From Word Embeddings To Large Language Models

Lecture 11-supp: Text Analytics for Big Data

Mark Keane, Insight/CSI, UCD

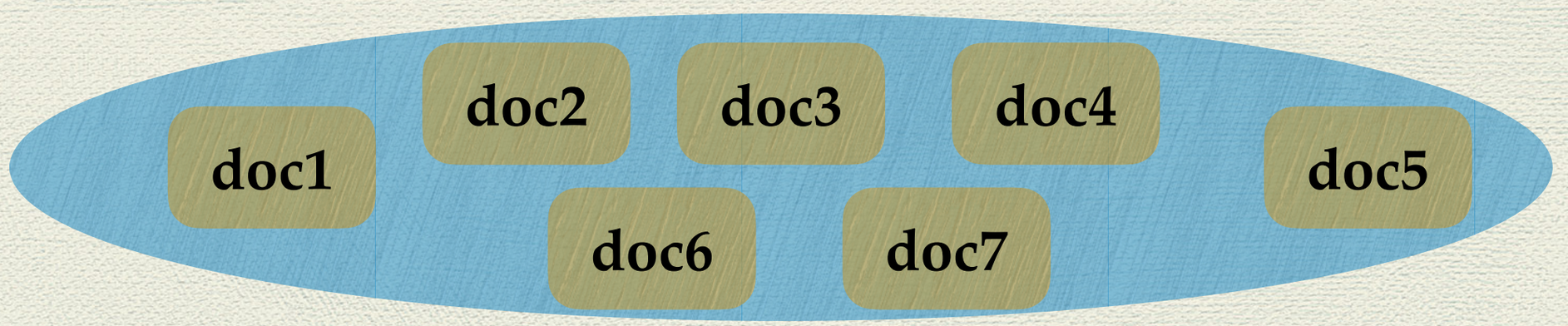
History

- Text Analytics discovers Latent Features
- *Word Embeddings* edge semantics problem
- Deep Learning and Sampling advances drive major breakthroughs
- Google gets in on the Act !
- Large Language Models (LLMs) are born

Historical Background

Latent Features (recall)

- A word's meaning is determined by the company it keeps…meaning comes from distributional info
- By analysing how words co-occur with one another, we can get semantics from text
- LSA, LSI all used windowing over TF-IDF-ed data for largish corpora
- All dimensionality reduction; PCA, SVD…

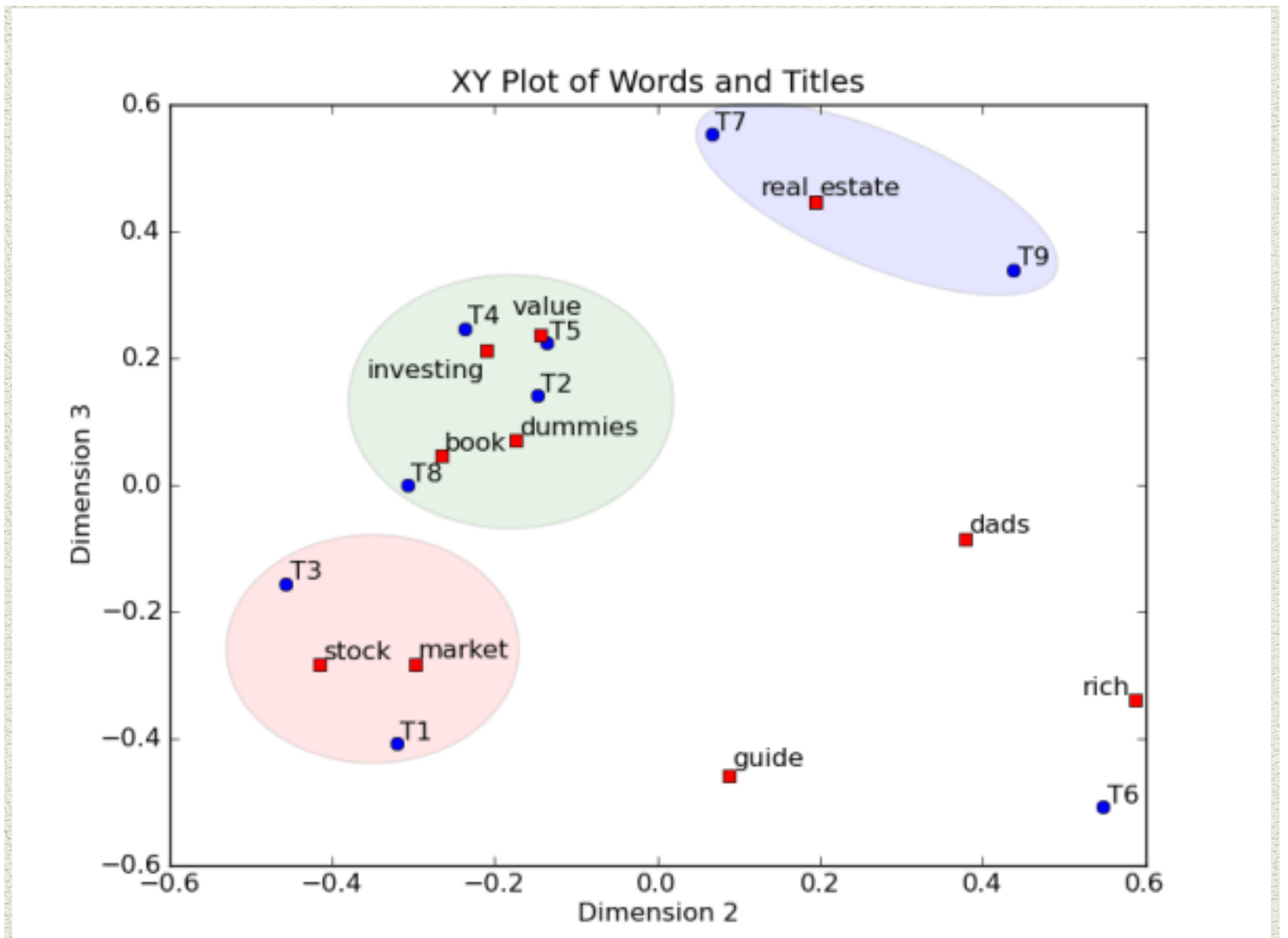


feel-good words will be correlated in their effects on the predicted outcome (**pos**), and **feel-bad words** in their effects on predicted outcome (**neg**); BUT the effects of **feel-good** and **feel-bad words** will be uncorrelated (because they are a different class of variables): Components/Factors = FEEL-GOOD/FEEL-BAD

From This to This…

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

book	0.15	-0.27	0.04
dads	0.24	0.38	-0.09
dummies	0.13	-0.17	0.07
estate	0.18	0.19	0.45
guide	0.22	0.09	-0.46
investing	0.74	-0.21	0.21
market	0.18	-0.3	-0.28
real	0.18	0.19	0.45
rich	0.36	0.59	-0.34
stock	0.25	-0.42	-0.28
value	0.12	-0.14	0.23



Better Embeddings
Modern Era
(~post-2010)

Embeddings: Prediction

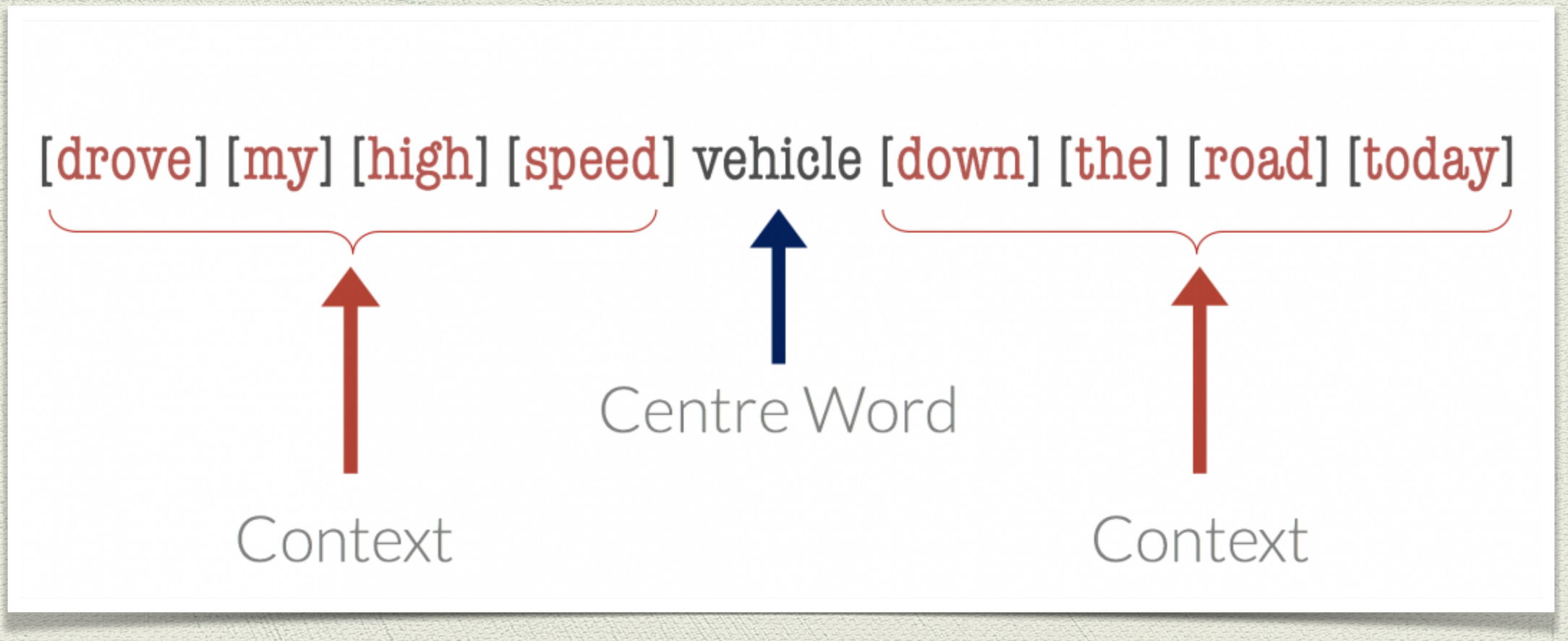
- Collobert & Weston (2008) using deep learning methods; archane, hard to compute
- Mikilova et al (2013) word2vec brings it to the masses (softmax, CBOW, Skip-gram)
- Vectors are better than ones from LSA-Count
- Area explodes with many different techniques

Mikolov, T., Chen, K., Corrado, G. and Dean J. (2013). Efficient estimation of word representations in vector space. CoRR, abs/1301.3781.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

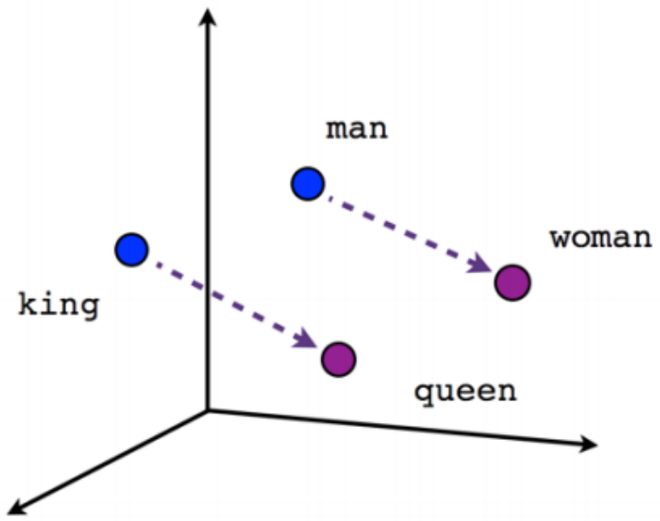
Embeddings: Predictions

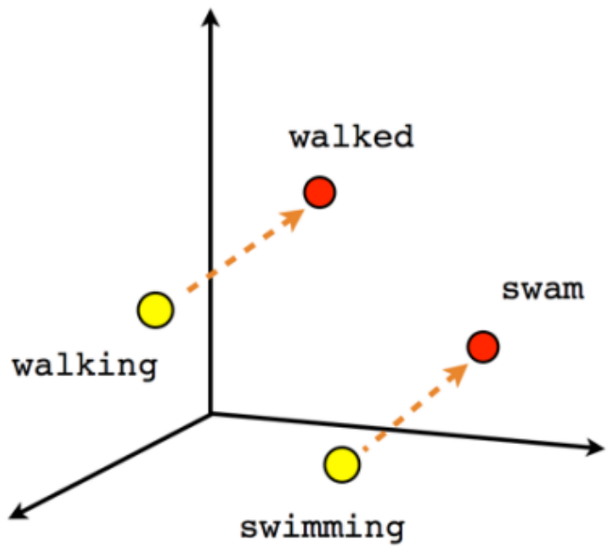
So, can we learn to predict a word from the context-words around it or predict the contexts from the word ?

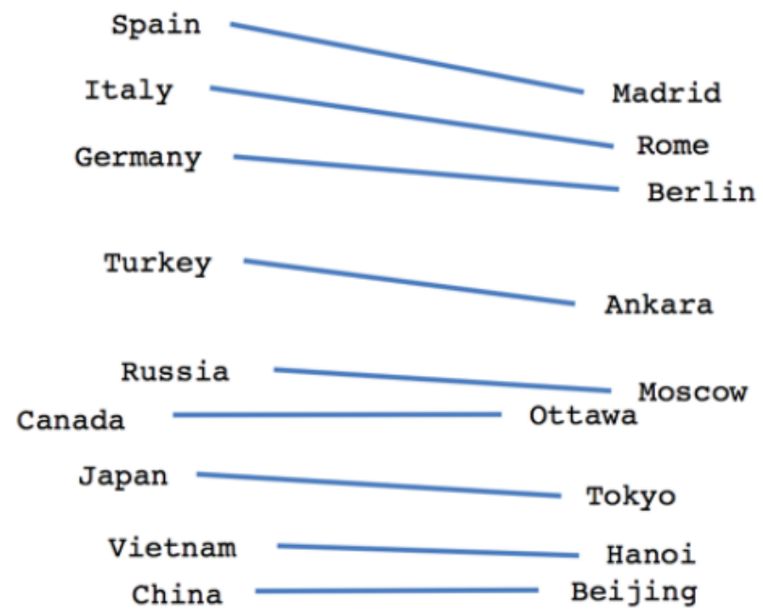


Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*(Aug), 2493-2537.

Using vectors: Trajectories..



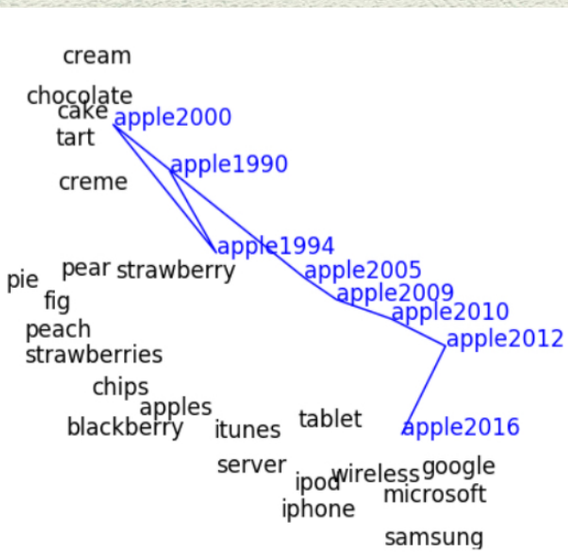




Male-Female

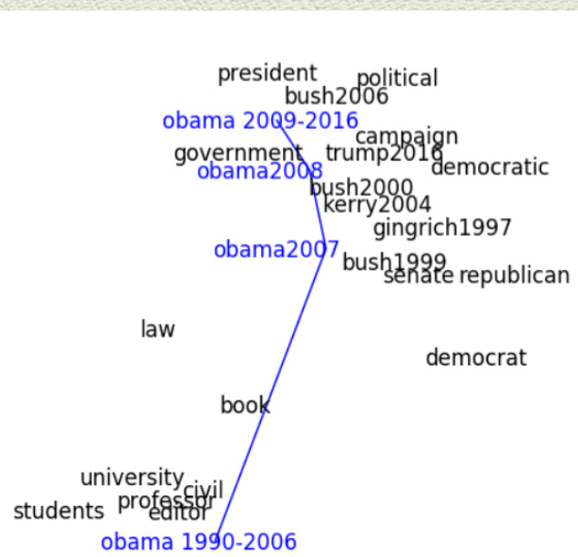
Verb tense

Country-Capital

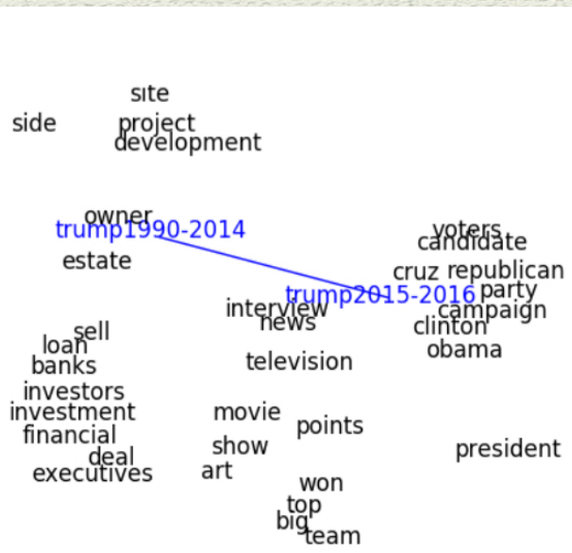


(a) apple

(b) amazon

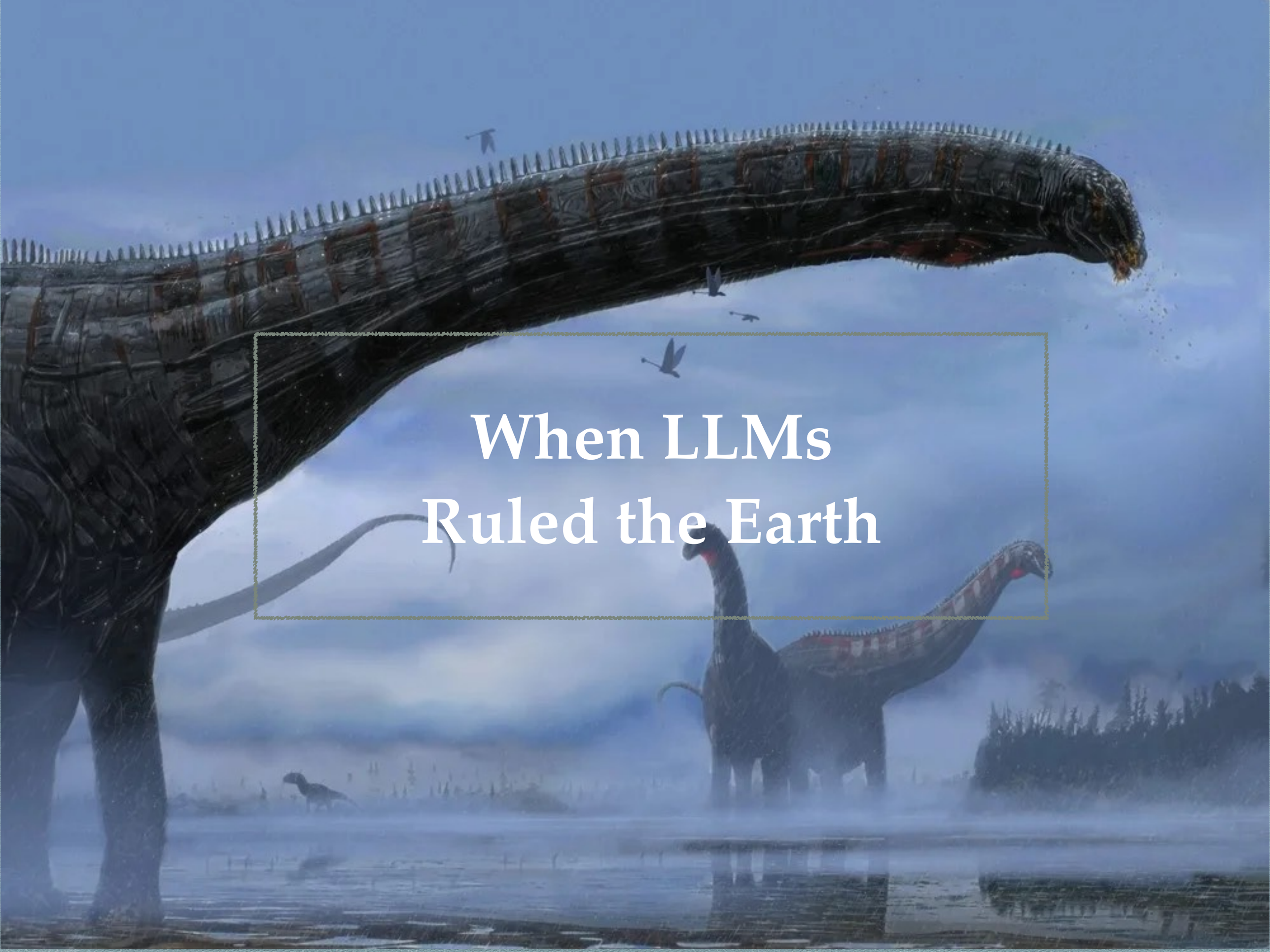


(c) obama



(d) trump

Figure 1: Trajectories of brand names and people through time: `apple`, `amazon`, `obama`, and `trump`.



