Jan. 17, 2020

CS 886 Deep Learning and NLP

Ming Li

CONTENT

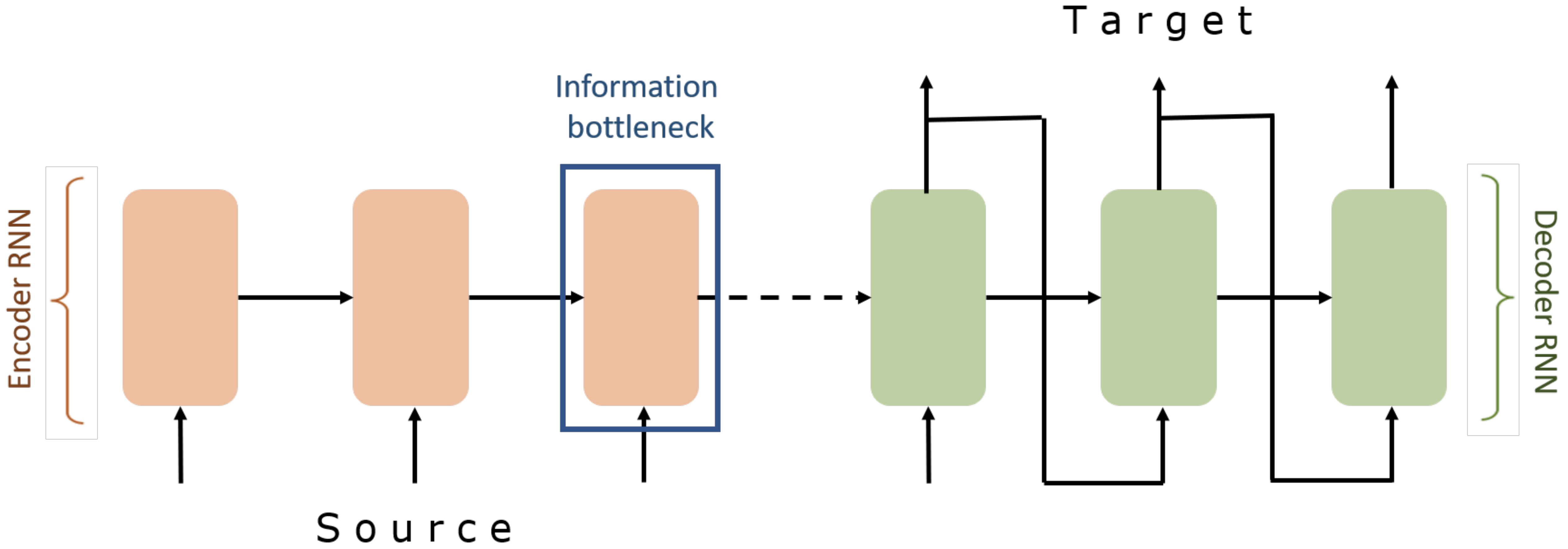
03

GPT-2 and BERT

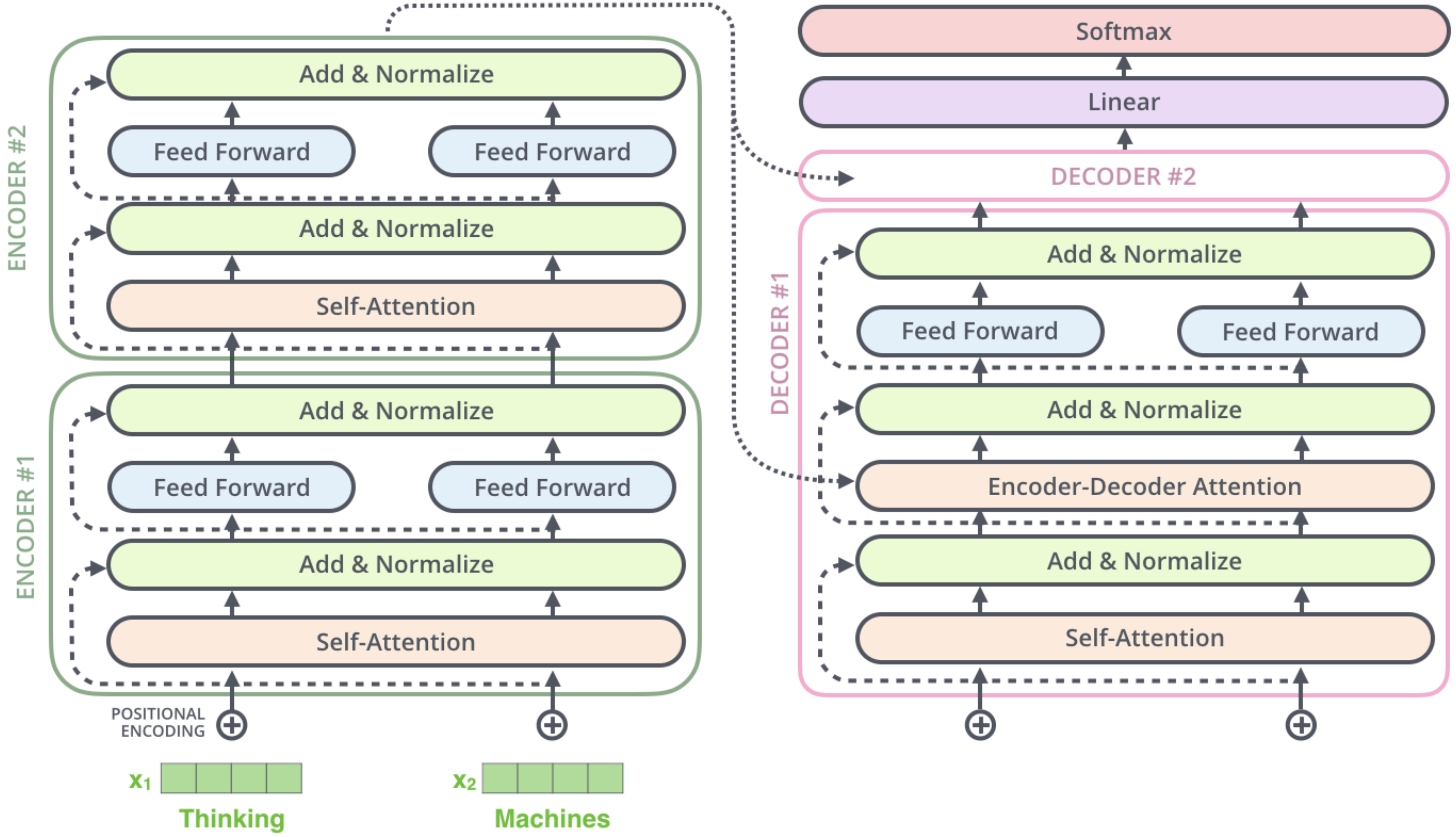
LECTURE THREE

Tying up loose ends from the last lecture, back to Lecture 2 notes.

