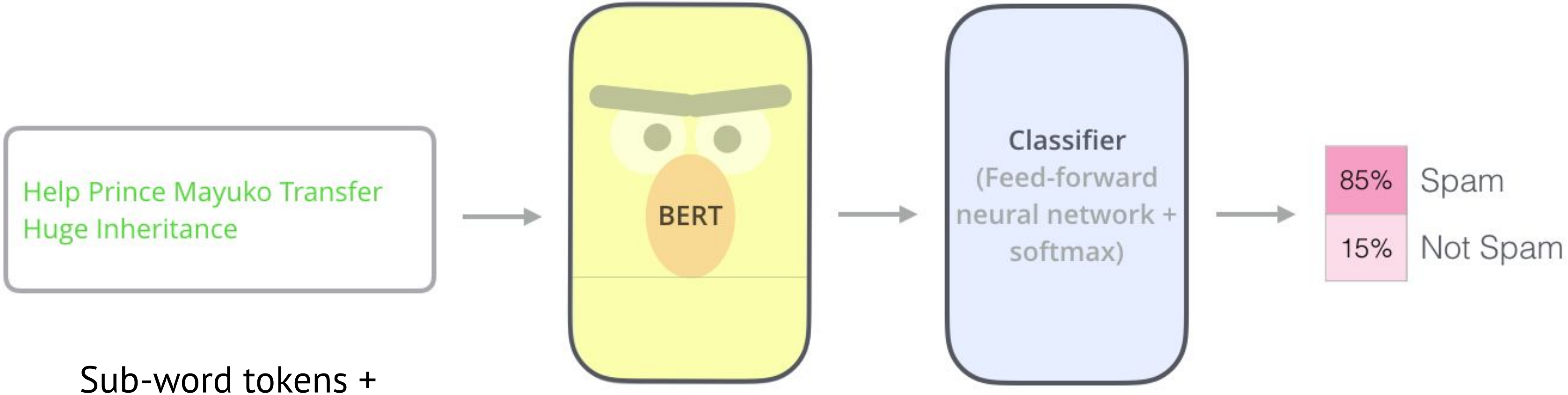
- Mathematical Model: P(out_seq_i | out_seq_0:i-1, in_seq_0:n)
- Tasks: Natural Language **Generation**, especially **Conditioned** Natural Generation (Seq2Seq)
- Prominent Models: T5, BART, Pegasus

Raffel et al. 2019, Zhang et al. 2020, Lewis et al. 2019

Sentence-level Objectives


Predict likelihood that sentence B belongs after sentence A
1%
IsNext
99%
NotNext
FFNN + Softmax
1
2
3
4
5
6
7
8
• • •
512
BERT
Tokenized Input
1
2
• • •
512
[CLS]
the
man
[MASK]
to
the
store
[SEP]
Input
[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]
Sentence A
Sentence B

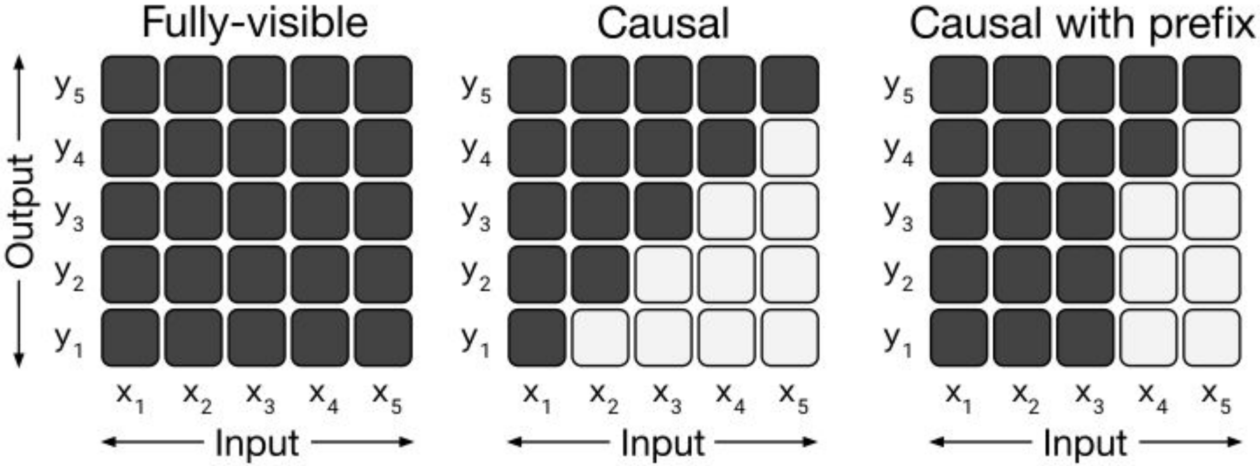
Example: Spam Detection


Help Prince Mayuko Transfer Huge Inheritance
BERT
Classifier
(Feed-forward neural network + softmax)
85% Spam
15% Not Spam

Sub-word tokens +
Attention mask +
Positional Embeddings +
Sentence Id

[CLS]
or
Avg. Pool

Attention Masking 😷



Fully-visible

y_5	■	■	■	■	■
y_4	■	■	■	■	■
y_3	■	■	■	■	■
y_2	■	■	■	■	■
y_1	■	■	■	■	■
	x_1	x_2	x_3	x_4	x_5

Causal

y_5	■	■	■	■	■
y_4	■	■	■	■	□
y_3	■	■	■	□	□
y_2	■	■	□	□	□
y_1	■	□	□	□	□
	x_1	x_2	x_3	x_4	x_5