When LLMs Ruled the Earth

New Tasks

- Word embeddings inspires encodings for Large Language Models (LLMs)
- Deep Learning meets Text Analytics
- Impressive results follow…

LLMs: 5 Innovations

1. *New Architectures*: RNNs, LSTM, Encoders,Decoders, Encoder-Decoders, Attention and so on…

2. *New Embeddings*: use more context (l->r, r->l) and combined vectors (word, positional/sentence, paras)

3. *Sequence Learning*: becomes task…predicting next token in sequence or predict seq_2 from seq_1

4. *Transfer Learning*: Pre-trained models on one task to do another task (re-use weights, not random)

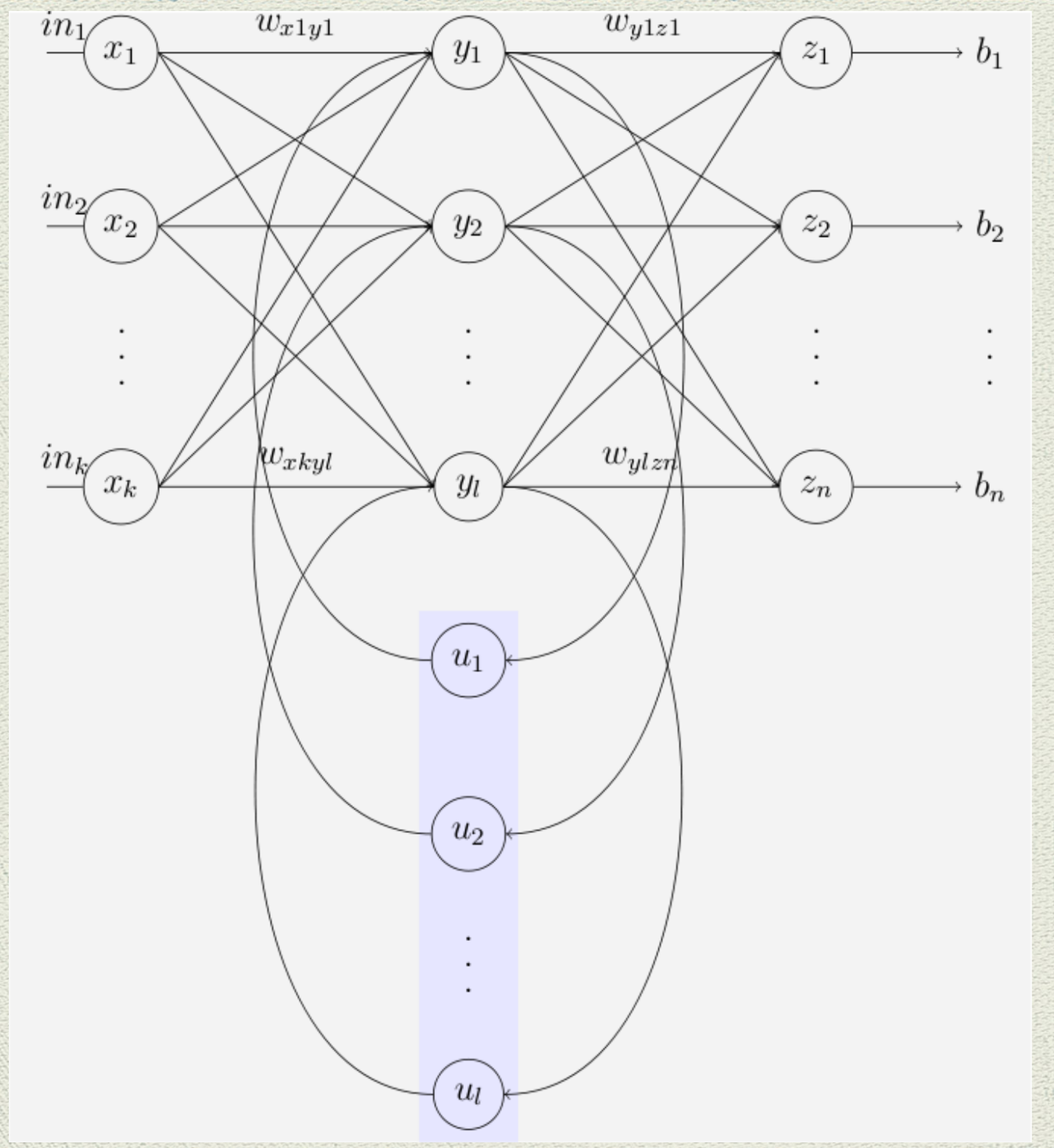
5. *Bigger*: More and more and more data and parameters

LLMs#1: New Architectures

- Traditionally, sequence learning was done by RNNs and LSTMs but encounter problems
- New models emerge with a lot of variations, as BIG money thrown at it…
- Transformers: Semi-supervised methods, training+tuning, self-attention and so on…

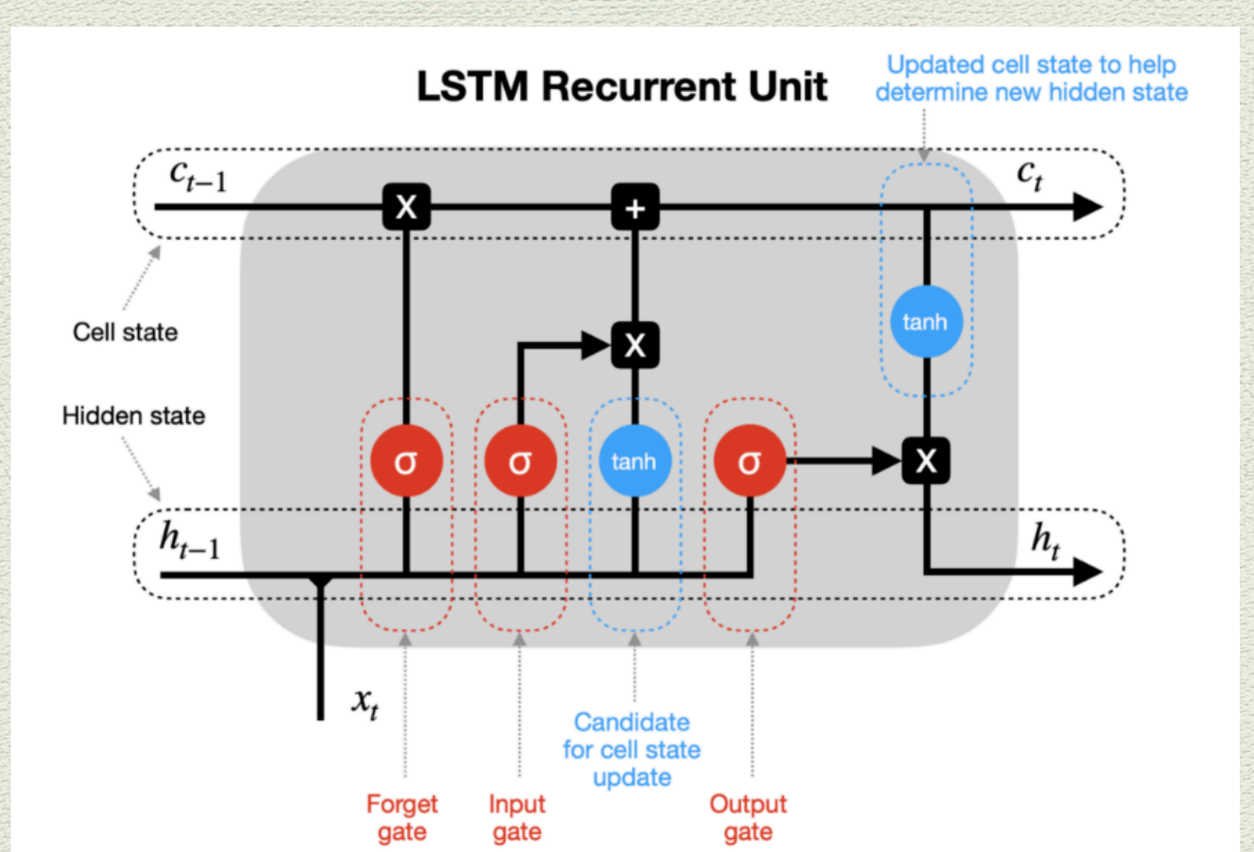
LLMs#1: Early Models:RNNs

Recurrent Neural Networks (RNNs) were used to learn sequences…context layer ($u_1 \ldots u_l$) provides temporal memory for network, so previous hidden-layer states "remembered"



LLMs#1: Early Models:LSTMs

Long Short Term Memory (LSTMs) have a STM for an RNN over 1000s of timesteps (i.e., long): LSTM unit has cell, input gate, output gate and forget gate…



LSTM cell remembers values over arbitrary time intervals and 3 gates regulate flow of information into and out of the cell.

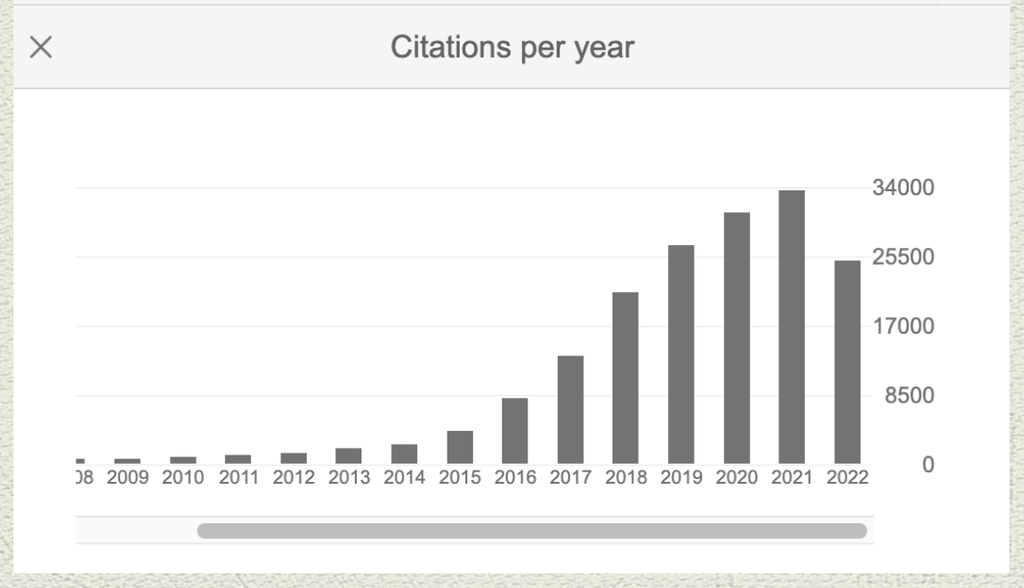
RNNs…LSTMs…

"the dog ran off, when *it* was chased by the cat down the street"

"the cat ran off, when the dog chased down the street after *it*"

- Long-term dependencies
- Models suffer **vanishing** or **exploding gradients**…
- Vanishing, gradients tend to zero, or exploding tend to infinity as they use finite-precision numbers
- LSTMs (1997->) improved on this somewhat; use extensively in smart phone speech recognition

Juergen Schmidhuber



LLMs#1: …Transformers

- Transformers, aimed to solve RNNs issues, without sequential recurrent processing

- They use attention mechanisms: transformer blocks learn word embeddings (good reps), with self-attention focussing on important parts

- Complex multi-layers models, with huge vectors and nos of parameters; encoder-decoder

LLMs#1: …Tranformers

Attention Is All You Need

55.5k cites, Oct-2022

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

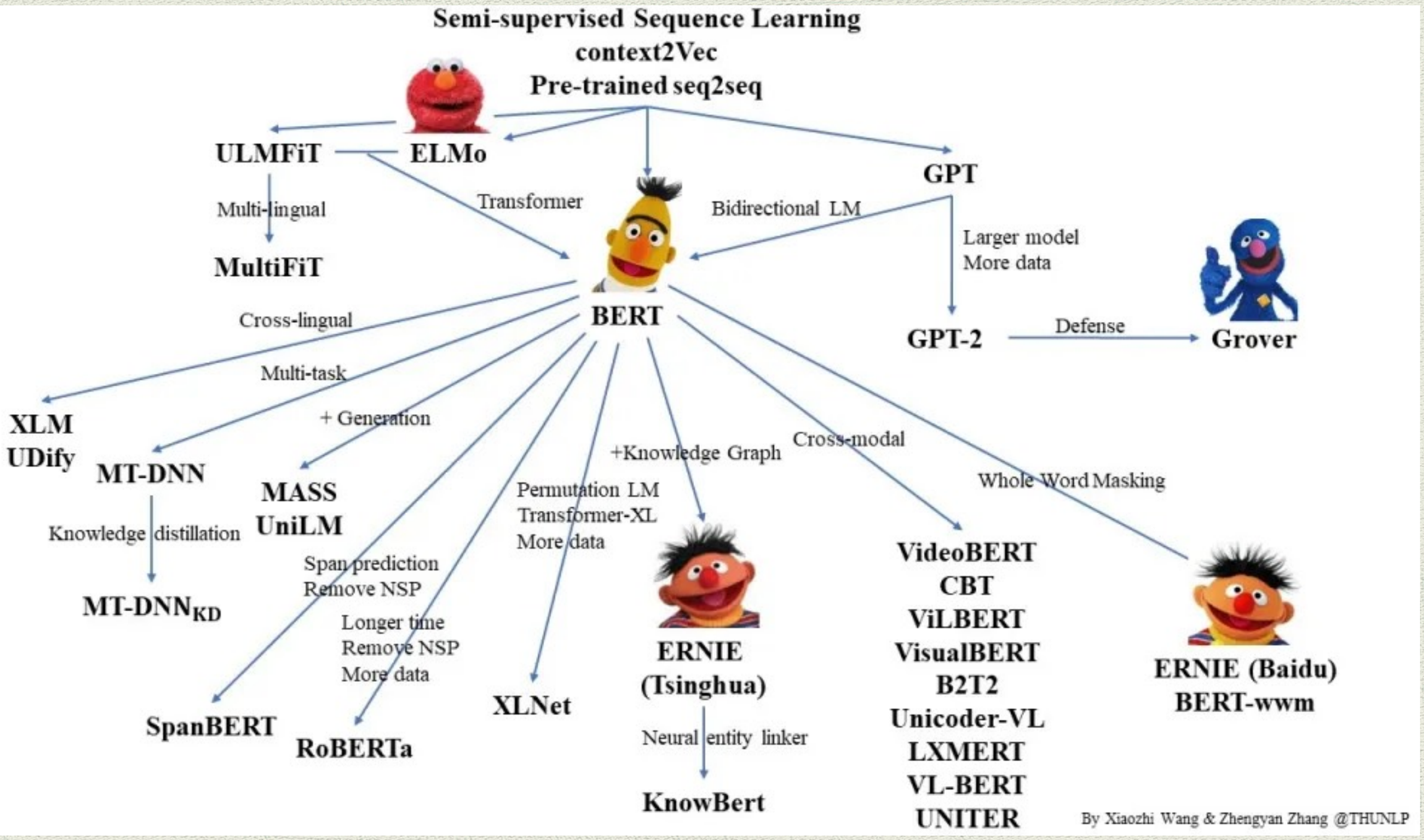
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Transformer: Encoder-Decoder



LLMs#2: Embeddings

- Word Embeddings in GloVe and word2vec are context free; vector for **bank** is computed over all instances “he robbed the bank” and “tales of the river bank” (same vector)

- Word Embeddings in LLMs take all context into account (512! word window / sentence, l-r and r-l); so learn different vectors for bank[1] and bank[2]

- Embeddings may be computed for words, sentences, paragraphs etc and are combined

LLMs#3: Sequence Learning

- Sequence learning & prediction is vimp
- From word-sequence in a sentence, predict next word: “dog bites <WHAT?>” => “dog bites *man*”
- Translation: predict french sequence from english sequence
- Q/A: Given a Q-seq + para-seq, get A-seq

LLMs#3: Predict a word

- Cloze Task (Taylor, 1953): Given X words in a sentence predict the next: "he shot the ball into the X" predict X (X= *goal, net, back-of-net*)