Avoiding Information bottleneck



Last time we introduced transformer

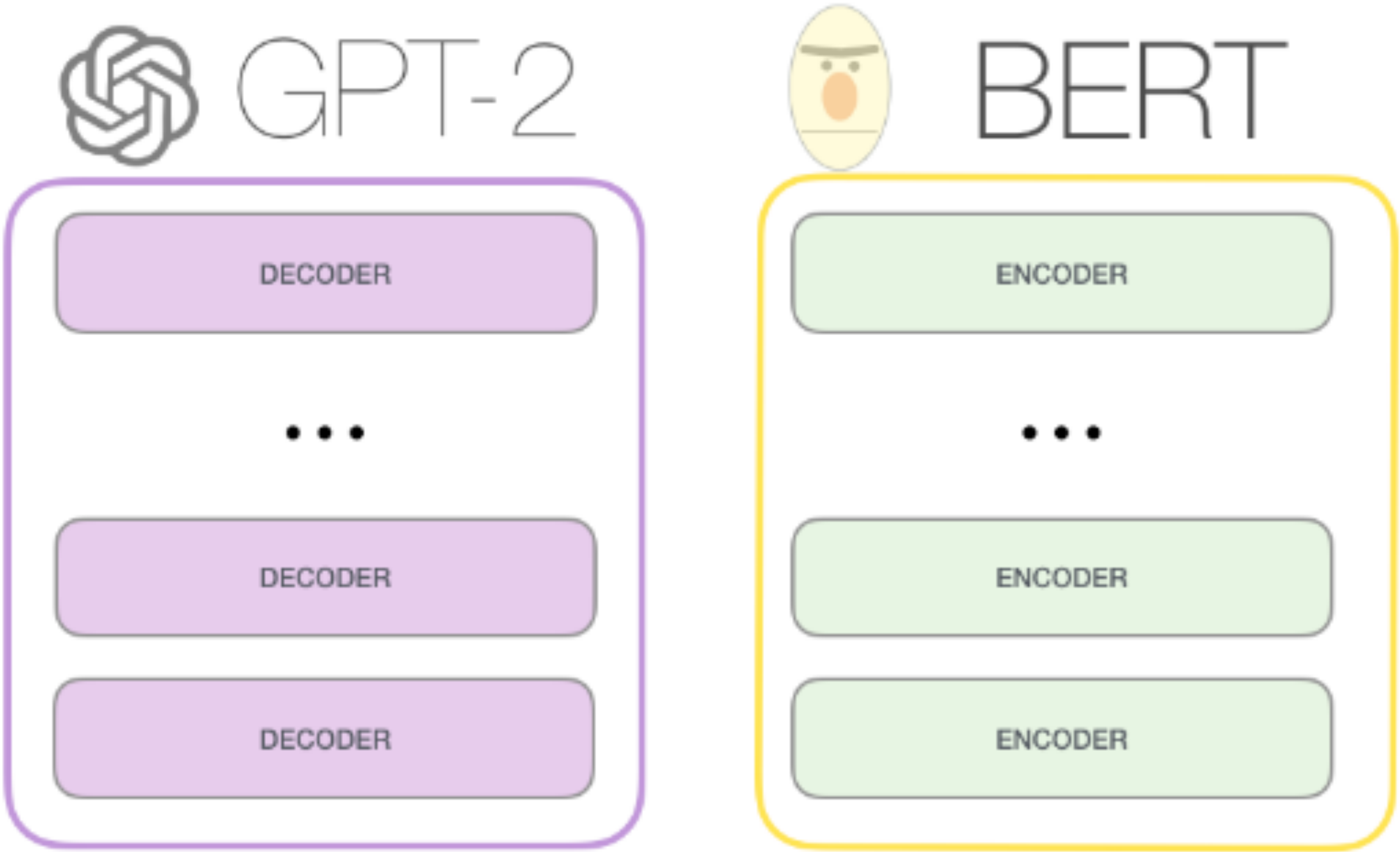


Transformers, GPT-2, and BERT

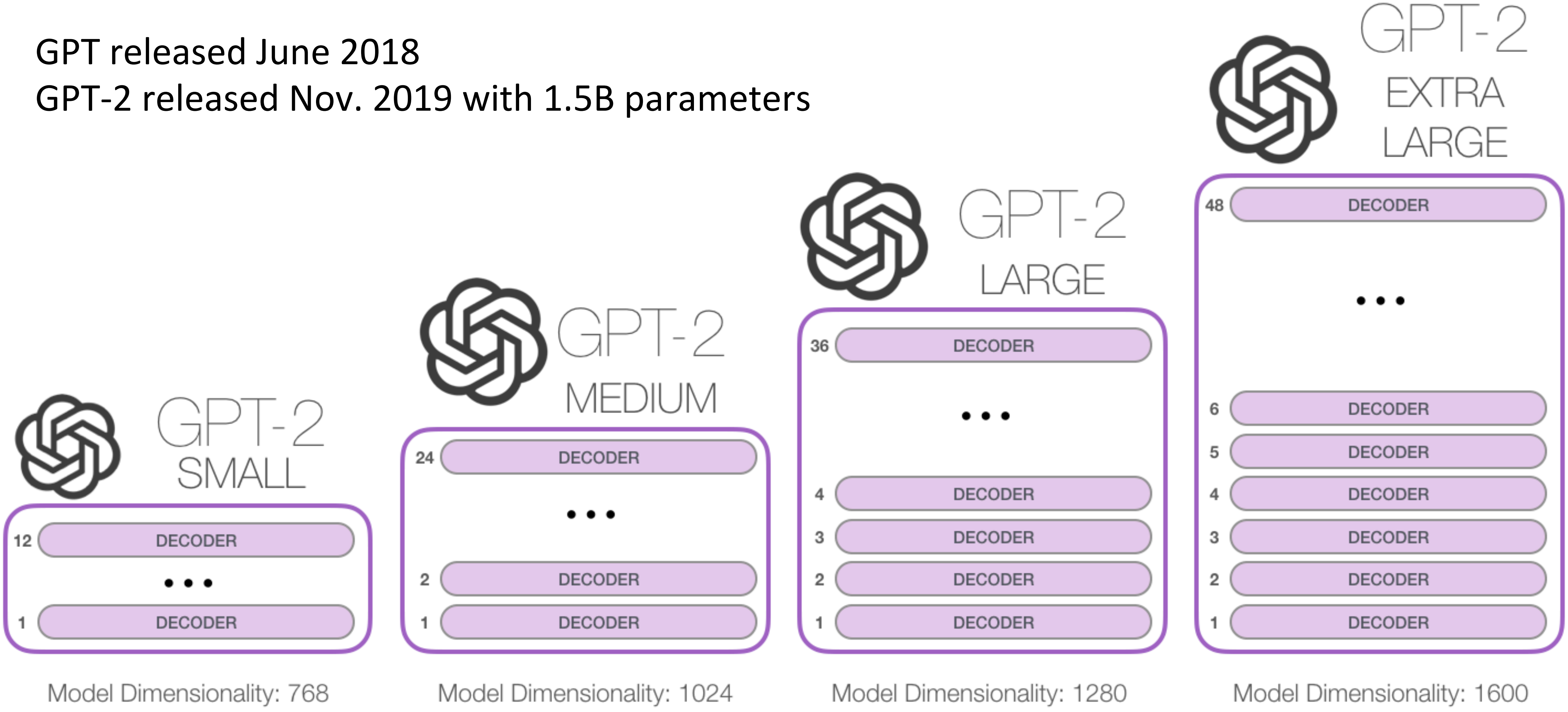
1. A transformer uses Encoder stack to model input, and uses Decoder stack to model output (using input information from encoder side).
2. But if we do not have input, we just want to model the “next word”, we can get rid of the Encoder side of a transformer and output “next word” one by one. This gives us GPT.
3. If we are only interested in training a language model for the input for some other tasks, then we do not need the Decoder of the transformer, that gives us BERT.

03

GPT-2, BERT

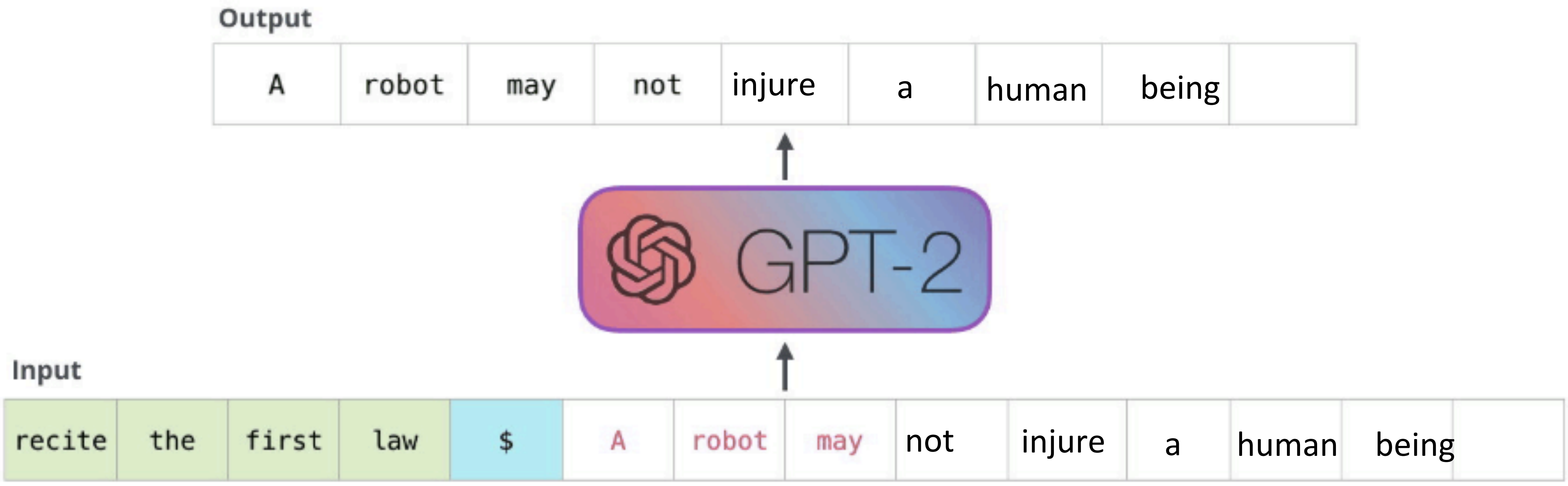

GPT-2
DECODER
DECODER
DECODER
BERT
ENCODER
ENCODER
ENCODER

GPT released June 2018
GPT-2 released Nov. 2019 with 1.5B parameters



117M parameters 345M 762M 1542M

GPT-2 in action



Byte Pair Encoding (BPE)

Word embedding sometimes is too high level, pure character embedding too low level. For example, if we have learned

old older oldest

We might also wish the computer to infer

smart smarter smartest

But at the whole word level, this might not be so direct. Thus the idea is to break the words up into pieces like er, est, and embed frequent fragments of words.

GPT adapts this BPE scheme.

Byte Pair Encoding (BPE)

GPT uses BPE scheme. The subwords are calculated by:

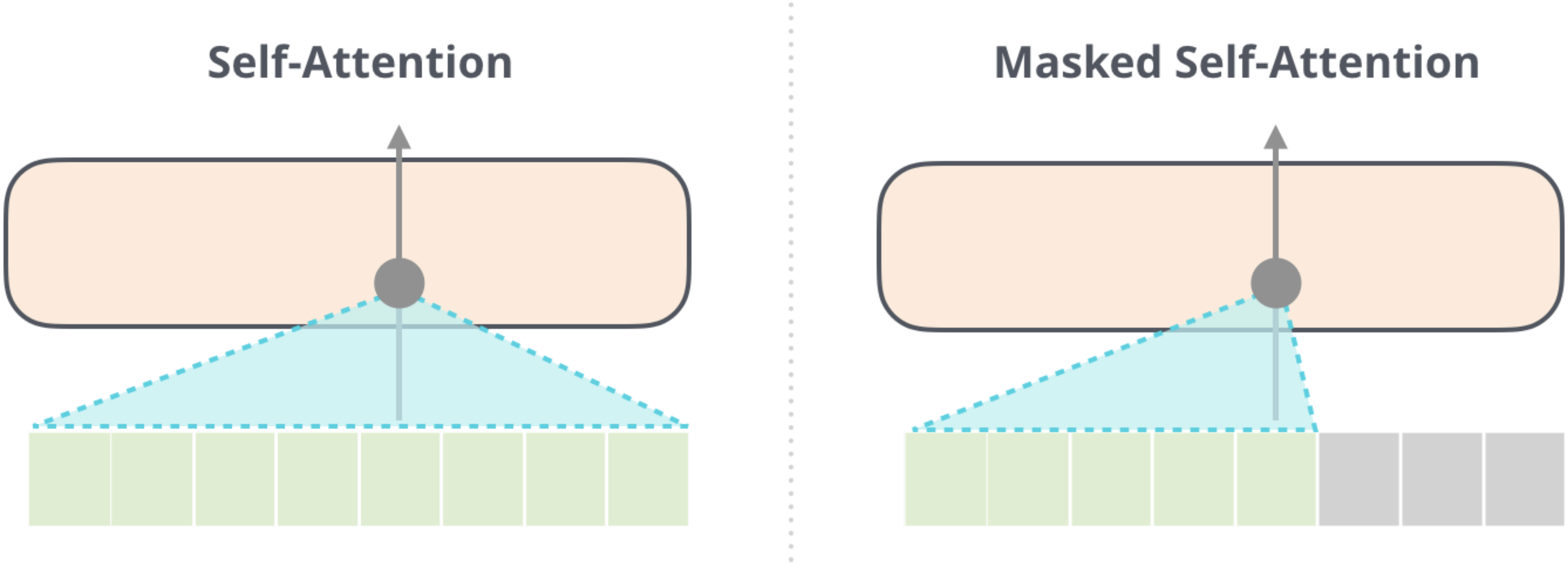
1. Split word to sequence of characters (add </w> char)
2. Joining the highest frequency pattern.
3. Keep doing step 2, until it hits the pre-defined maximum number of sub-words or iterations.