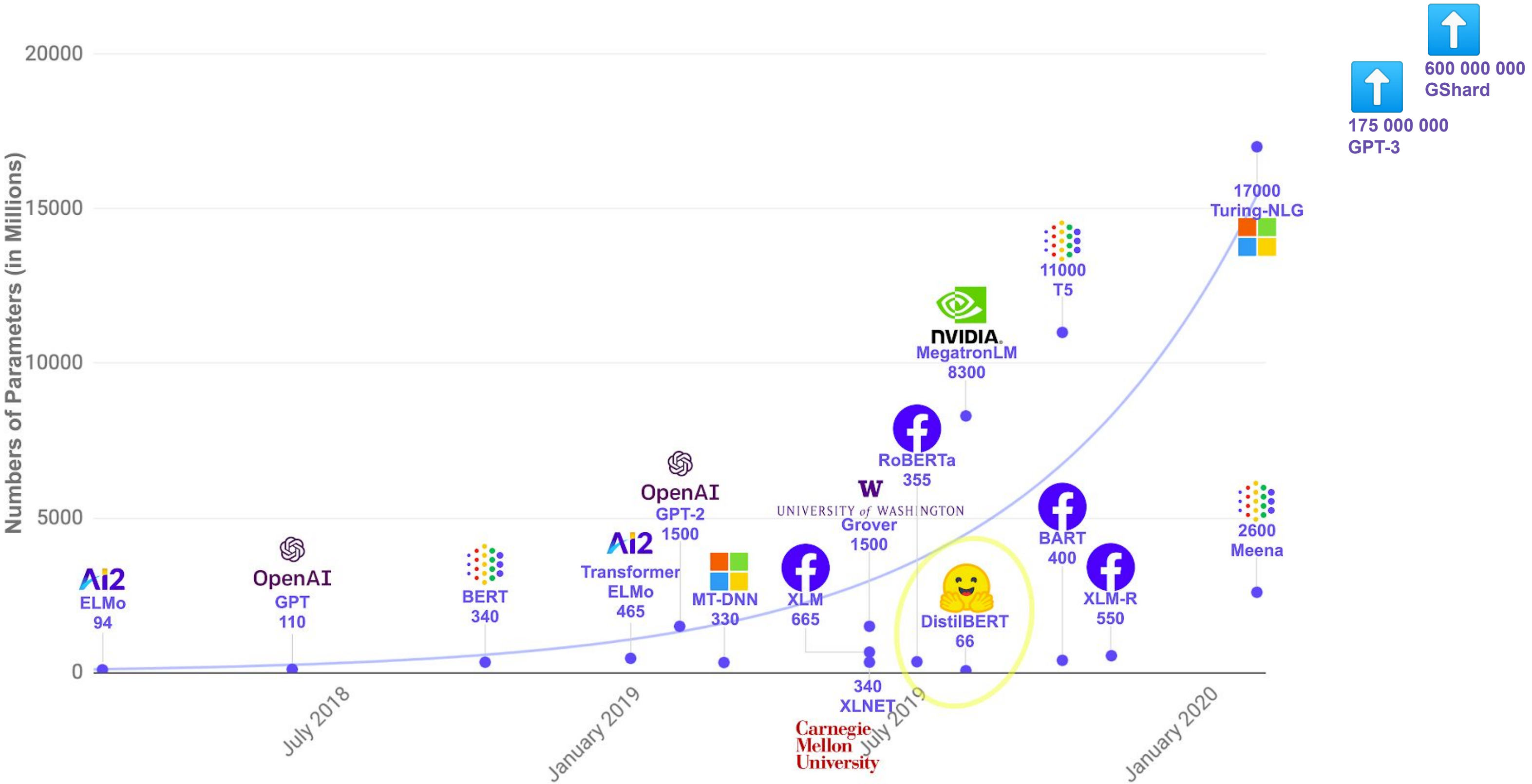
Causal with prefix

y_5	■	■	■	■	■
y_4	■	■	■	■	□
y_3	■	■	■	□	□
y_2	■	■	■	□	□
y_1	■	■	■	□	□
	x_1	x_2	x_3	x_4	x_5

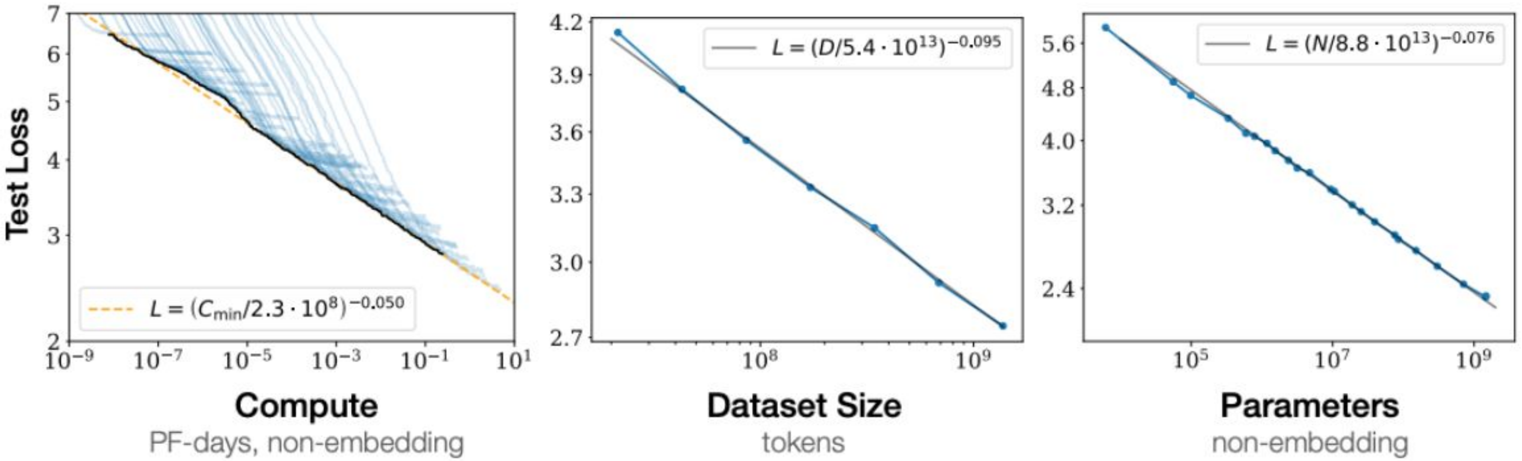
Purpose: Prevent the decoder to attend over future locations

