What we got so far …

- Each term has one learned embeddings vector, called “word type” embeddings.

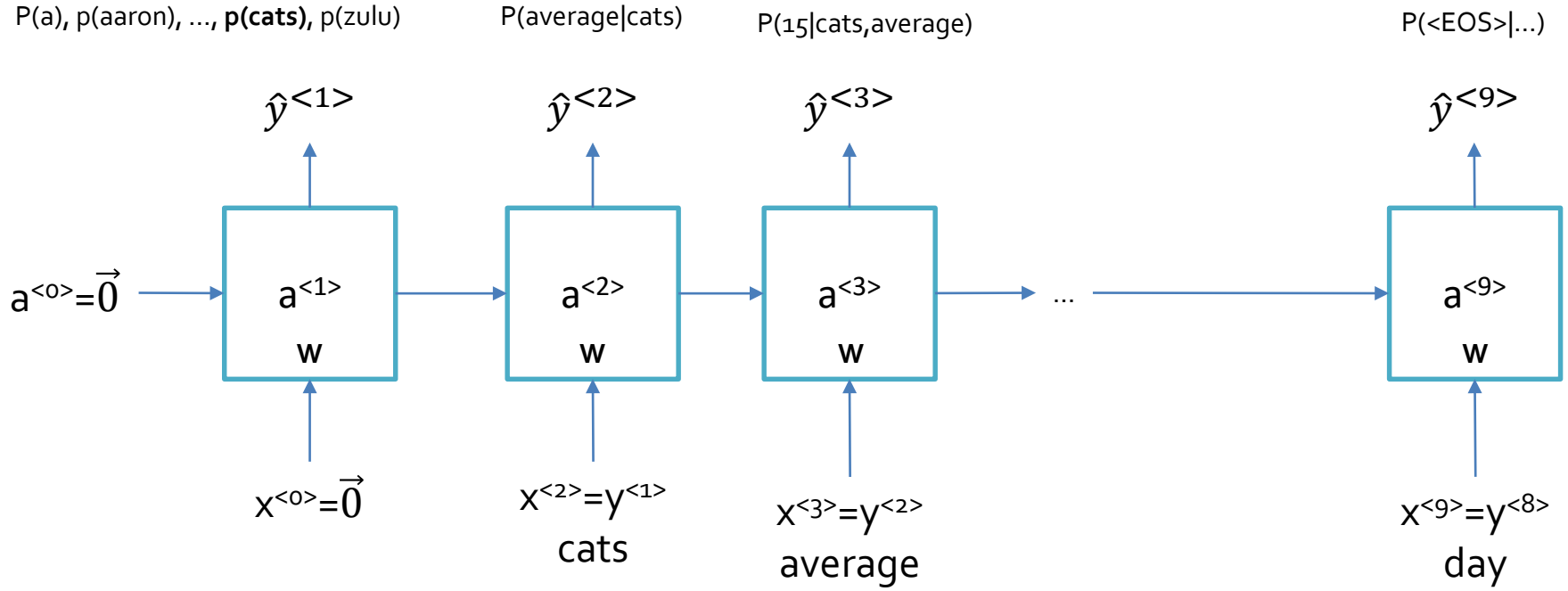
Problem?

- I passed over to the river *bank*
- I withdrew money from my *bank* account
- There is a shortage in the blood *bank*

Words have different meanings in different contexts!

- We need “Contextual Word Vectors”, called “word token” embeddings.

Learning a Language Model using RNN

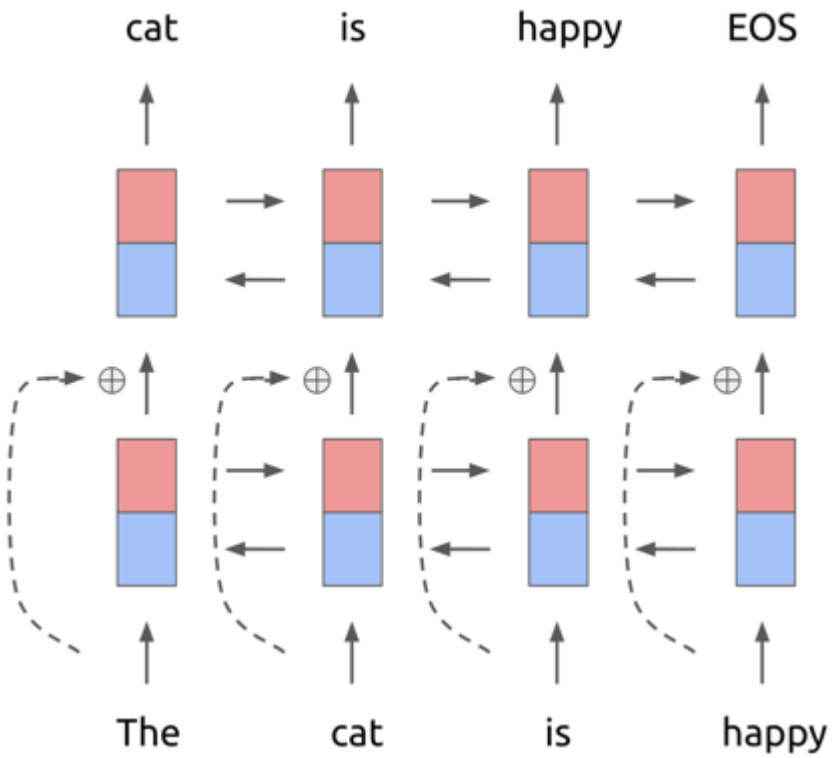


- Cats average 15 hours of sleep a day. <EOS>
  - P(sentence) = P(cats)P(average|cats)P(15|cats,average)…

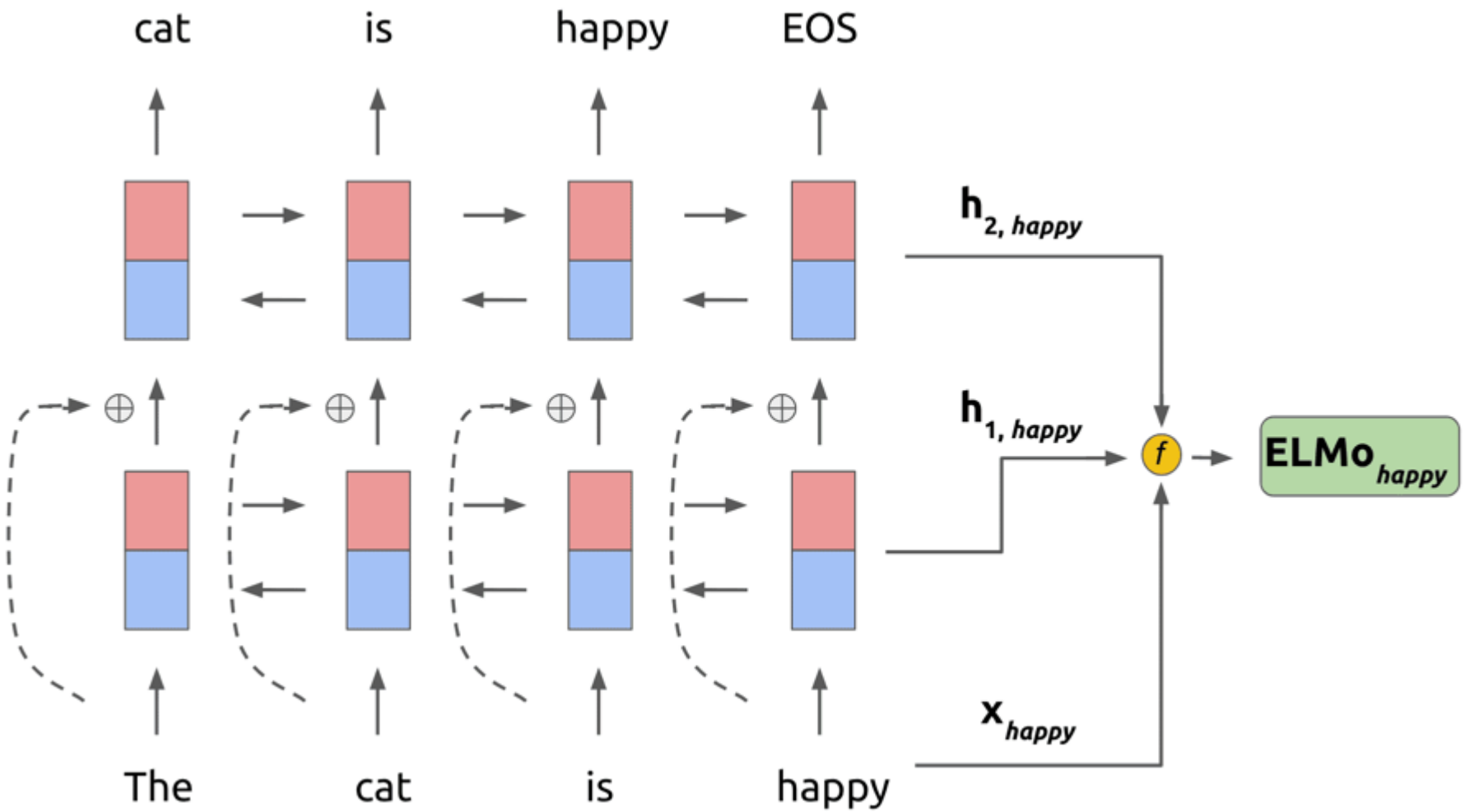
ELMo: Embeddings from Language Models

- Trains a language model using a 2-layer bi-directional LSTM (biLSTMs)

- Input: fixed-length word embedding.
  - One-hot encoding
  - Word2Vec
  - Glove
  - ….



ELMo: Embeddings from Language Models




24
POLL

24

ELMo provides different embeddings for the same word appearing in two different sentences.

- Yes
- No

To produce ELMo embeddings, we need ...