Example:

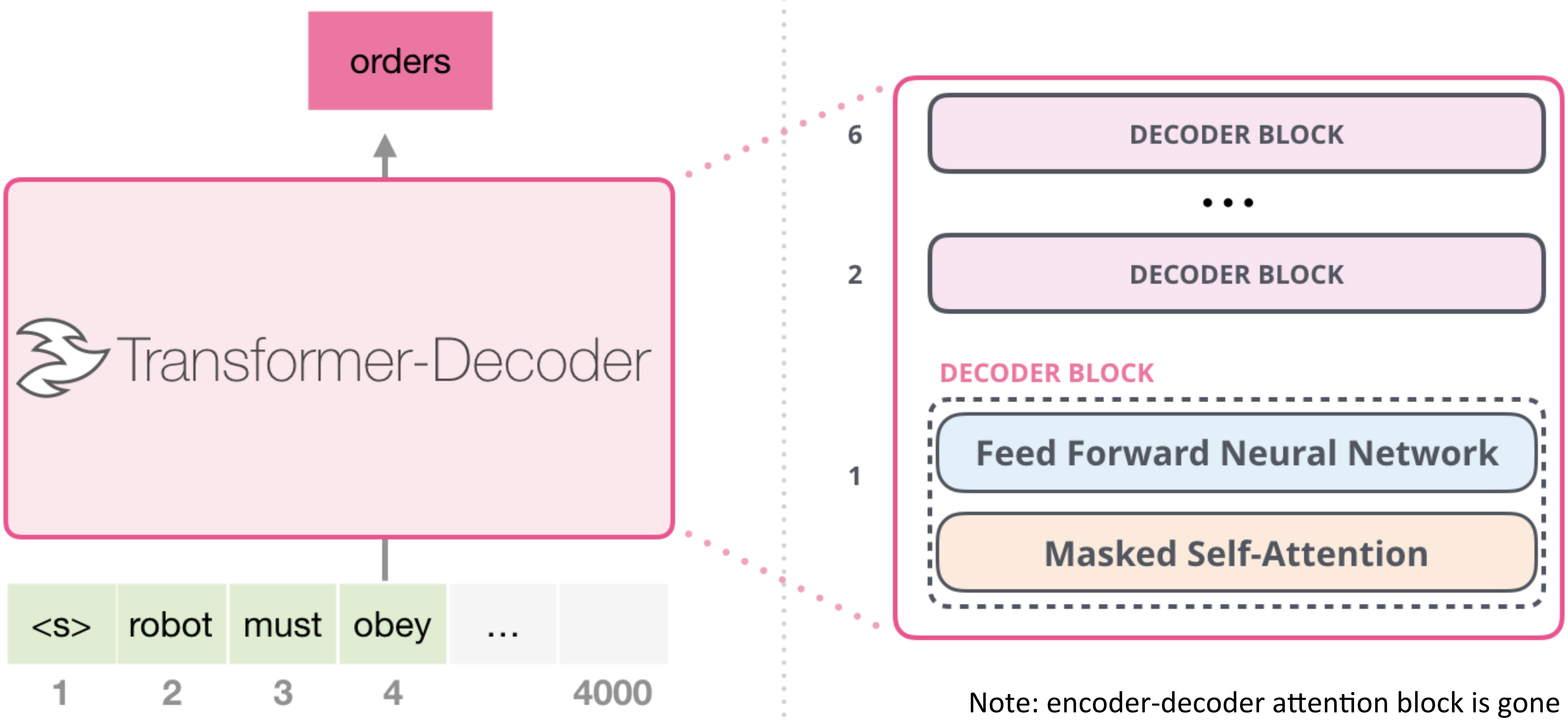
{‘l o w </w>’: 5, ‘l o w e r </w>’: 2, ‘n e w e s t </w>’: 6, ‘w I d e s t </w>’: 3 }
{‘l o w </w>’: 5, ‘l o w e r </w>’: 2, ‘n e w es t </w>’: 6, ‘w I d es t </w>’: 3 }
{‘l o w </w>’: 5, ‘l o w e r </w>’: 2, ‘n e w est </w>’: 6, ‘w I d est </w>’: 3 }
{‘l o w </w>’: 5, ‘l o w e r </w>’: 2, ‘n e w est</w>’: 6, ‘w I d est</w>’: 3 }
…..

Note that </w> is also an important character.

Masked Self-Attention



Masked Self-Attention



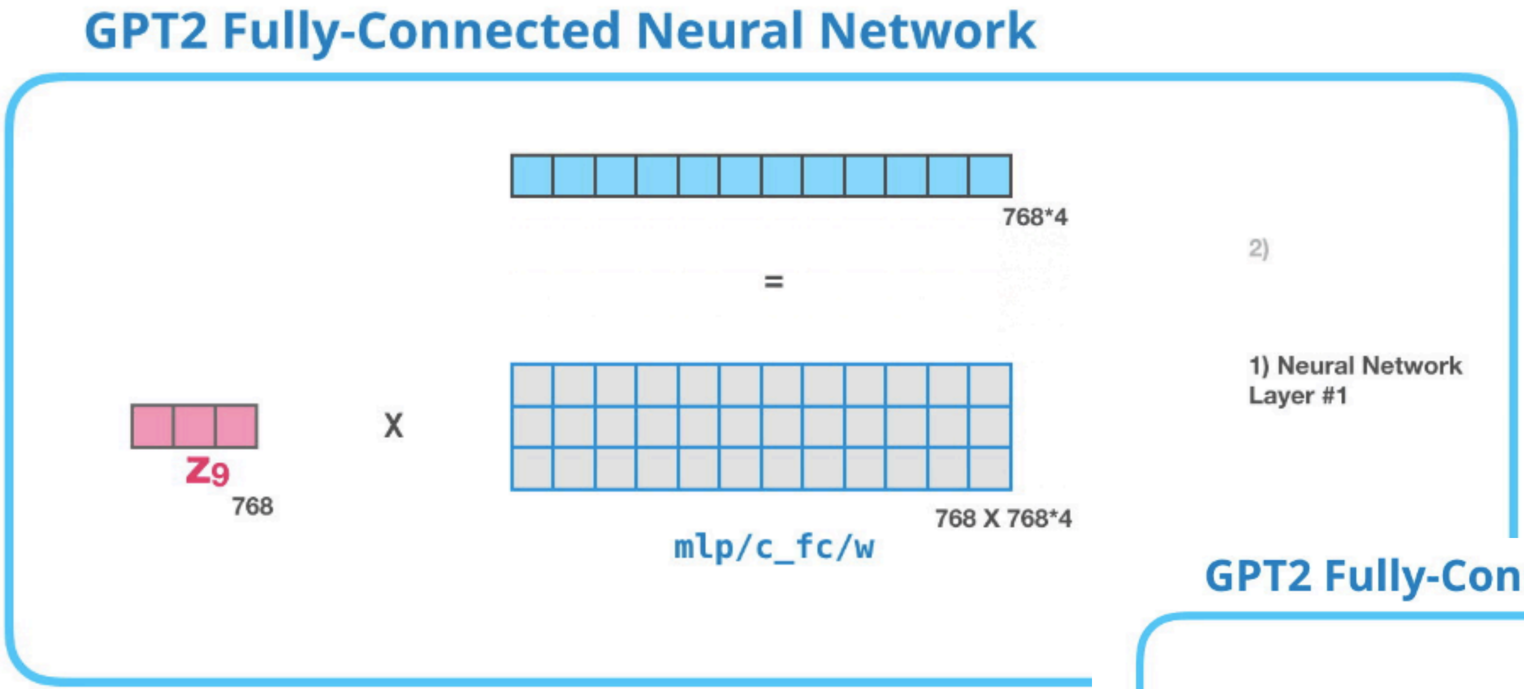
Note: encoder-decoder attention block is gone