- Yet language has a direction, we read from left or right or right to left…

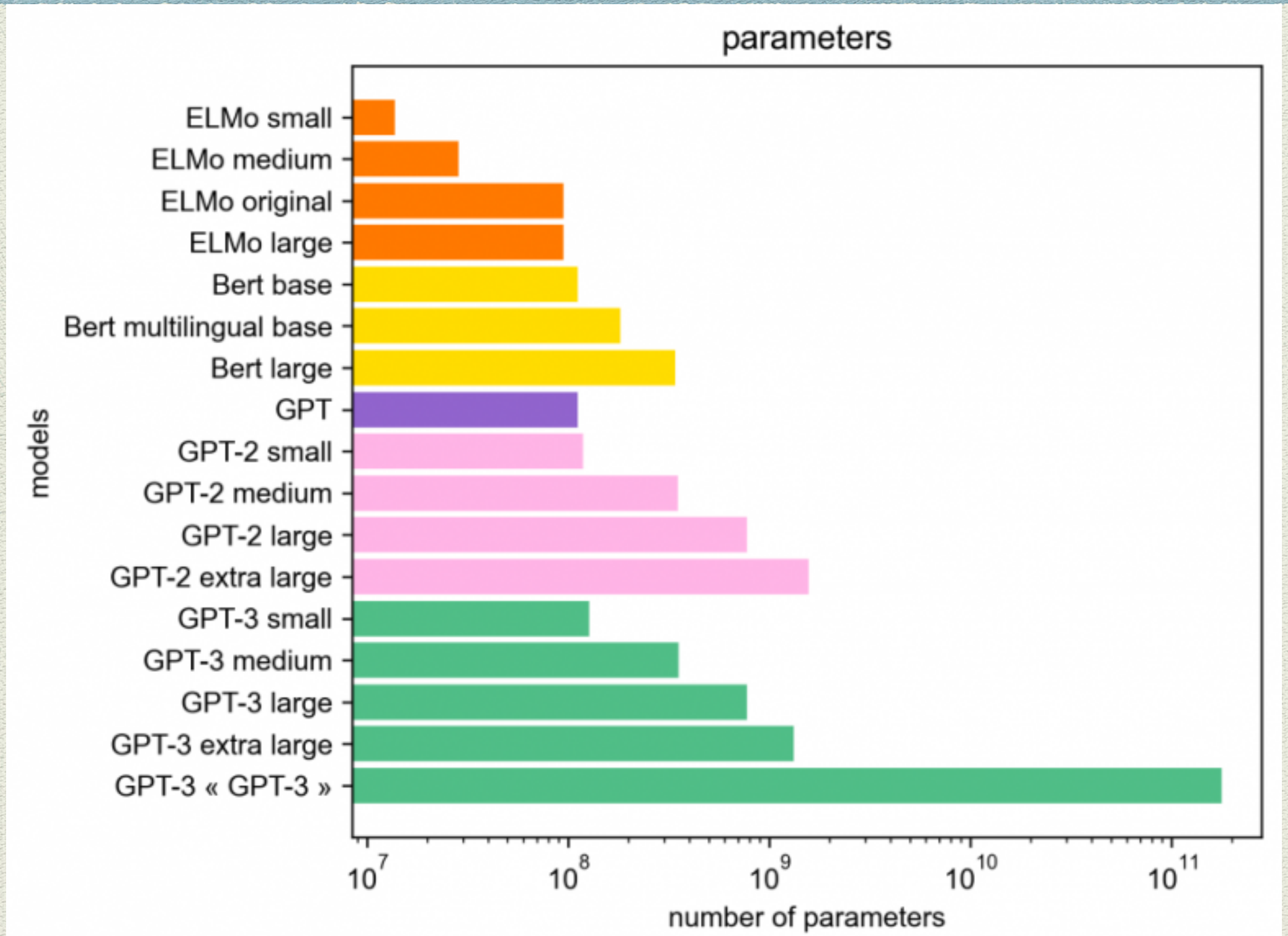
- Models being using left-to-right contexts and then right-to-left contexts; with a target at the centre of these contexts you have a lot to support a prediction

LLMs#4: Transfer Learning

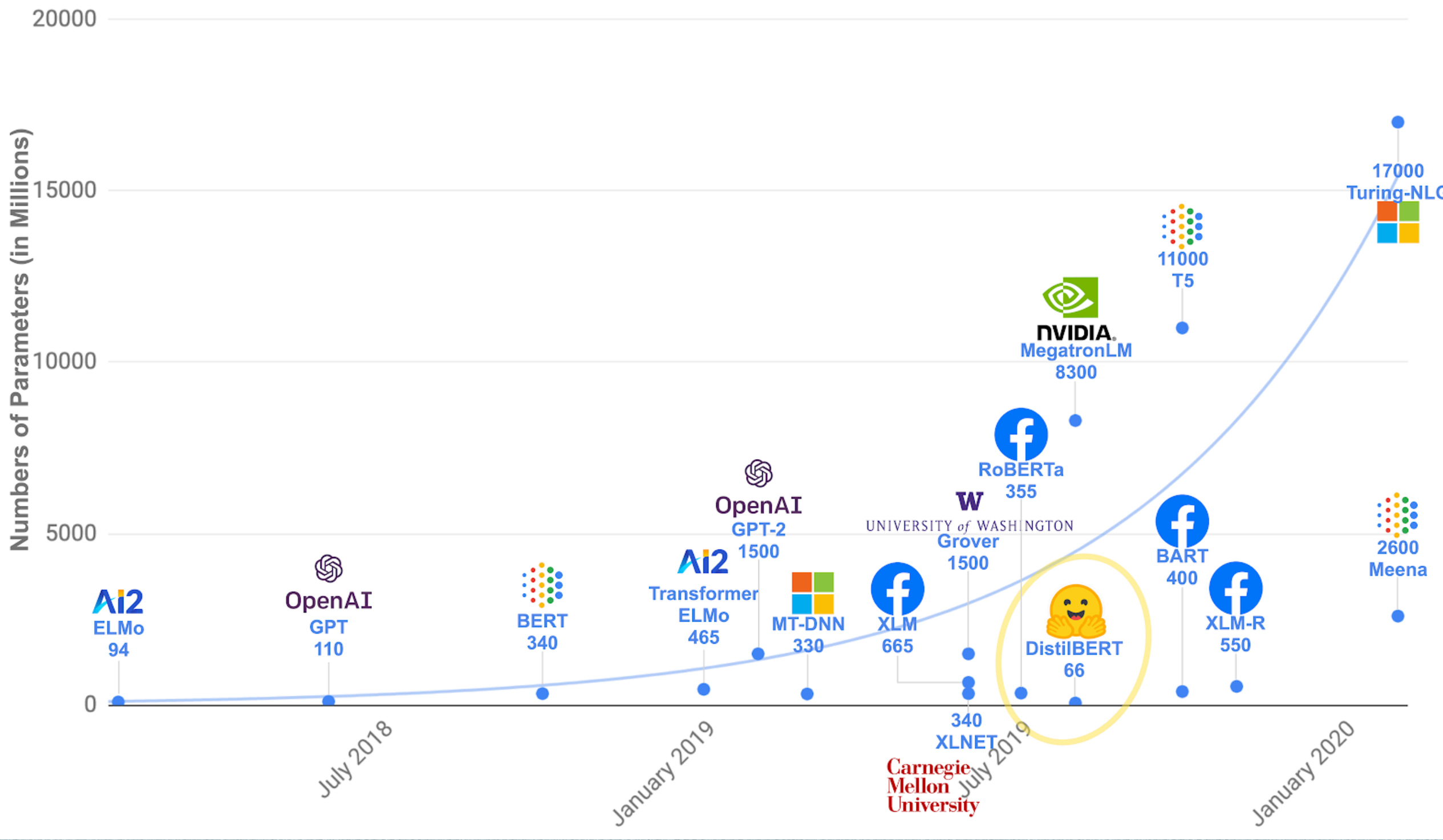
- *Transfer Learning* "focuses on storing knowledge gained while solving one problem and applying it to a different related problem" (wikipedia); so, showing that a crowd counting model, can also be applied to cutting apples, cells in MRIs etc

- BERT learns set of embeddings on one task and uses them on others (fine-tuning)

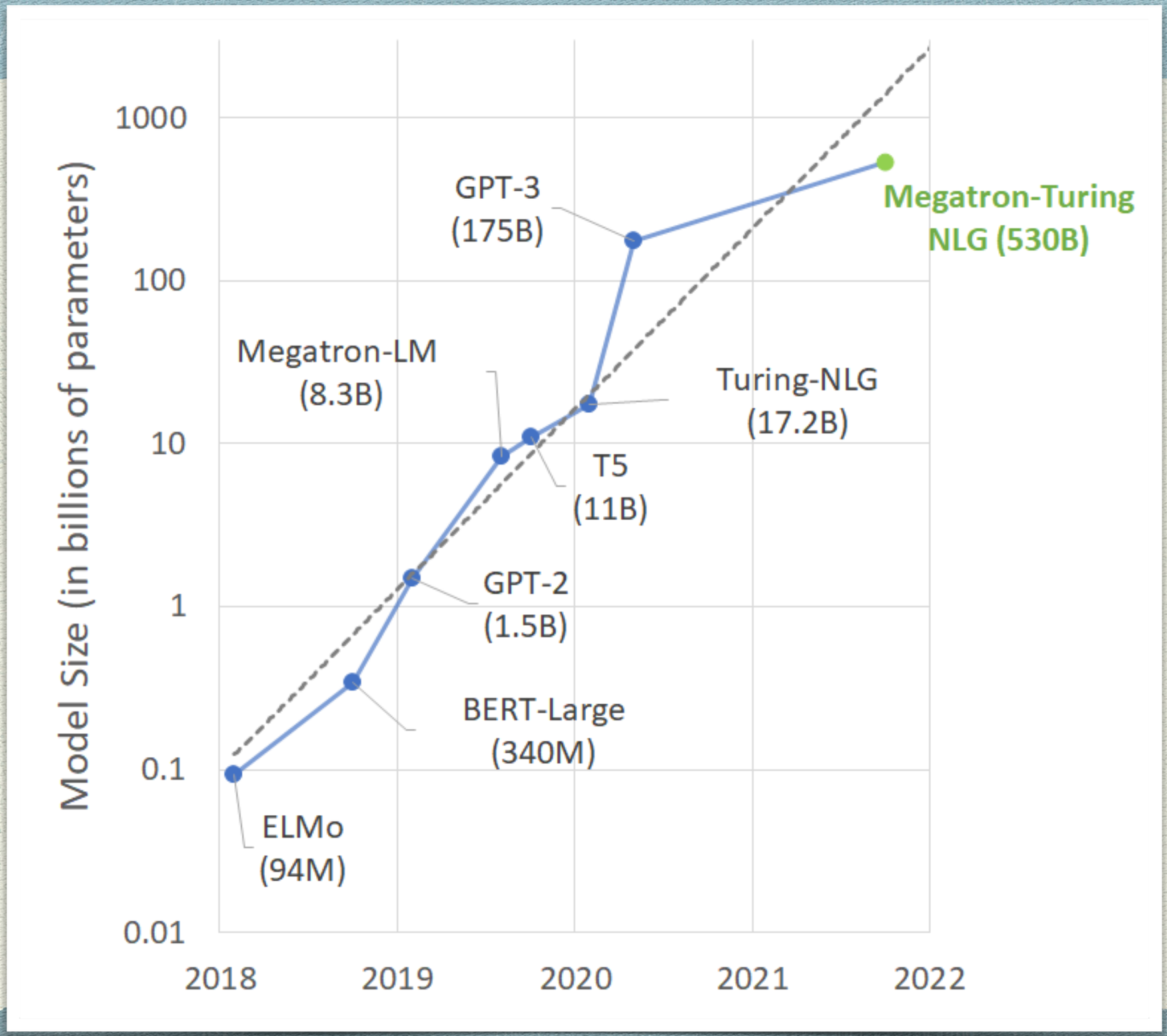
LLMs#5: BIGGER: Data



LLMs#5: BIGGER: Parameters



LLMs#5: BIGGER: Parameters



BERT: Overview

- BERT: Bidirectional Encoder Representations with Transformers (51k cites @ 2022)

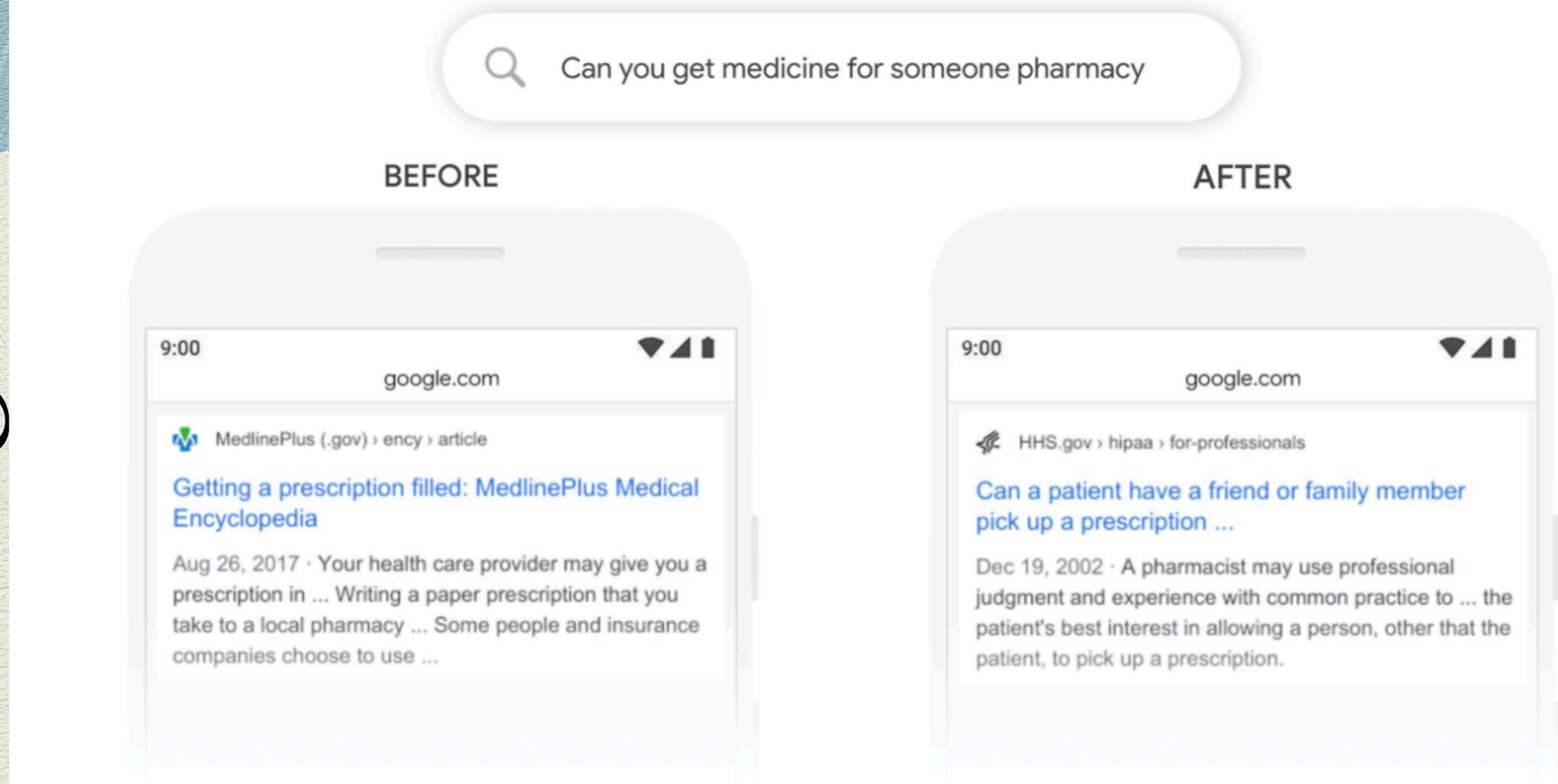
- Learning word embeddings without recurrent apparatus; become very influential

- Uses Transformer Arch. (Encoder), more complex encodings, self-attention and lots and lots of data

- In Nov. 2020, Google use it for queries (2018 Open)

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT: Eg



Source

Pre-BERT Google surfaced information about getting a prescription filled.

Post-BERT Google understands that “for someone” relates to picking up a prescription for someone else and the search results now help to answer that.

Screenshot

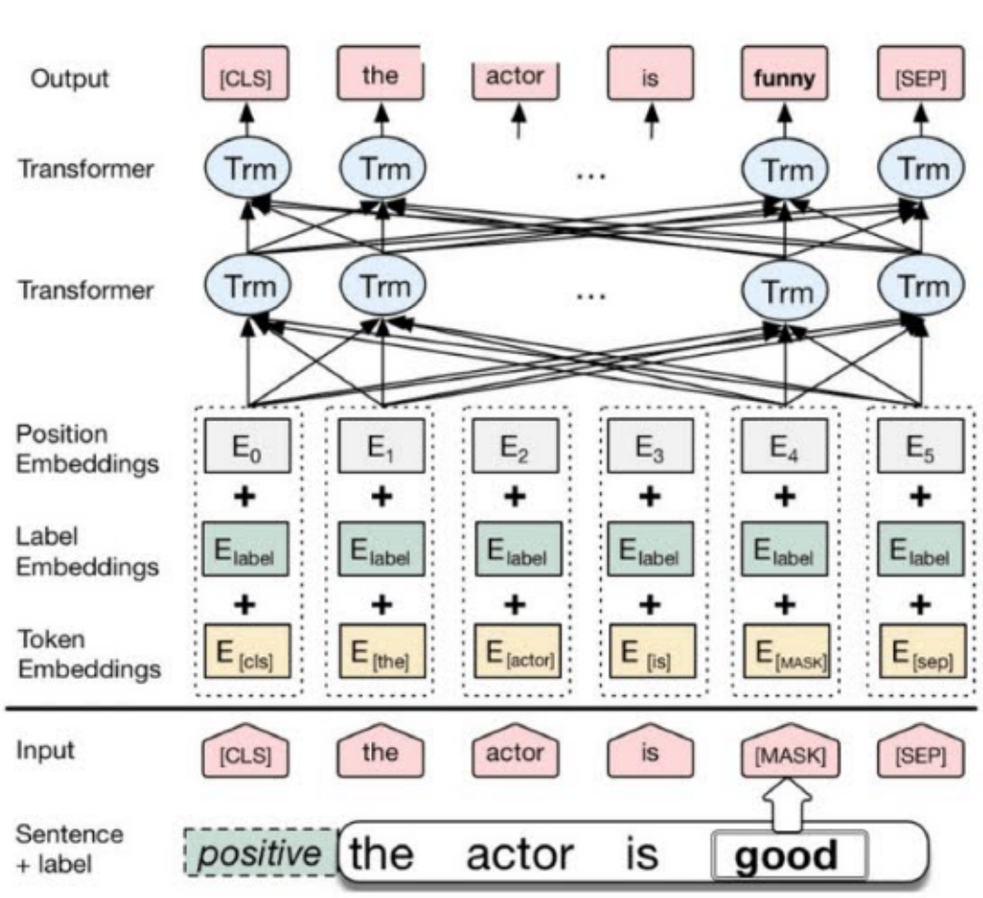
https://huggingface.co/blog/bert-101#1-what-is-bert-used-for

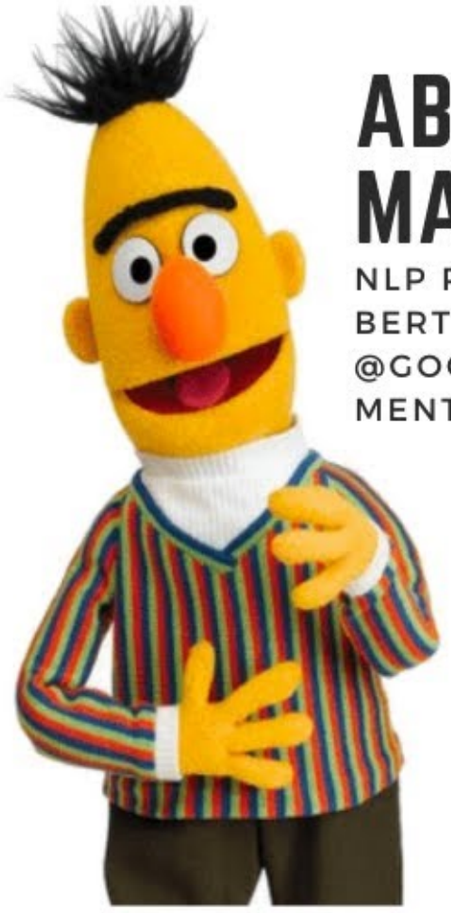
BERT: Thumbs Up !