Current Trends in NLP

Ramping up model & data sizes 🚀



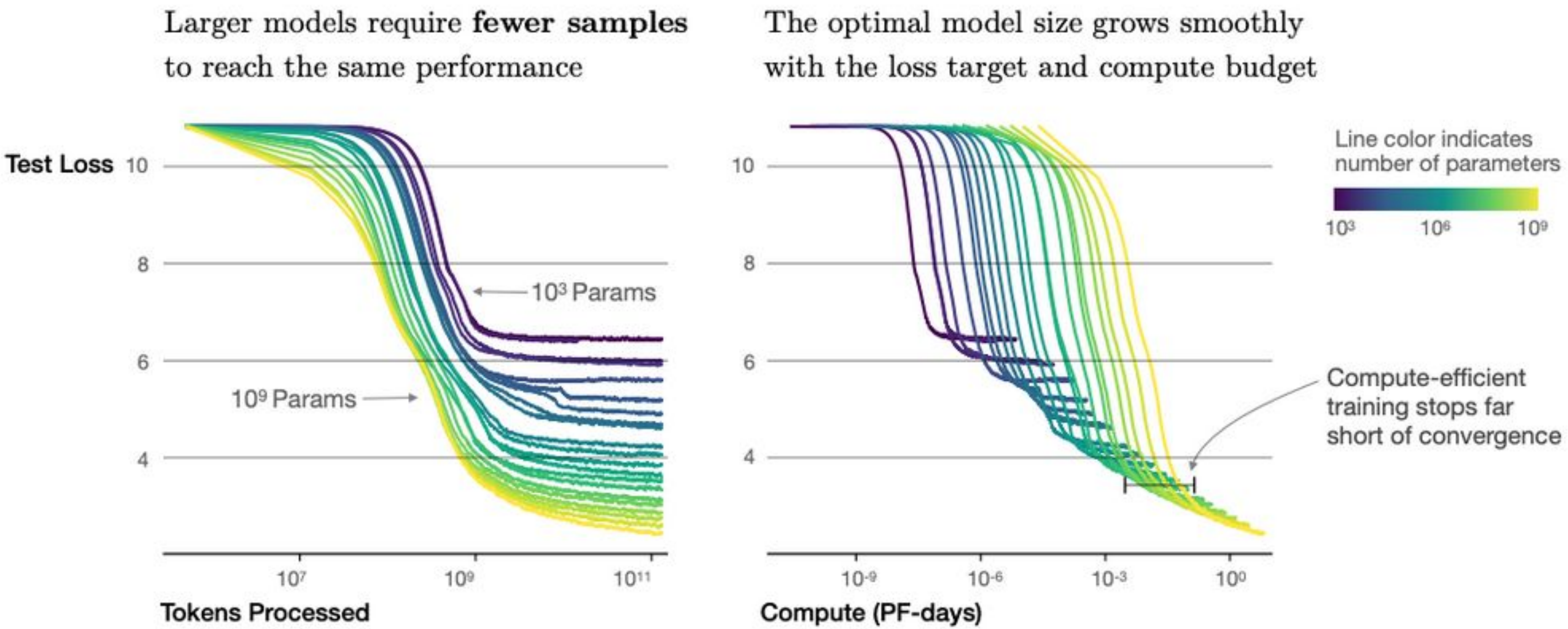
Size Drives Performances



Estimated training costs: **~2M US$** for T5-11B, **>10M US$** for GPT-3

Brown et al. 2020

Large Models are Data Efficient



Larger models require **fewer samples** to reach the same performance

The optimal model size grows smoothly with the loss target and compute budget

Brown et al. 2020

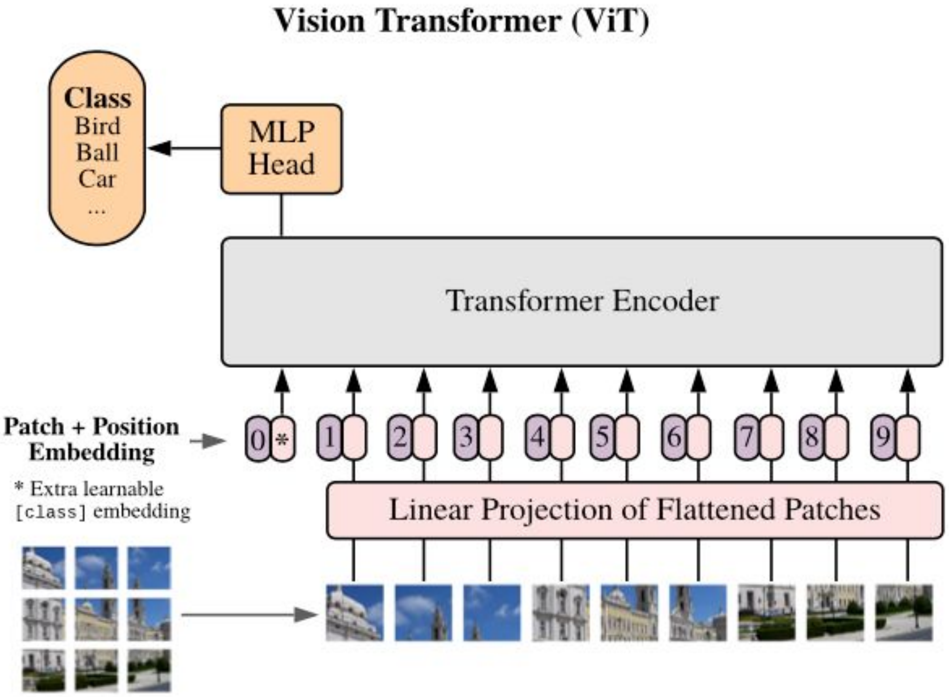
Everything is Text-to-Text 🔤➡️🔤

T5

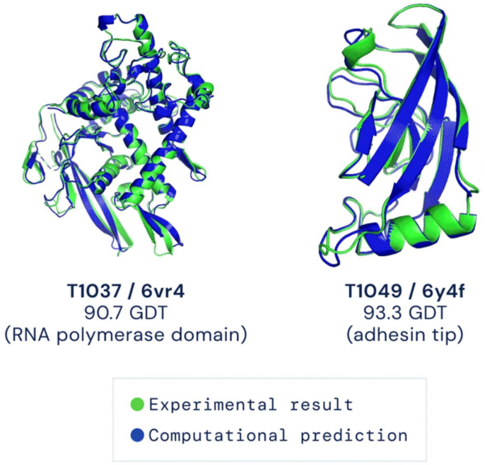
At Aindo we are currently using **UnifiedQA**, a variant of T5 built for unifying different QA formats, for performing structured inference over clinical reports.

