

Jan. 17, 2020



# CS 886 Deep Learning and NLP



Ming Li

# CONTENT

---

- 01. Word2Vec
- 02. Attention / Transformers
- 03. GPT / BERT
- 04. Simplicity, ALBERT, Single headed attention RNN
- 05. Student presentations Starting Feb. 3
- 06. Student presentations ending March 30
- 07. Student short presentations of research projects



# GPT-2 and BERT

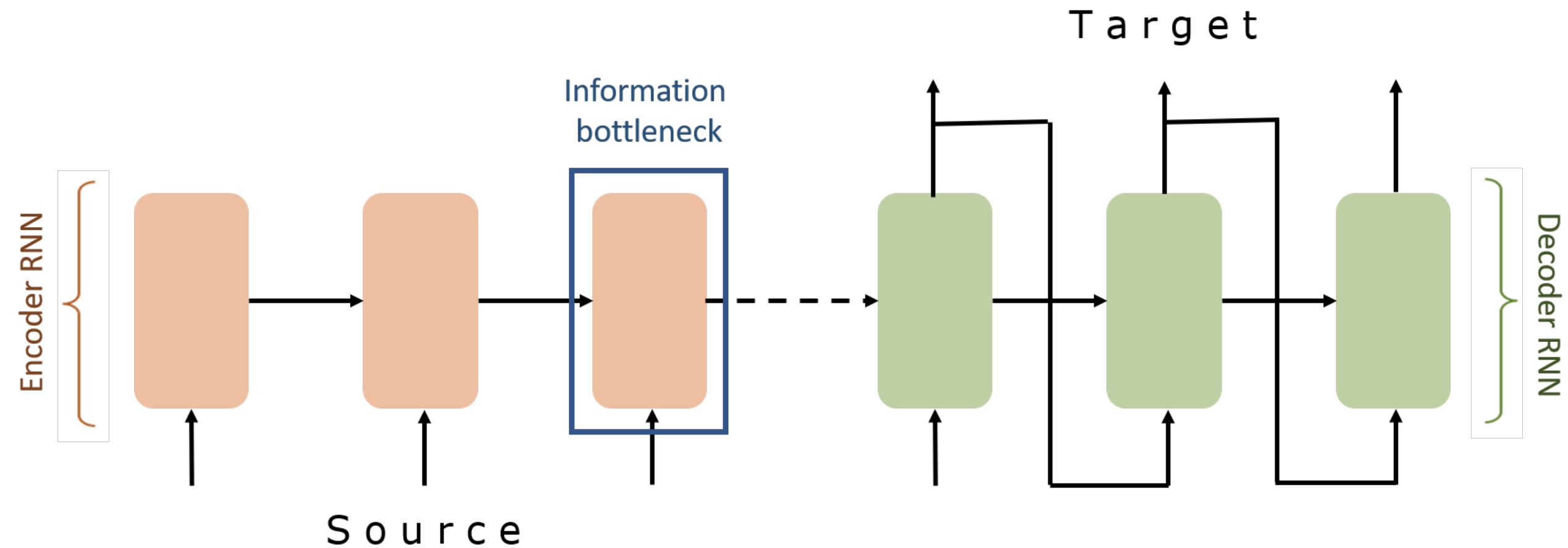
---

LECTURE THREE

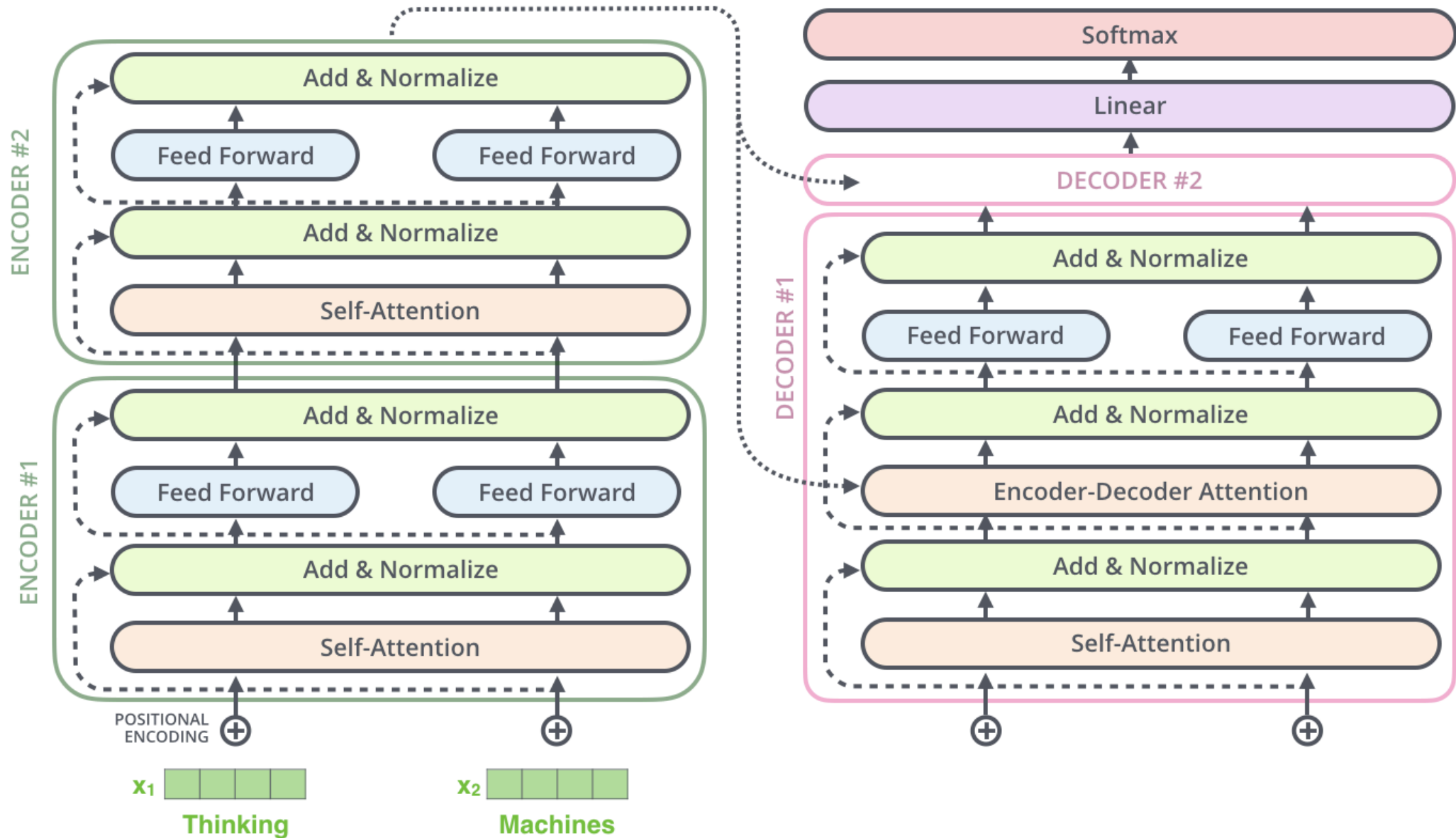


Tying up loose ends from the last lecture, back to  
Lecture 2 notes.