- the static embeddings of terms
- the learned ELMo model
- both

بسم الله الرحمن الرحيم

دورة "استرجاع المعلومات" باللغة العربية - صيف ٢٠٢١

Information Retrieval – Summer 2021

8. Introduction to Transformer & BERT

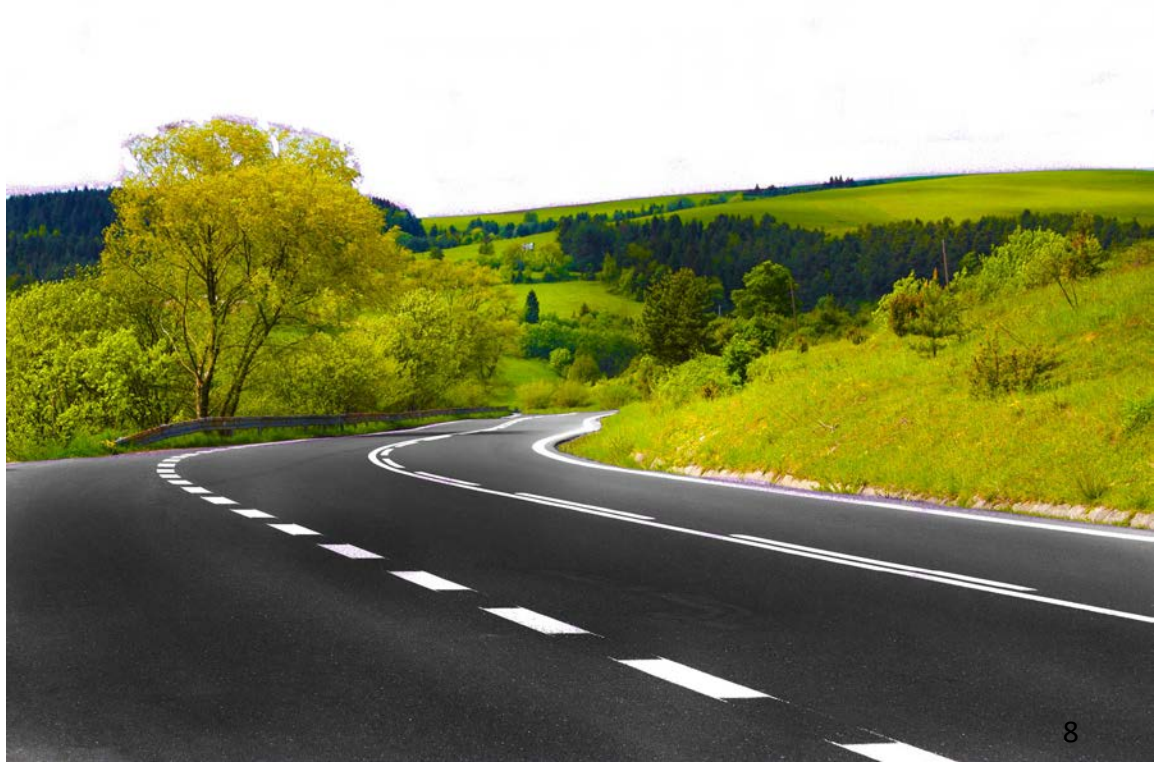
Tamer Elsayed

Qatar University

Today's Roadmap

- Transformer
- BERT

- *BERT for reranking?*

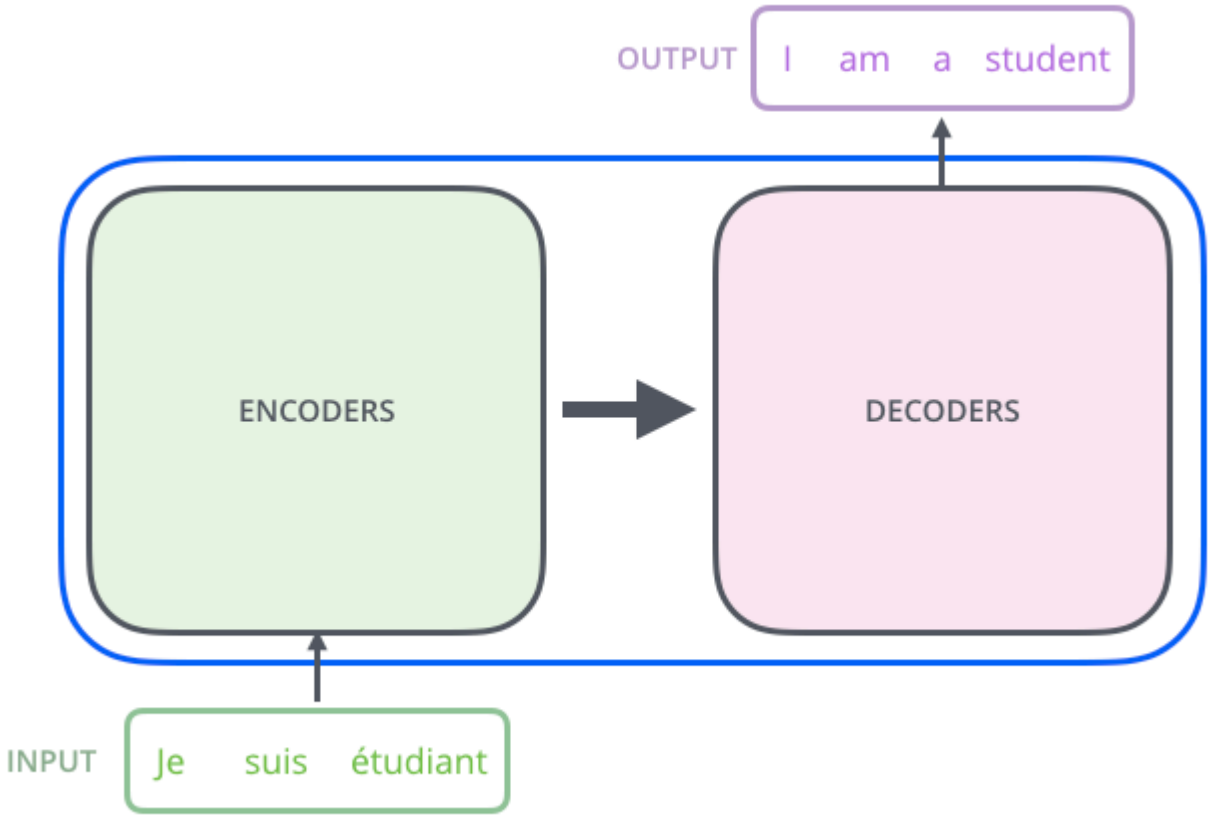


TRANSFORMER

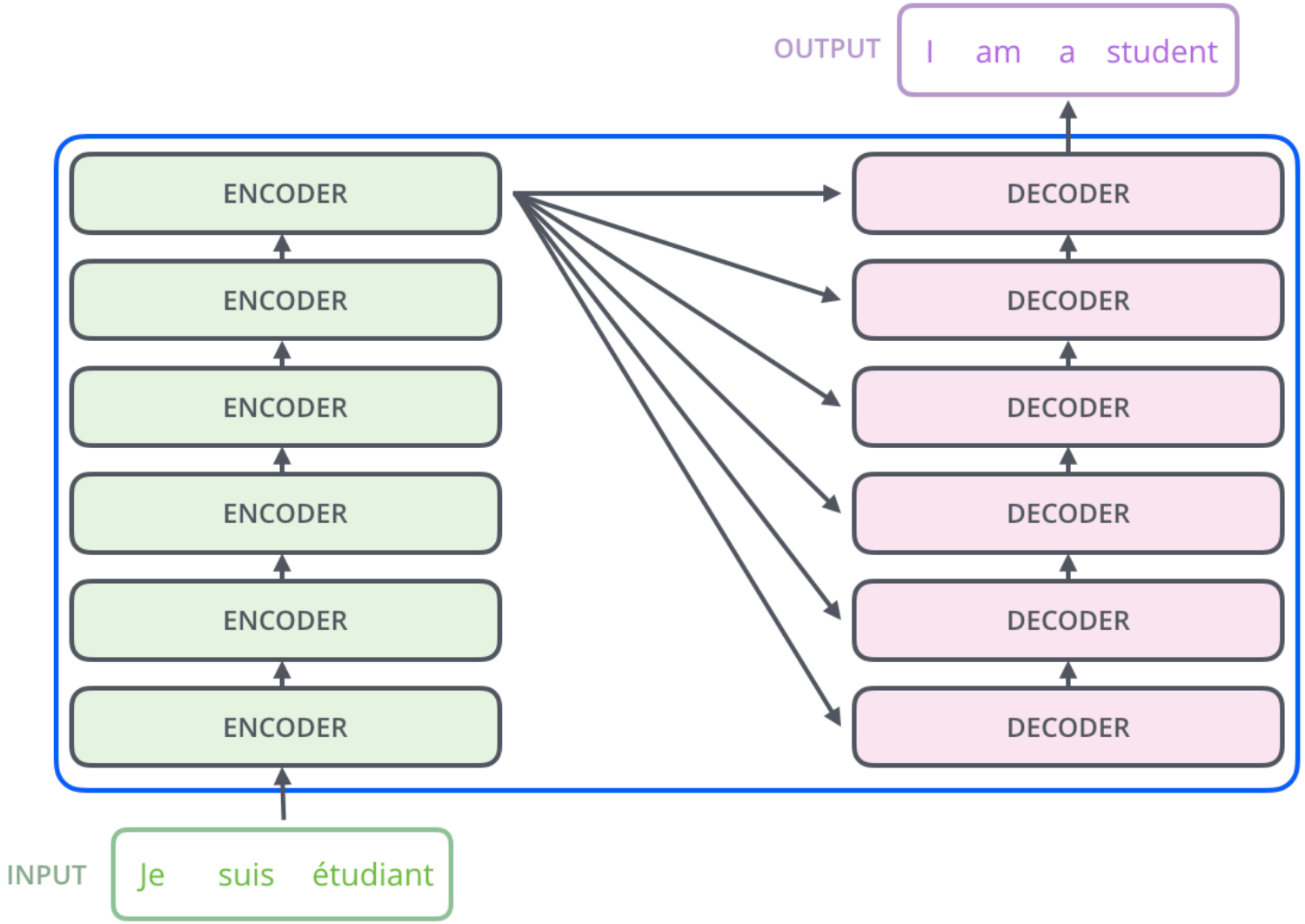
Transformer

Sequence-to-Sequence tasks, e.g., machine translation

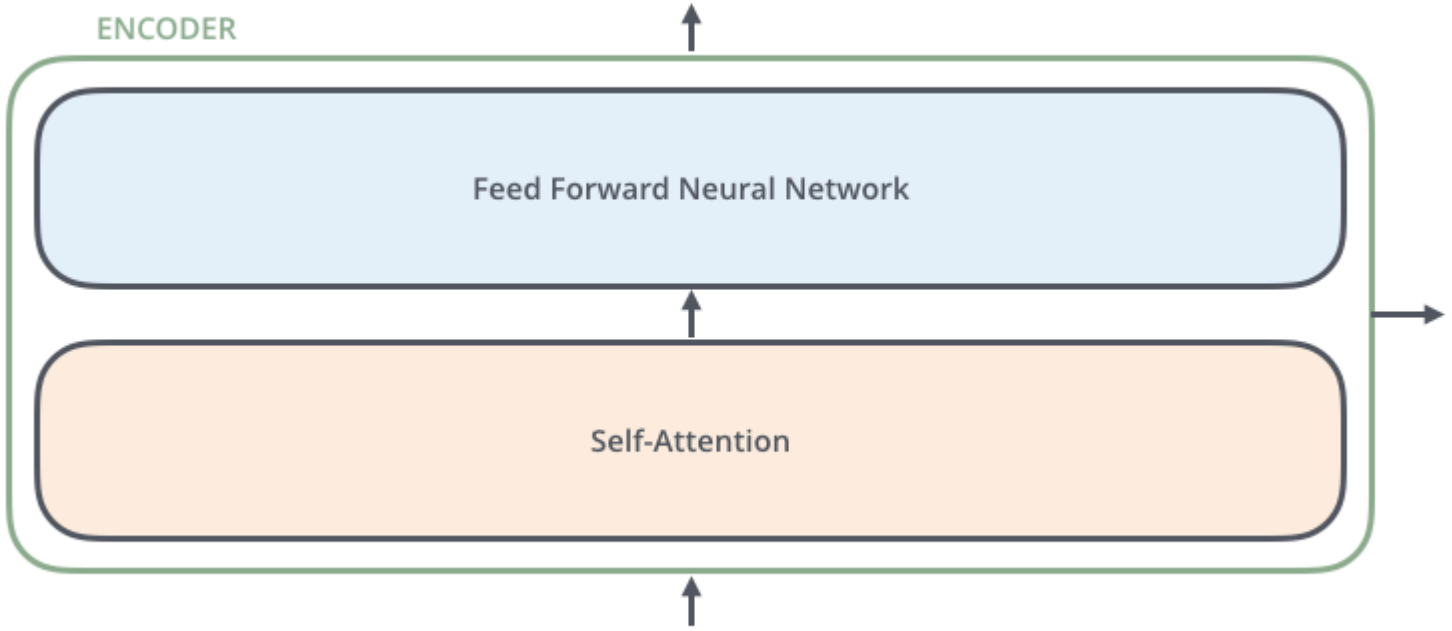
Transformer



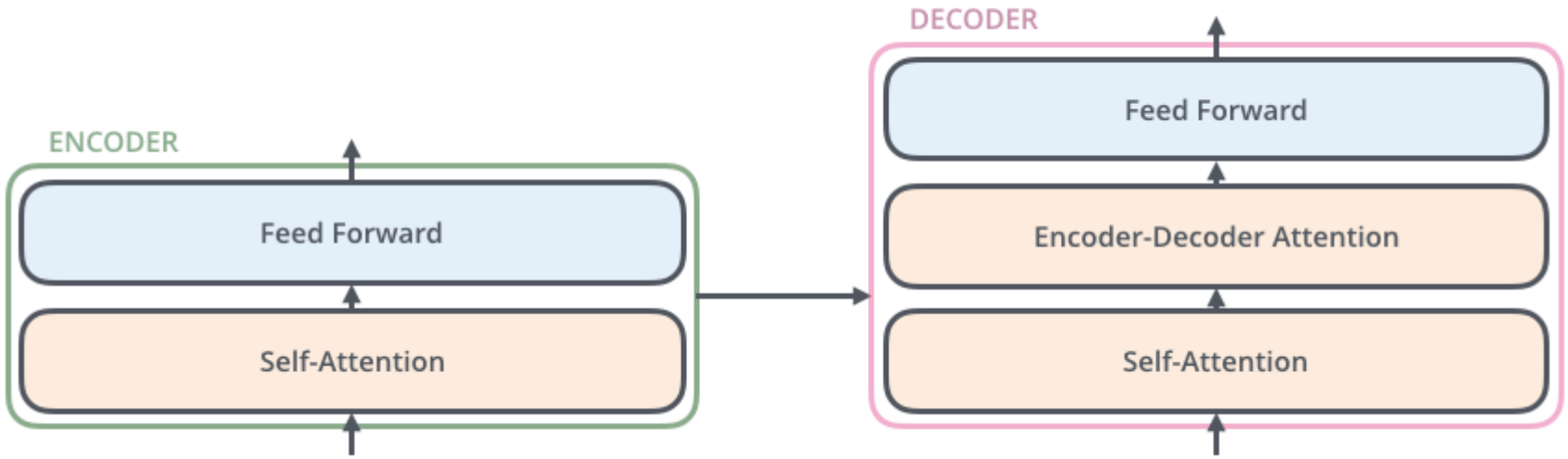
Stacks …


OUTPUT
I am a student
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
INPUT
Je suis étudiant