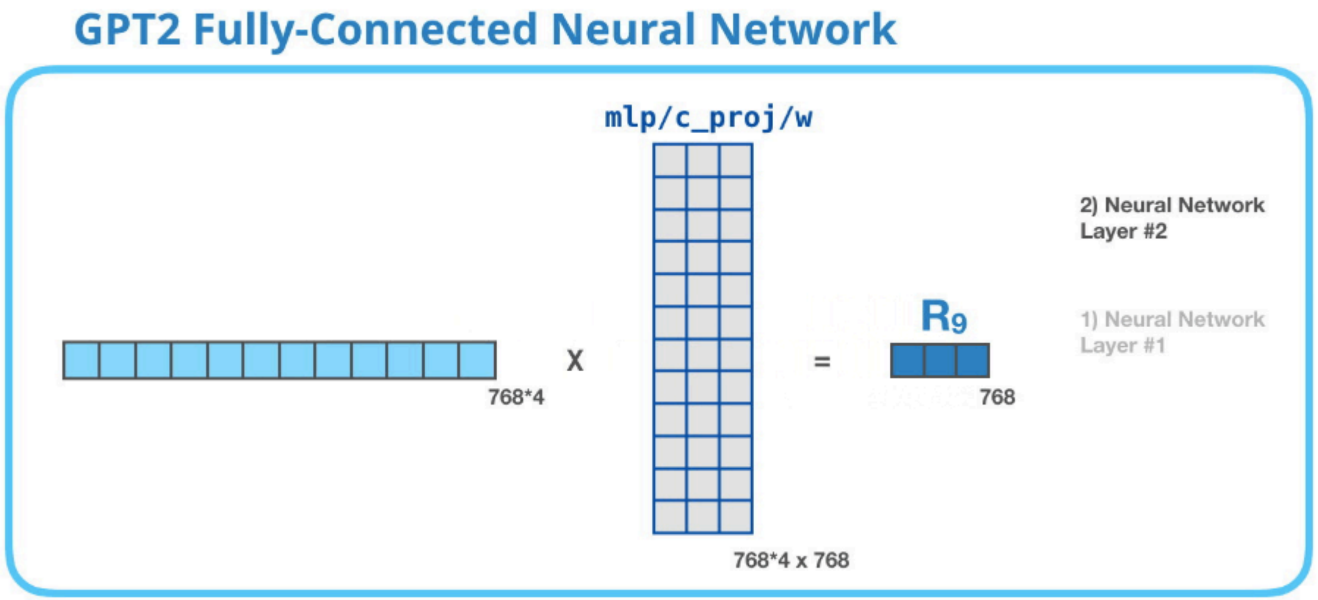
Masked Self-Attention Calculation

Re-use previous computation results: at any step, only need to results of q, k , v related to the new output word, no need to re-compute the others. Additional computation is linear, instead of quadratic.

Note: encoder-decoder attention block is gone

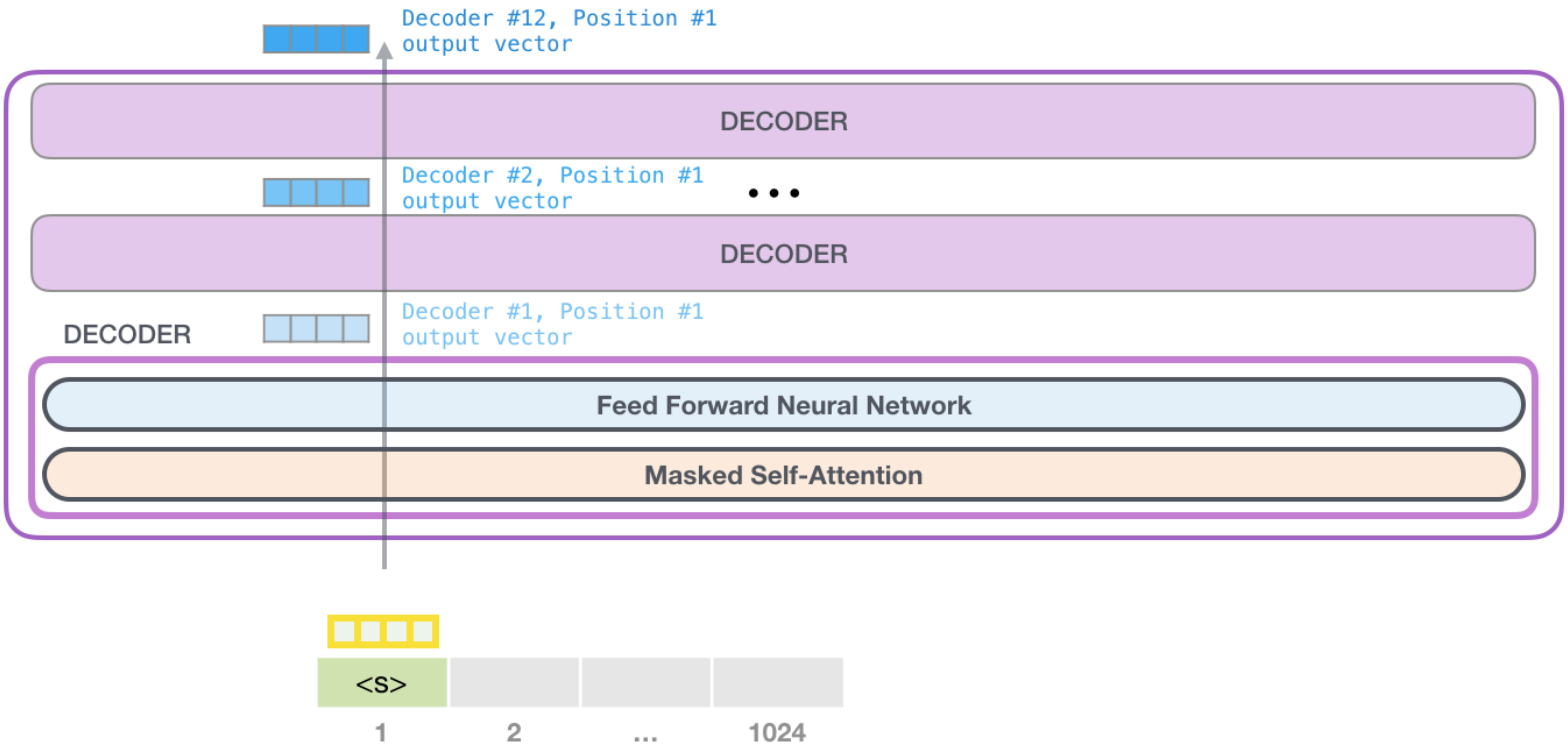
GPT-2 fully connected network has two layers (Example for GPT-2 small)





GPT-2, BERT

GPT-2 has a parameter top-k, so that we sample works from top k (highest probability from softmax) words for each each output



This top-k parameter, if k=1, we would have output like:

The first time I saw the new version of the game, I was so excited. I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of

GPT Training

GPT-2 uses unsupervised learning approach to training the language model.

There is no custom training for GPT-2, no separation of pre-training and fine-tuning like BERT.

A story generated by GPT-2

“The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

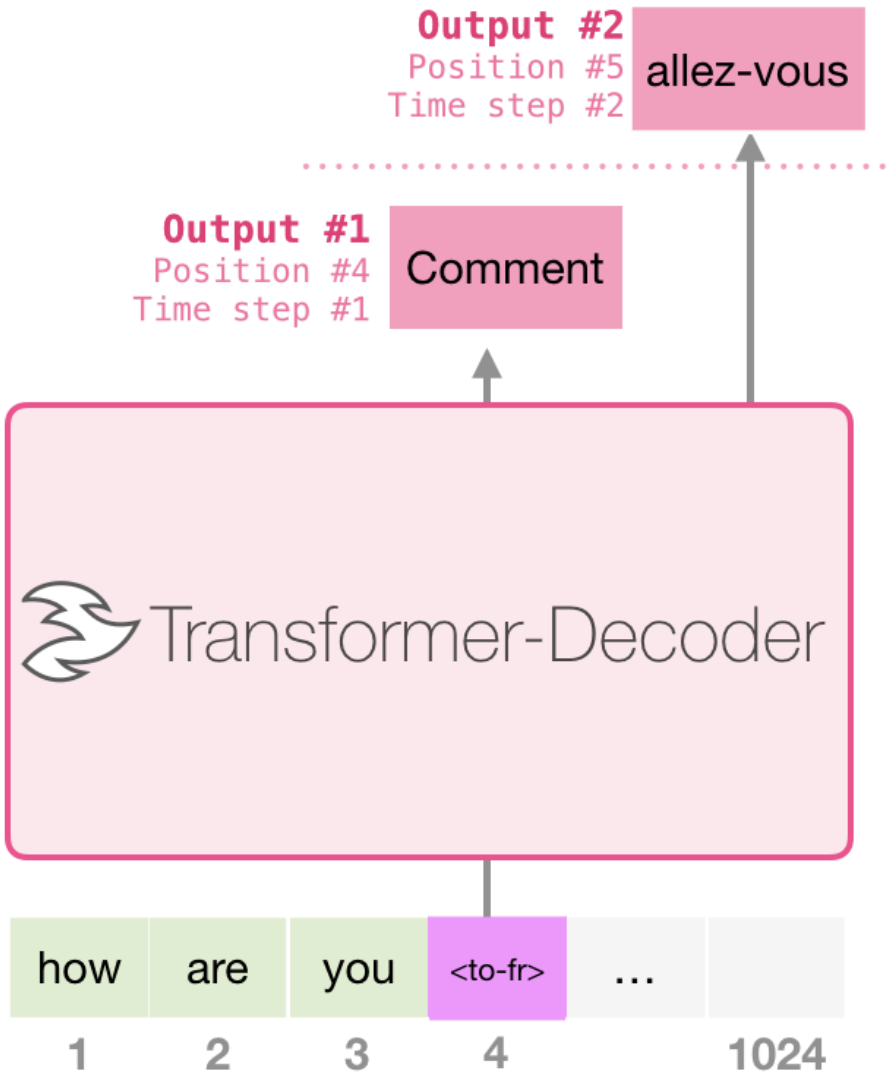
Pérez and the others then ventured further into the valley. `By the time we reached the top of one peak, the water looked blue, with some crystals on top,’ said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns."