# Avoiding Information bottleneck



# Last time we introduced transformer



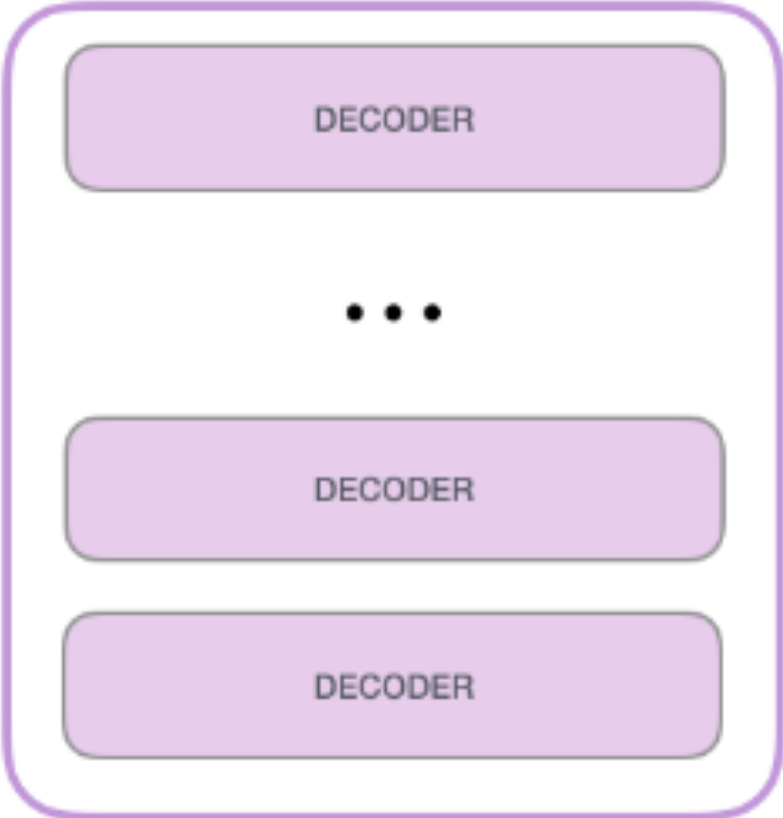


### Transformers, GPT-2, and BERT

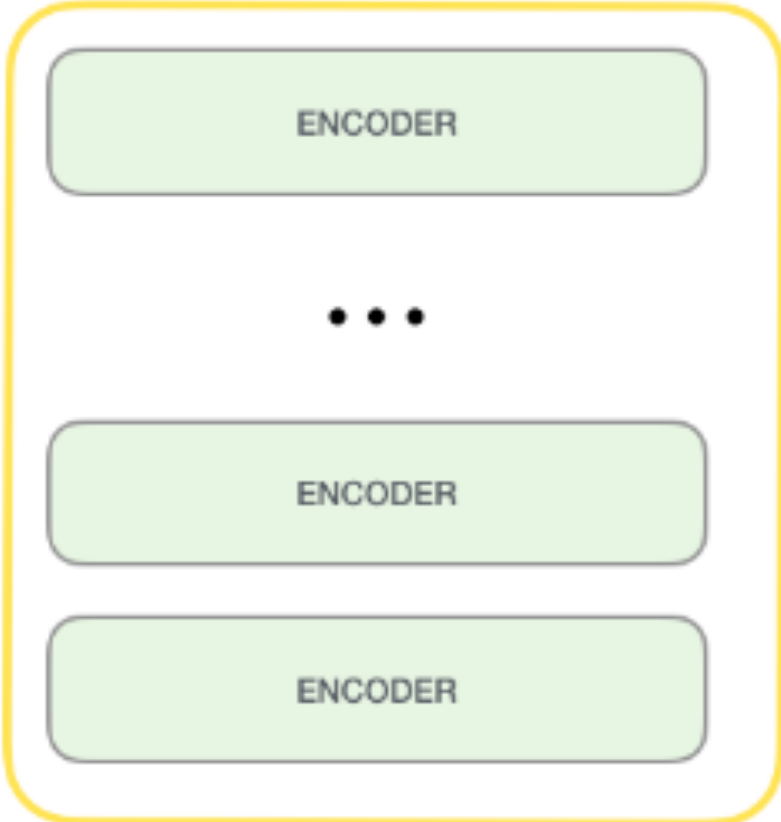
1. A transformer uses Encoder stack to model input, and uses Decoder stack to model output (using input information from encoder side).
2. But if we do not have input, we just want to model the “next word”, we can get rid of the Encoder side of a transformer and output “next word” one by one. This gives 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.