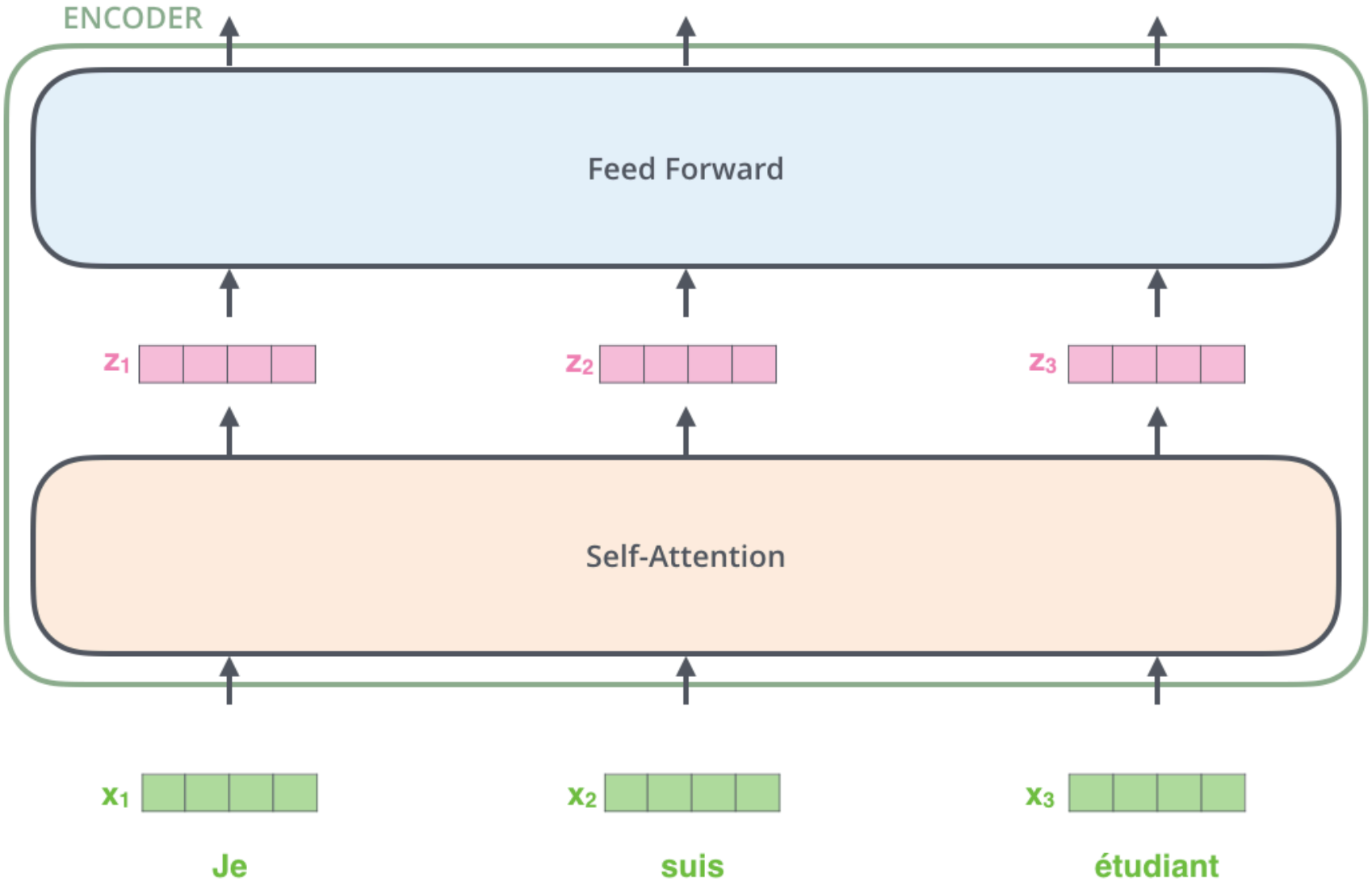
Encoders …



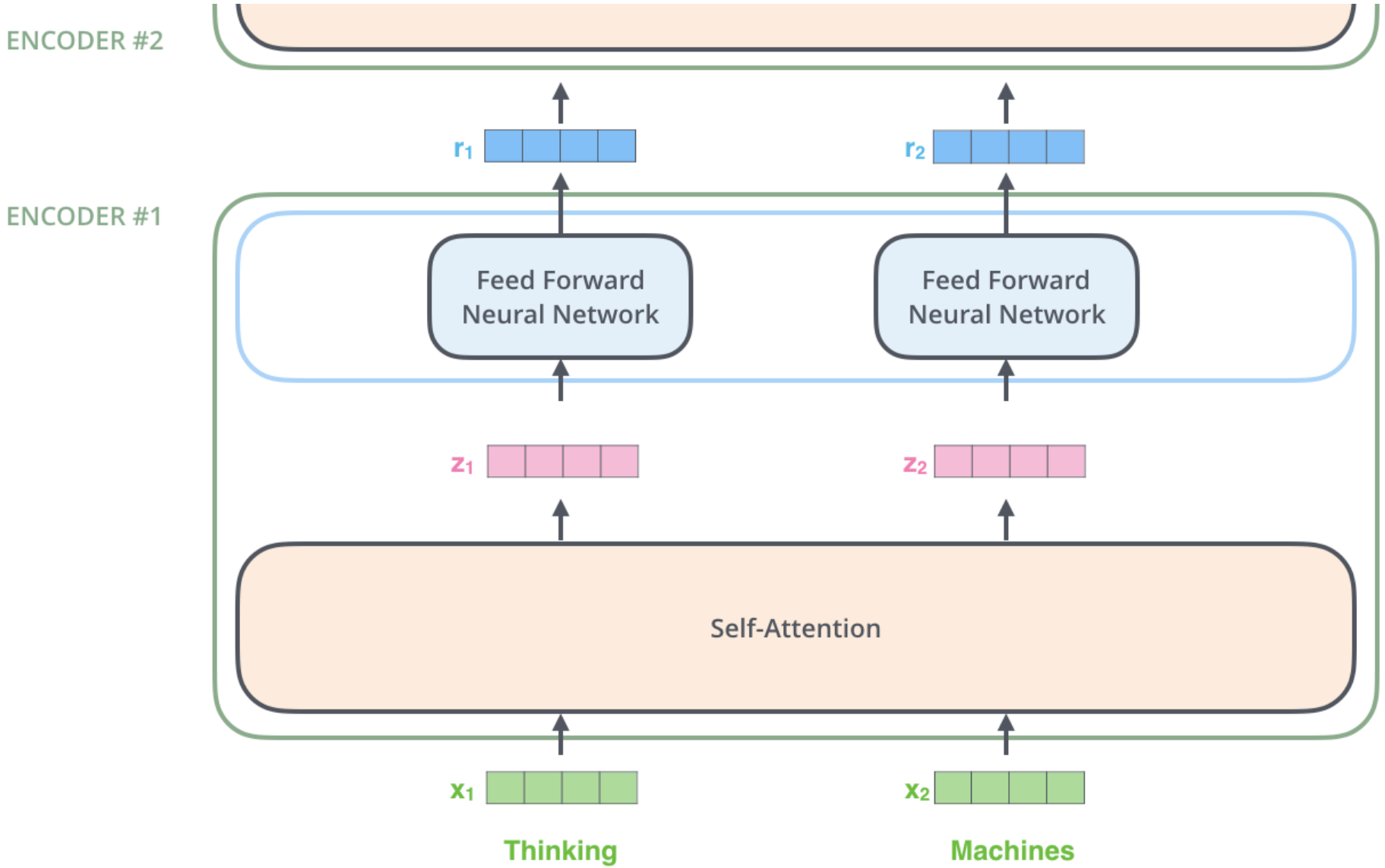
and Decoder …


ENCODER
Feed Forward
Self-Attention
DECODER
Feed Forward
Encoder-Decoder Attention
Self-Attention

Input Embeddings



Dependencies



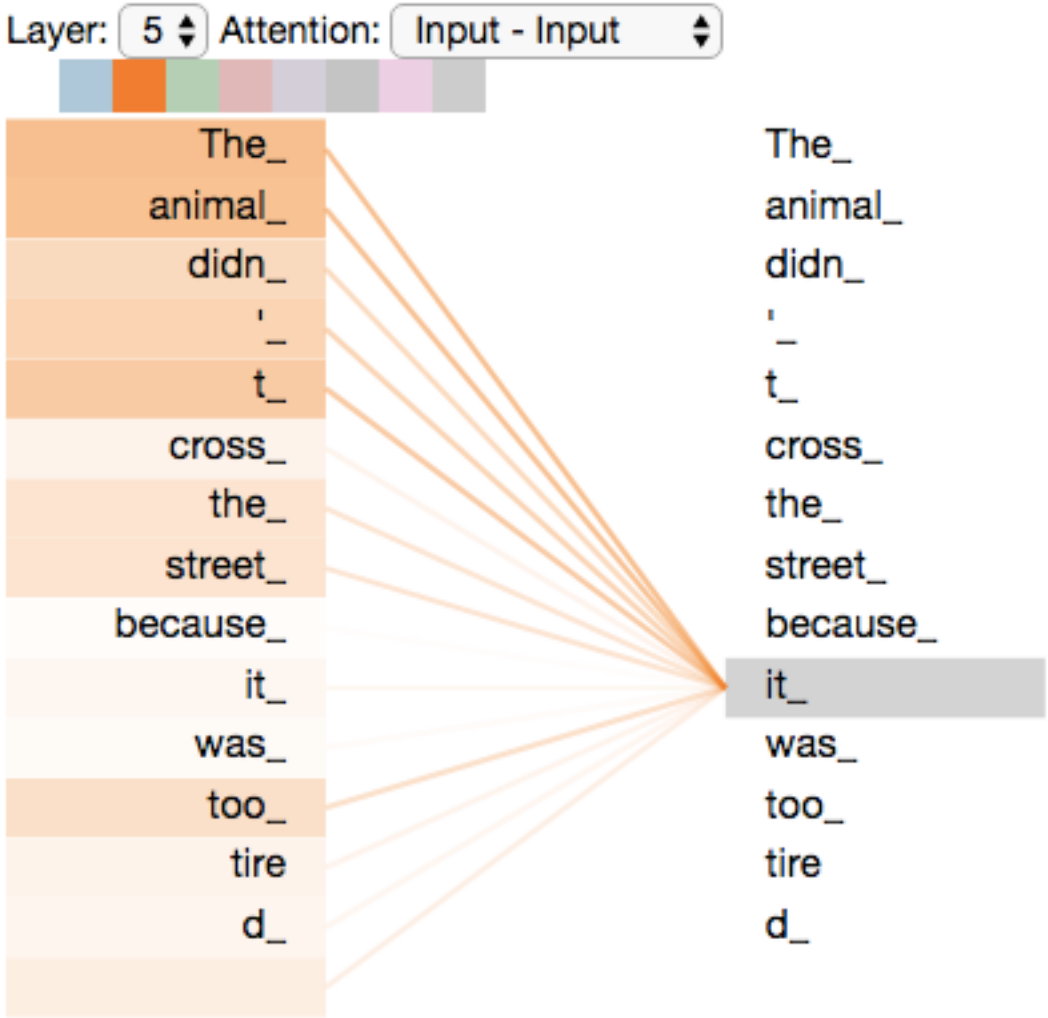
Self Attention

- Input: ”The animal didn't cross the street because it was too tired”

What does “it” in this sentence refer to?

- When the model is processing the word “it”, self-attention allows it to associate “it” with “animal”.

Self Attention

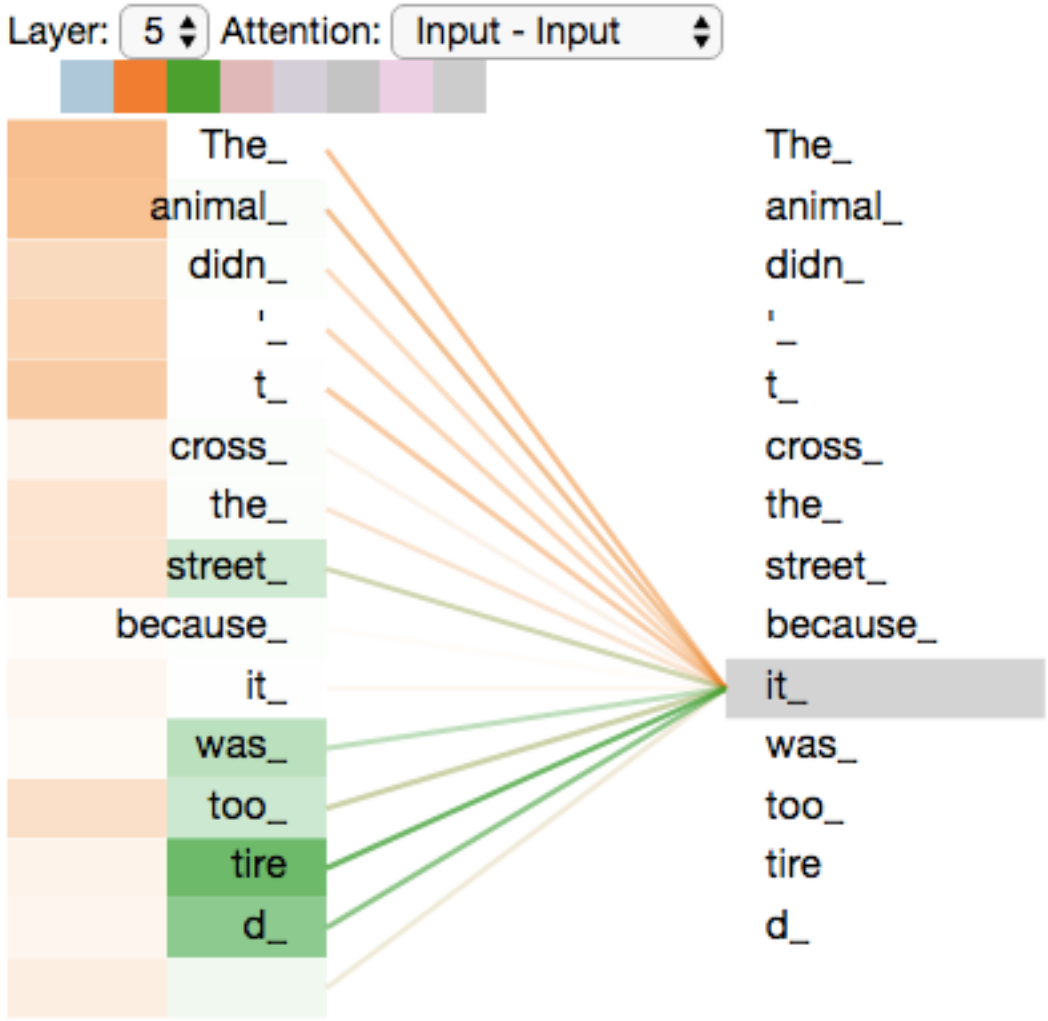


Self Attention

- As the model processes each word, self attention allows it to ***look at other positions in the input sequence for clues that can help lead to a better encoding*** for this word.

- **Self-attention**: The mechanism the Transformer uses to bake "understanding" of other relevant words into the one we're currently processing.

Multi-headed Attention




25
POLL

25

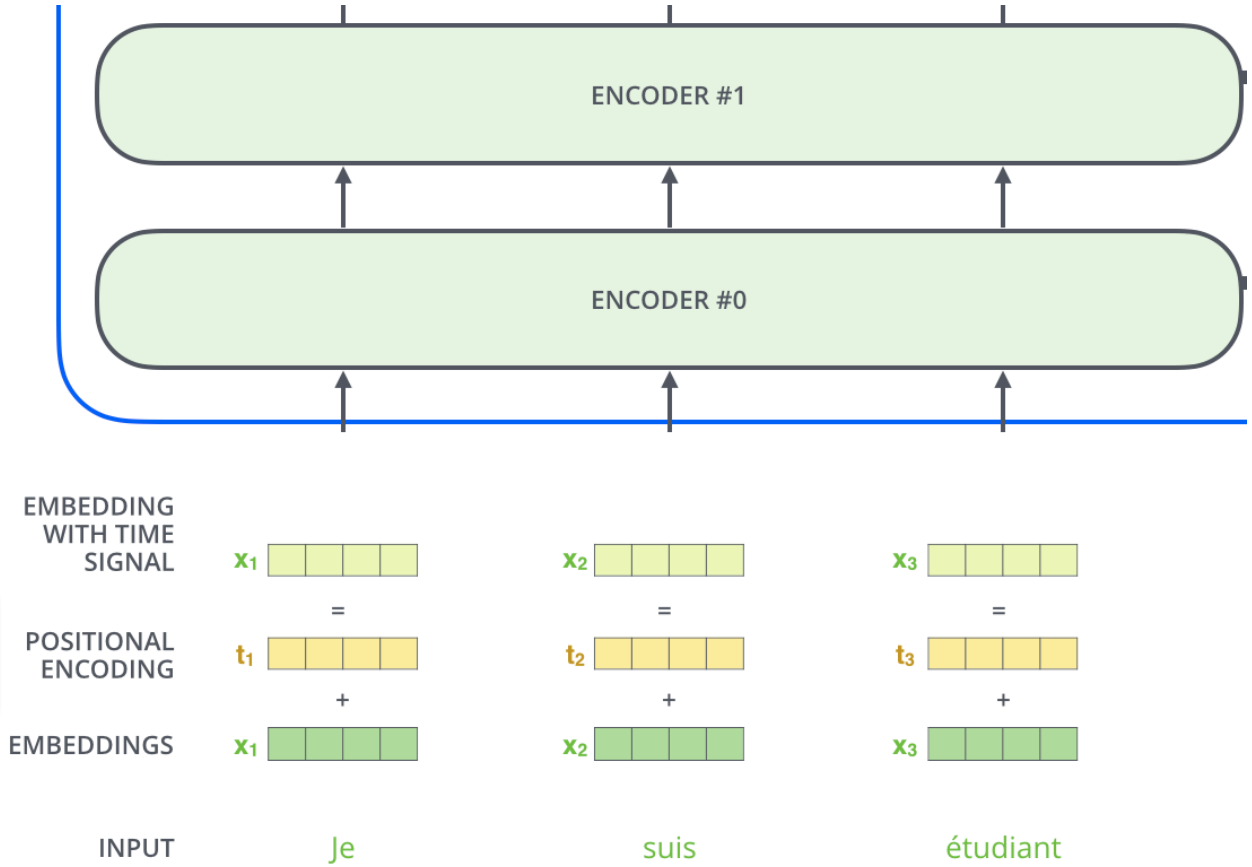
Self-attention sub-layer relates words from different input sentences.

- Yes
- No

In the Transformer architecture, the encoder and decoder have encoder-decoder attention sub-layer in each block (layer).