GPT-2 Application: Translation

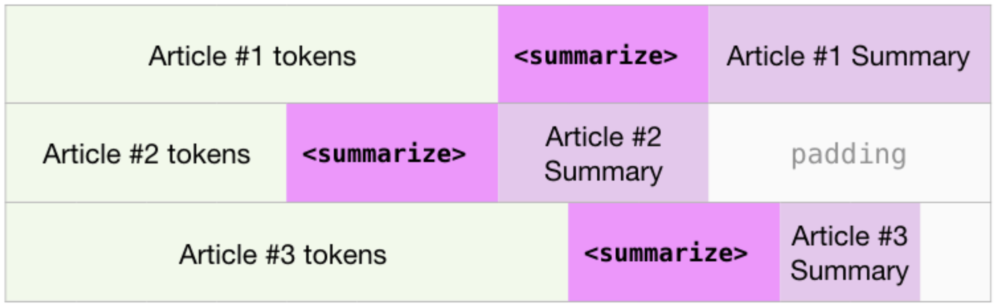
Training Dataset

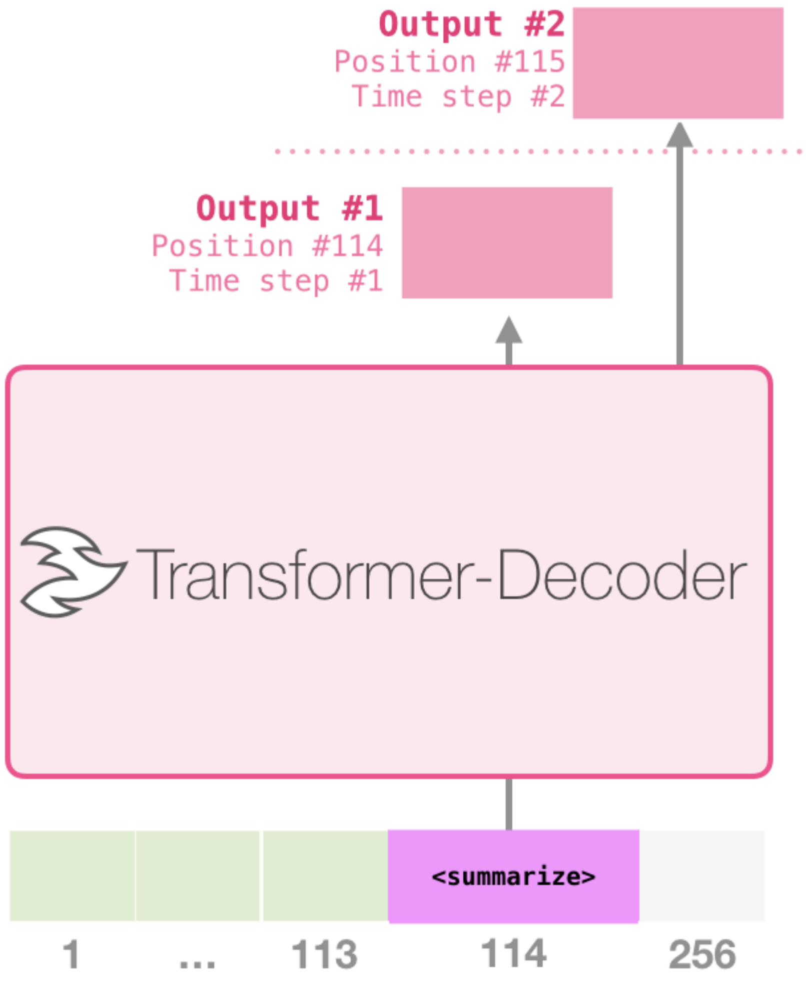
I	am	a	student	<to-fr>	je	suis	étudiant
let	them	eat	cake	<to-fr>	Qu'ils	mangent	de
good	morning	<to-fr>	Bonjour				