Raffel et al. 2019

Transformers Beyond NLP

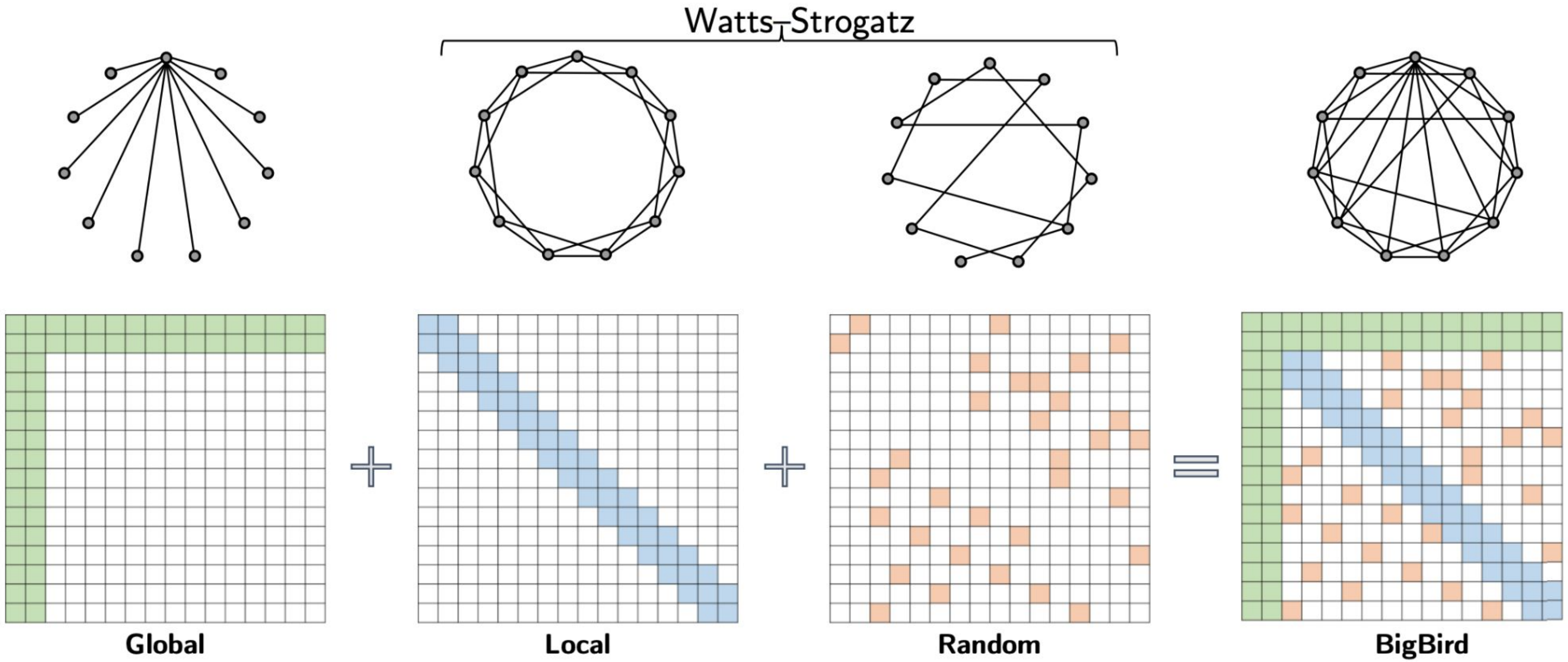

Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

DeepMind



Dosovitskiy et al. 2020, Jumper et al. 2020

Making the Attention Computation Efficient



Zaheer et al. 2020

Multilingual Neural Language Models 🔤

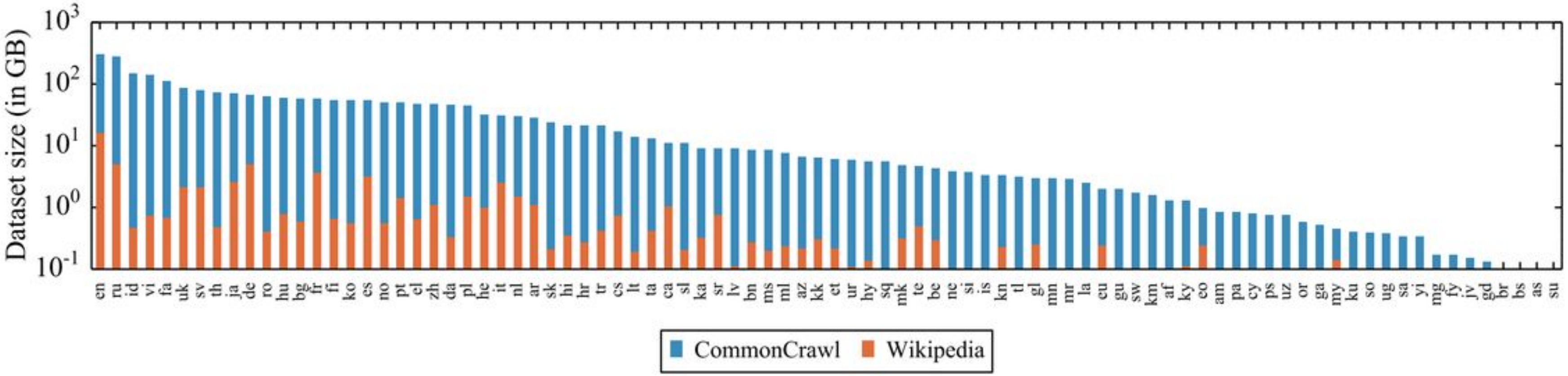
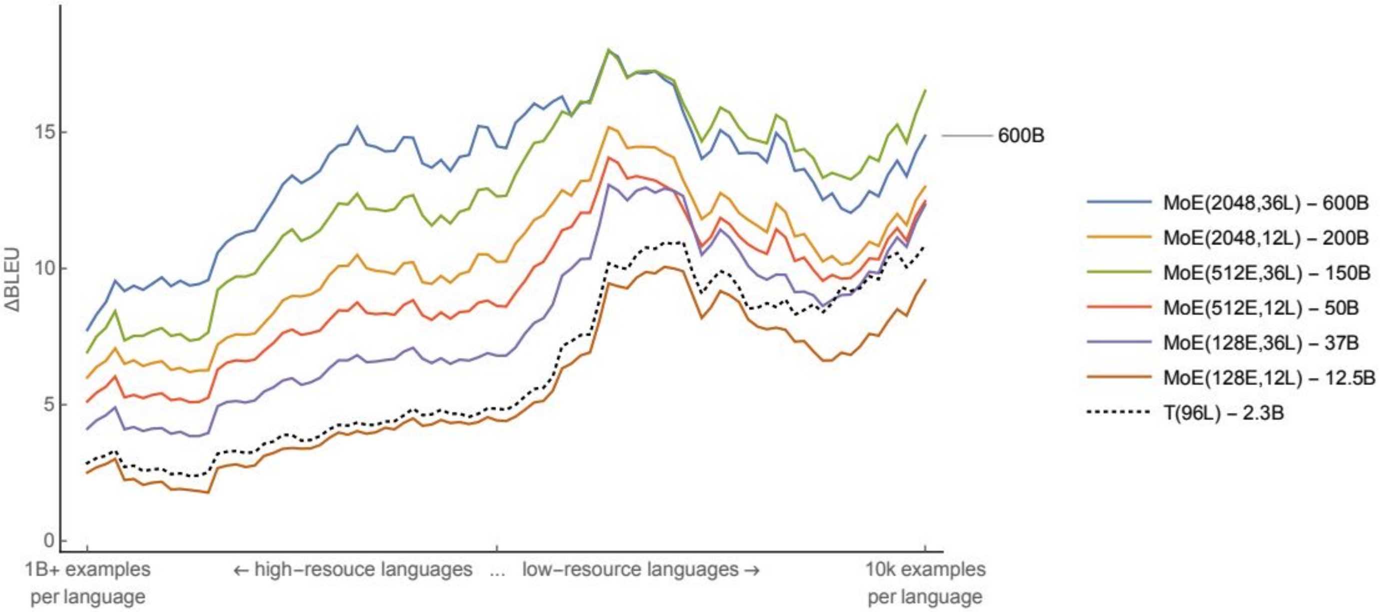