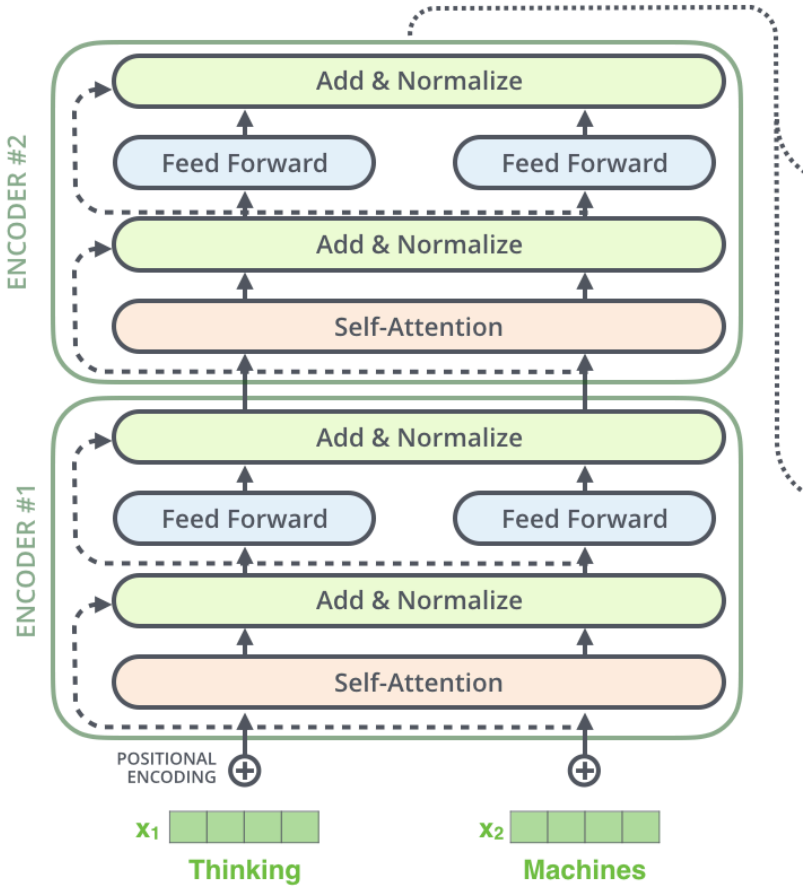
- Yes
- No

Positional Encoding

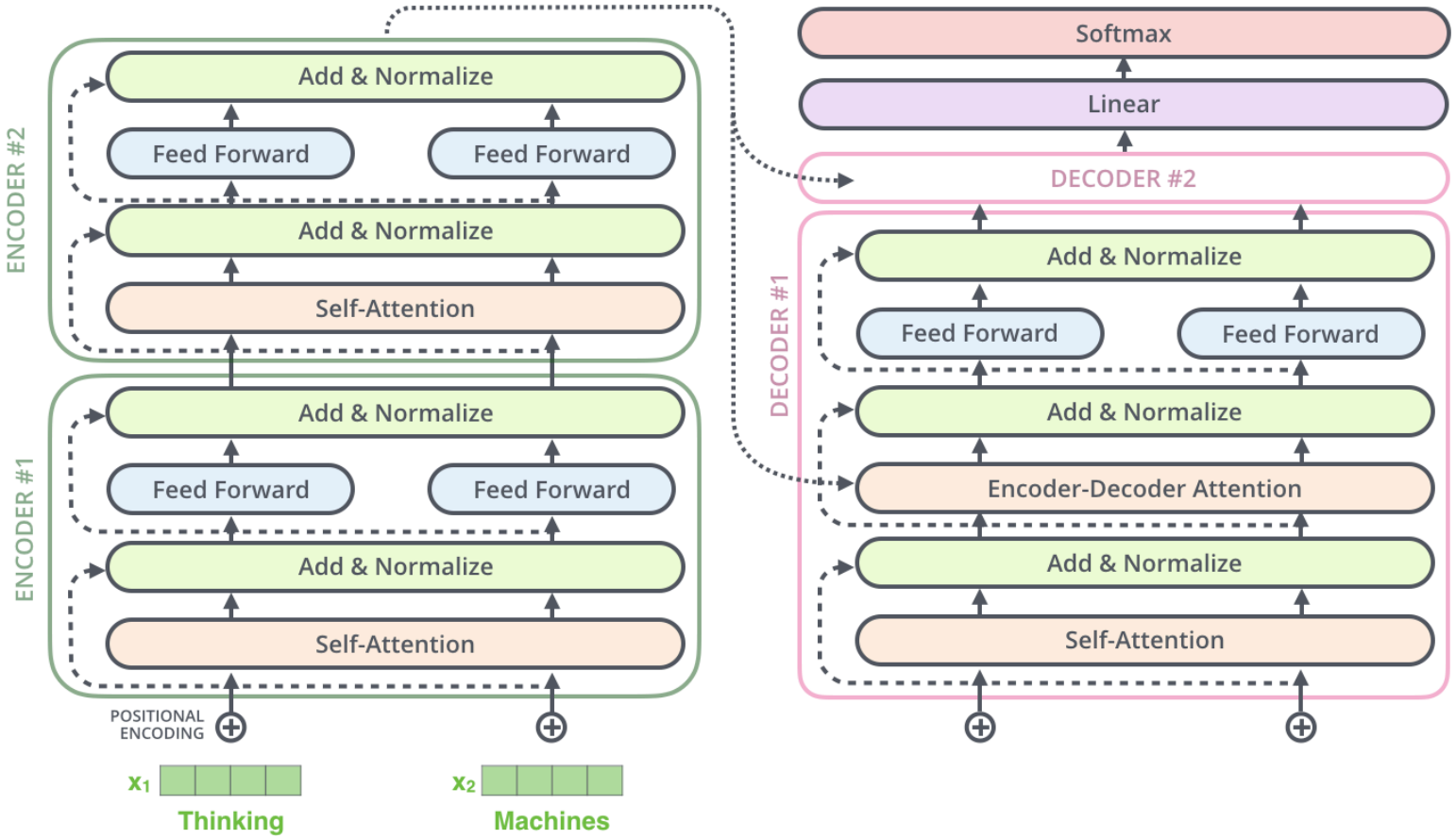


Representing the order of the sequence

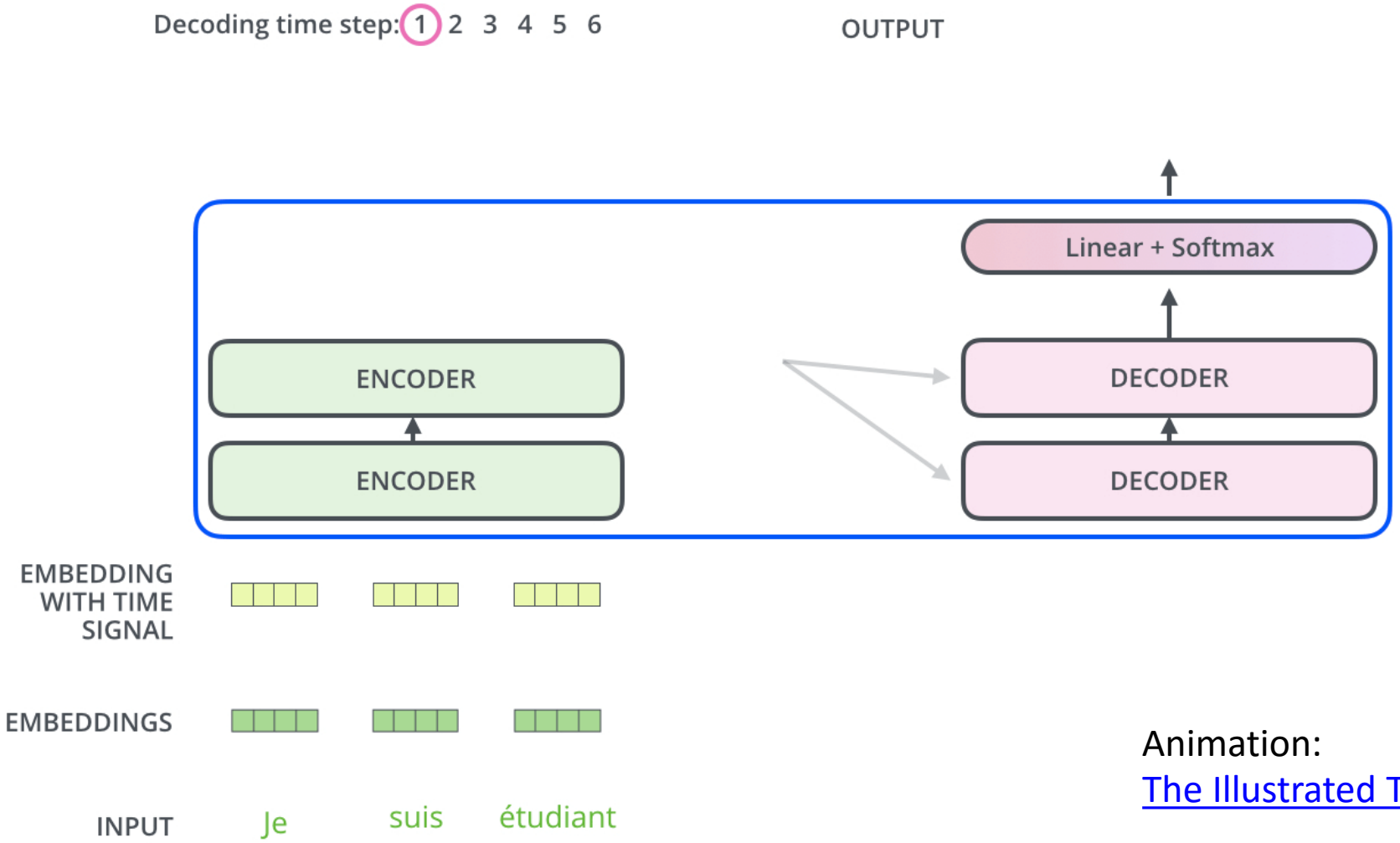
more …


ENCODER #2
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines

Decoder

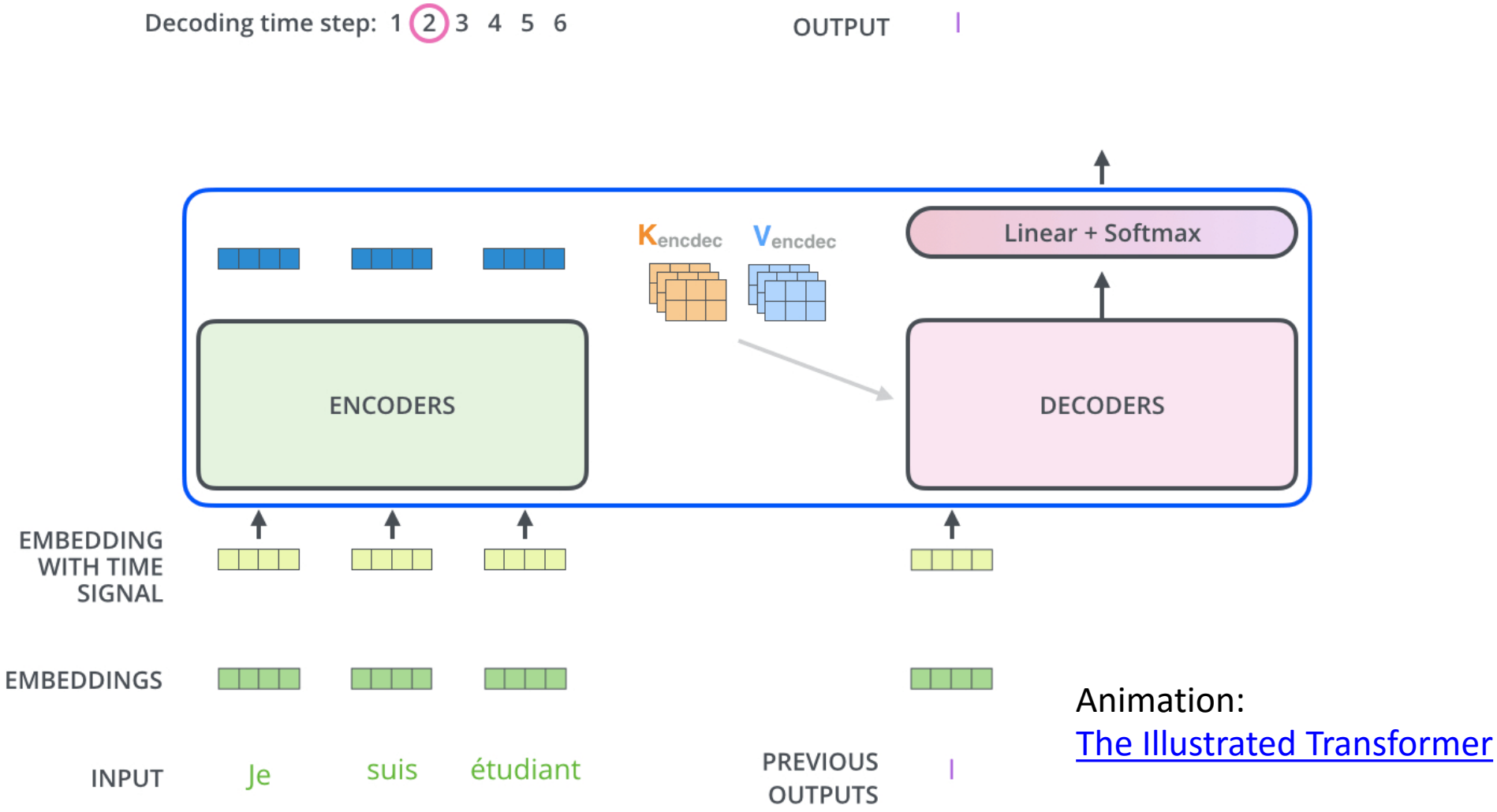

Softmax
Linear
DECODER #2
DECODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Encoder-Decoder Attention
Add & Normalize
Self-Attention
ENCODER #2
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines

Decoder



Animation:
[The Illustrated Transformer]

Decoder



Animation:
[The Illustrated Transformer]

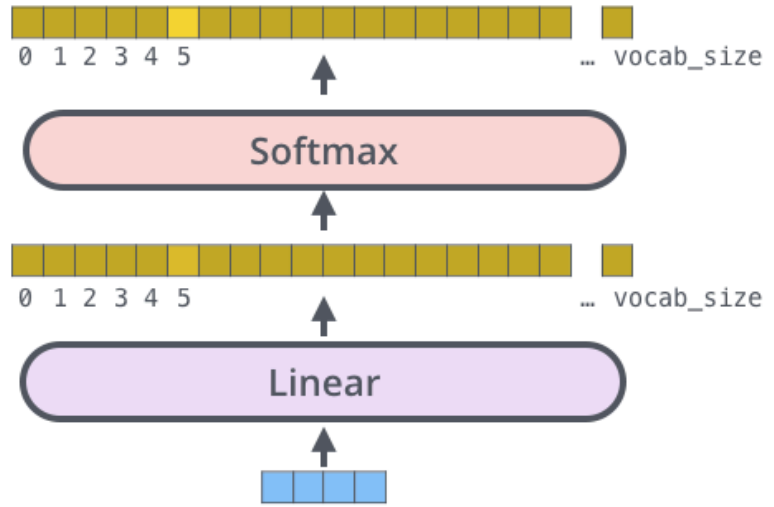
Final Linear and Softmax Layer

Which word in our vocabulary is associated with this index?