GPT-2 Application: Summarization

Training Dataset

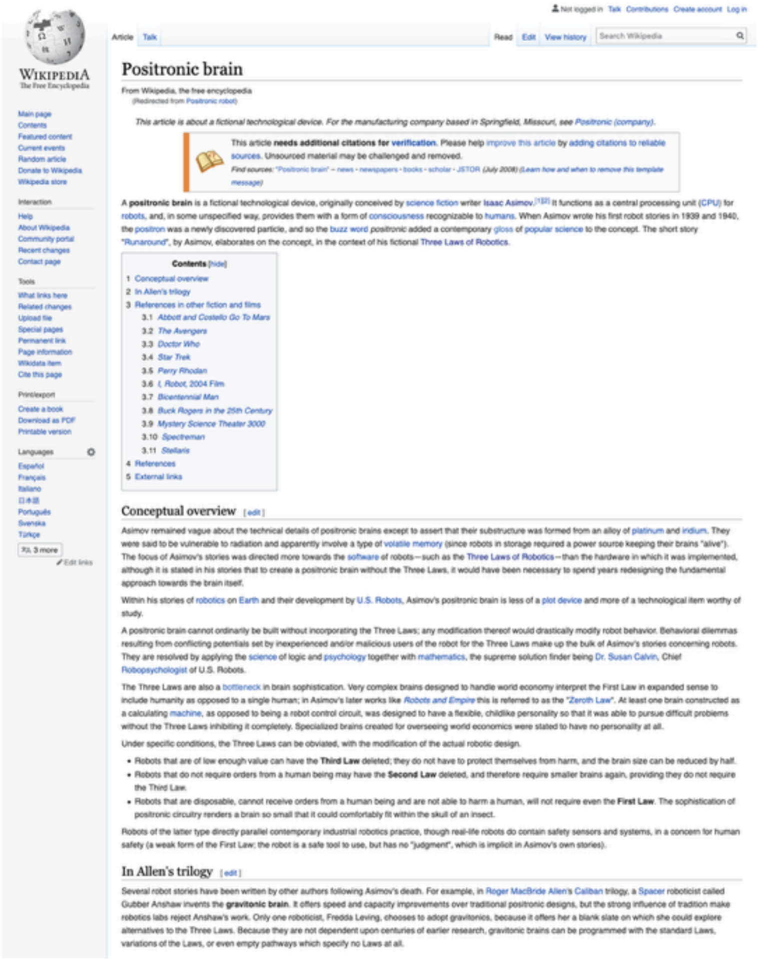




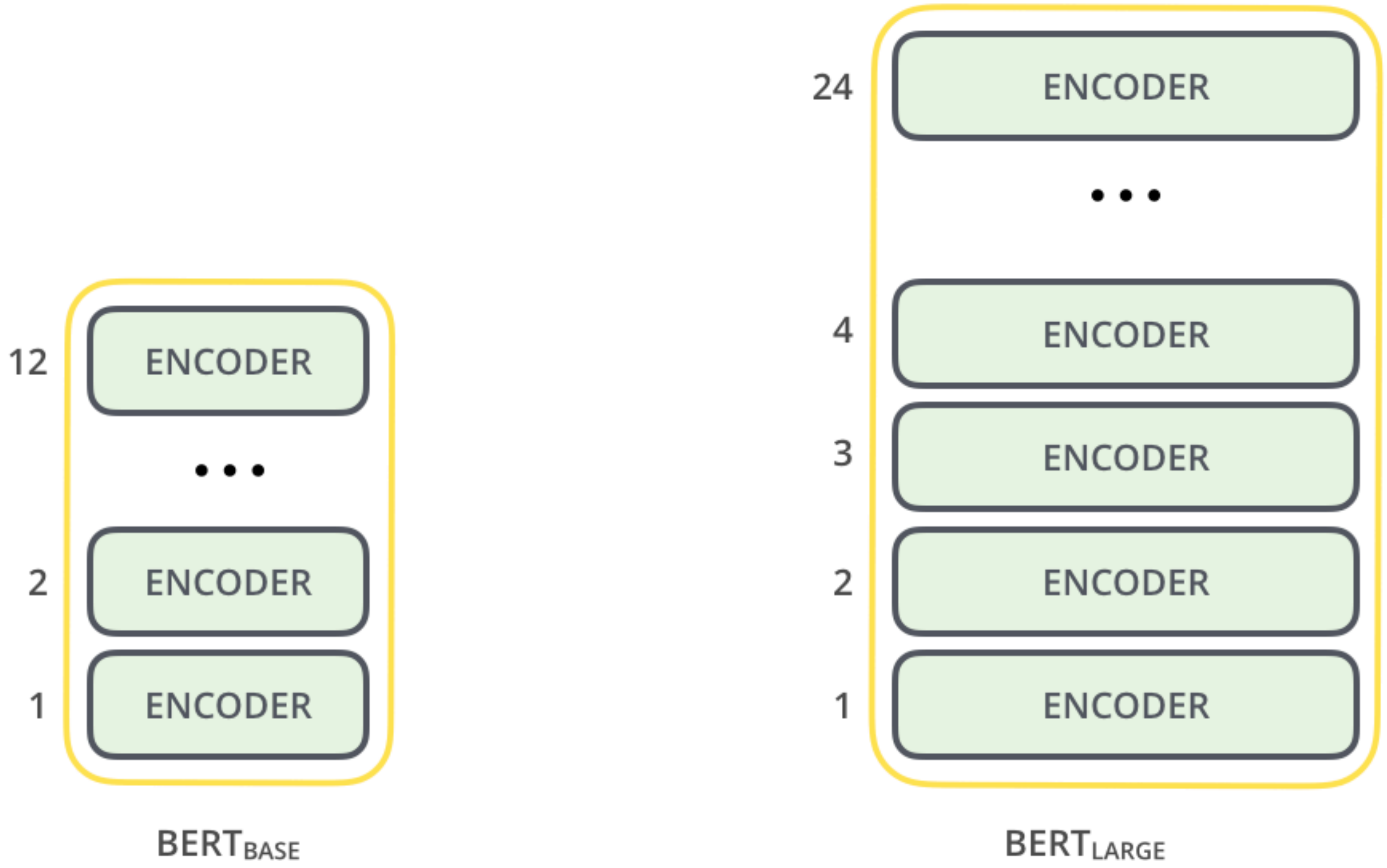
03

GPT-2, BERT

Using wikipedia data



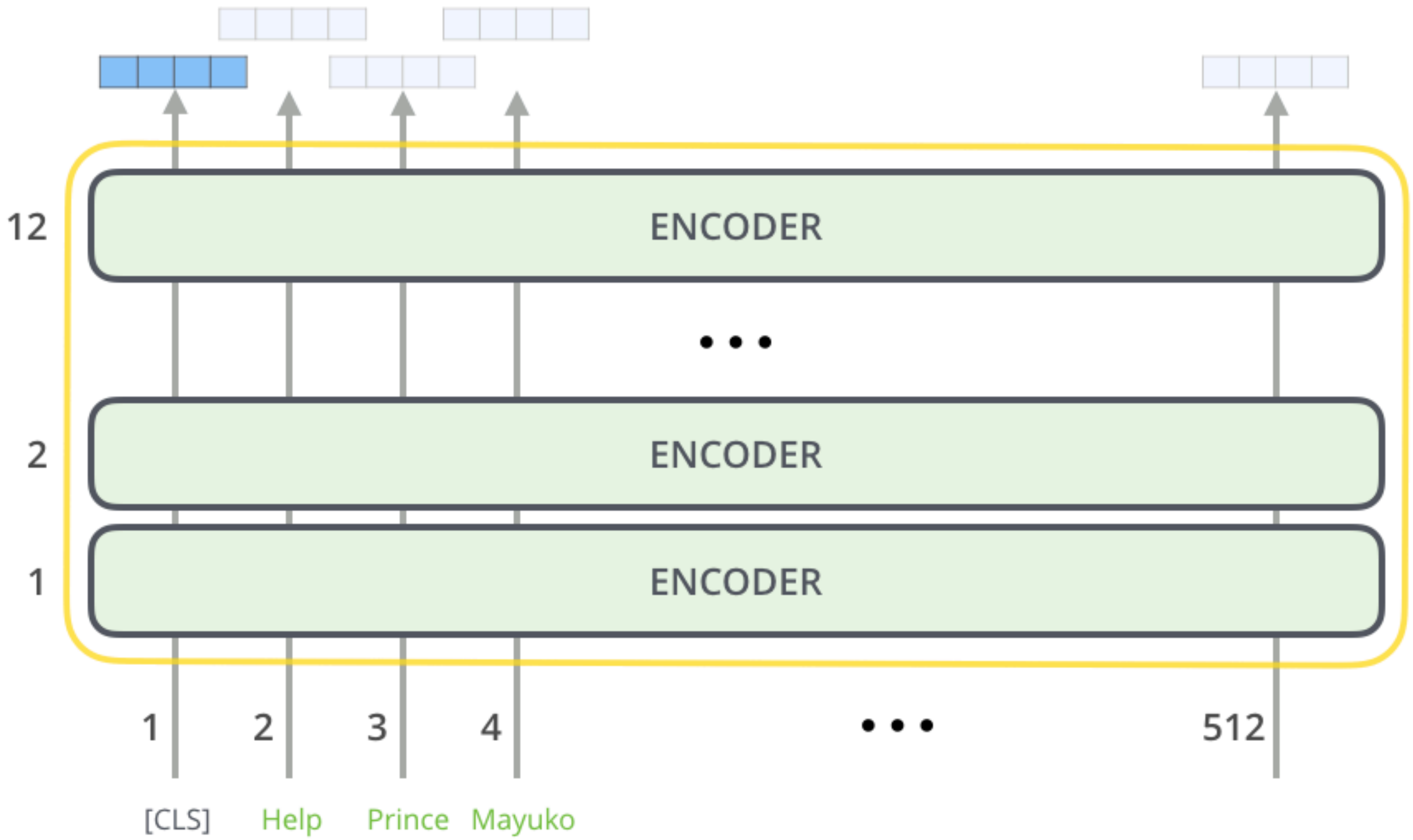
BERT (Bidirectional Encoder Representation from Transformers)



GPT-2, BERT

Model input dimention 512

Input and output vector size (Also 768, and 1024)

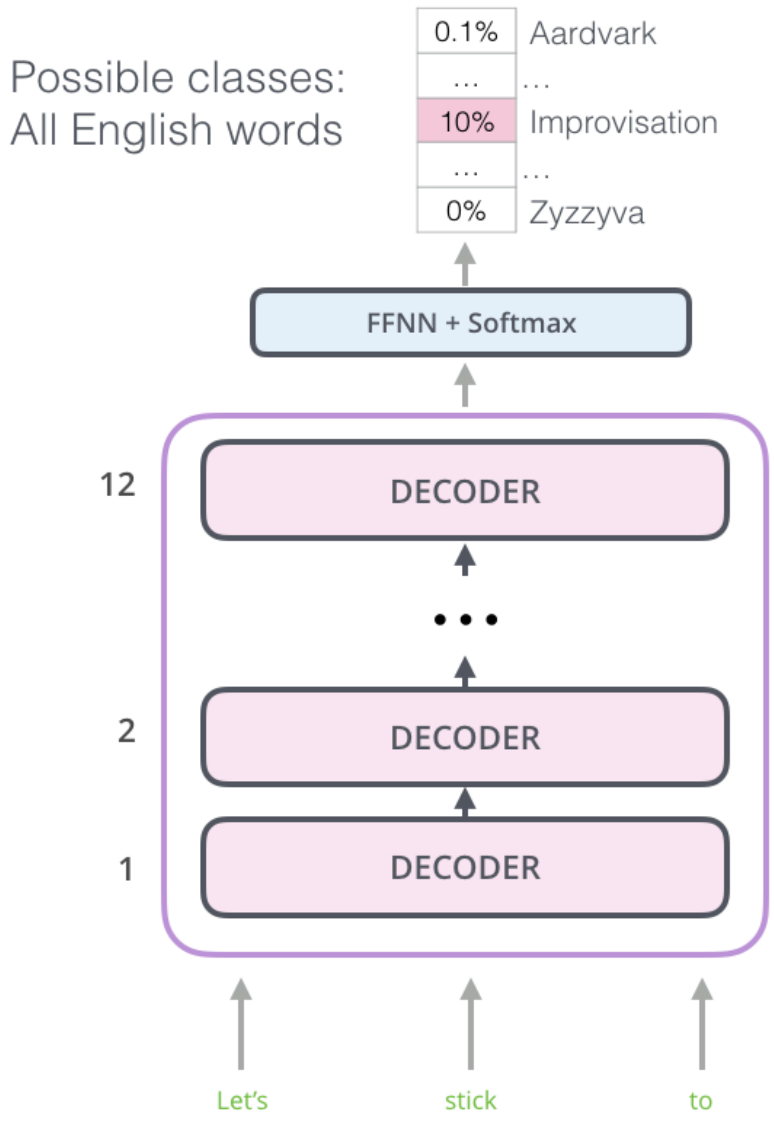


BERT

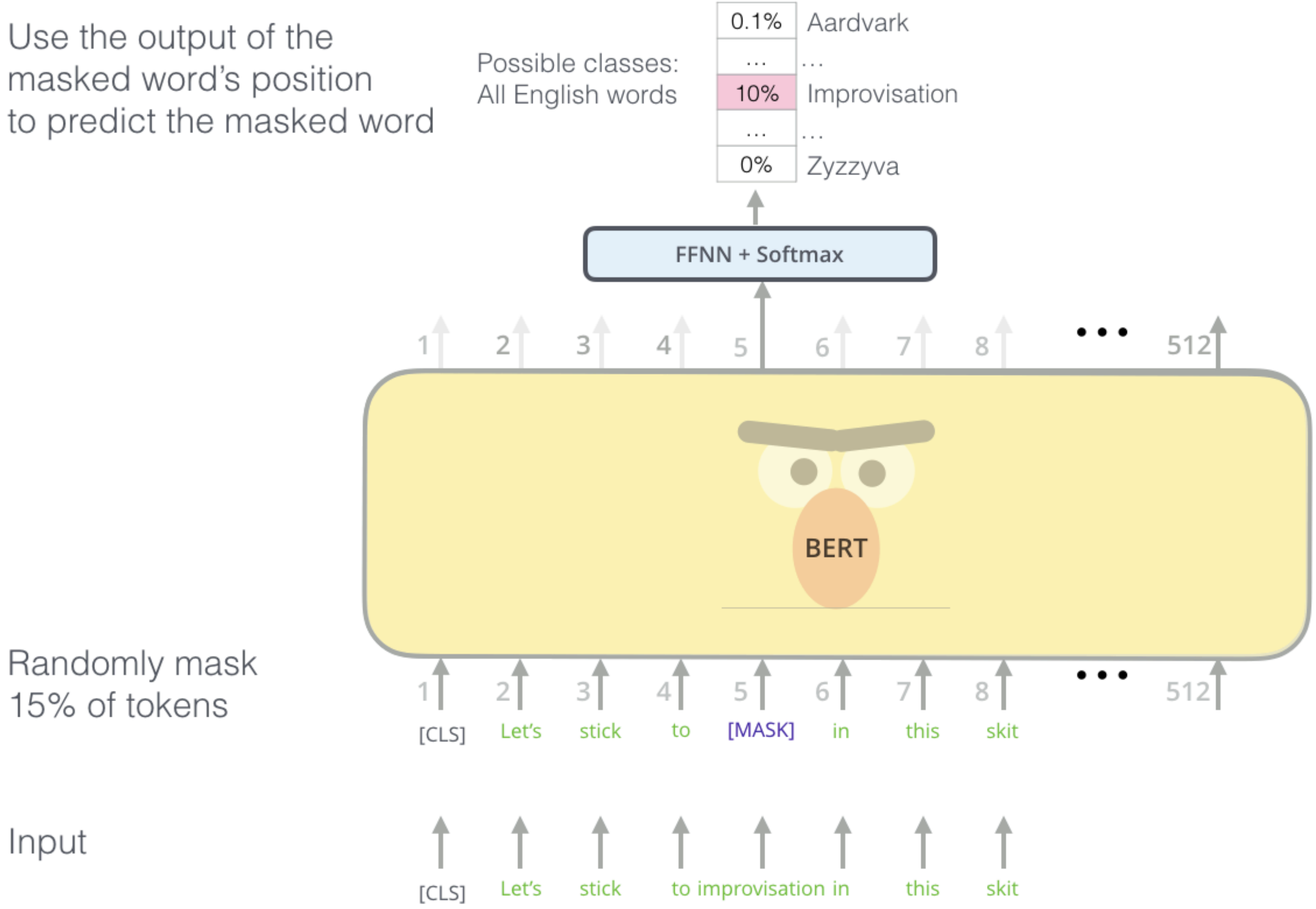
BERT pretraining

ULM-FiT (2018): Pre-training ideas, transfer learning in NLP.
ELMo: Bidirectional training (LSTM)
Transformer: Although used things from left, but still missing from the right.
GPT: Use Transformer Decoder half.
BERT: Switches from Decoder to Encoder, so that it can use both sides in training and invented corresponding training tasks: masked language model