GPT-2



BERT





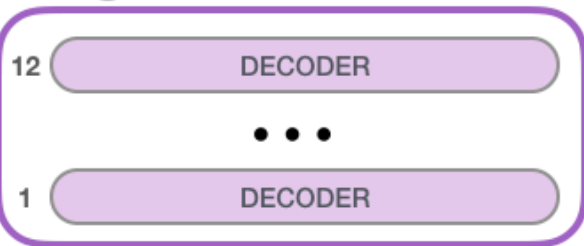


GPT released June 2018

GPT-2 released Nov. 2019 with 1.5B parameters



GPT-2  
SMALL

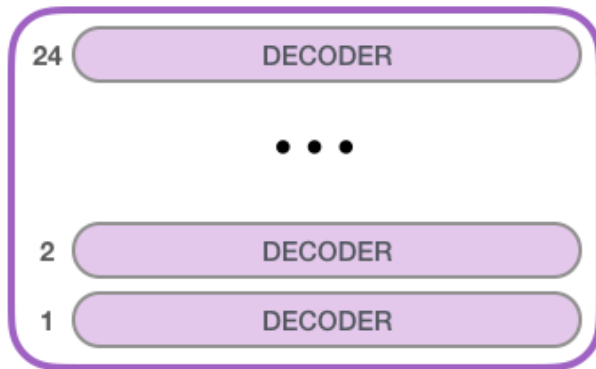


Model Dimensionality: 768

117M parameters



GPT-2  
MEDIUM

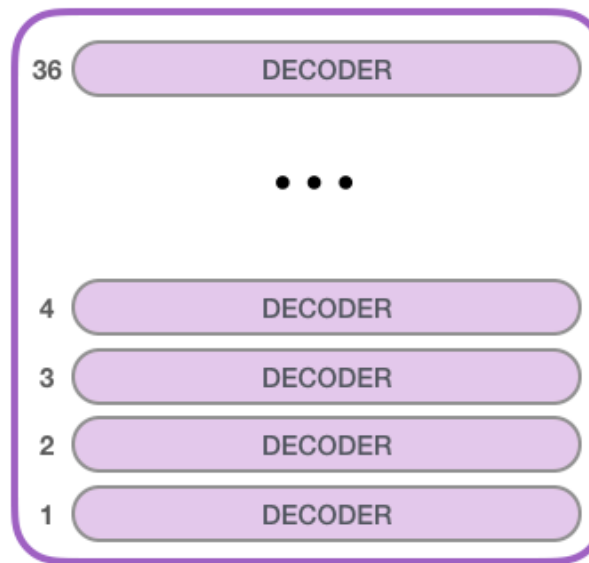


Model Dimensionality: 1024

345M