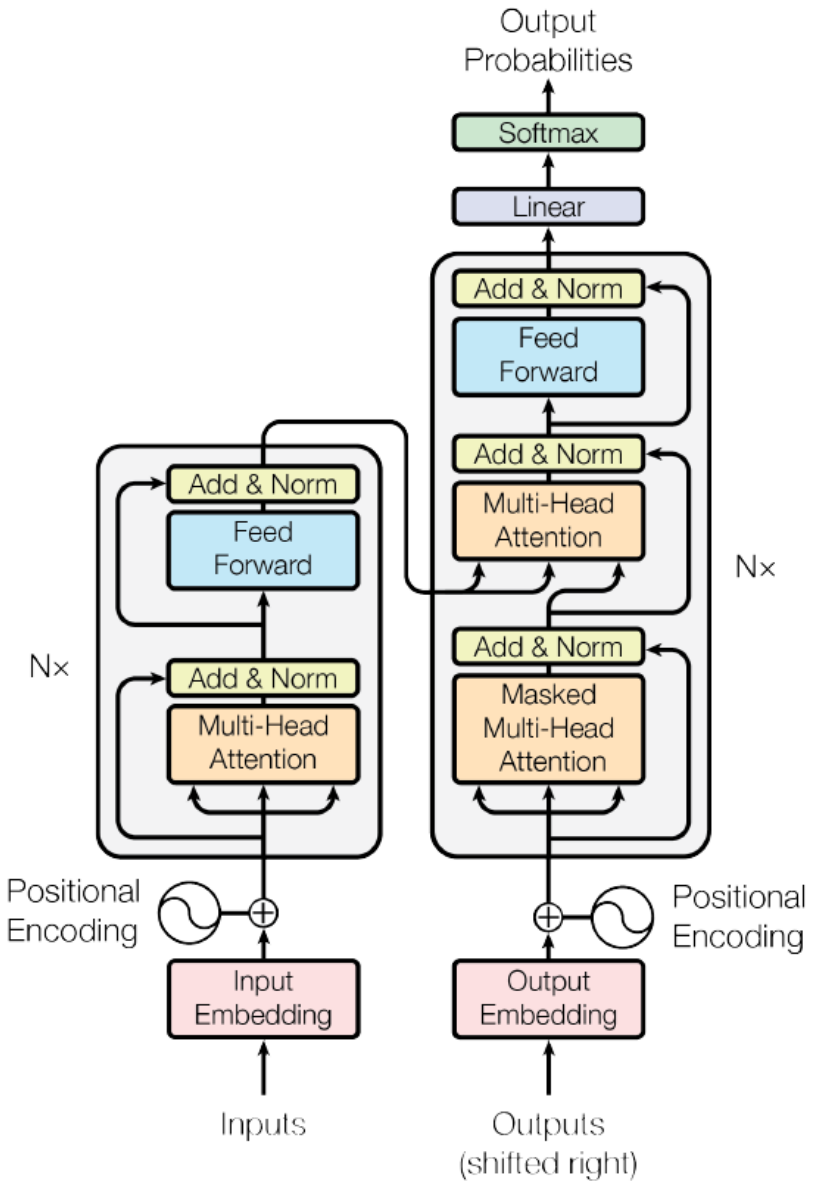
am

Get the index of the cell with the highest value **(argmax)**

5



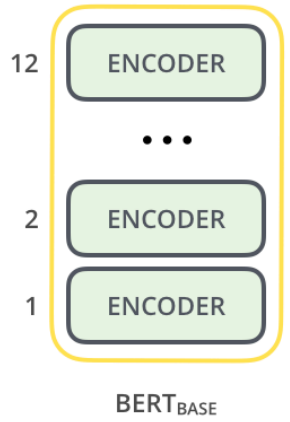
“Attention is All You Need”



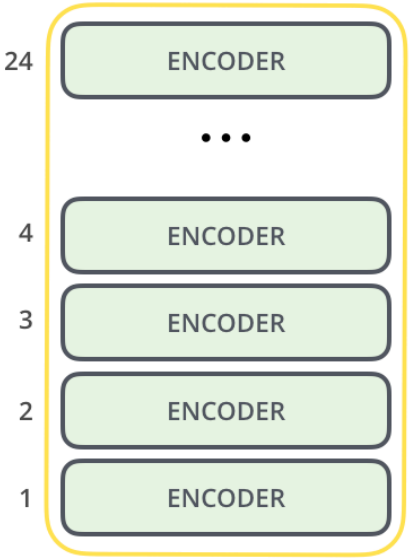
BERT

What is BERT?

- **Language representation model**:
  Bidirectional Encoder Representations from Transformers
- A trained Transformer Encoder stack.
- Larger feedforward-networks (768 and 1024 hidden units), and more attention heads (12 and 16) than the default (6 encoder layers, 512 hidden units, and 8 attention heads).