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

Conneau et al. 2020

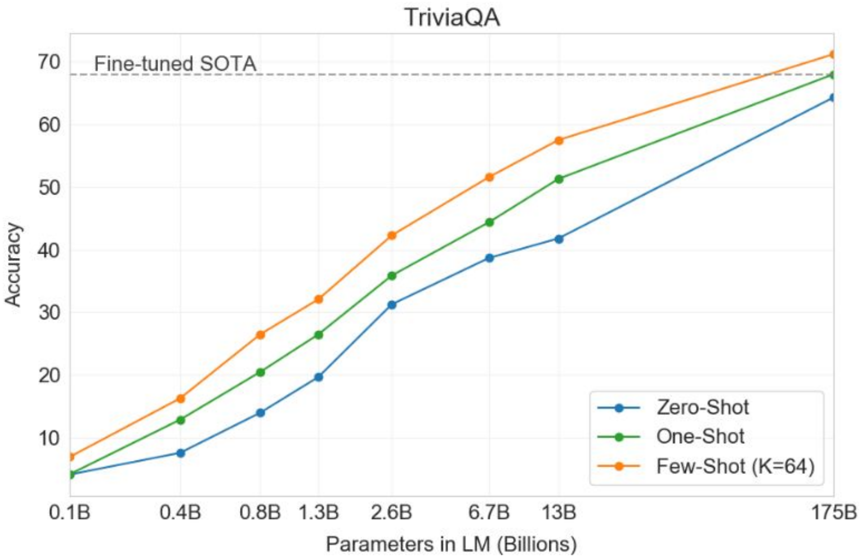
Multilingual Neural Language Models 🔡



While by no means low-resource, Italian is very lacking in terms of datasets. Our research project **TransQA** is aimed at building a model translation pipeline to create new Italian NLMs without retraining.

Conneau et al. 2020

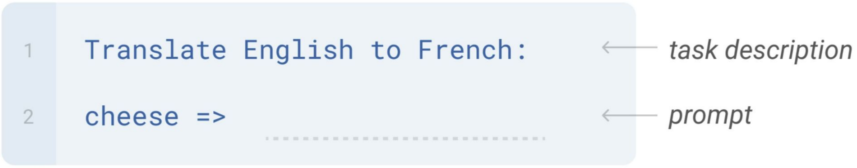
Prompting 📝



Brown et al. 2020, Schick et al. 2020

Zero-shot

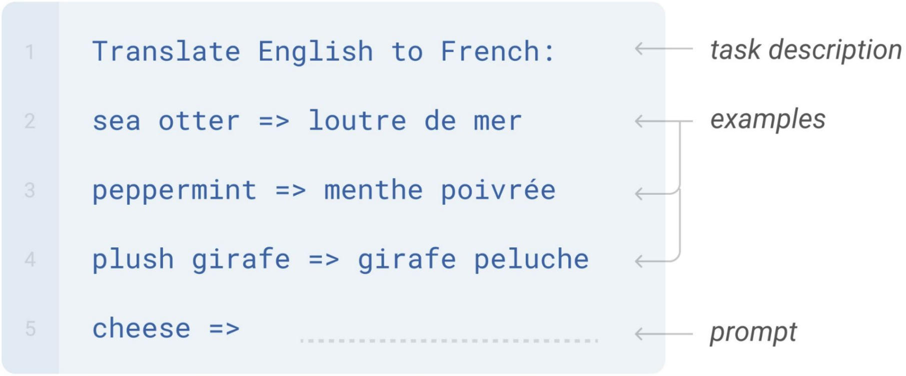
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



```
Translate English to French:
cheese =>
```

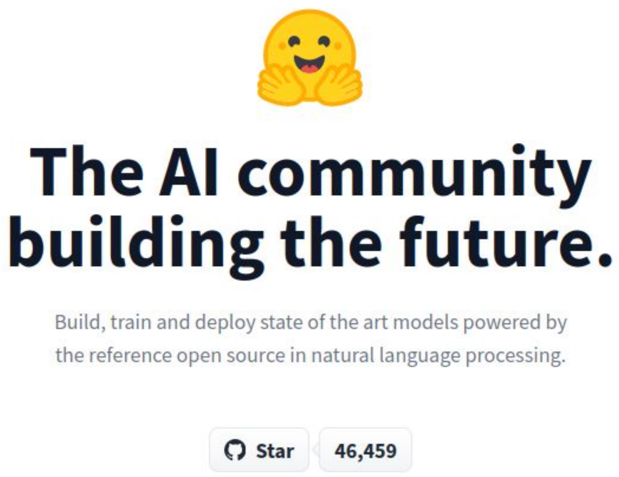
Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