GPT-2  
LARGE

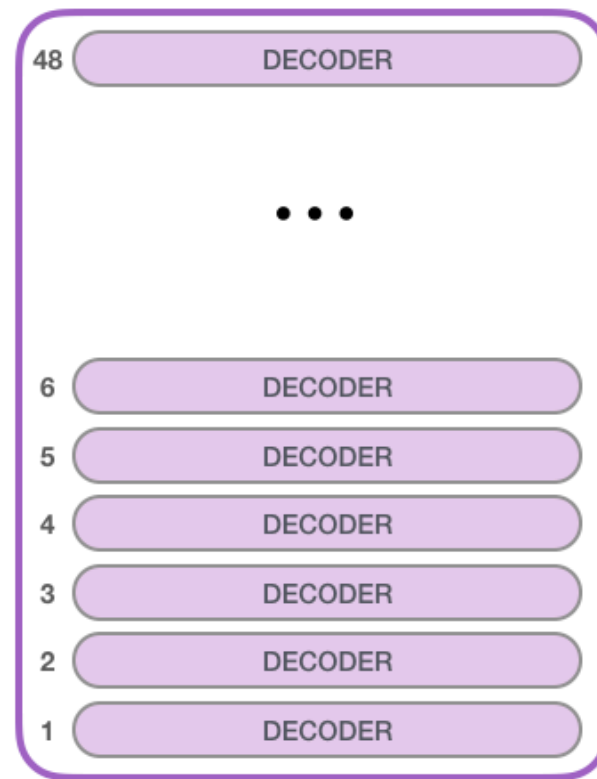


Model Dimensionality: 1280

762M



GPT-2  
EXTRA  
LARGE

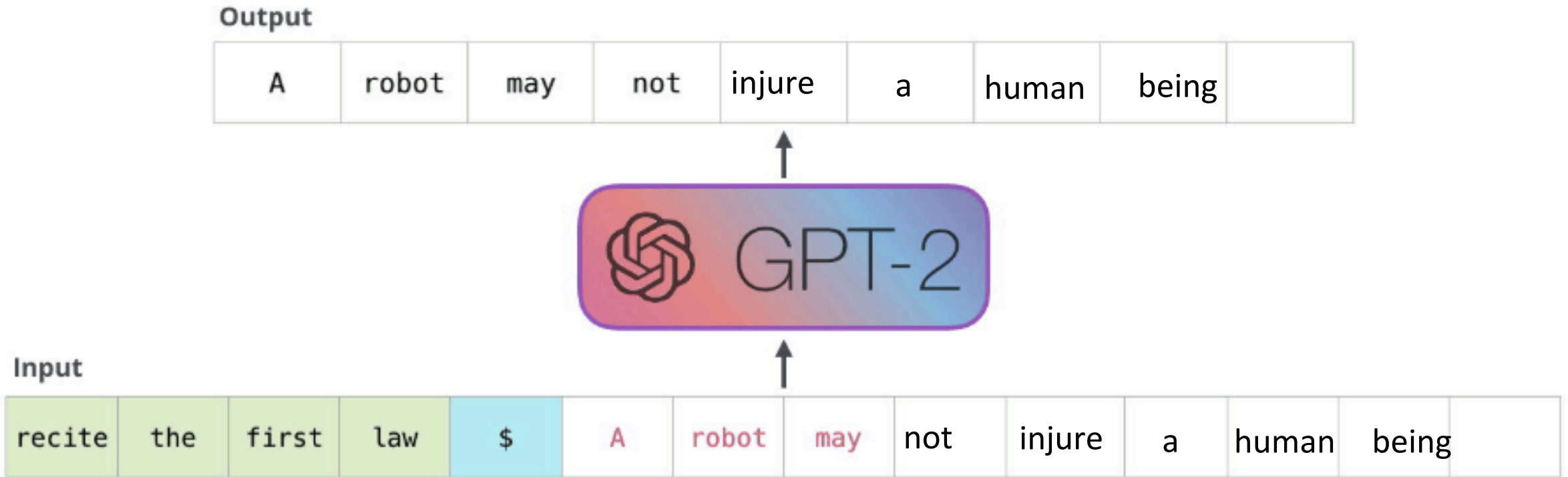


Model Dimensionality: 1600

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 subwords 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 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 