BERT Explained - State of the art language model for NLP

Co - Learning Lounge

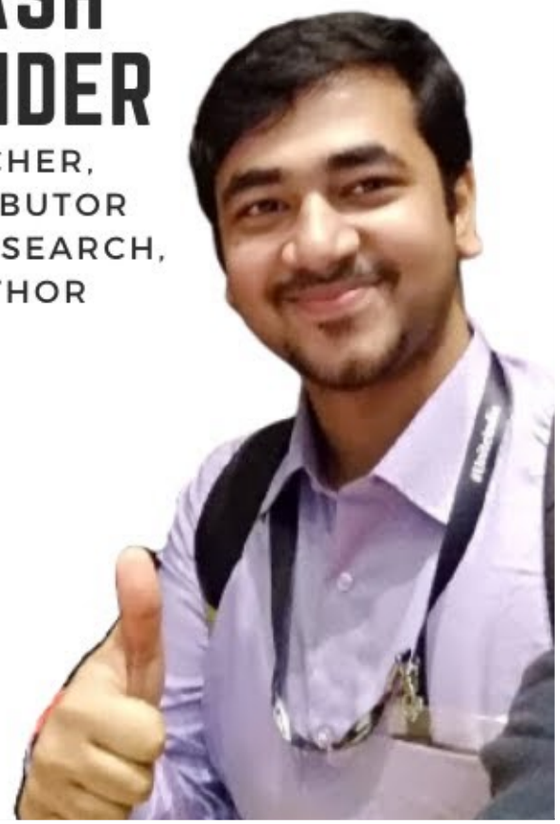
Deep Bidirectional Transformers for Language Understanding





ABHILASH MAJUMDER

NLP RESEARCHER,
BERT CONTRIBUTOR
@GOOGLE-RESEARCH,
MENTOR, AUTHOR

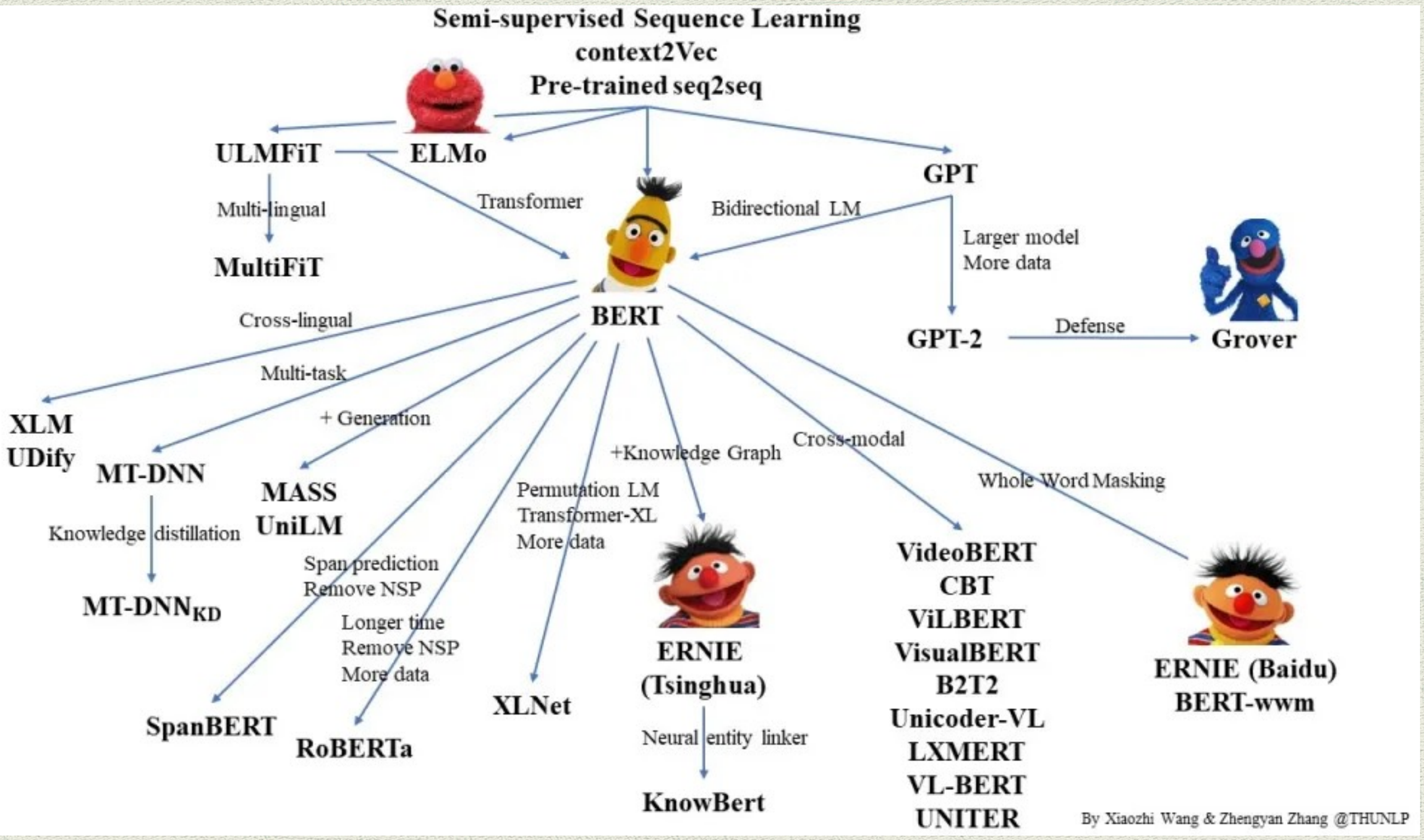


BERT: Impressives

- Can do multiple tasks: sentiment, Q-A, email sentence completion, legal contract summarisation, writing articles, polysemy
- Record breaking, 11 Benchmark NLP Tasks
- 50% of top-10 models built on BERT

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Transformer: Encoder-Decoder



BERT: Bits & Pieces

1. *Architecture*: Encoder (not Decoder) of stacked Transformer Blocks

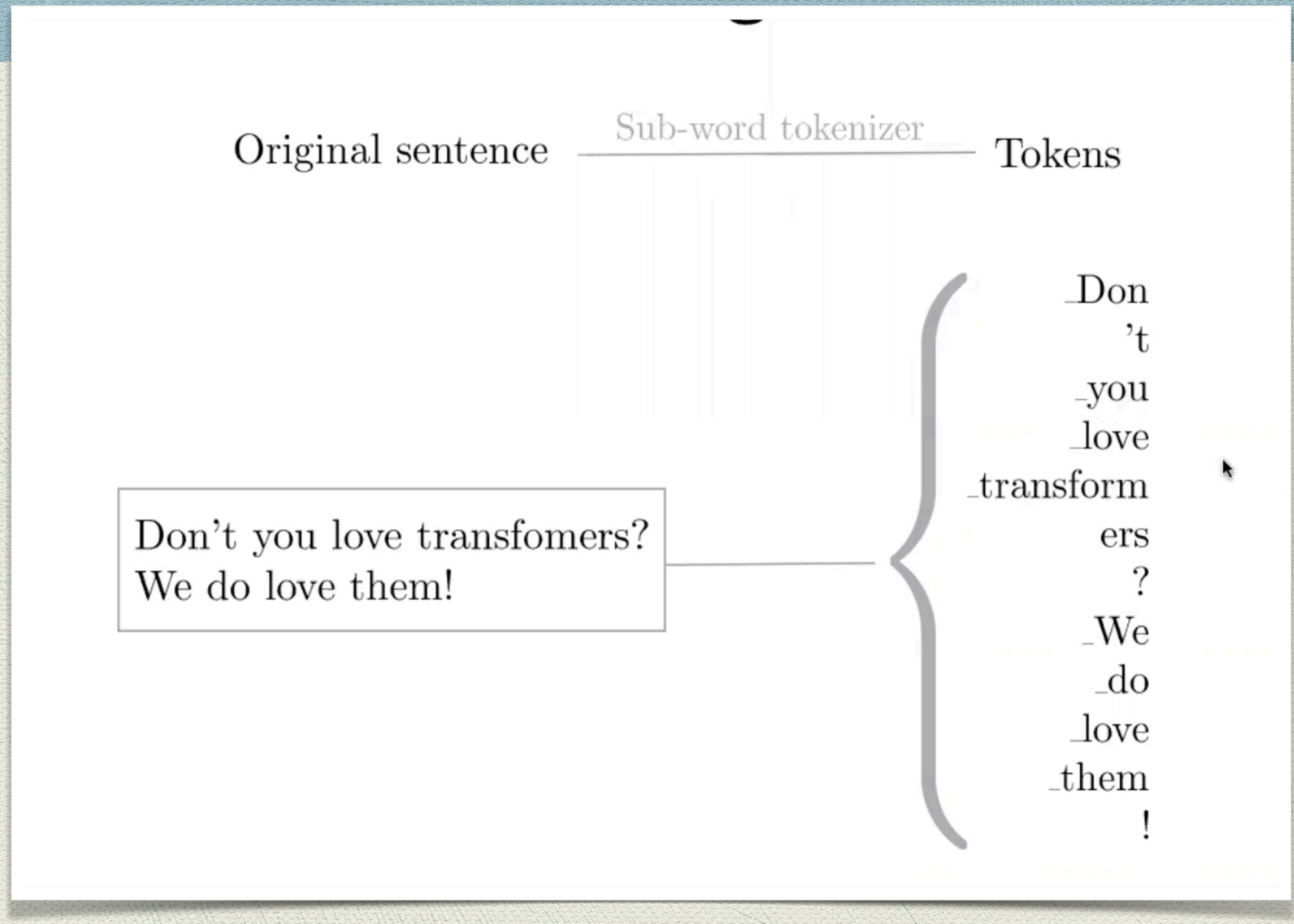
2. *Embeddings*: Smart tokenising and learns own embeddings

3. *Transfer Learning*: Pre-trained models on one task then fine-tuning for other task (re-using weights)

4. *Many Tasks*: predict seq_{t1} from seq_{t-1} or predicting [mask] token in sequence…

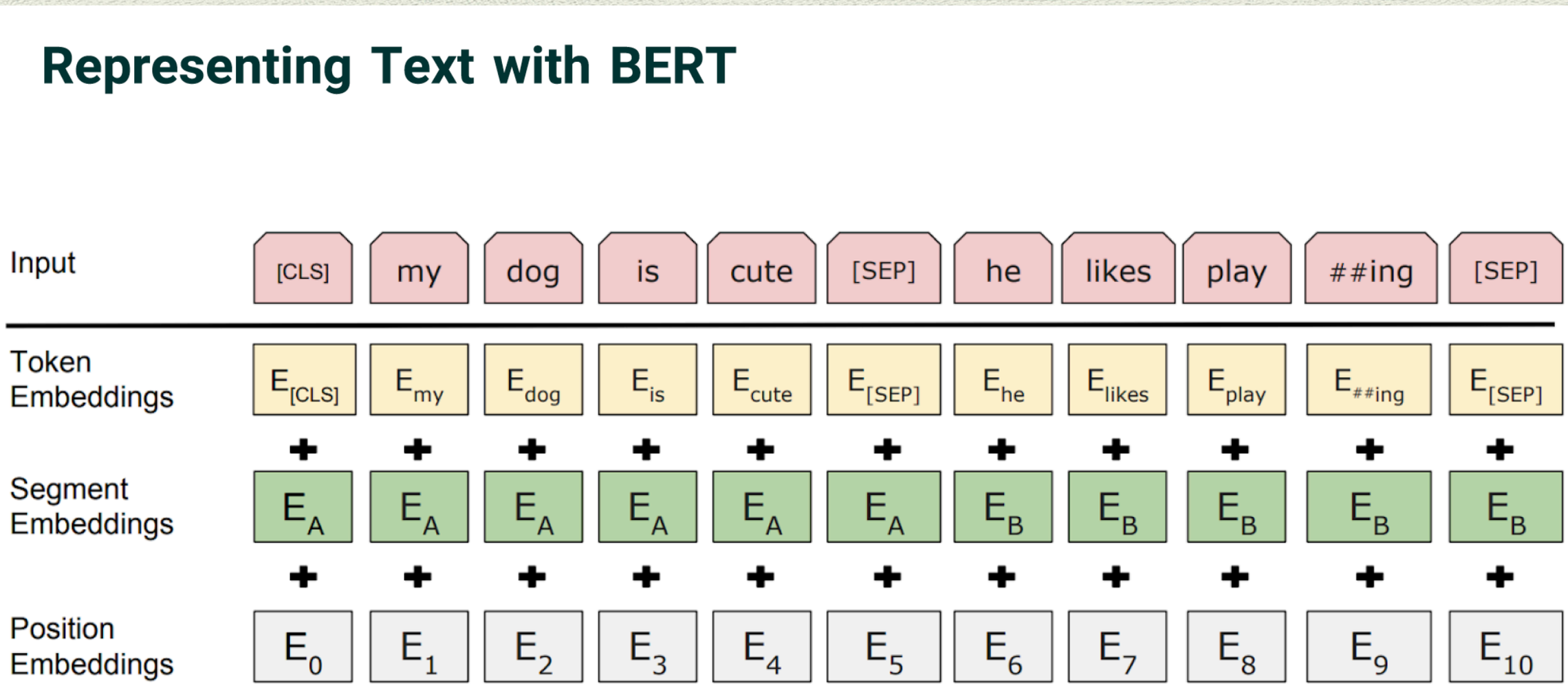
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT: Token Inputs



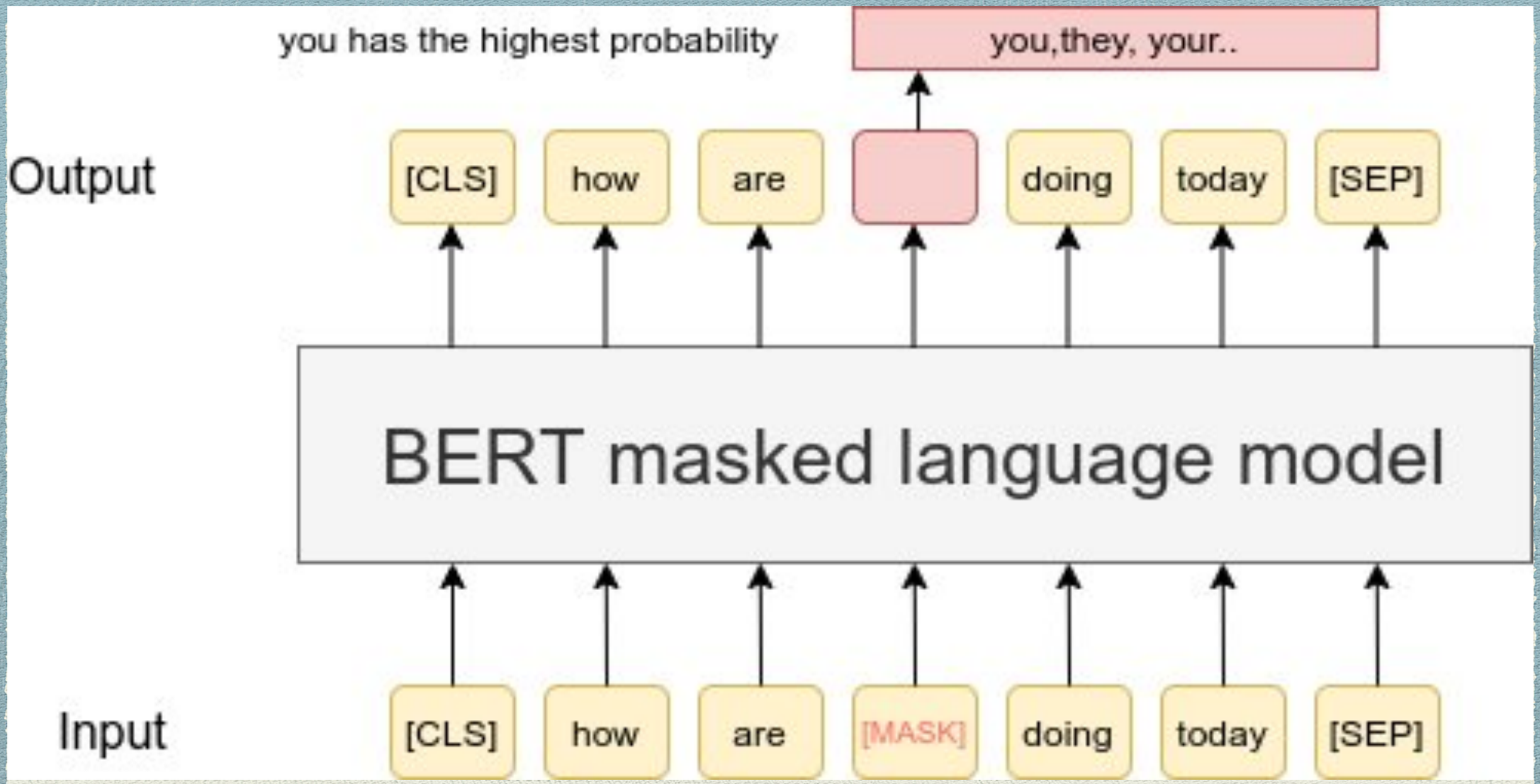
But, uses own specific tokeniser: HuggingFace & WordPiece