Transformer / GPT prediction

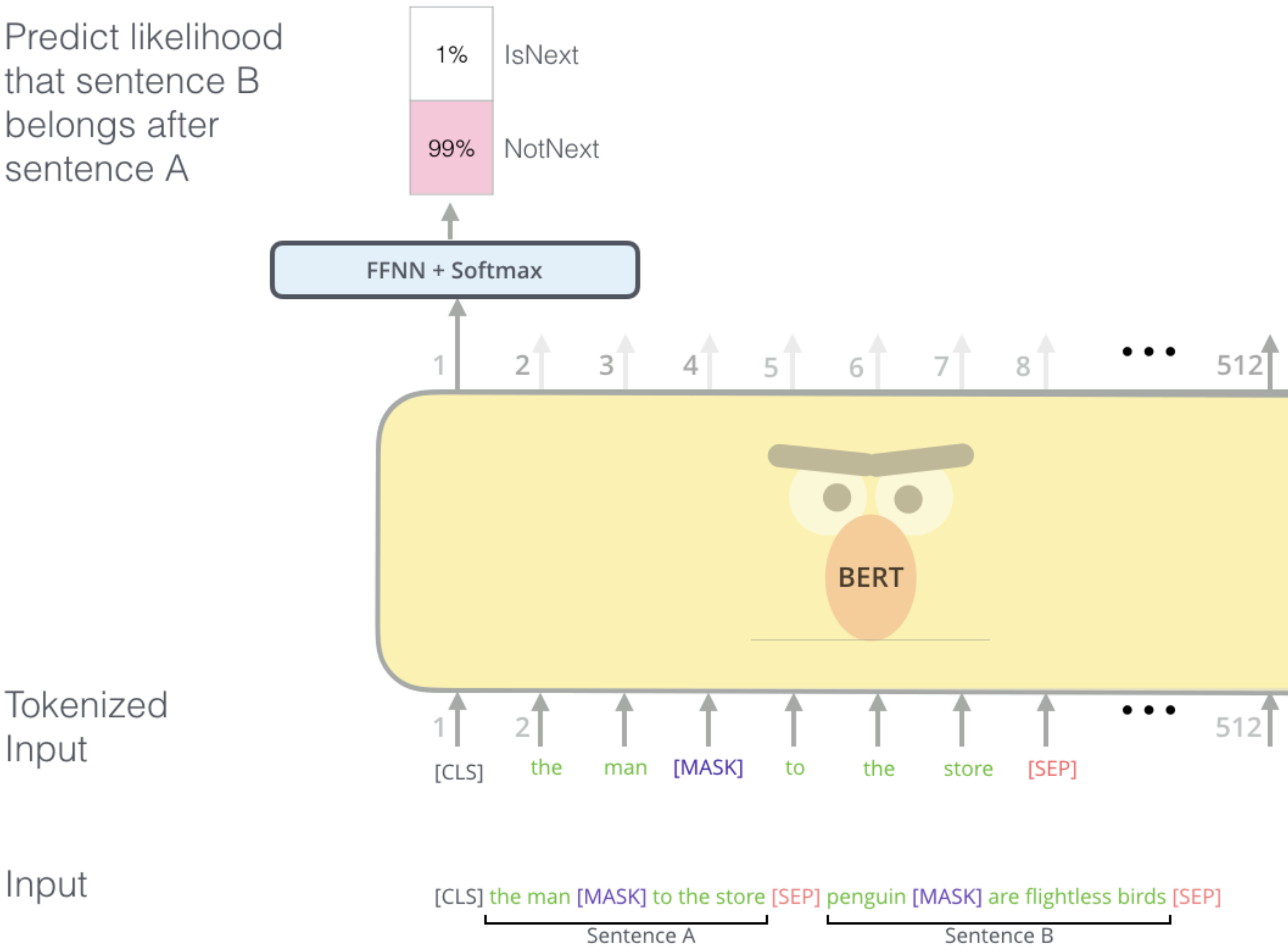


BERT Pretraining Task 1: masked words

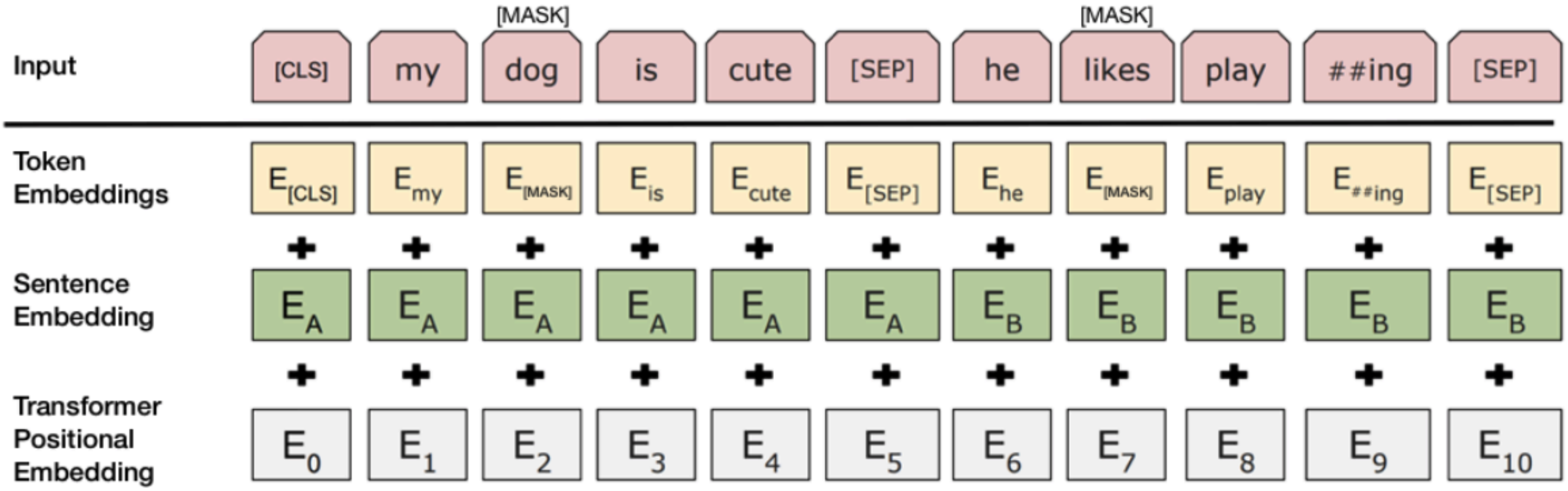


Out of this 15%,
80% are [Mask],
10% random words
10% original words

BERT Pretraining Task 2: two sentences



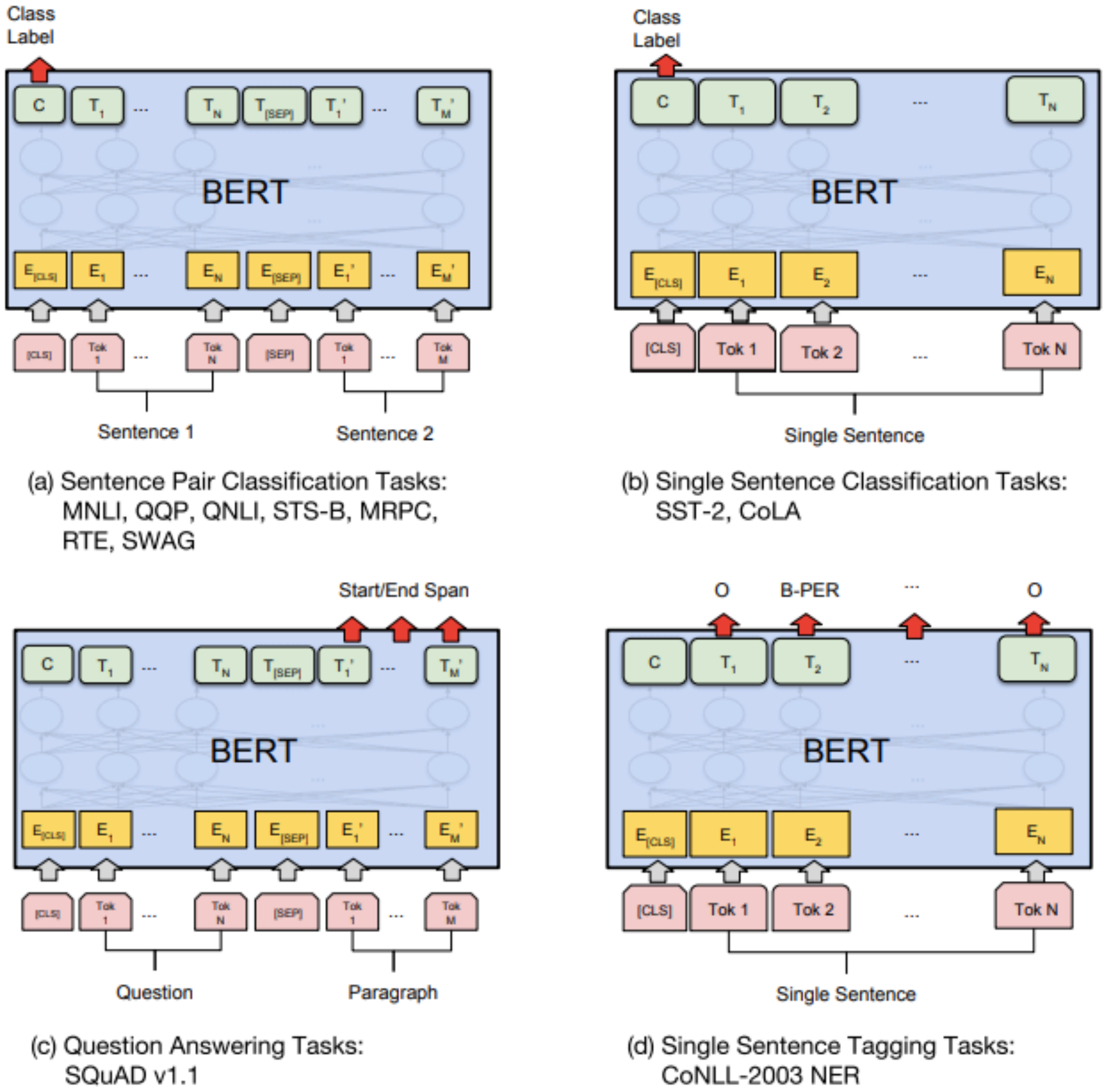
BERT Pretraining Task 2: two sentences



50% true second sentences

50% random second sentences

Fine-tuning BERT for other specific tasks



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

SST (Stanford sentiment treebank): 215k phrases with fine-grained sentiment labels in the parse trees of 11k sentences.

NLP Tasks: Multi-Genre Natural Lang. Inference

MNLI: 433k pairs of examples, labeled by entailment, neutral or contraction

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** N N N N	The 8 million dollars for emergency housing was still not enough to solve the problem.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE **neutral** N N E N	I enjoy eating cheese too much to abstain from dairy.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	TELEPHONE **contradiction** C C C C	No one noticed and it wasn't funny at all.

Table 1: Randomly chosen examples from the development set of our new corpus, shown with their genre labels, their selected gold labels, and the validation labels (abbreviated E, N, C) assigned by individual annotators.

NLP Tasks (SQuAD -- Stanford Question Answering Dataset):

Sample: Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

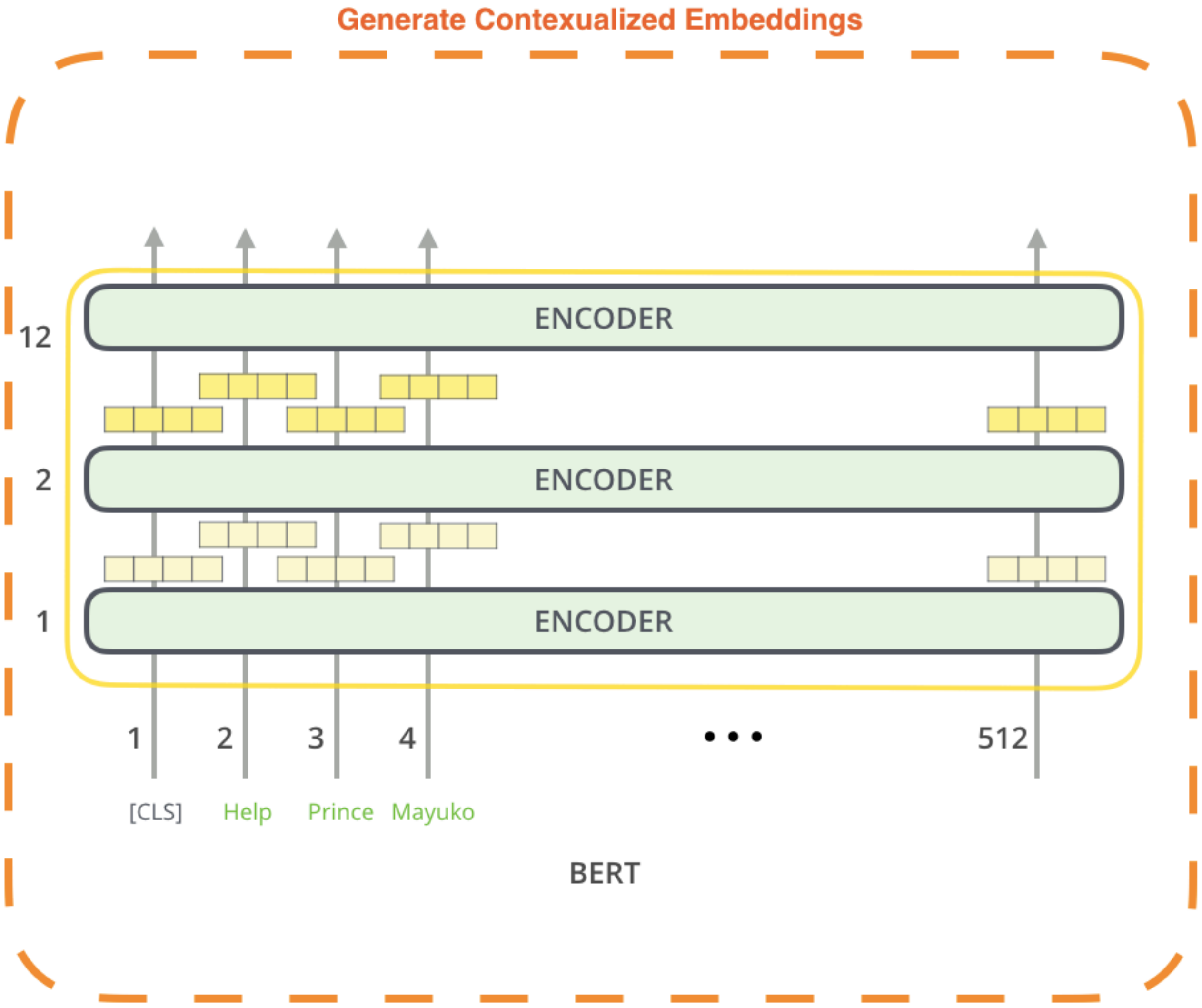
Which NFL team represented the AFC at Super Bowl 50?

Ground Truth Answers: Denver Broncos

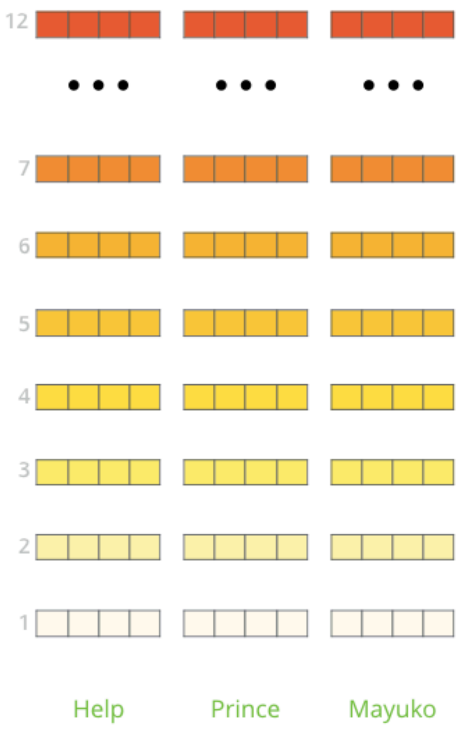