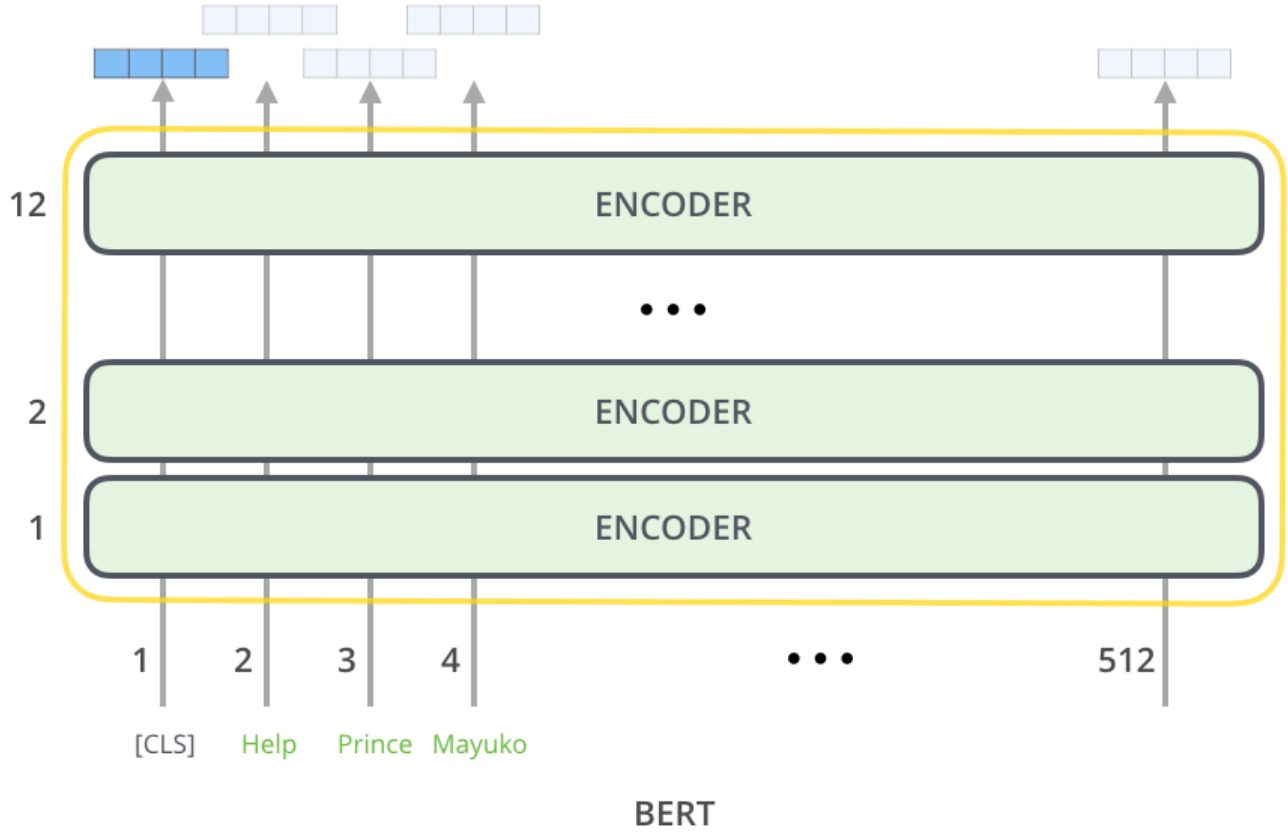


BERT$_{BASE}$

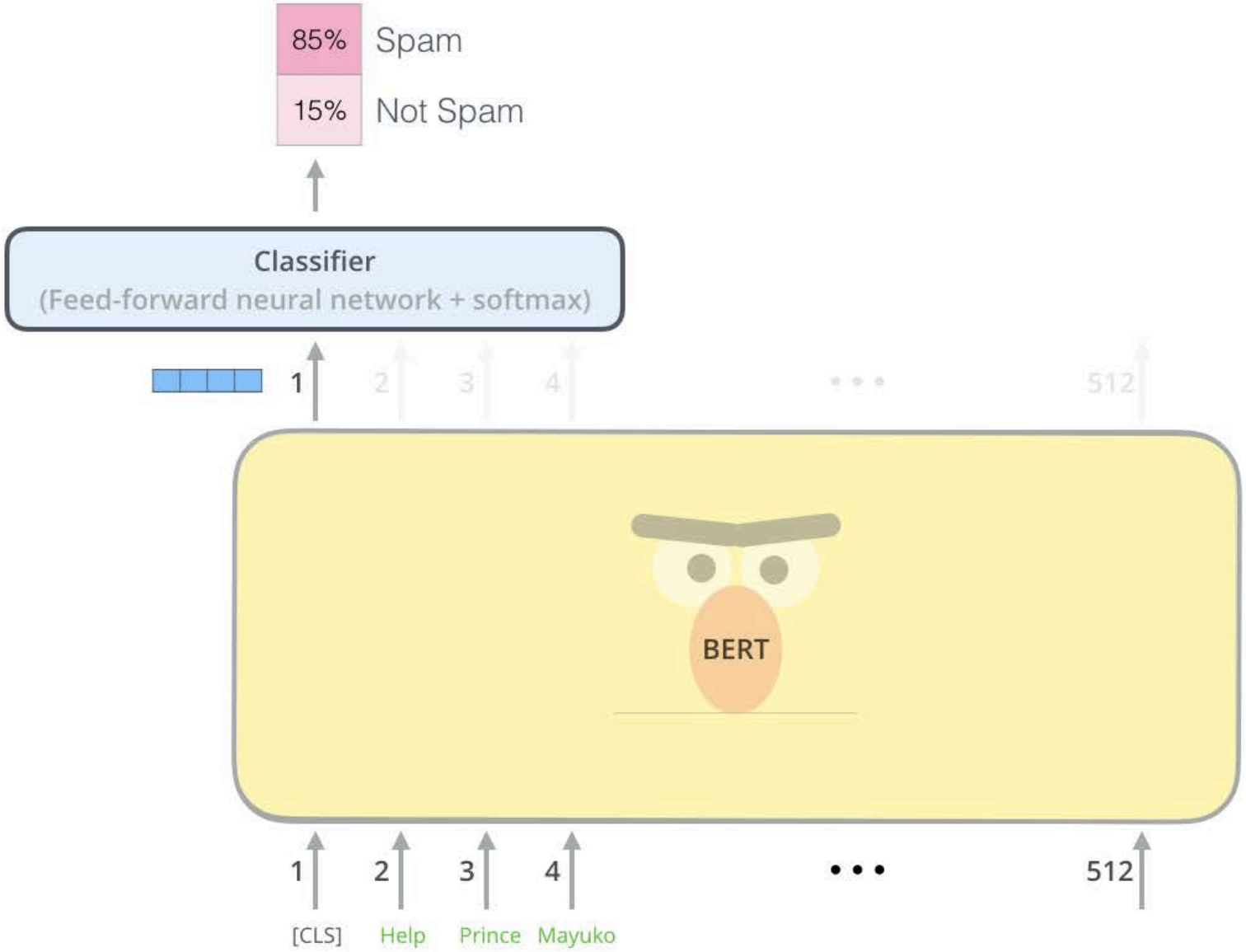


BERT$_{LARGE}$

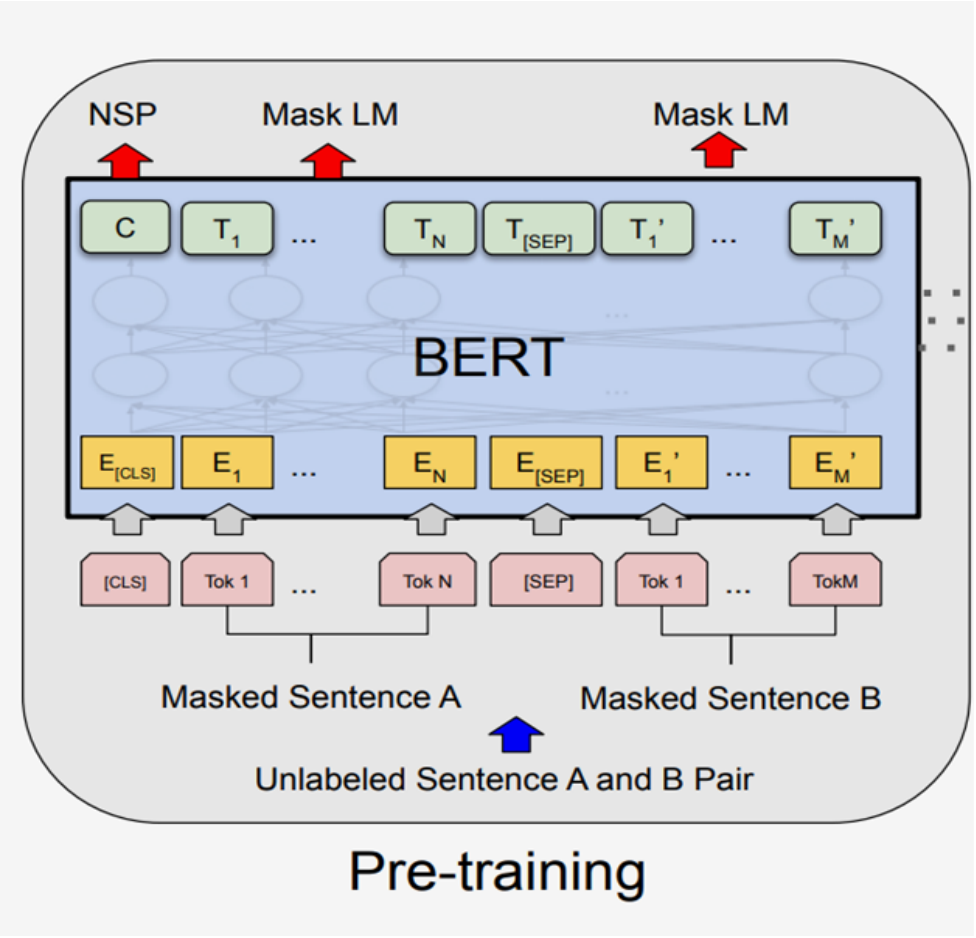
Model Inputs & Outputs



Classification



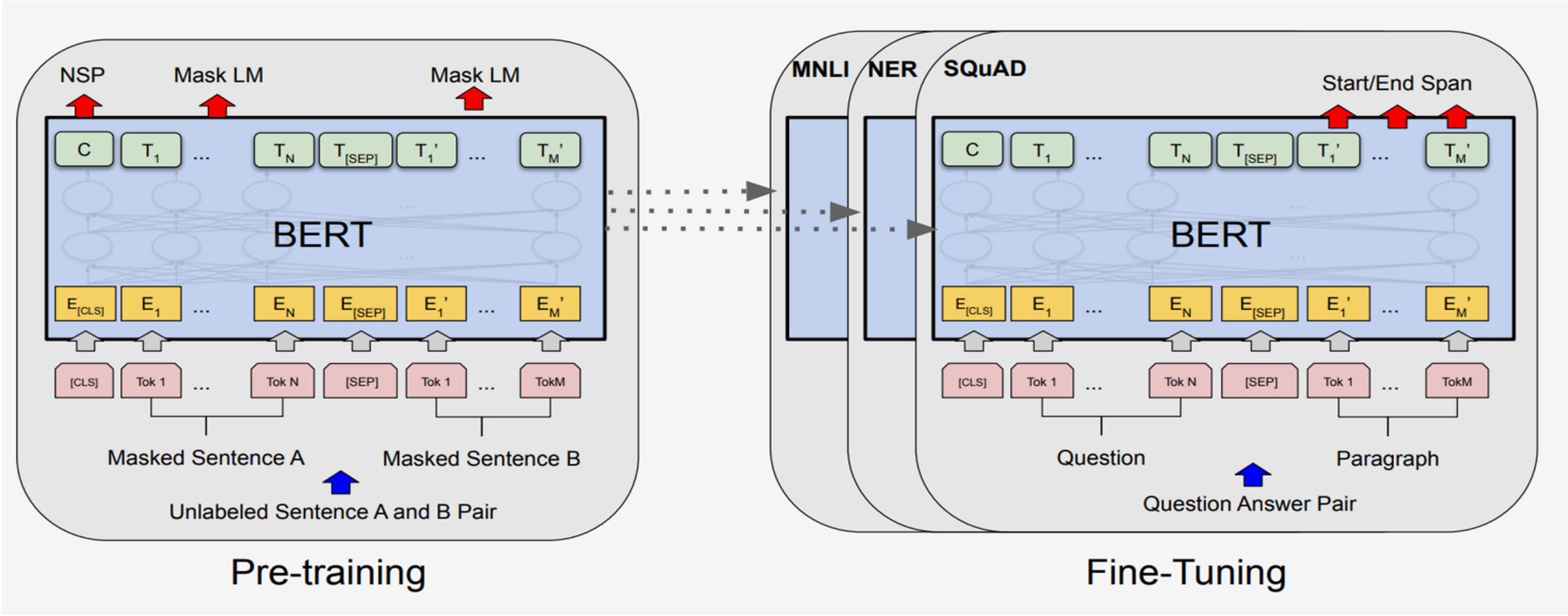
Training BERT



Self-supervised: ∞ training data

Devlin, Chang, Lee, Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

Training BERT

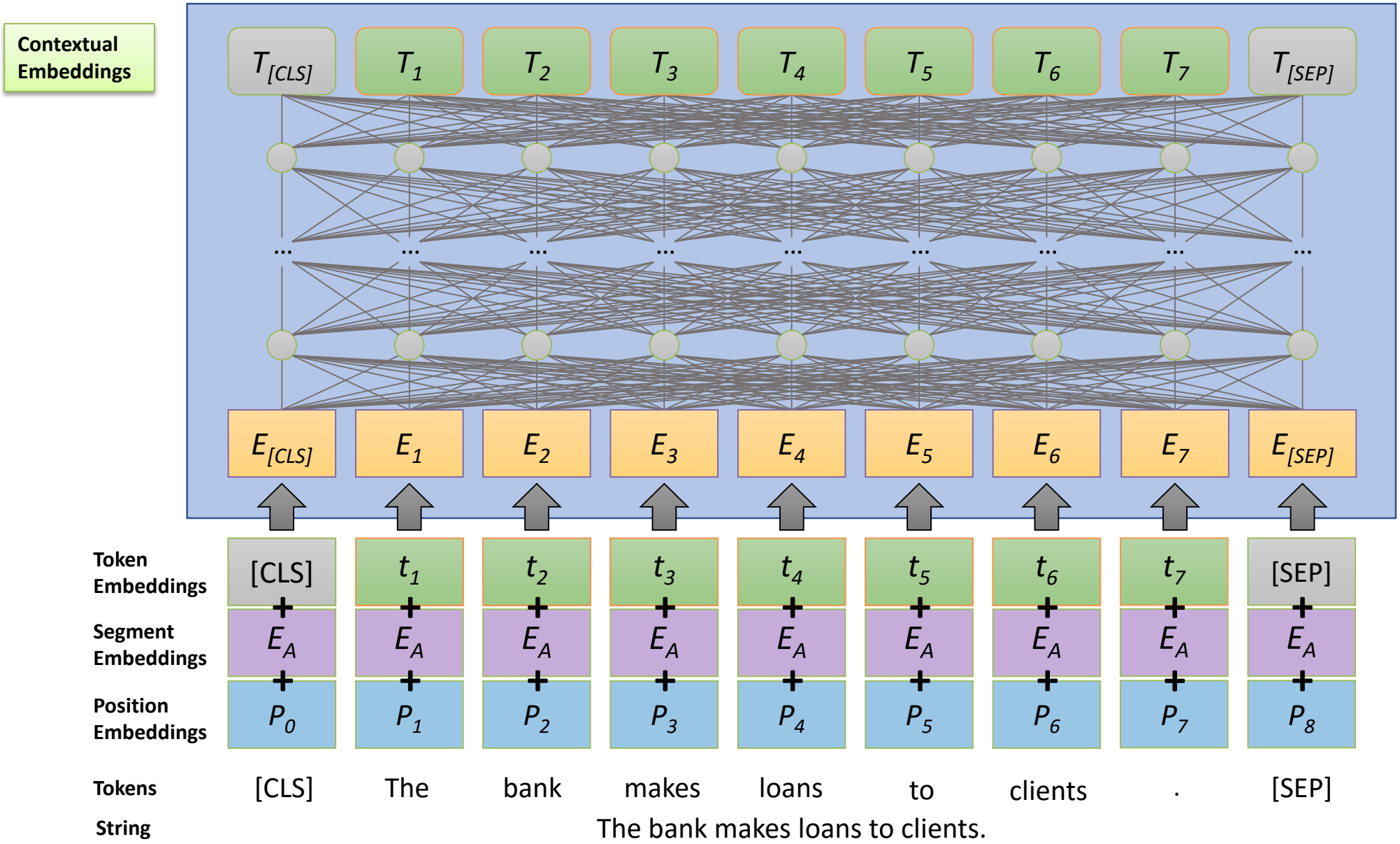


Self-supervised: ∞ training data

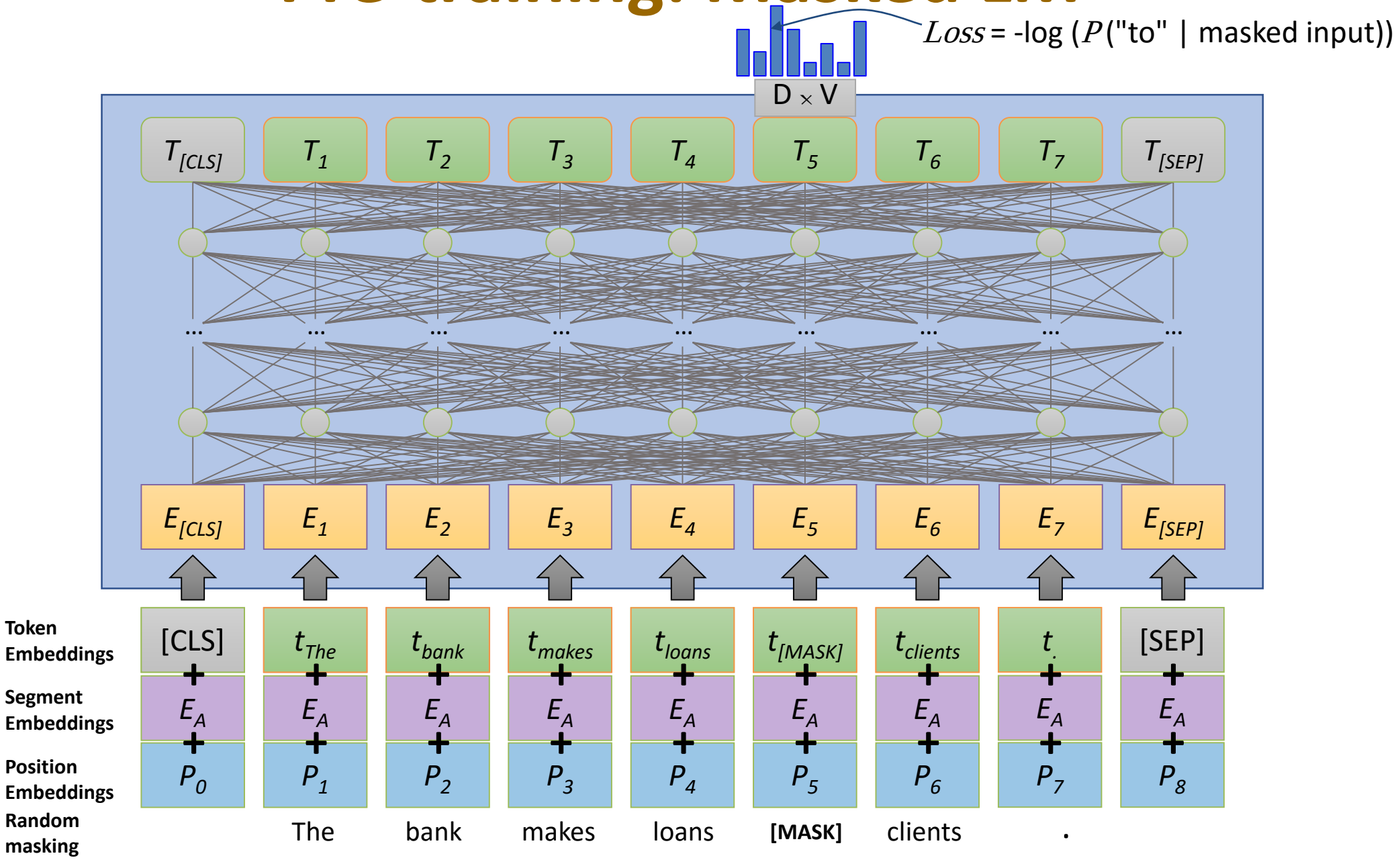
Supervised: (few) labeled examples

Devlin, Chang, Lee, Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

BERT (in more detail)



Pre-training: Masked LM



Pre-training

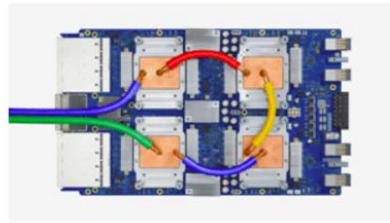
Token Probability		
	year 0.02	land 0.01
	century 0.94	**Europe 0.97**
	car 0.00	is 0.00
	...	...

softmax | softmax

linear | linear

... vec i ... vec j

BERT



Cloud TPU v3

420 teraflops

128 GB HBM

Cloud TPU v3 Pod

100+ petaflops

32 TB HBM

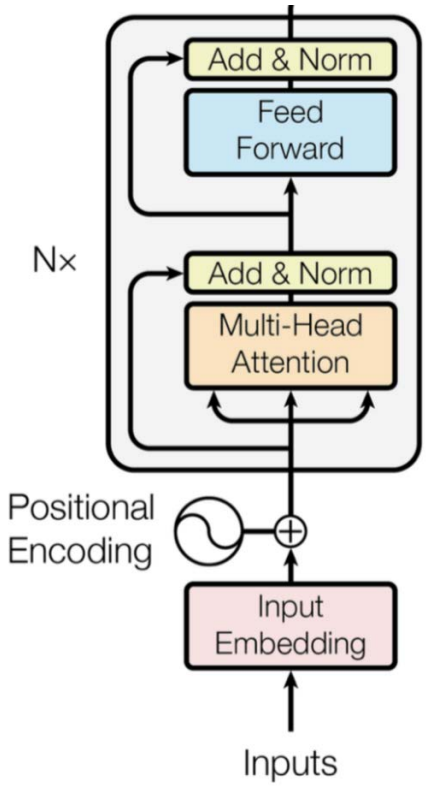
The Mongol invasion of Europe in the 13th [MASK] was the conquest of much of [MASK] by the Mongol Empire.

Random Masking

The Mongol invasion of Europe in the 13th century was the conquest of much of Europe by the Mongol Empire.

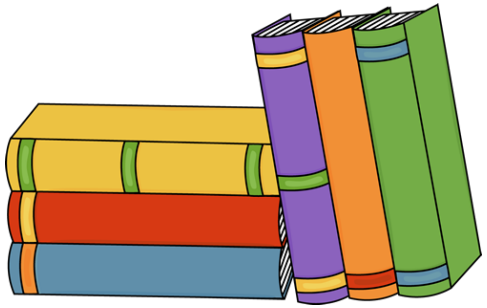
Input: a document

BERT's Pre-training



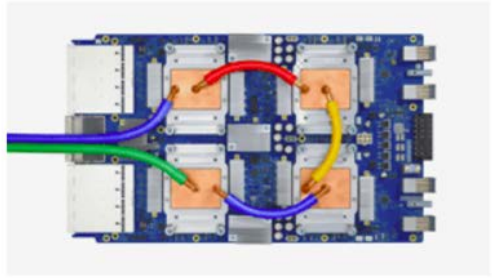
Transformer (encoder-only) with lots of parameters