BERT: Sentence Inputs



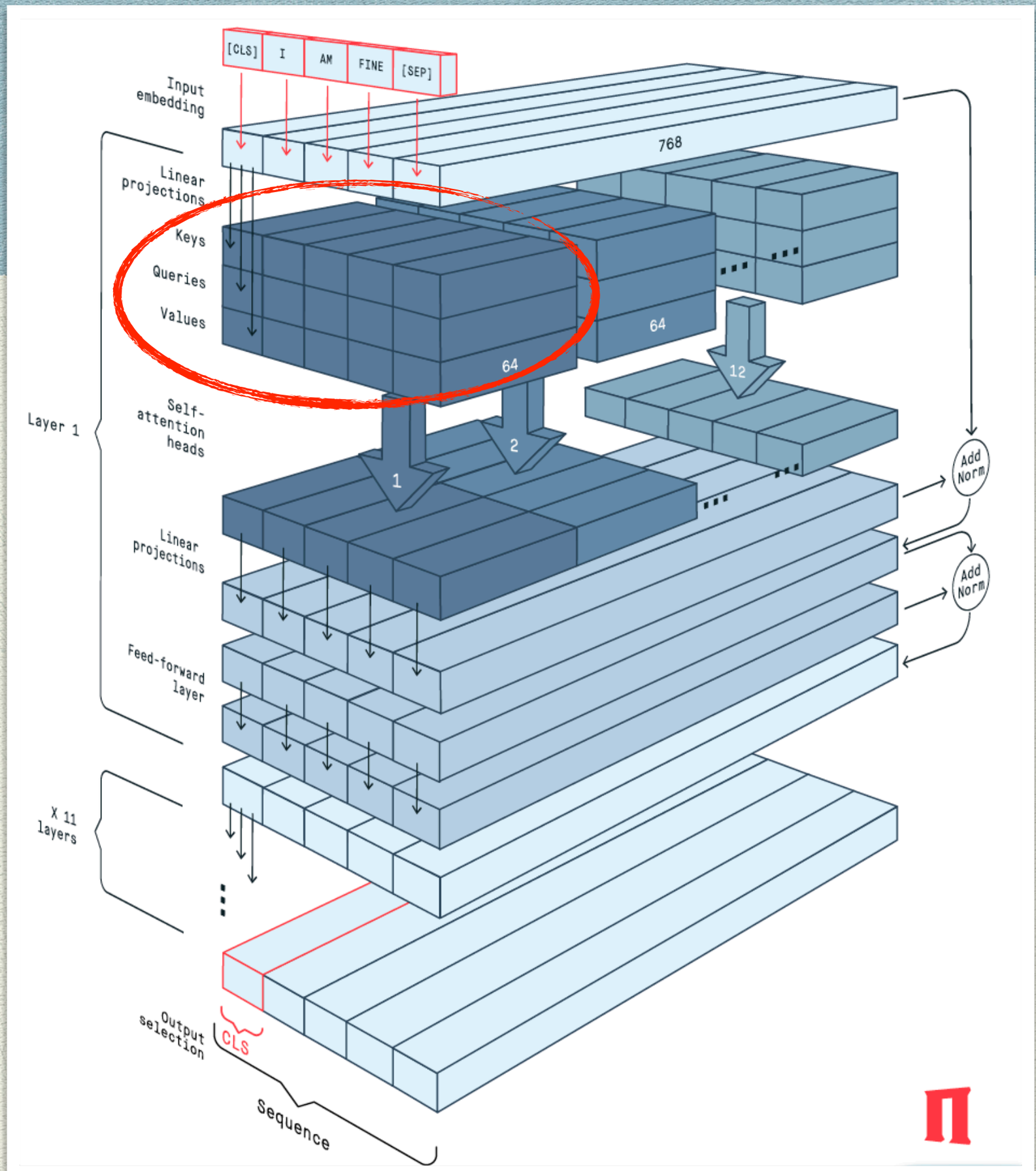
But, uses own specific tokeniser: HuggingFace & WordPiece

BERT: Inputs & Outputs



BERT enables context-sensitive embeddings….beyond word2vec

BERT: Architecture



Keys

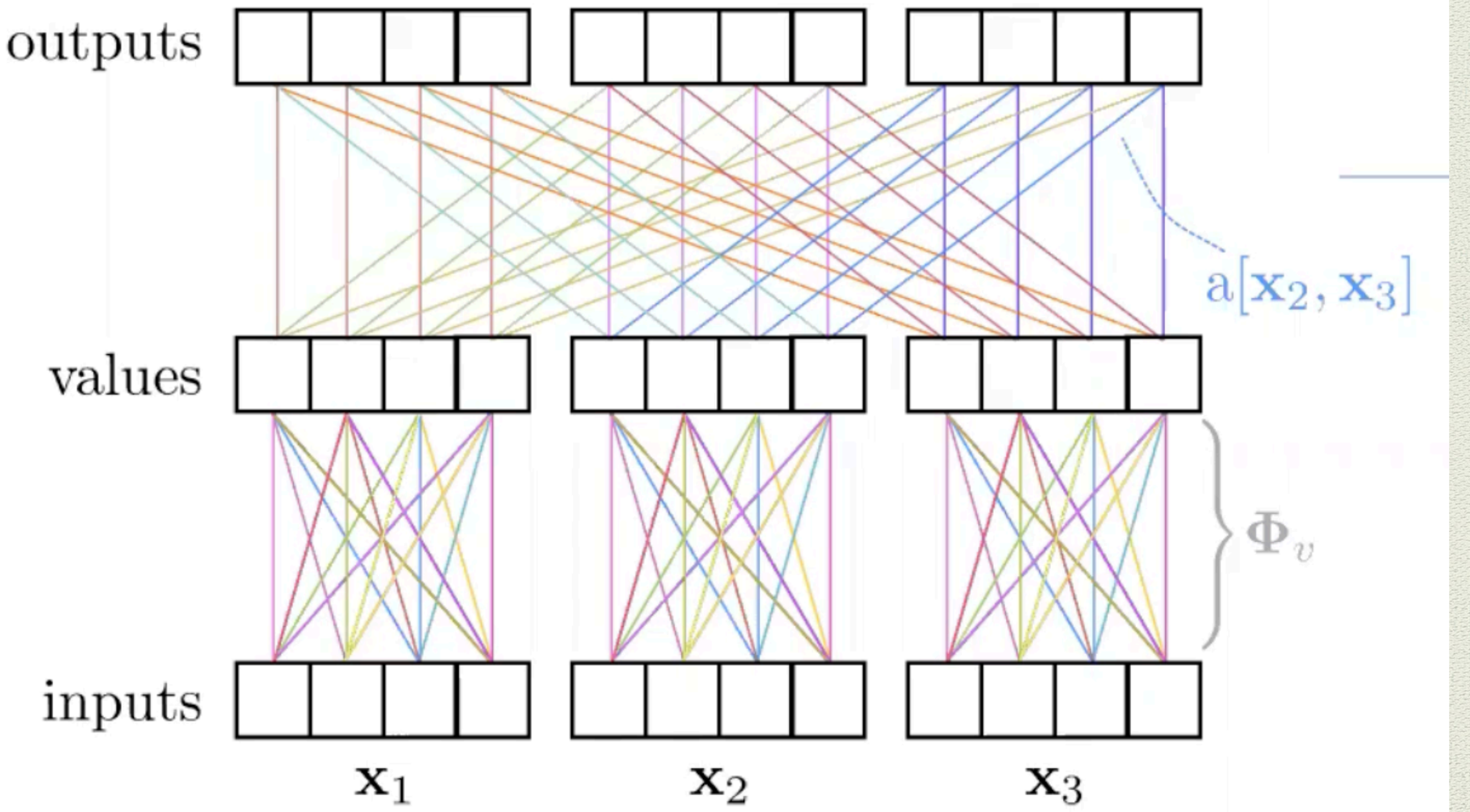
Queries

Values

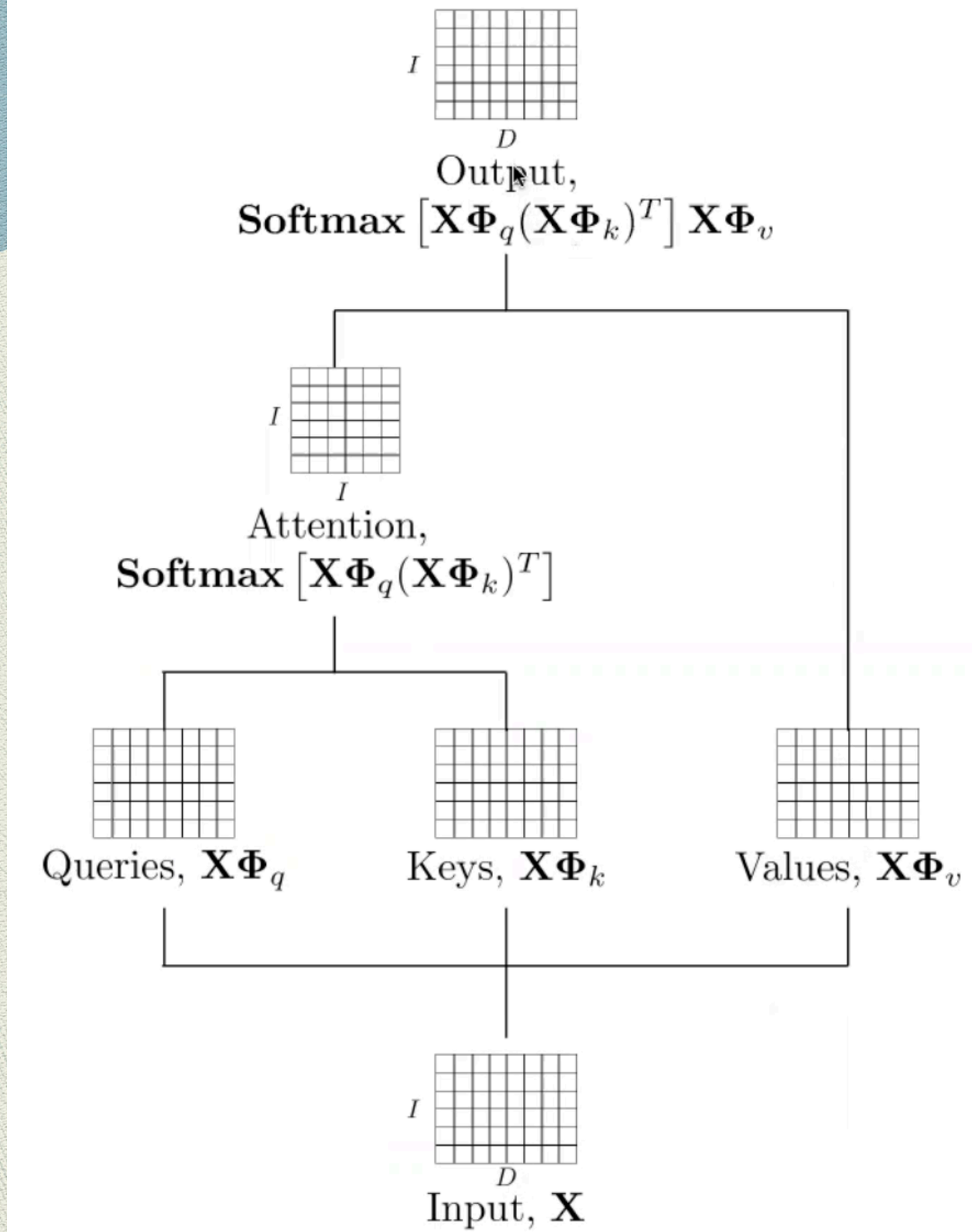
https://peltarion.com/knowledge-center/modeling-view/build-an-ai-model/blocks/bert-encoder

BERT: Self-Attention

$$\mathbf{sa}[\mathbf{x}_i] = \sum_{j=1}^{I} a[\mathbf{x}_i, \mathbf{x}_j](\mathbf{x}_j \boldsymbol{\Phi}_v)$$



Simon Prince (2022) Intro to Transformers & Self-Attention

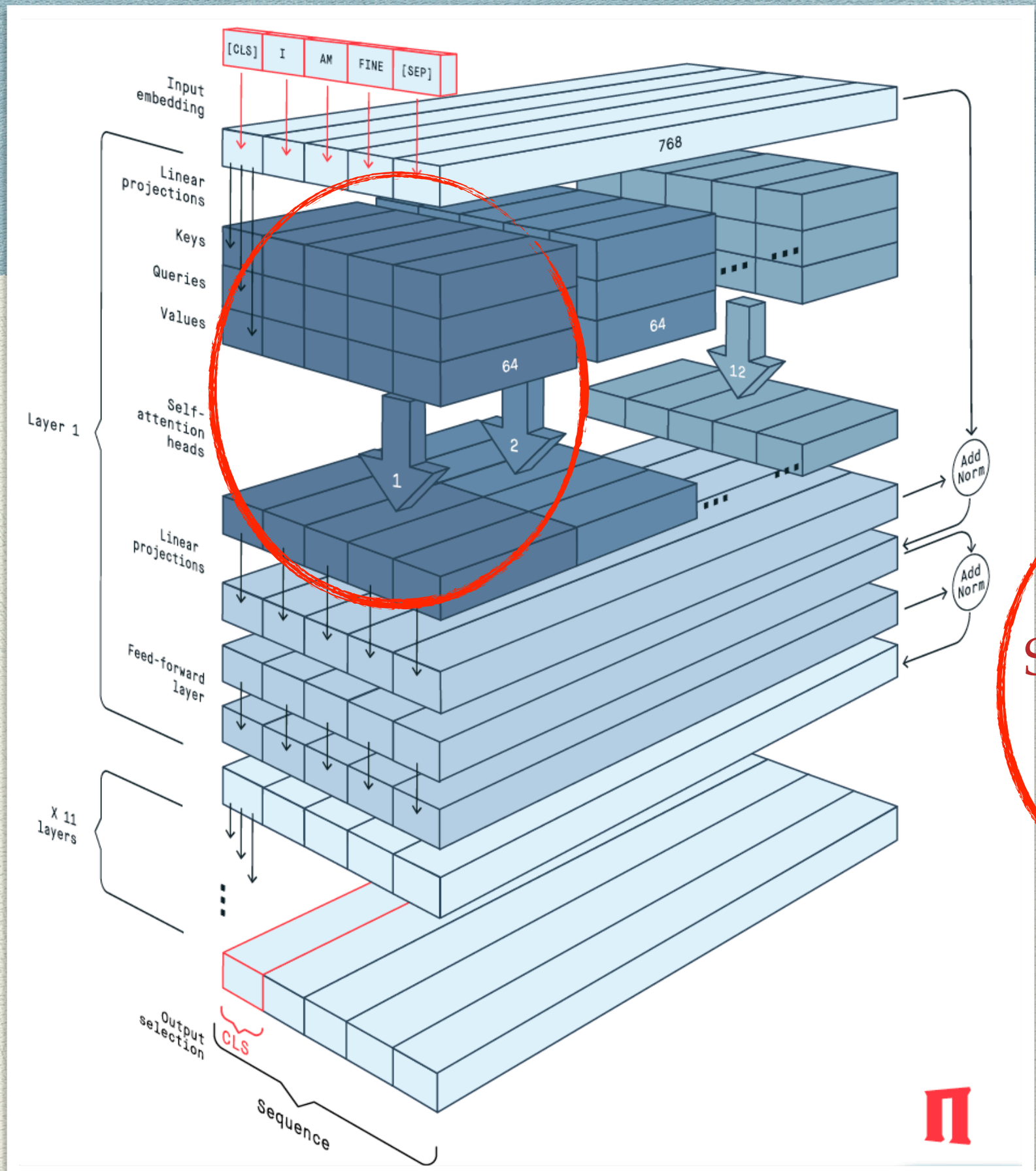


Keys

Queries

Values

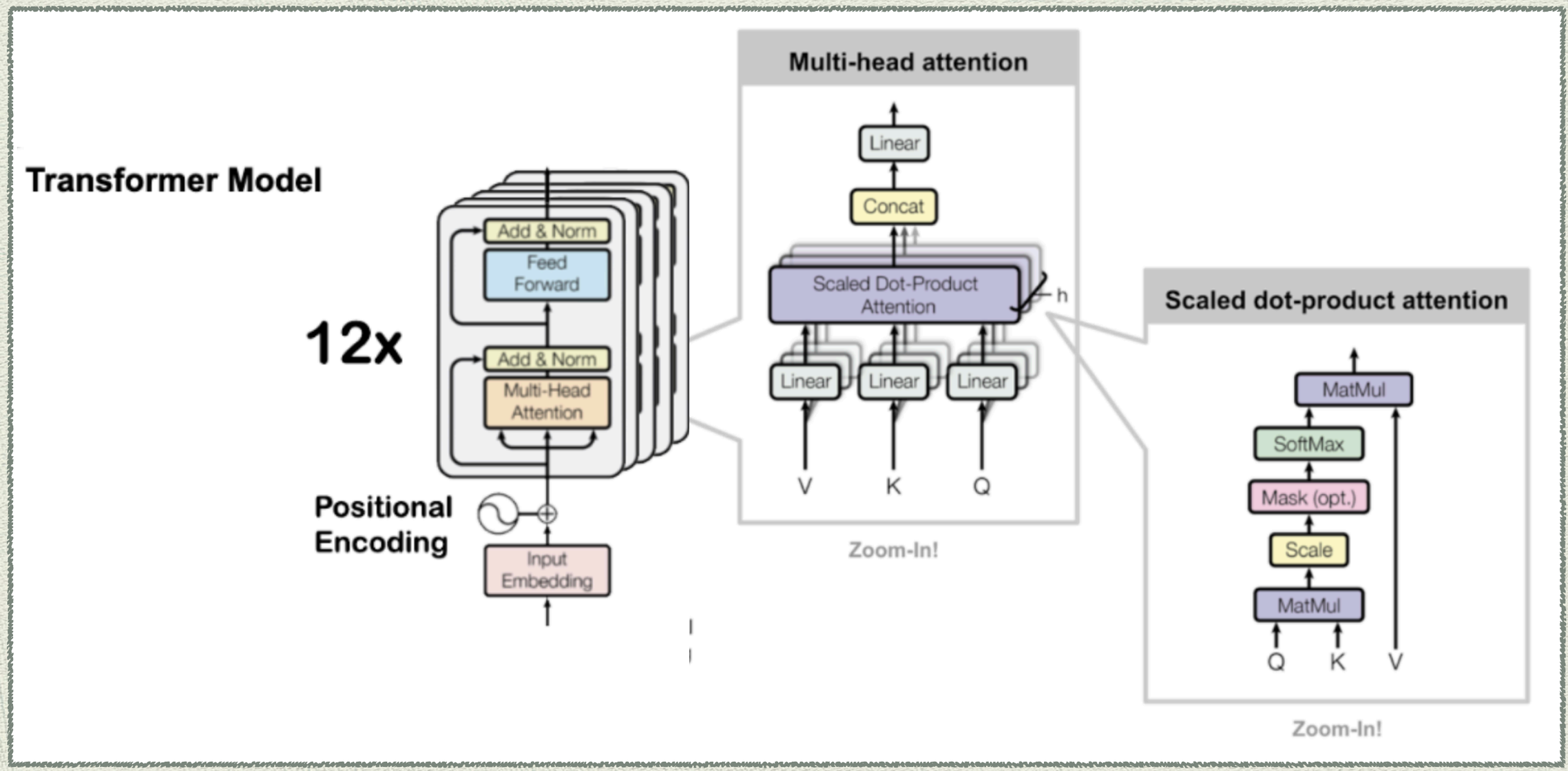
BERT: Architecture



Self-Attention Head #1

https://peltarion.com/knowledge-center/modeling-view/build-an-ai-model/blocks/bert-encoder

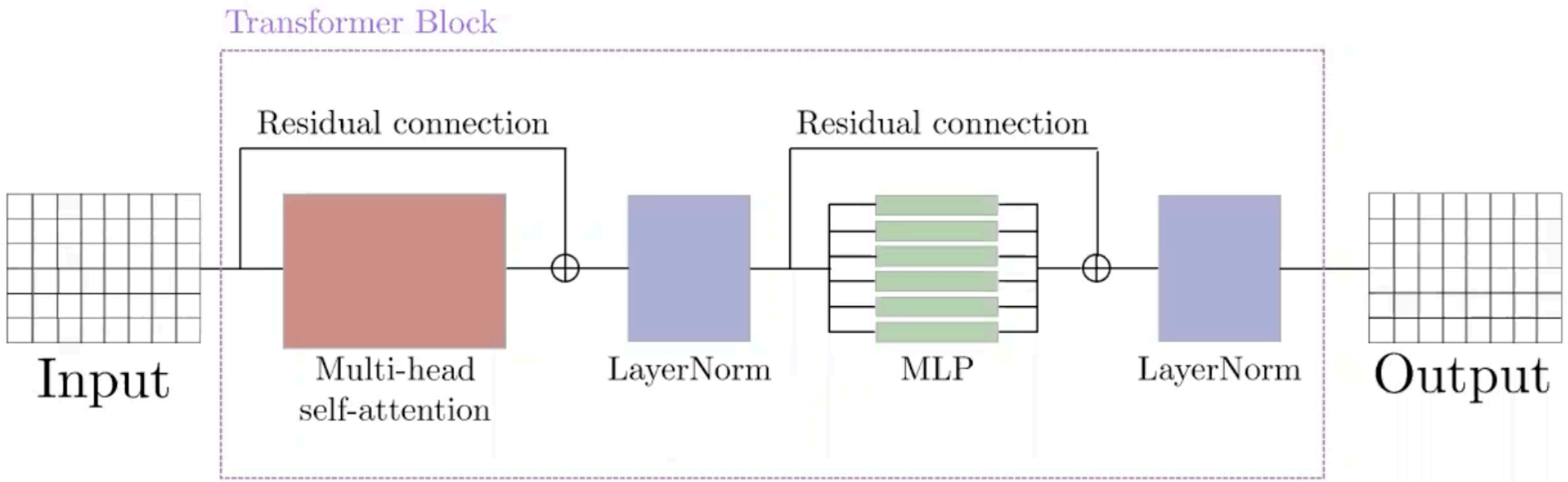
BERT: Self-Attention Heads



Khalid, U., Beg, M. O., & Arshad, M. U. (2021). Rubert: A bilingual roman urdu bert using cross lingual transfer learning. *arXiv preprint arXiv:2102.11278*.