Which NFL team represented the NFC at Super Bowl 50?

Ground Truth Answers: Carolina Panthers

Feature Extraction



The output of each encoder layer along each token's path can be used as a feature representing that token.



We end up with some embedding for each word related to current input

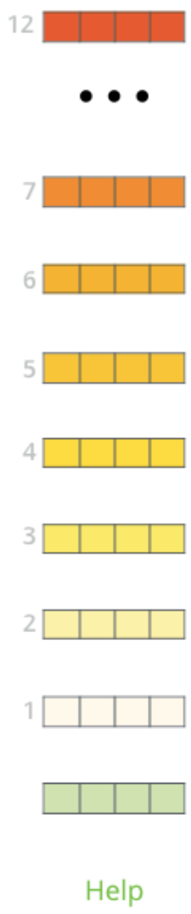
We start with independent word embedding at first level

But which one should we use?

GPT-2, BERT Feature Extraction, which embedding to use?

What is the best contextualized embedding for “Help” in that context?

For named-entity recognition task CoNLL-2003 NER



	Dev F1 Score
First Layer (Embedding)	91.0
Last Hidden Layer (12)	94.9
Sum All 12 Layers (12 + … + 2 + 1)	95.5
Second-to-Last Hidden Layer (11)	95.6
Sum Last Four Hidden (12 + 11 + 10 + 9)	95.9
Concat Last Four Hidden (9, 10, 11, 12)	96.1

Summary of some facts

1. Model size matters (345 million parameters is better than 110 million parameters).

2. With enough training data, more training steps implies higher accuracy

3. BERT's bidirectional approach converges slower than left-to-right approaches but outperforms left-to-right training after a small number of pre-training steps.

4. What do all these mean?

Literature & Resources for Transformers

Resources:

OpenAI GPT-2 implementation: https://github.com/openai/gpt-2

BERT paper: J. Devlin et al, BERT, pretraining of deep bidirectional transformers for language understanding. Oct. 2018.

ELMo paper: M. Peters, et al, Deep contextualized word representation, 2018

ULM-FiT paper: Universal language model fine-tuning for text classification. J. Howeard, S. Ruder., 2018

Jay Alammar, The illustrated GPT-2, https://jalammar.github.io/illustrated-gpt2/