+



Lots of texts

+



Cloud TPU v3

420 teraflops

128 GB HBM

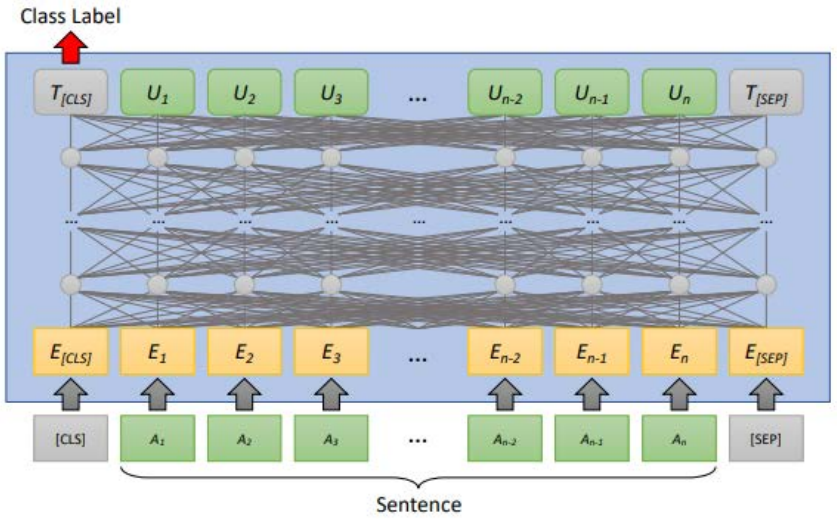
Cloud TPU v3 Pod

100+ petaflops

32 TB HBM

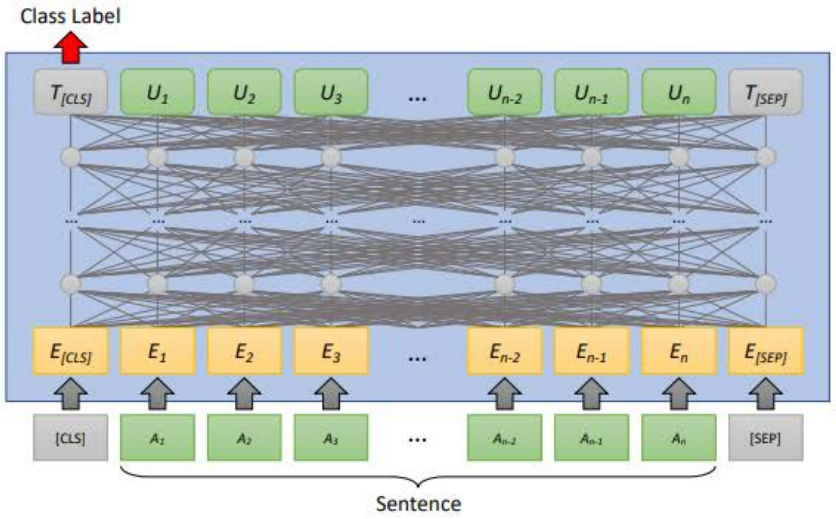
Lots of Compute

Fine-tuning

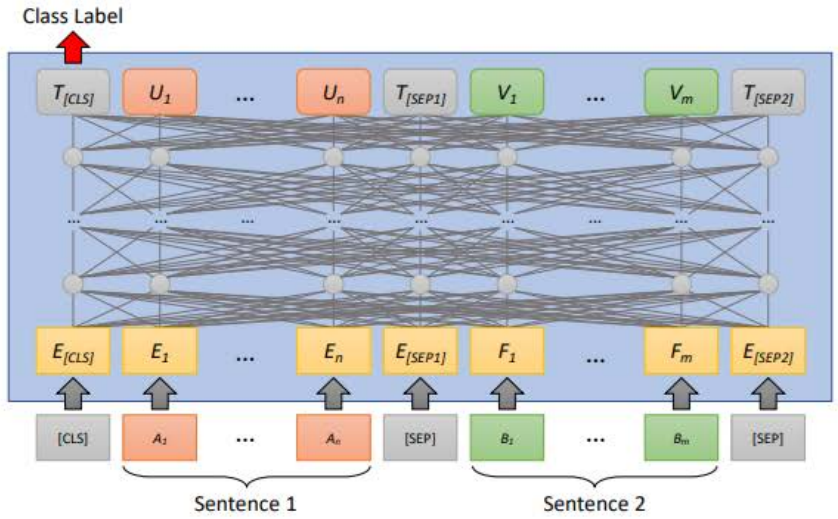


(a) Single-Input Classification Tasks

Fine-tuning

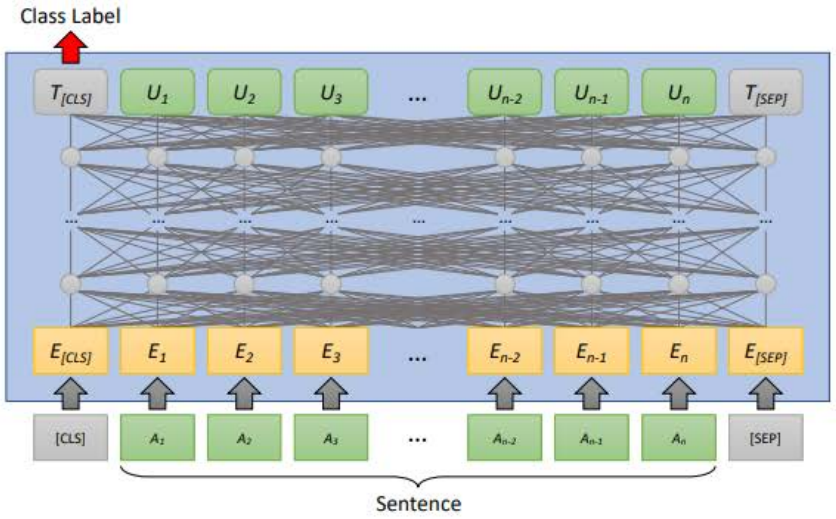


(a) Single-Input Classification Tasks

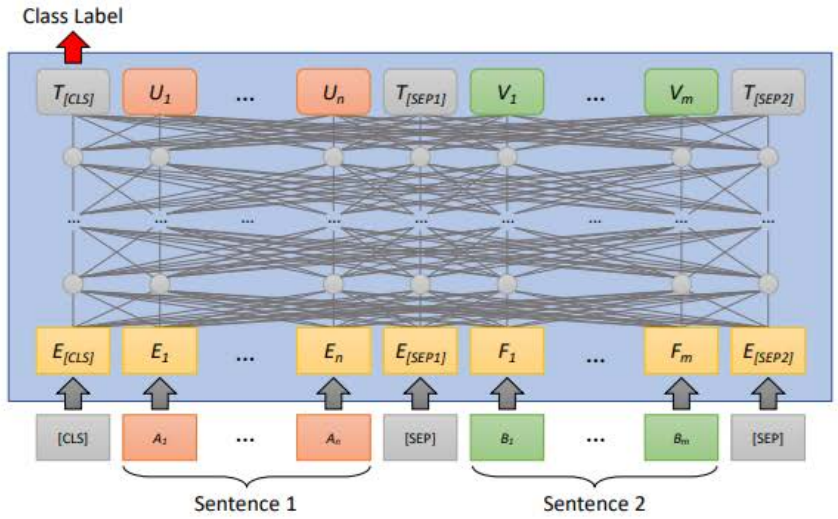


(b) Two-Input Classification Tasks

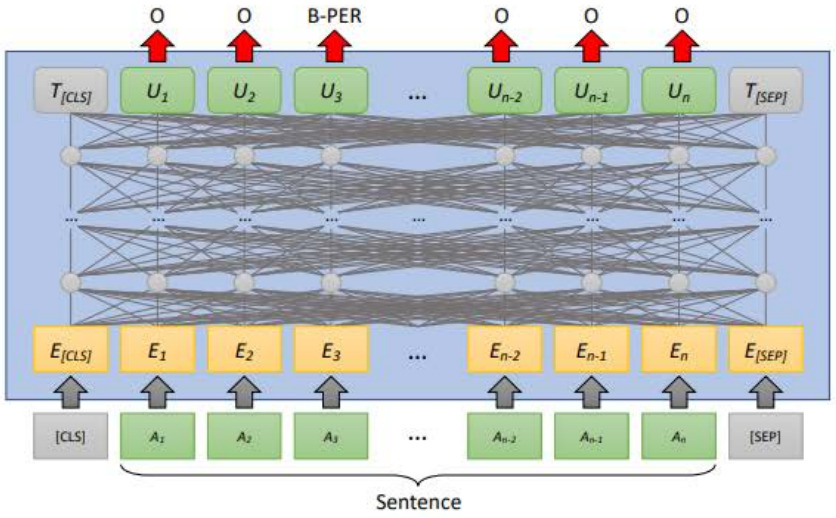
Fine-tuning



(a) Single-Input Classification Tasks

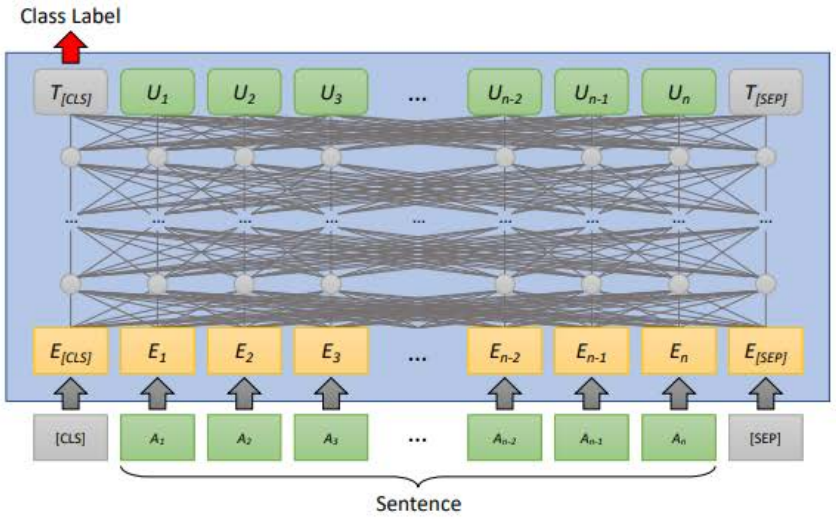


(b) Two-Input Classification Tasks

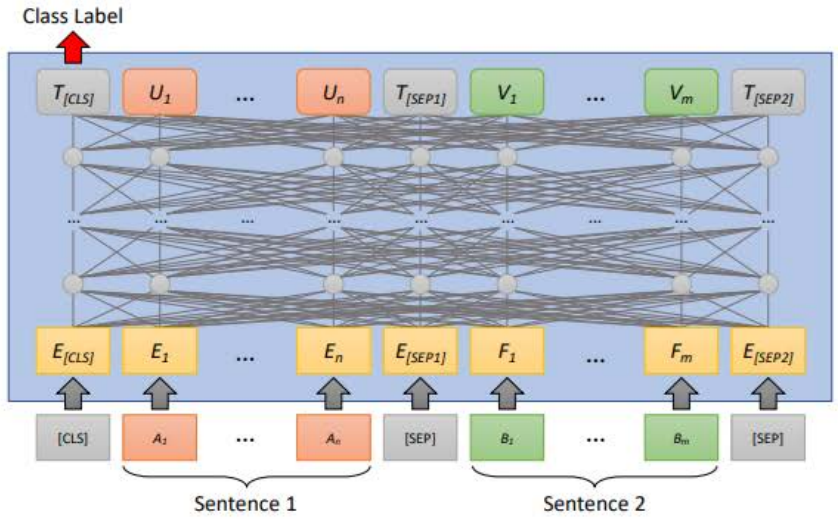


(c) Single-Input Sequence Labeling Tasks

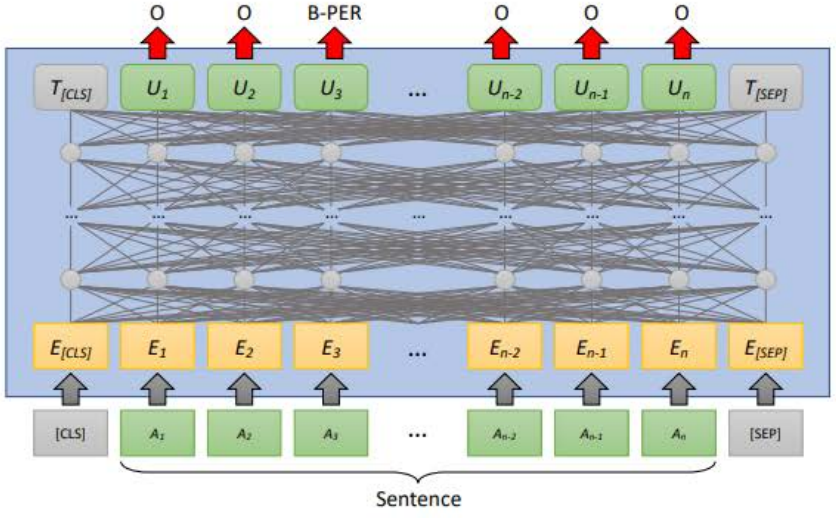
Fine-tuning



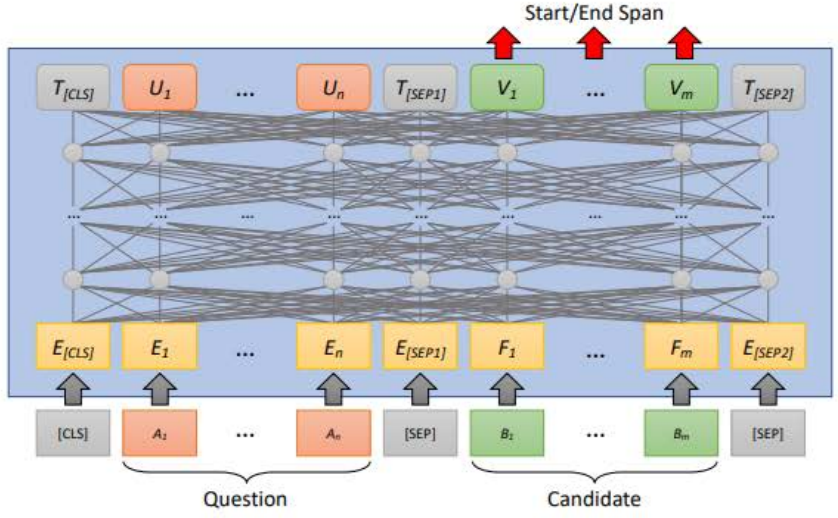
(a) Single-Input Classification Tasks



(b) Two-Input Classification Tasks



(c) Single-Input Sequence Labeling Tasks



(d) Two-Input Sequence Labeling Tasks


26
POLL

26

BERT can be used for the task of finding an answer to a questions within a passage.

- Yes
- No

When we fine-tune BERT, the pre-trained model changes.

- Yes
- No

Fine-tuning BERT is a self-supervised task.

- Yes
- No

MS MARCO Passage Ranking Leaderboard on January 2019

Rank	Model	Submission Date	MRR@10 On Eval	MRR@10 On Dev
1	**BERT + Small Training** Rodrigo Nogueira and Kyunghyun Cho - New York University	January 7th, 2019	35.87	
2	**IRNet (Deep CNN/IR Hybrid Network)** Dave DeBarr, Navendu Jain, Robert Sim, Justin Wang, Nirupama Chandrasekaran – Microsoft	January 2nd, 2019	28.061	
3	**Neural Kernel Match IR (Conv-KNRM)**(1)Yifan Qiao, (2)Chenyan Xiong, (3)Zhenghao Liu, (4)Zhiyuan Liu-Tsinghua University(1, 3, 4); Microsoft Research AI(2) [Dai et al. '18]	Novmeber 28th, 2018	27.12	29.02
4	**Neural Kernel Match IR (KNRM)** ((1)Yifan Qiao, (2)Chenyan Xiong, (3)Zhenghao Liu, (4)Zhiyuan Liu-Tsinghua University(1, 3, 4); Microsoft Research AI(2) [Xiong et al. '17]	December 10th, 2018	19.82	21.84
5	**Feature-based LeToR: simple-feature based RankSVM**(1)Yifan Qiao, (2)Chenyan Xiong, (3)Zhenghao Liu, (4)Zhiyuan Liu-Tsinghua University(1, 3, 4); Microsoft Research AI(2)	December 10th, 2018	19.05	19.47
6	**BM25**	Novmeber 1st, 2018	16.49	16.70

~8 points!