BERT: Is Small…!

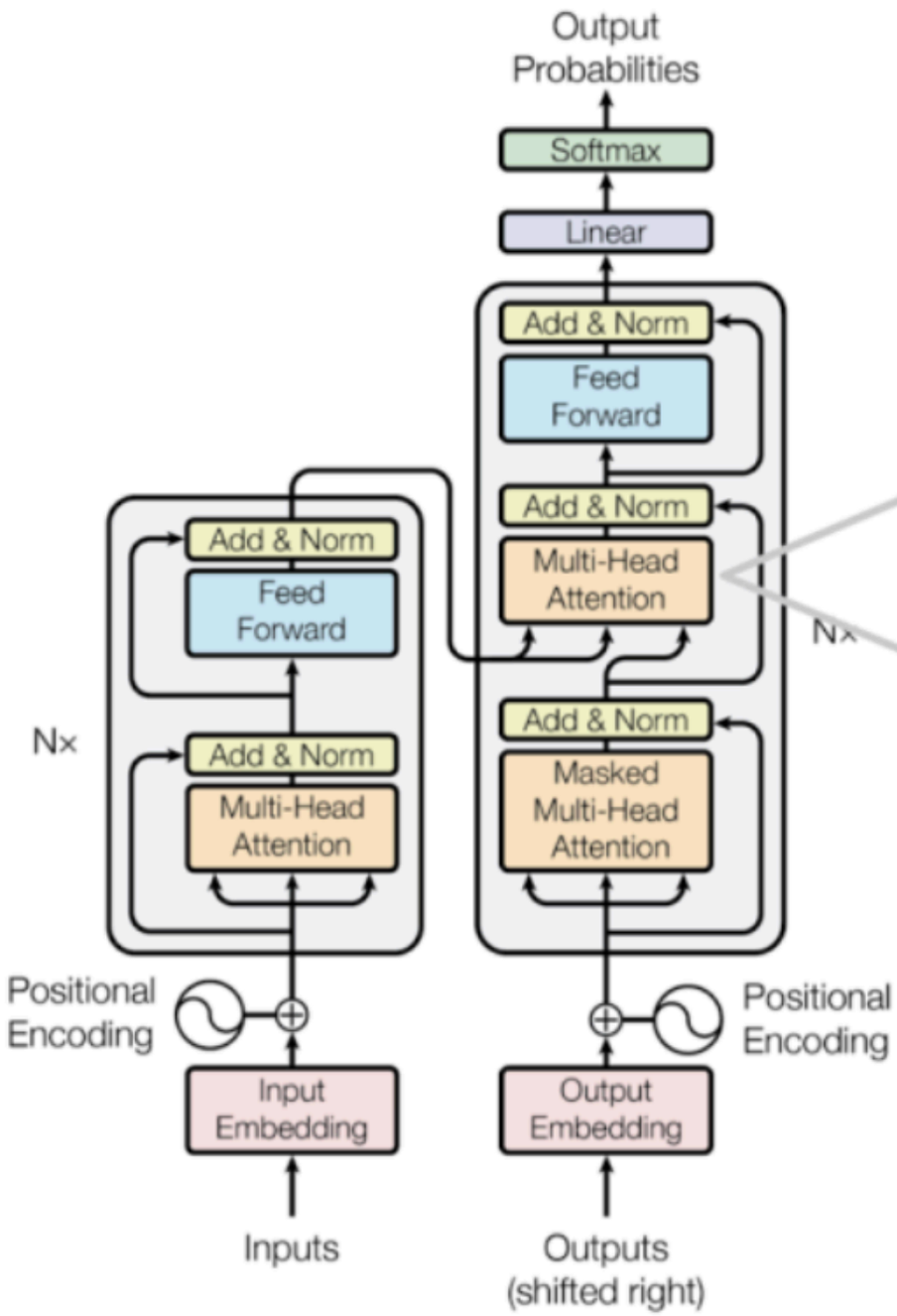
Encoder model example: BERT

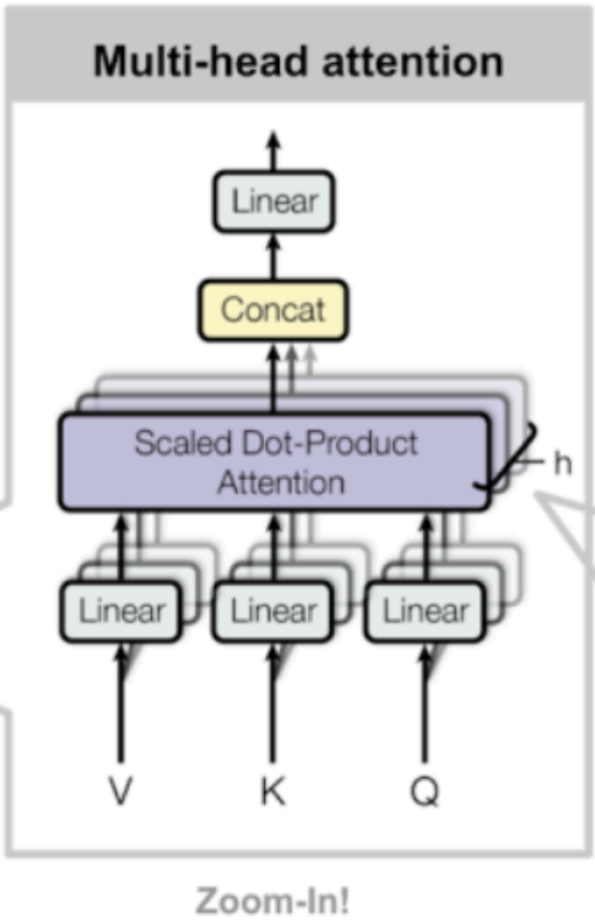


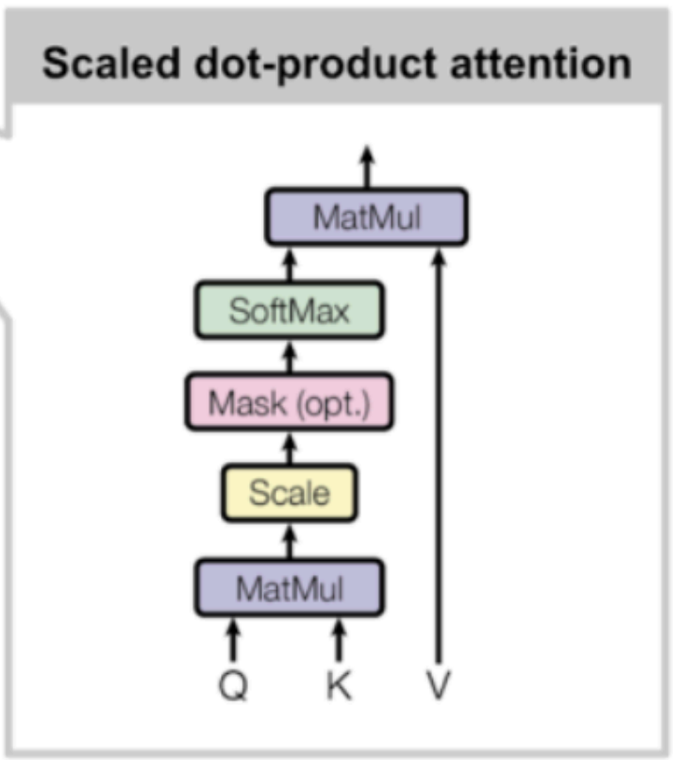
- Bi-directional Encoder Representations from Transformers (Devlin et al. 2018)
  - Vocabulary 30,000 tokens
  - Word embeddings length 1024
  - 24 transformer layers
  - 16 heads per layer, each dimension 64
  - Neural network hidden layer 4096
  - 340 million parameters

Encoder-Decoder: Architectures

Encoder **Decoder**



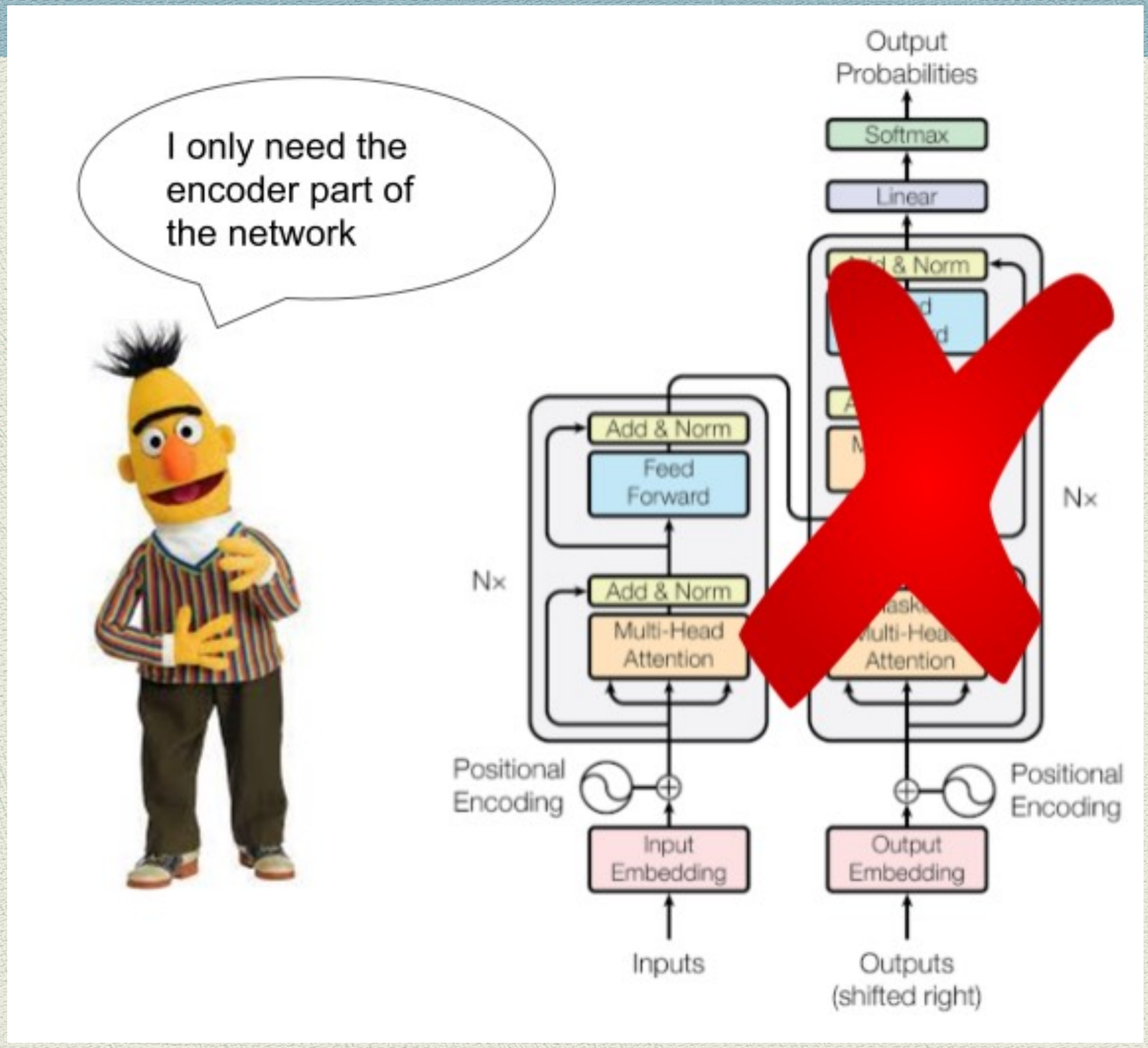




https://lilianweng.github.io/posts/2018-06-24-attention/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, ….& Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Encoder-Decoder: Architectures



https://neptune.ai/blog/bert-and-the-transformer-architecture

BERT:Training & Tuning

- Transfer Learning; broad learnings transfer
- Train model in BIG corpus to cover WHOLE domain, then specialise with fine-tuning
- Pre-trained models off-the-shelf, passed around
- So, “poorer” players can play (GTP $20M)
- Note, dominance of large companies

Training & Tuning

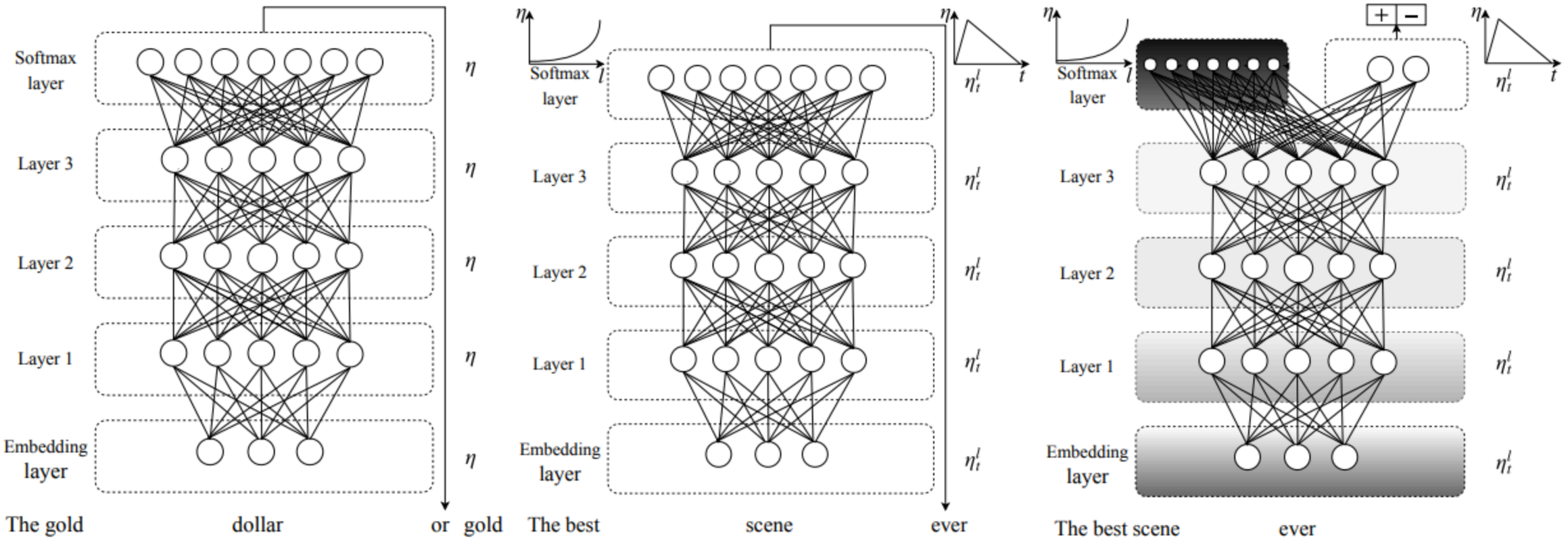


Figure 1: ULMFiT consists of three stages: a) The LM is trained on a general-domain corpus to capture general features of the language in different layers. b) The full LM is fine-tuned on target task data using discriminative fine-tuning ('*Discr*') and slanted triangular learning rates (STLR) to learn task-specific features. c) The classifier is fine-tuned on the target task using gradual unfreezing, '*Discr*', and STLR to preserve low-level representations and adapt high-level ones (shaded: unfreezing stages; black: frozen).