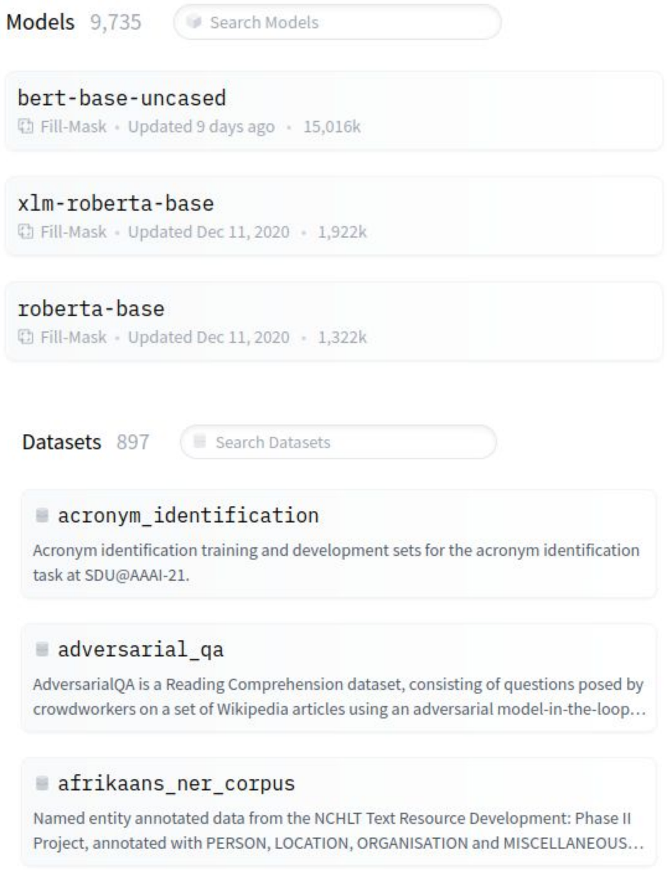
Open Source Communities 🤝



The AI community building the future.

Build, train and deploy state of the art models powered by the reference open source in natural language processing.

Star 46,459



Models 9,735 Search Models

bert-base-uncased
Fill-Mask · Updated 9 days ago · 15,016k

xlm-roberta-base
Fill-Mask · Updated Dec 11, 2020 · 1,922k

roberta-base
Fill-Mask · Updated Dec 11, 2020 · 1,322k

Datasets 897 Search Datasets

acronym_identification
Acronym identification training and development sets for the acronym identification task at SDU@AAAI-21.

adversarial_qa
AdversarialQA is a Reading Comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles using an adversarial model-in-the-loop...

afrikaans_ner_corpus
Named entity annotated data from the NCHLT Text Resource Development: Phase II Project, annotated with PERSON, LOCATION, ORGANISATION and MISCELLANEOUS...

BigScience Workshop 🌸

The Summer of Language Models ‘21

Applications to Software Development 👨‍💻 👩‍💻

Example of using GPT-3 to build React.js apps on the fly.

Other use cases:

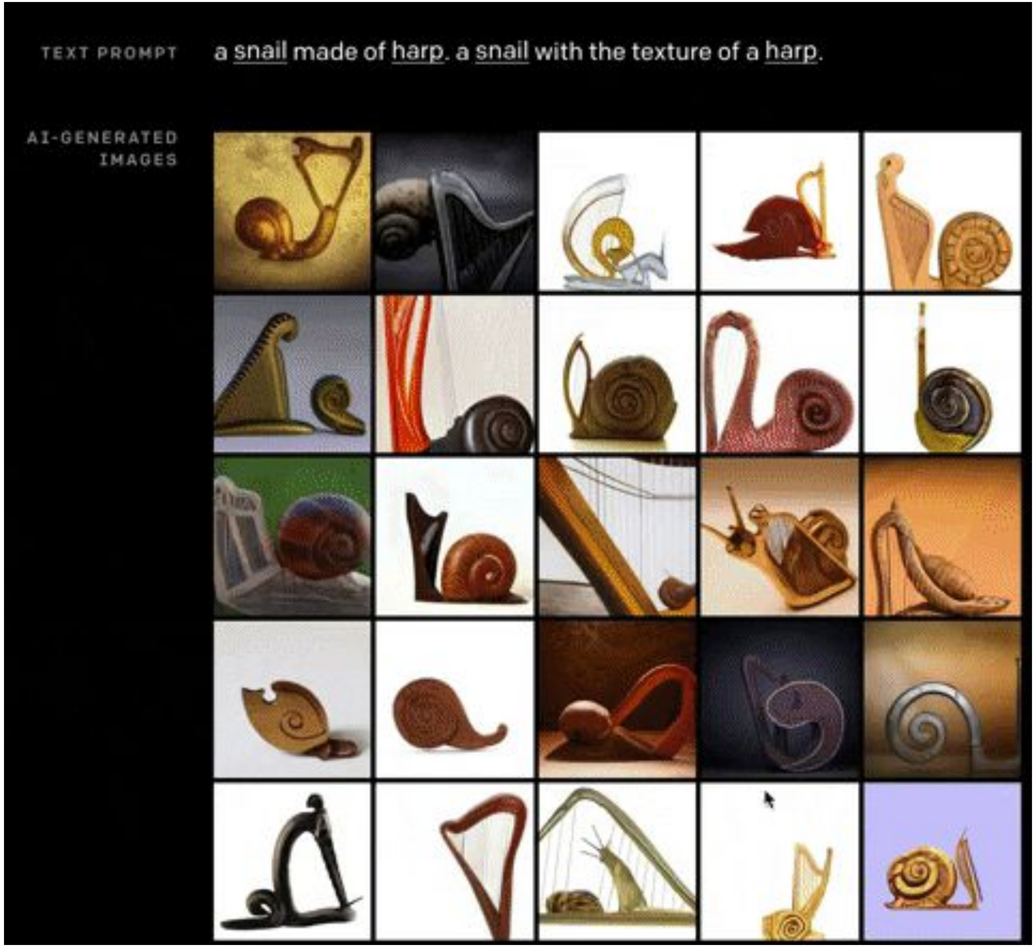
- Debugging
- Programming Language Translation

Language Meets Vision

DALL-E is a 12B version of GPT-3 trained to generate images from text descriptions, using a dataset of text–image pairs.