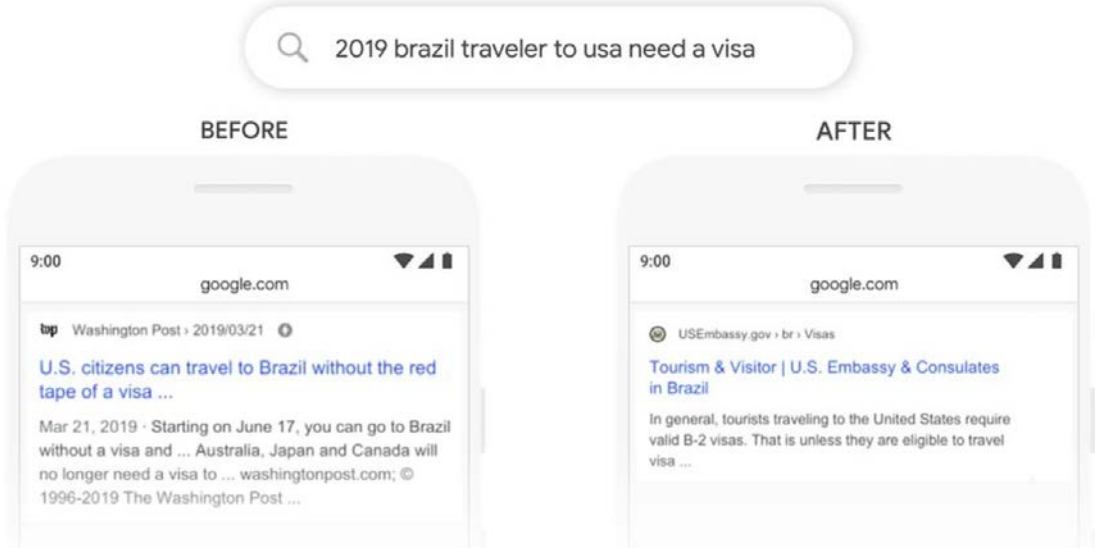
MS MARCO Passage Ranking Leaderboard on March 2021

Passage Ranking Leaderboard(10/26/2018-Present) ranked by MRR on Eval

6 **UED** Anonymous Full Ranking

Adoption by Commercial Search Engines

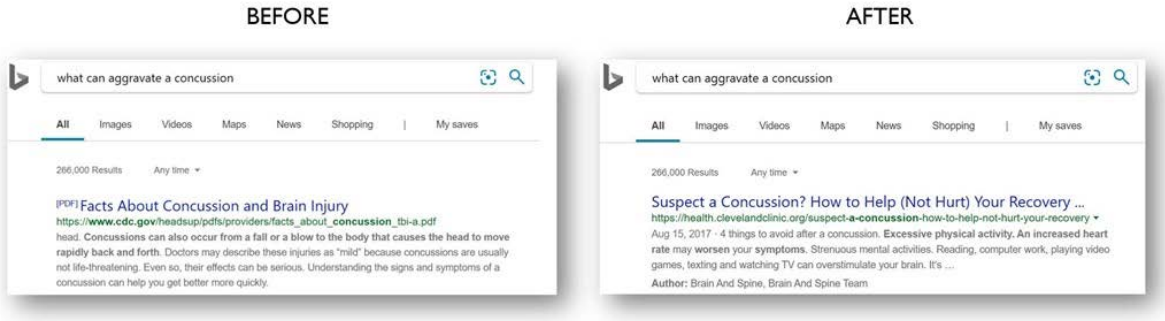
Google Search



We're making a significant improvement to how we understand queries, representing the biggest leap forward in the past five years, and one of the biggest leaps forward in the history of Search.

source

MS Bing



Starting from April of this year (2019), we used large transformer models to deliver the largest quality improvements to our Bing customers in the past year.

source

for *Ranking*?

for *Ranking*?

BERT for Ranking?

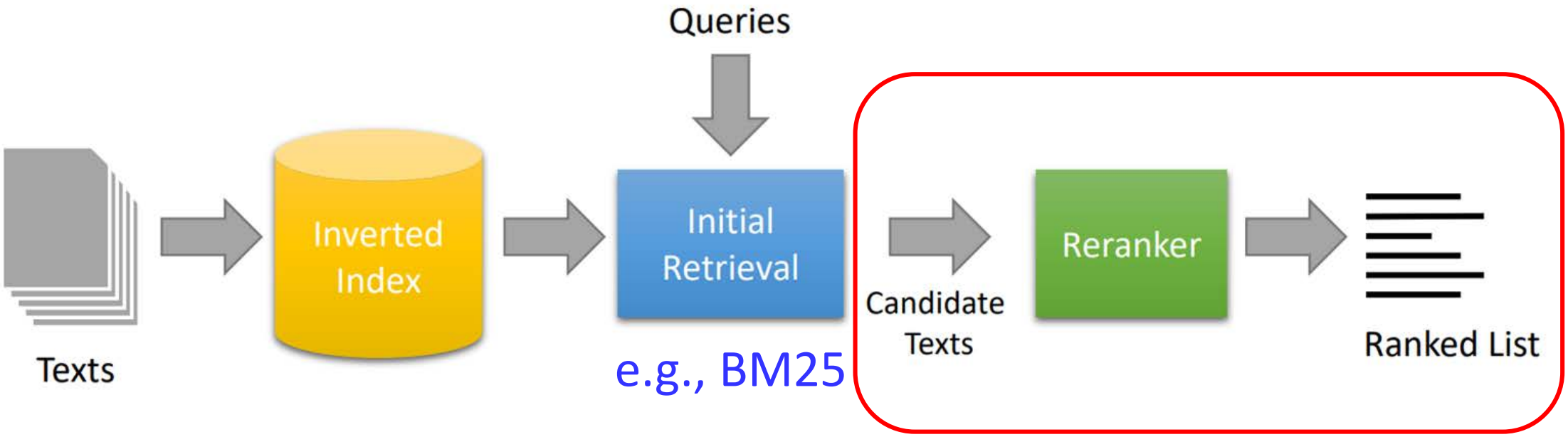
1. Ranking ➔ classification problem
2. Sort the texts to be ranked based on the probability that each "item" belongs to the relevance class.

$$P(\text{Relevant} = 1 | d_i, q)$$

Relevance Classification

L2R?

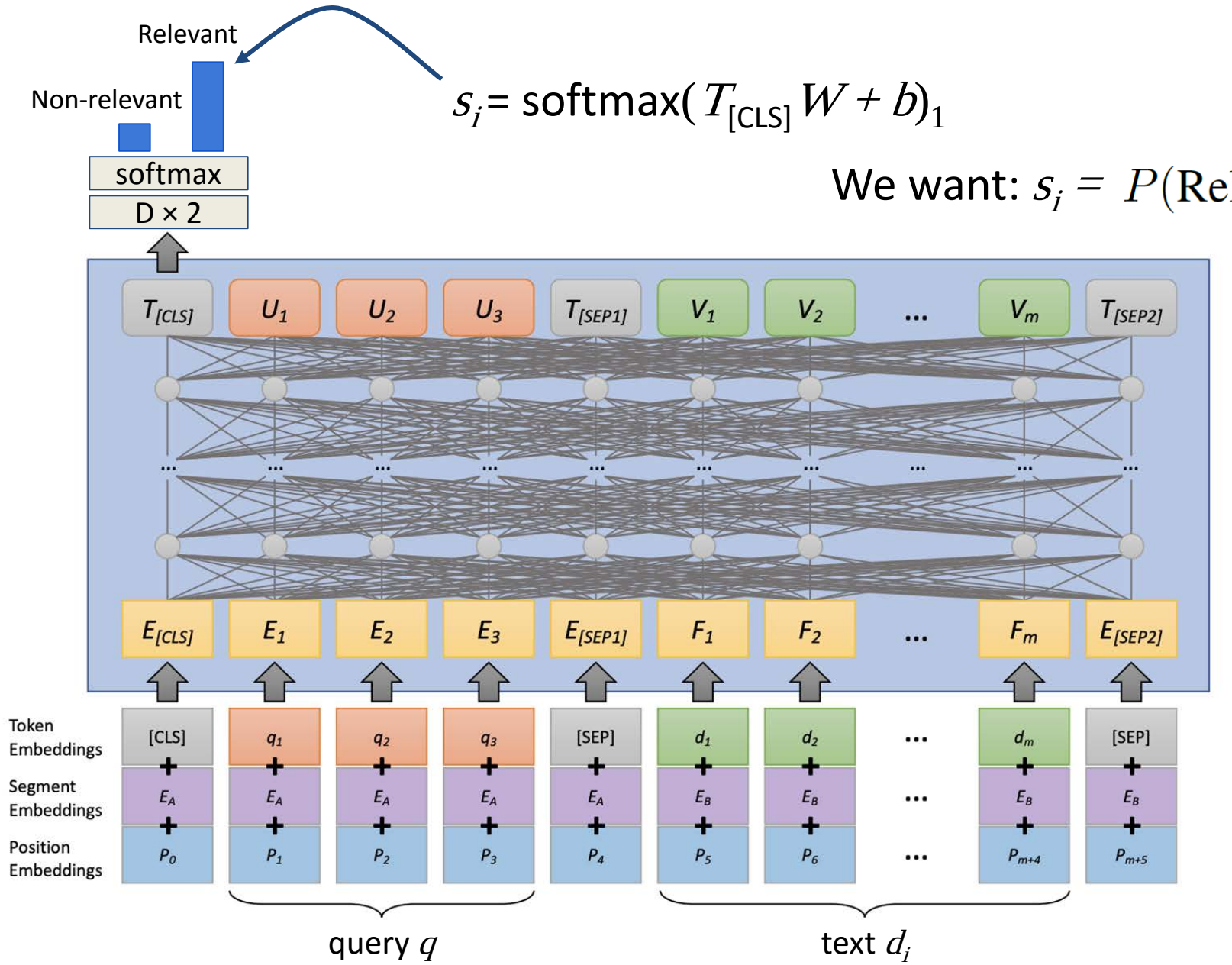
A Simple Search Engine




DATA
CONFUSING
MATHEMATICAL
DOCUMENTS
DIFFICULT
CORPUS
QUERY
SEARCH
RANK
INDEX
WEB
INFORMATION
LARGE
IMPORTANT
HOW
COOL
AMAZING
FOR
TEXT
LIST
FAST
FUN
SCIENCE
PAGE
CRAWL
RETRIEVE
METADATA
DOCUMENT
A
GOOGLE
SEARCHING
CRAWLER
RETRIEVAL
DATABASE
USEFUL
DISTRIBUTED
GATHER

MONOBERT

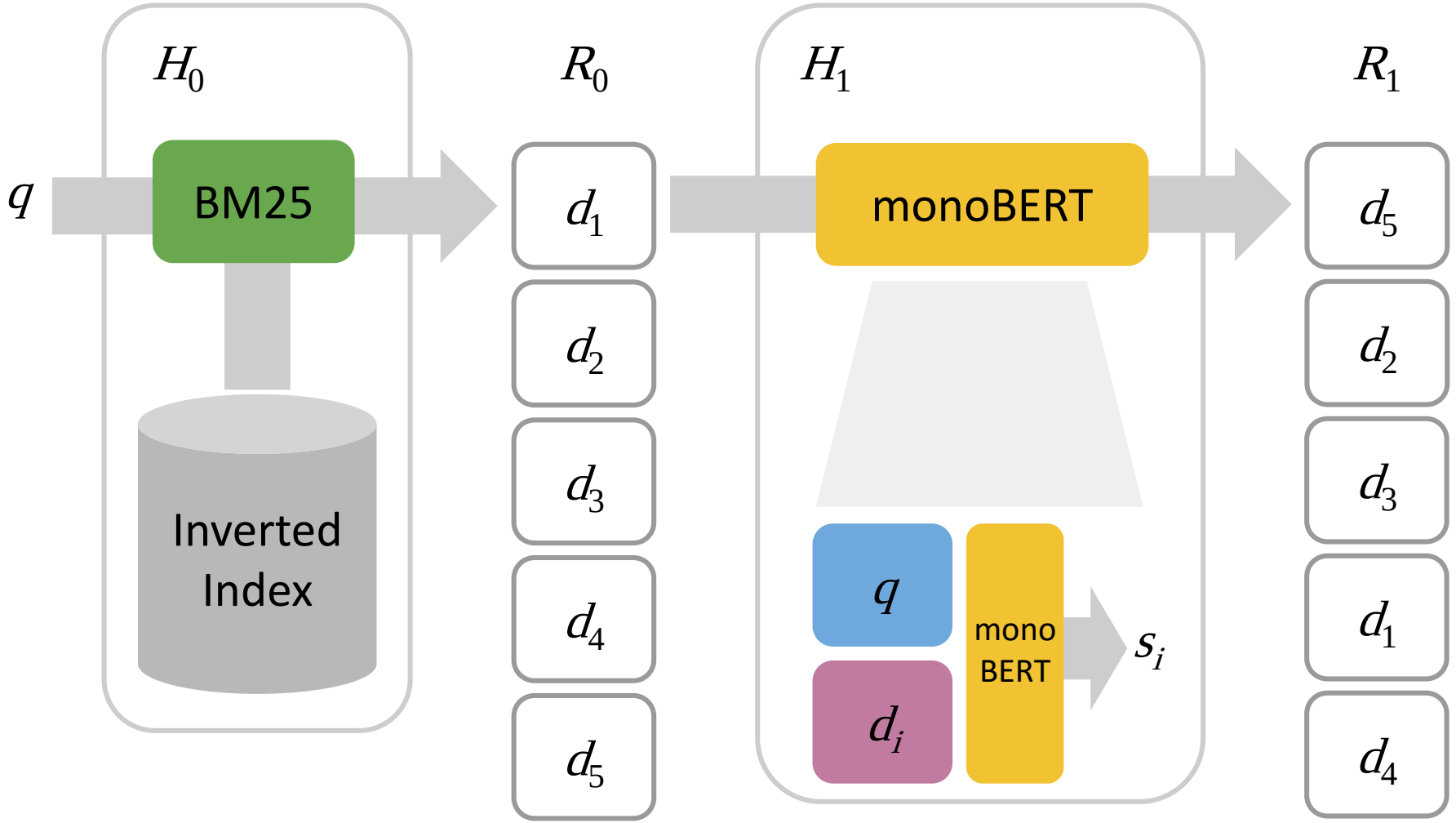
monoBERT: BERT reranker



Training monoBERT

Loss: $L = -\sum_{j \in J_{\text{pos}}} \log(s_j) - \sum_{j \in J_{\text{neg}}} \log(1 - s_j)$

Once monoBERT is trained...



TREC 2019 - Deep Learning Track - Passage

	nDCG@10	MAP	Recall@1k
BM25	0.506	0.377	0.739
+ monoBERT	**0.738**	**0.506**	0.739
BM25 + RM3	0.518	0.427	0.788
+ monoBERT	**0.742**	**0.529**	0.788

for Ranking?