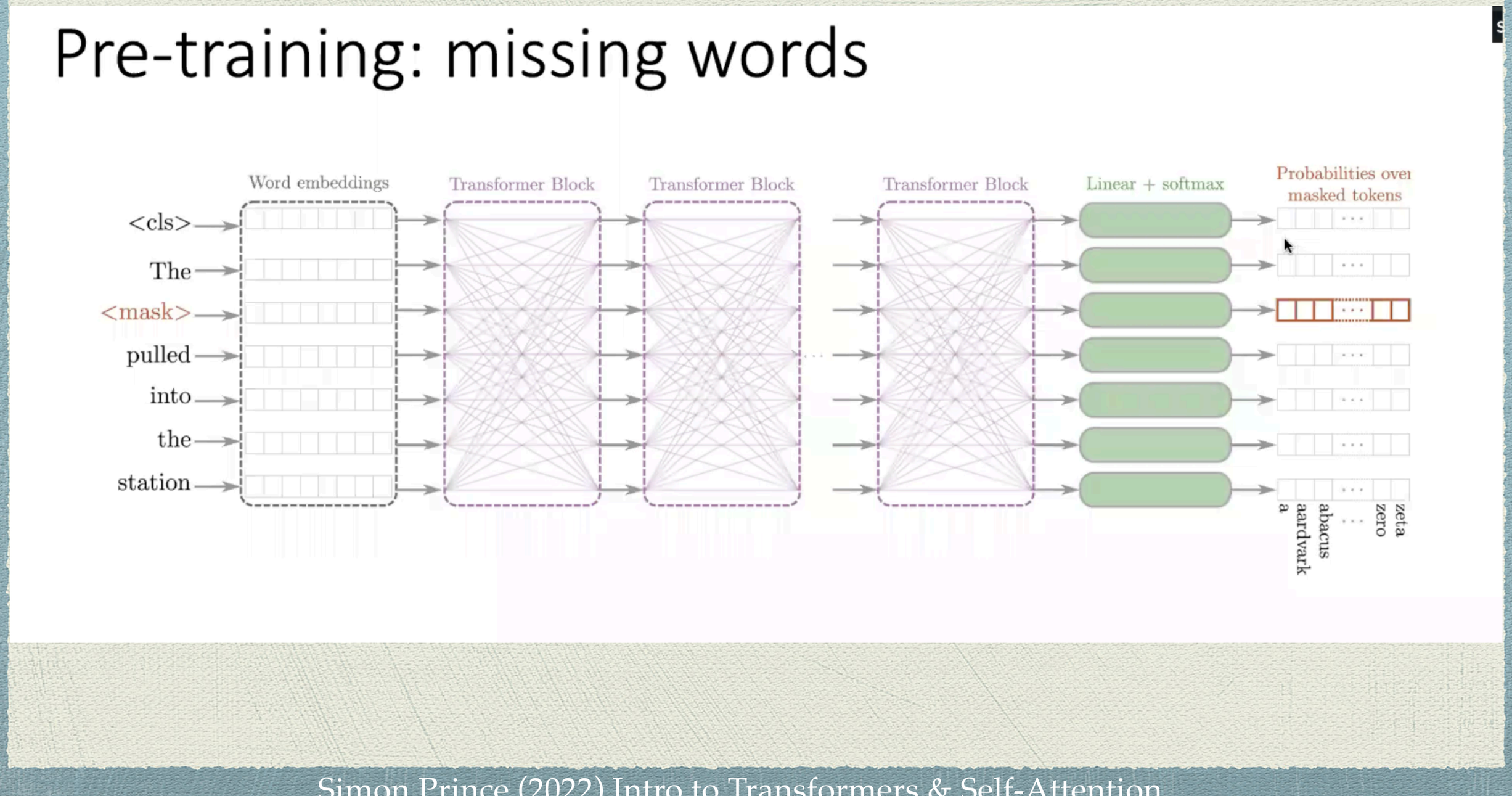
https://neptune.ai/blog/bert-and-the-transformer-architecture

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

BERT: Pre-Training

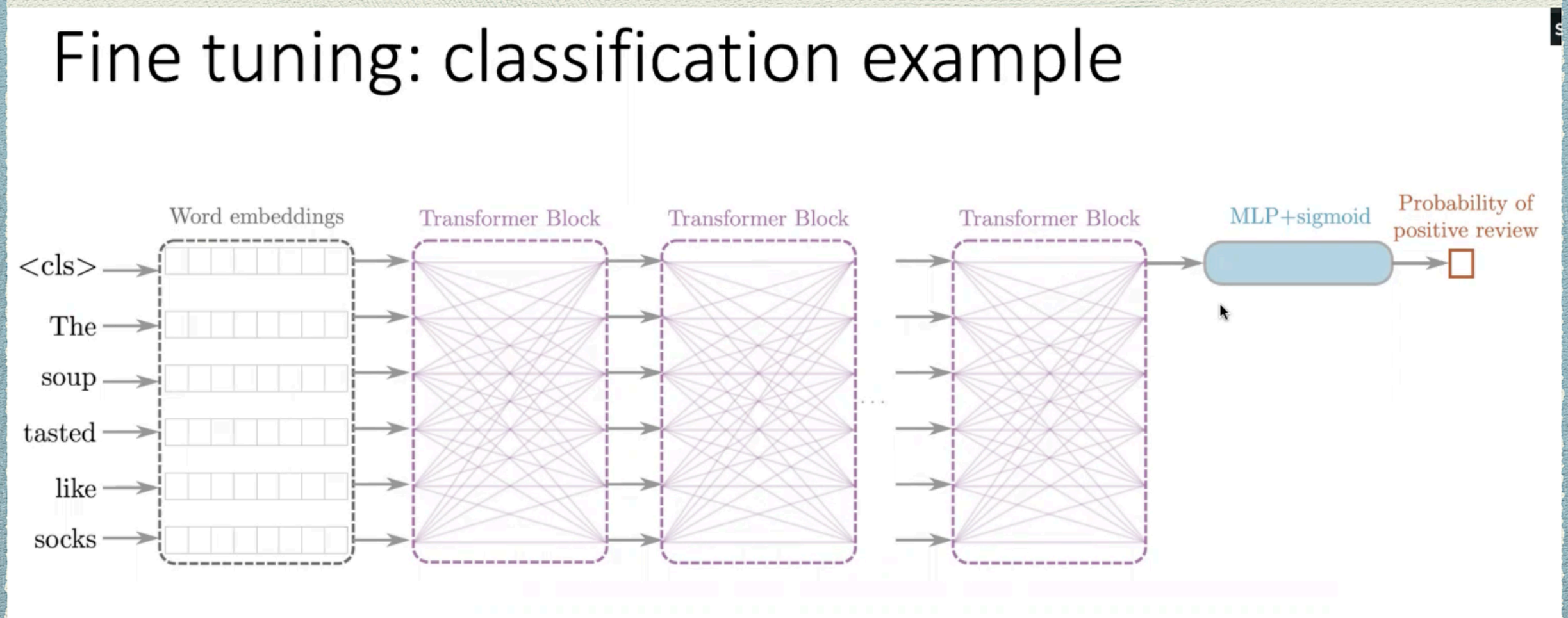
- Done on two tasks: Masked LM (MLM) and Next Sentence Prediction (NSP)
- Provides two core ways to consider data for other tasks: single sent. versus pair of sents.
- Fine-tuning works from these *generic* tasks

BERT: Pre-Training MLM



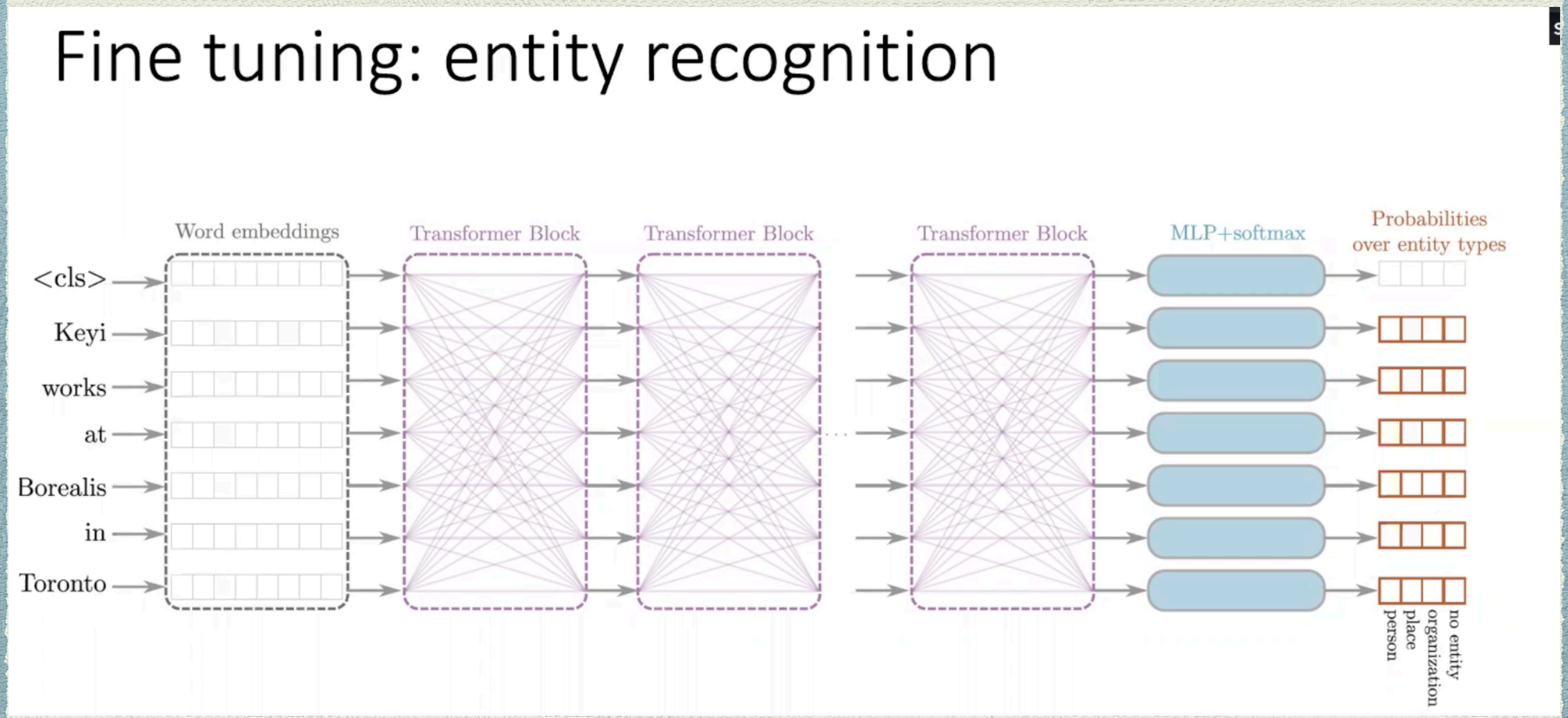
Simon Prince (2022) Intro to Transformers & Self-Attention

BERT: Tuning: Classification



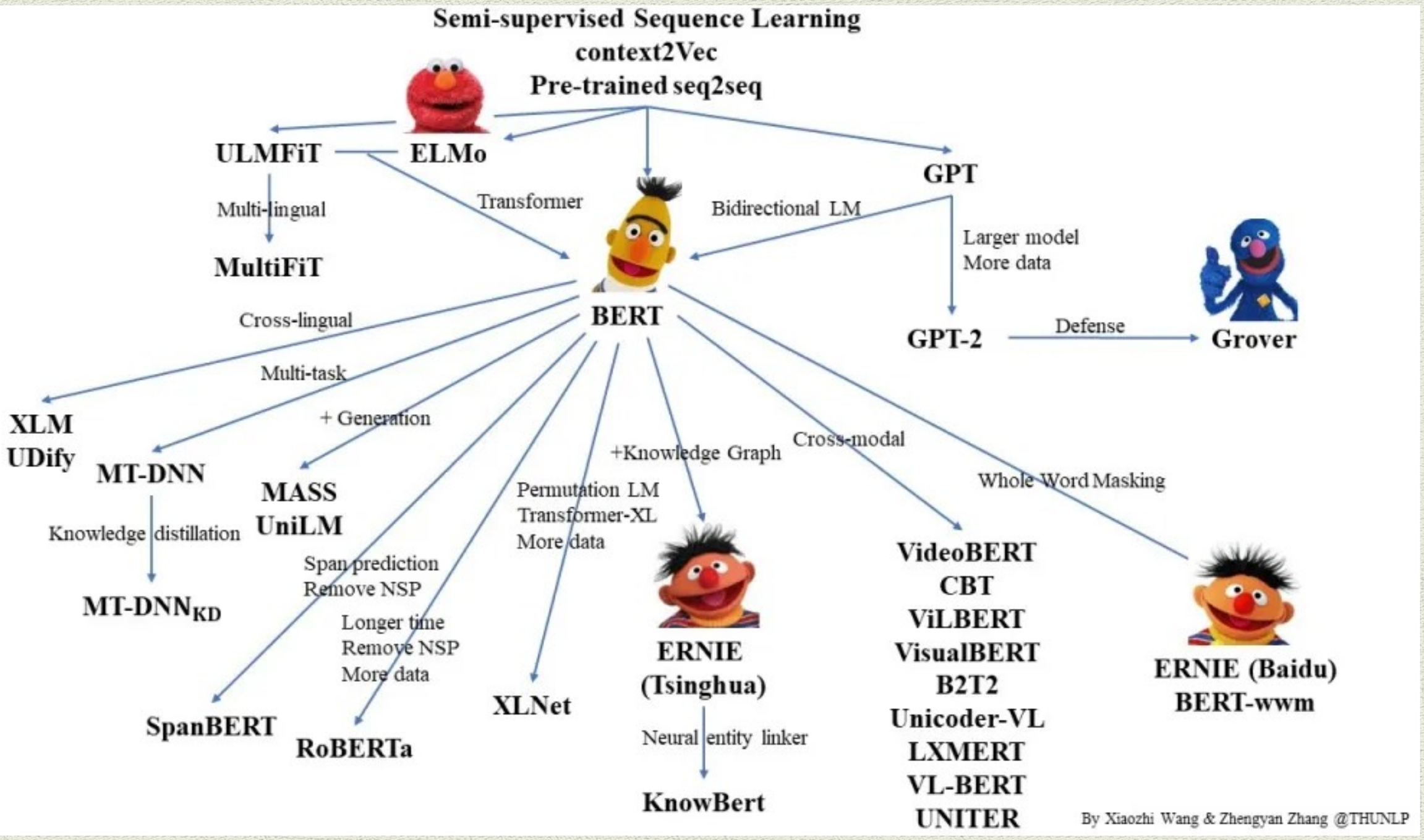
Simon Prince (2022) Intro to Transformers & Self-Attention

BERT: Tuning: Entity Recog.

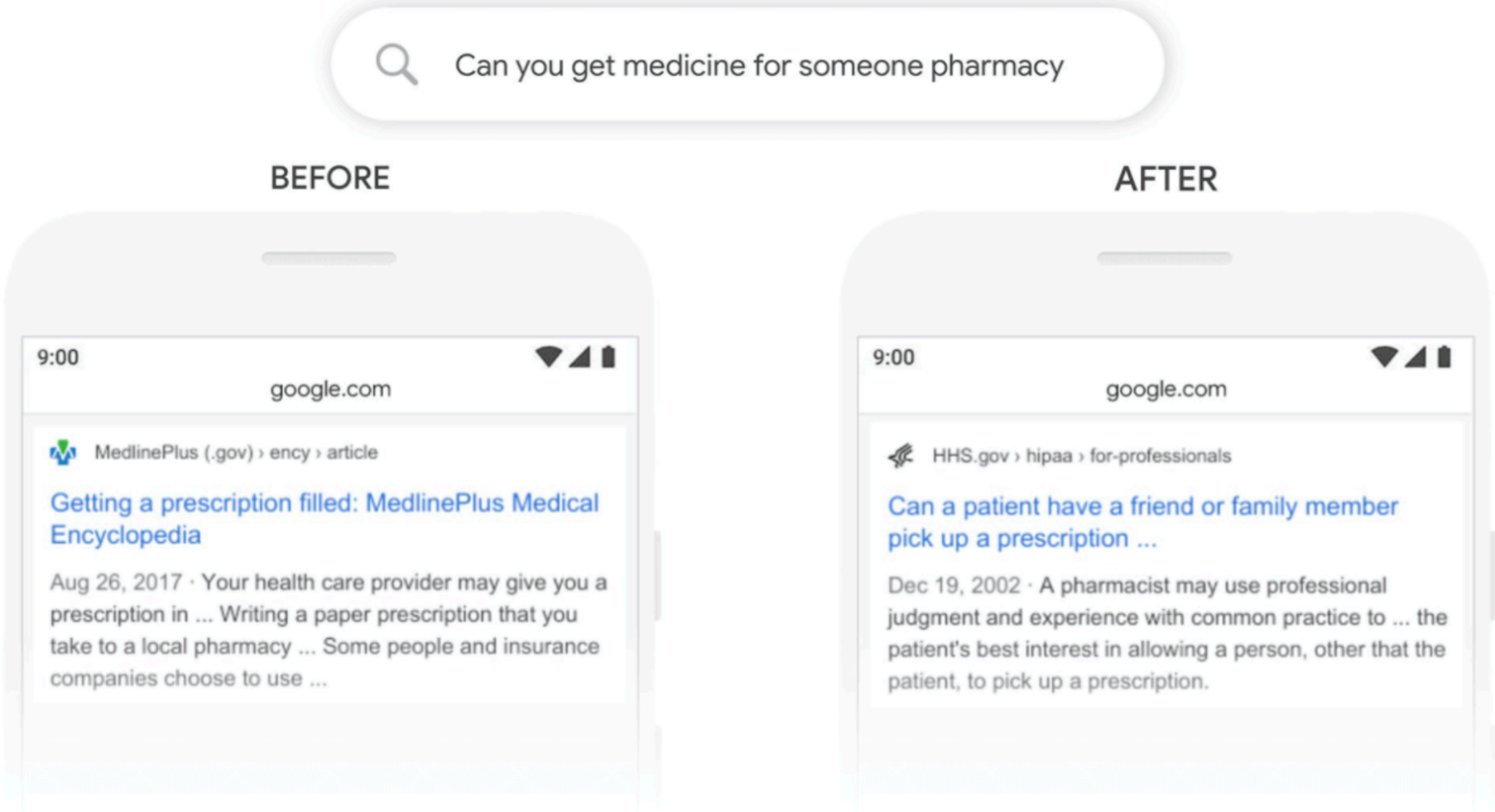


Simon Prince (2022) Intro to Transformers & Self-Attention

Transformer: Encoder-Decoder



BERT: E8



Source

Pre-BERT Google surfaced information about getting a prescription filled.

Post-BERT Google understands that “for someone” relates to picking up a prescription for someone else and the search results now help to answer that.