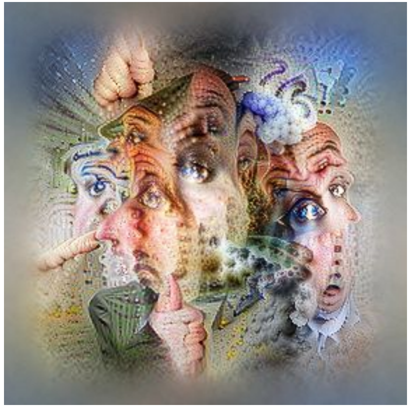
OpenAI

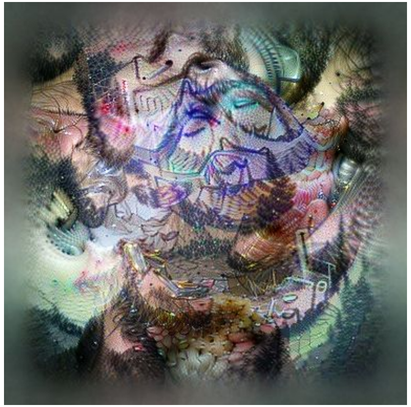
Ramesh et al. 2021

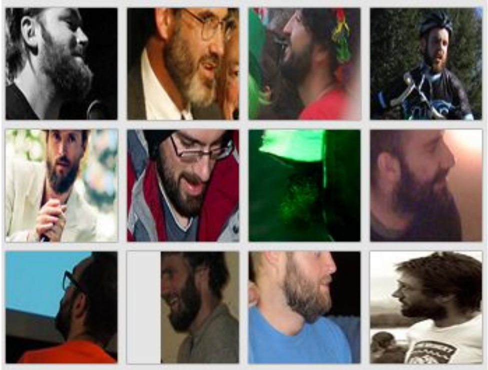


TEXT PROMPT a snail made of harp. a snail with the texture of a harp.

Multimodal Neurons for Language and Vision







CLIP ResNet 50 4x
“Comedy neuron”

0.40 the comedy circus geek challenge ! – lori spicer ,
0.40 the comedy circus geek challenge ! – lori spicer ,
0.40 new review :@ funlens - duh !
0.40 new review :@ funlens - duh !
0.40 new review :@ funlens duh !
0.39 # tax lien comedy faq - the big one !
0.39 # tax lien double comedy : the big one !

CLIP ResNet 50 16x
Unit 2,298
“Beard neuron”

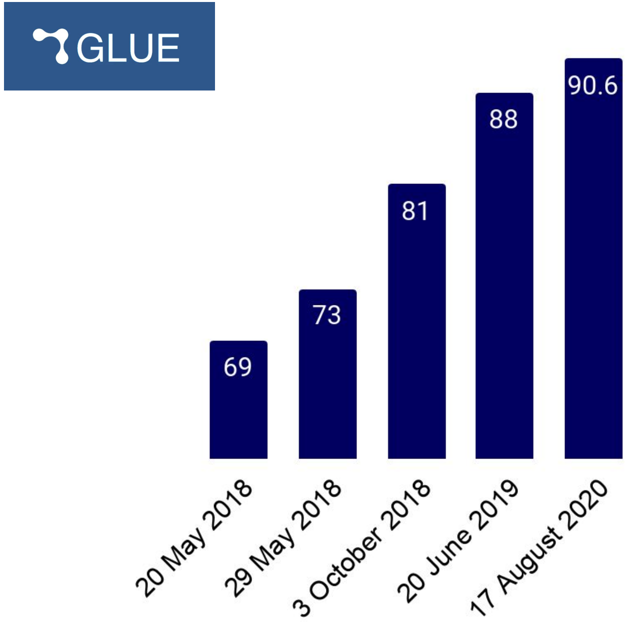
0.42 strong beard dynamo ! # iter # pler 4
0.42 truebeardchampionship , torpemiento , wpf
0.42 truebeardchampionship , torpemiento , wpf
0.41 # tempe imam salah mirza gani 's dispositions are
0.41 strong beards # wethepeople family love to keep
0.41 strong beards # wethepeople family love to keep
0.41 beard dynamo ! # iter # kepler 4
0.41 truebeardroad . facebook en movimiento

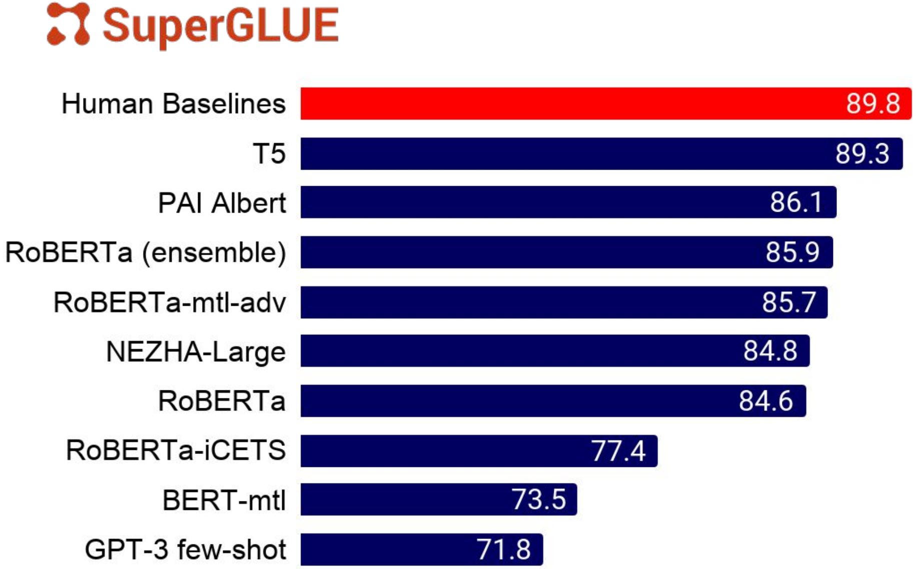
OpenAI

Goh et al. 2021

Current Limitations

Build Better Benchmarks 🏋️





Wang et al. 2018, 2019

The Dangers of Stochastic Parrots 🦜

1. Massive data, inscrutable models

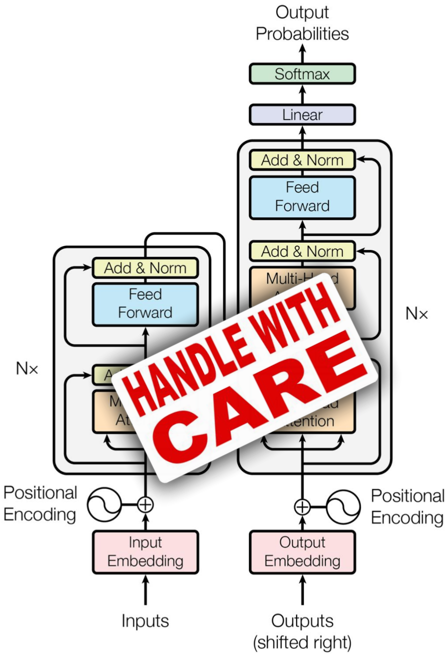
Models reflect the biases present in their training data. Undocumented data are risky.

2. Manipulating language is not understanding it

The financial interest in NLP is only in producing the best model. More effort should be devoted to **curation**, **interpretability** and **efficiency.**

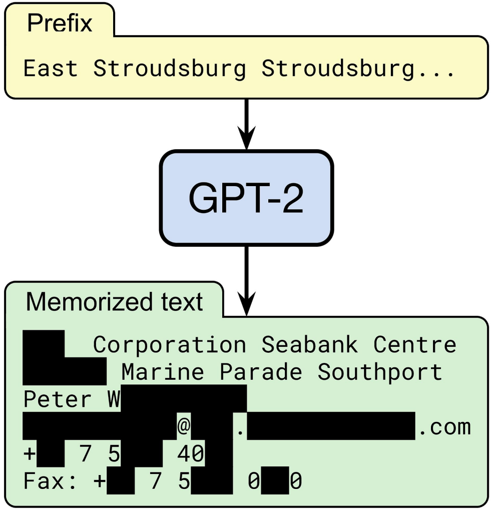
3. The illusion of meaning

Models fluent in generating language are at best morally dubious, at worst a threat to our society and our democracy.

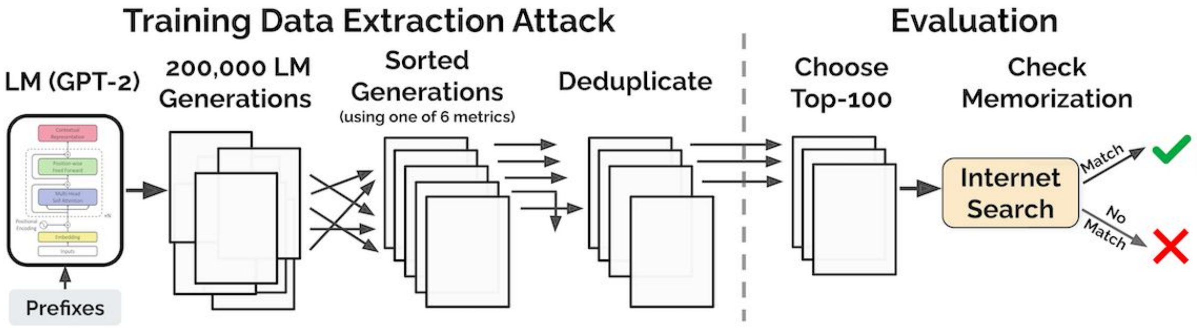


Bender et al. 2021

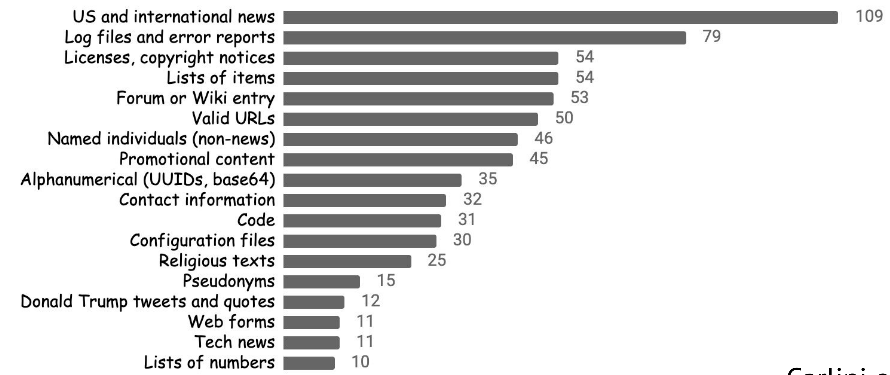
Generalization or Memorization?



```
[redacted] Corporation Seabank Centre
 Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```



Categorization of memorized data



Carlini et al. 2020

AINDO

Thanks for the

$$softmax(\frac{QK^T}{\sqrt{d_k}})V \quad !$$

@gsarti_

gabrielesarti

gsarti.com

gsarti

gabriele.sarti996@gmail.com

References

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

Mikolov, Tomas, et al. "Advances in pre-training distributed word representations." *arXiv preprint arXiv:1712.09405* (2017).

Peters, Matthew E., et al. "Deep contextualized word representations." *arXiv preprint arXiv:1802.05365* (2018).

Hooker, Sara. "The hardware lottery." *arXiv preprint arXiv:2009.06489* (2020).

Vaswani, Ashish, et al. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." *arXiv preprint arXiv:1801.06146* (2018).

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." *arXiv preprint arXiv:1909.11942* (2019).

Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019).

Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

Brown, Tom B., et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).

Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *arXiv preprint arXiv:1910.10683* (2019).

Zhang, Jingqing, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." *International Conference on Machine Learning*. PMLR, 2020.

Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019).

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

Jumper, John, et al. "High accuracy protein structure prediction using deep learning." *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)* 22 (2020): 24.

Zaheer, Manzil, et al. "Big bird: Transformers for longer sequences." *arXiv preprint arXiv:2007.14062* (2020).

Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." *arXiv preprint arXiv:1911.02116* (2019).

Schick, Timo, and Hinrich Schütze. "Few-Shot Text Generation with Pattern-Exploiting Training." *arXiv preprint arXiv:2012.11926* (2020).

Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *arXiv preprint arXiv:2102.12092* (2021).

Goh, Gabriel, et al. "Multimodal neurons in artificial neural networks." *Distill* 6.3 (2021): e30.

Wang, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." *arXiv preprint arXiv:1804.07461* (2018).

Wang, Alex, et al. "Superglue: A stickier benchmark for general-purpose language understanding systems." *arXiv preprint arXiv:1905.00537* (2019).

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021.

Carlini, Nicholas, et al. "Extracting Training Data from Large Language Models." *arXiv preprint arXiv:2012.07805* (2020).