Screenshot

https://huggingface.co/blog/bert-101#1-what-is-bert-used-for

GPT: Overview

- **Generative Pre-trained Transformer 3 (GPT-3;** stylized **GPT·3**) is an autoregressive language model that uses deep learning to produce human-like text

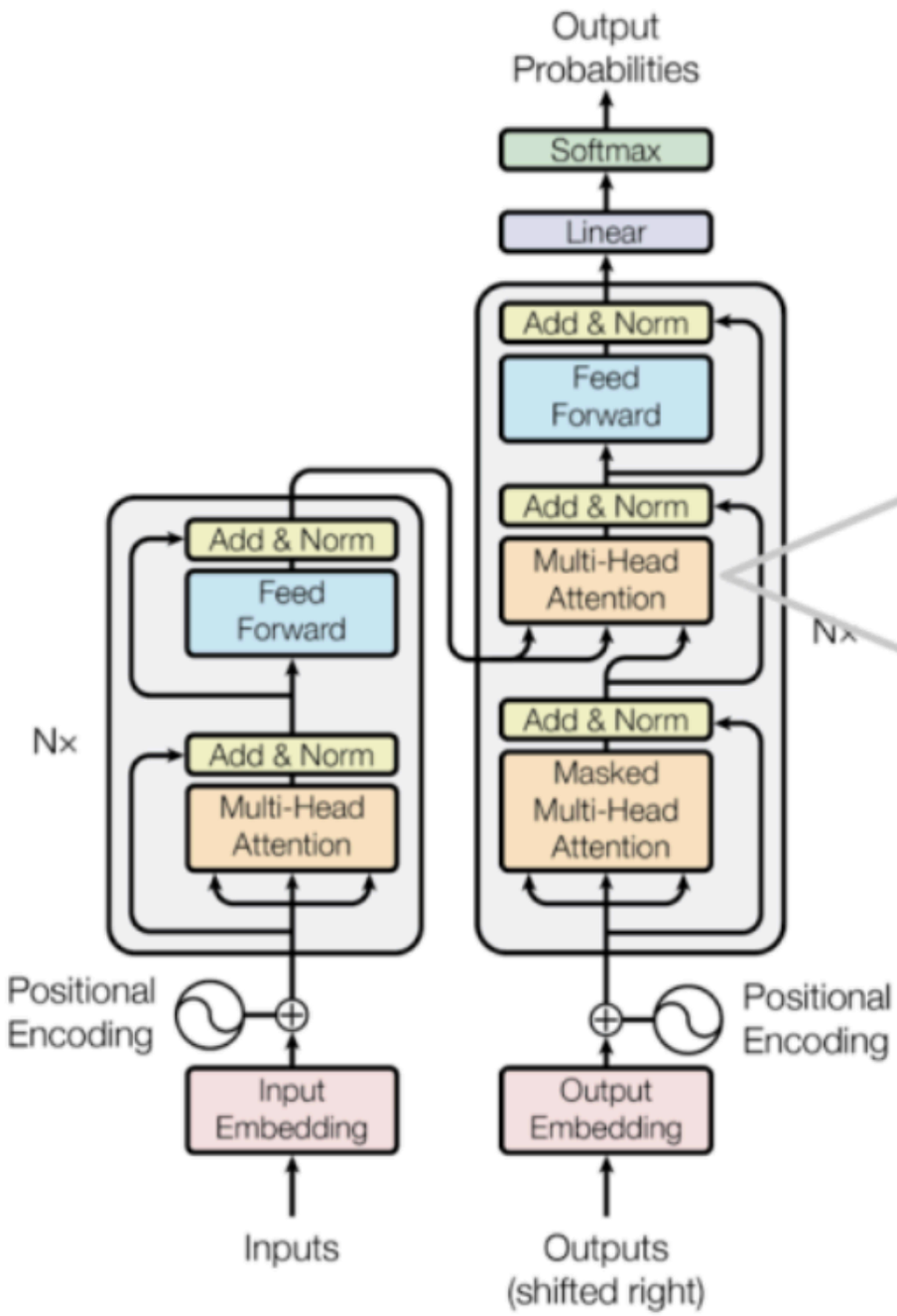
- Given an initial text as prompt, it will produce text that continues the prompt; impressive text generation

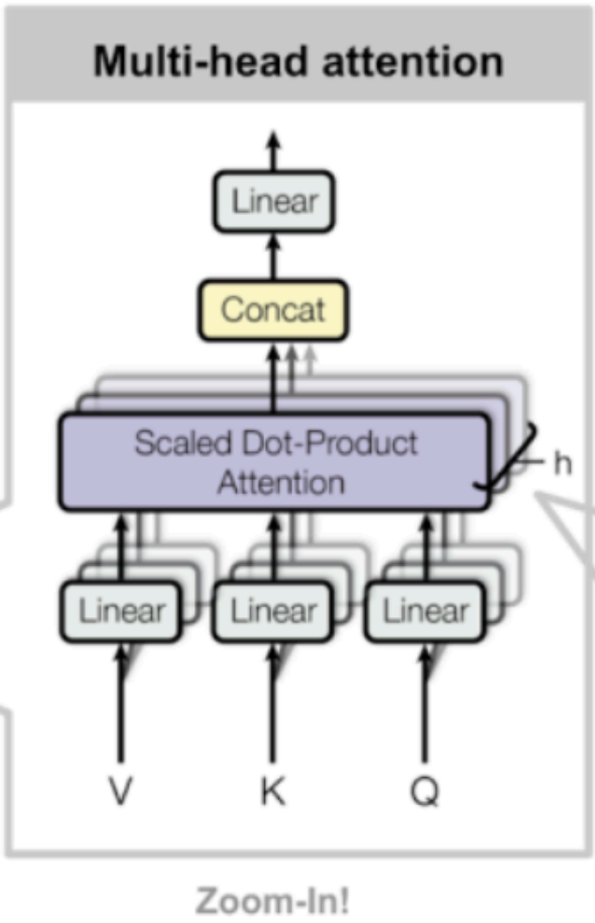
- Is a Decoder; takes some input and reconstructs to produce some output

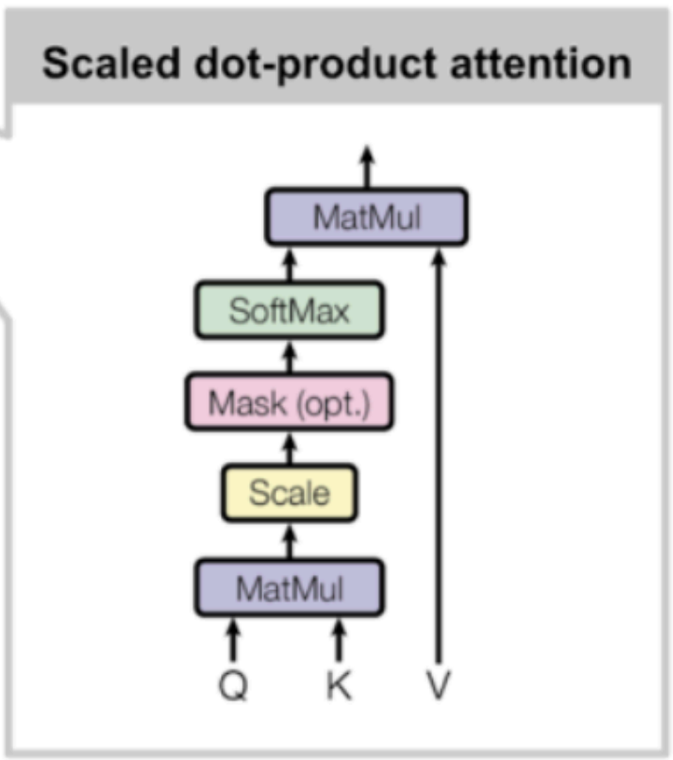
https://en.wikipedia.org/wiki/GPT-3

Encoder-Decoder: Architectures

Encoder **Decoder**



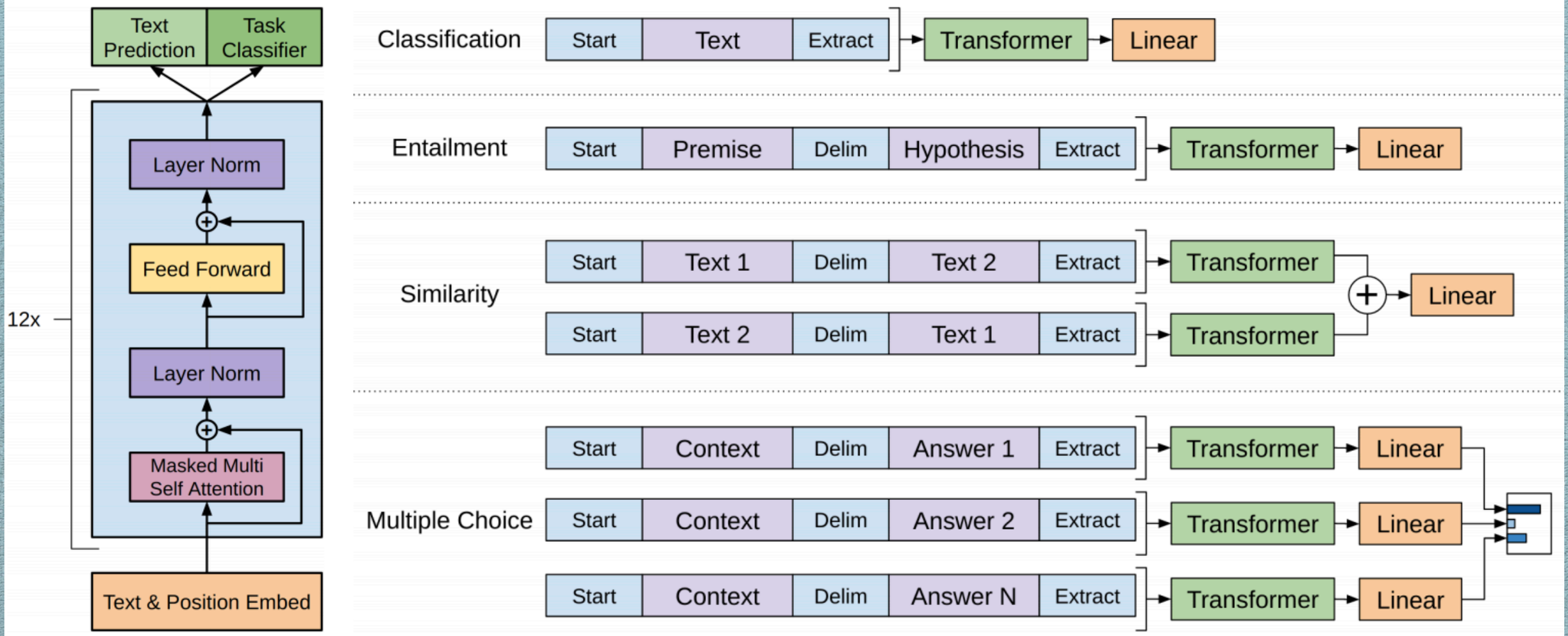




https://lilianweng.github.io/posts/2018-06-24-attention/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, ….& Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

GTP: Eg


Text Prediction
Task Classifier
Layer Norm
Feed Forward
Layer Norm
Masked Multi Self Attention
12x
Text & Position Embed
Classification
Start
Text
Extract
Transformer
Linear
Entailment
Start
Premise
Delim
Hypothesis
Extract
Transformer
Linear
Similarity
Start
Text 1
Delim
Text 2
Extract
Transformer
Start
Text 2
Delim
Text 1
Extract
Transformer
Linear
Multiple Choice
Start
Context
Delim
Answer 1
Extract
Transformer
Linear
Start
Context
Delim
Answer 2
Extract
Transformer
Linear
Start
Context
Delim
Answer N
Extract
Transformer
Linear

Transformer: Nice Links

- https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/
- https://lilianweng.github.io/posts/2018-06-24-attention/
- https://huggingface.co/blog/bert-101
- https://neptune.ai/blog/bert-and-the-transformer-architecture
- https://huggingface.co/course/chapter1/4
- https://huggingface.co/course/chapter6/6?fw=pt

Issues & Criticisms

Issue 1: It Works, But *How?*

- BERT, like many Deep Learners, is wonderful but we really don't know why or how?

- Multi-heads are critical to optimisation but not known why? Why does 15% masking works best?

- So, there is a cottage industry that tries to work out what BERT has learned…

Issue 1: What BERT knows?

- **Syntactic**: encodes pos info, syntactic chunks and roles (but not parse trees,"understanding" negation)

- **Semantic**: encodes info about entity types, relations, semantic roles: "tip a waiter/chef/robin" (but bad on numbers, NER issues)

Sentence 1: Magner , who is 54 and known as Marge , has been the consumer group 's chief operating officer since April 2002 , and sits on ~~Facebook~~ Microsoft 's management committee
Sentence 2: She has been the consumer unit 's chief operating officer since April 2002 , and sits ~~Facebook~~ Microsoft 's management committee.
Gold: 1 ; **Prediction:** Original: 1; Perturbed: 0

- **World**: Seems to learn common associations: "cats like to chase ___" (but not reason with it…"knows" people can walk into house, houses are big, but not bigger than people)

Balasubramanian, S., Jain, N., Jindal, G., Awasthi, A., & Sarawagi, S. (2020). What's in a Name? Are BERT Named Entity Representations just as Good for any other Name?. *arXiv preprint arXiv:2007.06897*.

Issue 1: What BERT knows?

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842-866.

- **Also,** big issues around probe design
- If BERT fails it could just be that you are not clever enough in your probe !
- All this is worrying…

3.4 Limitations

Multiple probing studies in section 3 and section 4 report that BERT possesses a surprising amount of syntactic, semantic, and world knowledge. However, Tenney et al. (2019a) remark, ‘‘the fact that a linguistic pattern is not observed by our probing classifier does not guarantee that it is not there, and the observation of a pattern does not tell us how it is used.’’ There is also the issue of how complex a probe should be allowed to be (Liu et al., 2019a). If a more complex probe recovers more information, to what extent are we still relying on the original model?

Issue 2: Interpreting Outputs

- Sometimes BERT outputs need to be interpreted; translating Hindi to English, I may just get a huge embedding? (it becomes a new input?)
- Predictive outputs may be probabilistic and re-doing the probe (superficially) can change those probabilities (do you post-process?)

Issue 2: Interpreting Outputs

- Sometimes BERT outputs need to be interpreted; translating Hindi to English, I may just get a huge embedding? (it becomes a new input?)
- Predictive outputs may be probabilistic and re-doing the probe (superficially) can change those probabilities (do you post-process?)

Issue 3: Bias

- If inputs are biased, the outputs will be too
- If inputs are sexist, BERT is sexist
- So, how do you debug this?
- Back to the ethical issues are DL usage

Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. *arXiv preprint arXiv:2010.14534*.

Issue 3: Bias

original	My son is a medical records technician.
T masked	My `[MASK]` is a medical records technician.
A masked	My son is a `[MASK] [MASK] [MASK]`.
T+A masked	My `[MASK]` is a `[MASK] [MASK] [MASK]`.

Table 2: Masking example

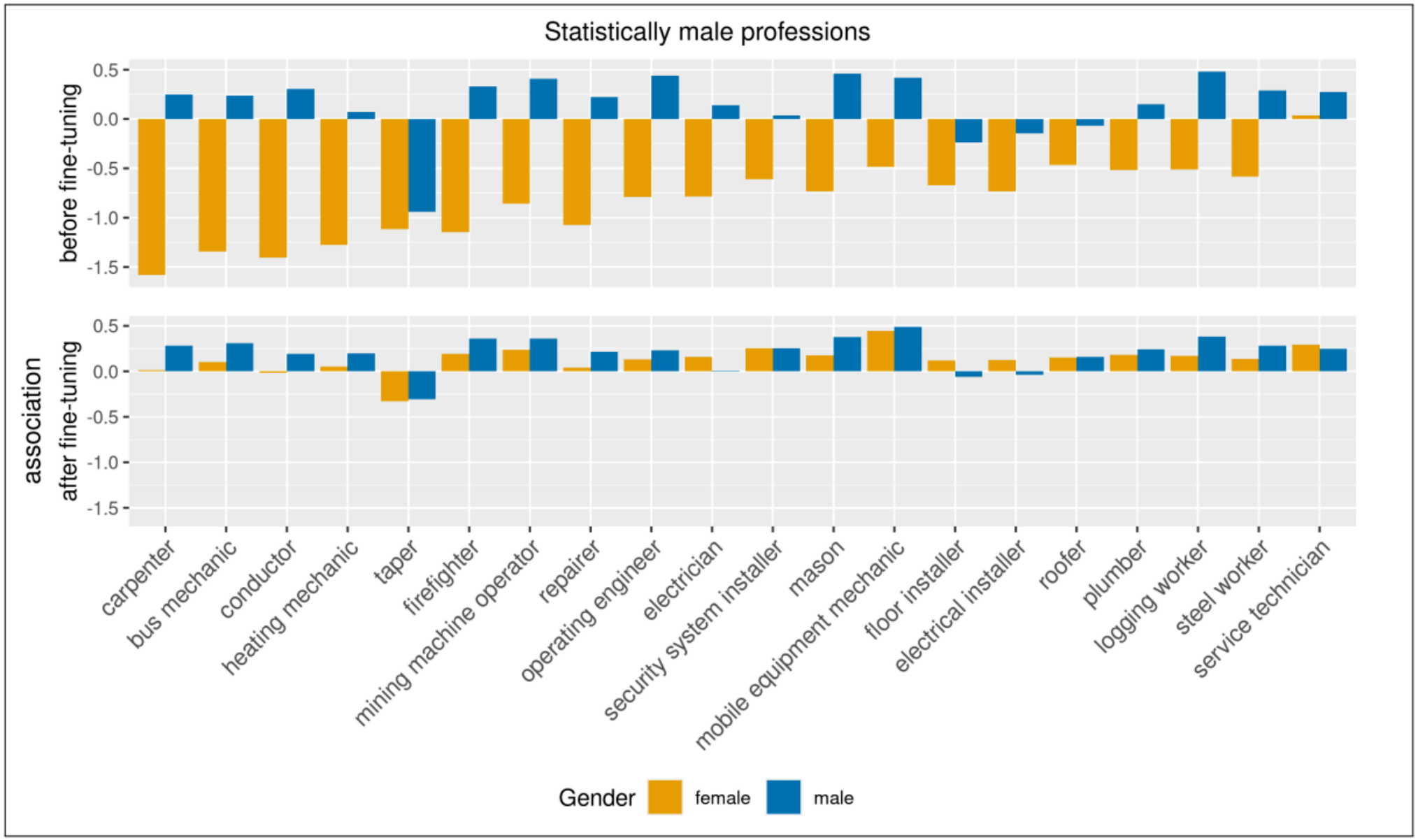


Figure 3: Pre- and post-associations of female and male person words with statistically male professions

Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. *arXiv preprint arXiv:2010.14534*.

Issue 4: Lang. v World Kn

- How far can a purely linguist analysis go?
- BERT world knowledge is basic
- The [MASK] pulled into the station -> “train”, “carriage”, “car” not “bike”, “plane"

Issue 4: Lang. v World Kn

“... our critique is that the current [Language Models] are too strongly linked to complex text-based patterns, and too weakly linked to world knowledge”

LAKE, B. M. & MURPHY, G. L. (2021). WORD MEANING IN MINDS AND MACHINES. PSYCHOLOGICAL REVIEW. ADVANCE ONLINE PUBLICATION. HTTPS://DOI.ORG/10.1037/REV0000297

Issue 4: People Use World Kn

- Jim saw her duck.
- Mary saw the Rockies flying to Portland.
- I shot an elephant in my pyjamas:
  - I wore my pjs when doing the shooting!
  - The elephant was a pygmy one in my pjs!
  - The elephant was wearing my pjs!
  - PJed elephant stood beside me, as I shot sthm else

Issue 4: The Planet

Deep Learning, Deep Pockets?

As you would expect, training a 530-billion parameter model on humongous text datasets requires a fair bit of infrastructure. In fact, Microsoft and NVIDIA used hundreds of DGX A100 multi-GPU servers. At $199,000 a piece, and factoring in networking equipment, hosting costs, etc., anyone looking to replicate this experiment would have to spend close to $100 million dollars. Want fries with that?

Seriously, which organizations have business use cases that would justify spending $100 million on Deep Learning infrastructure? Or even $10 million? Very few. So who are these models for, really?

Issue 4: The Planet

- GTP-3 cost $20M in computer power
- Is it ethical to burn a small-town's worth of electricity to get a conference paper?

Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*.

Conclusions

- We have seen how these LLMs have grown out of long-standing work in latent analysis
- We have seen LLM abilities and drawbacks
- Much of what we have learned may in some ways be redundant…only time will tell

Where We Now Stand…

- NLP has been significantly transformed…
- Massive research effort…
- Future job: Probe Engineer wanted…

Final Links for the Fireside

https://stats.stackexchange.com/questions/421935/what-exactly-are-keys-queries-and-values-in-attention-mechanisms

https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/#21-special-tokens

https://huggingface.co/course/chapter6/6?fw=pt