e s t </w>': 6, 'w i d e s t </w>': 3 }
```

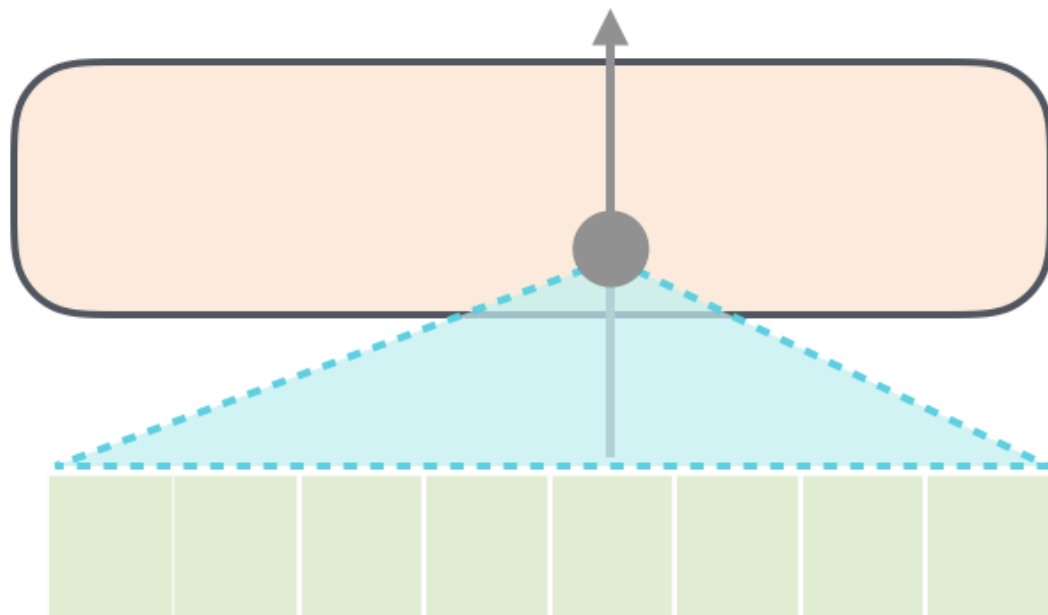
```
{'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w e s t </w>': 6, 'w i d e s t </w>': 3 }
```

.....

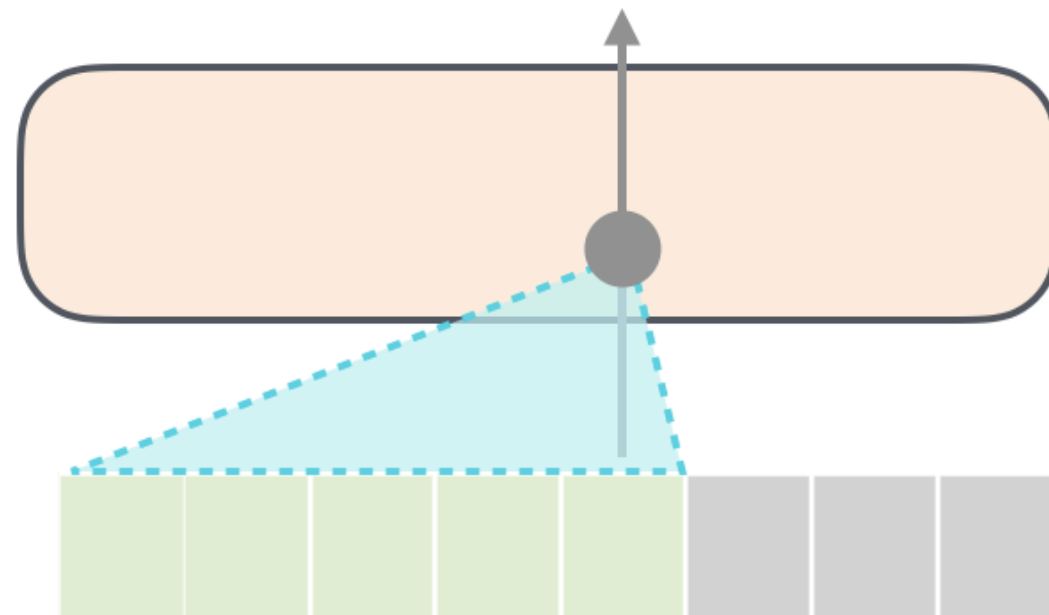
Note that `</w>` is also an important character.

# Masked Self-Attention

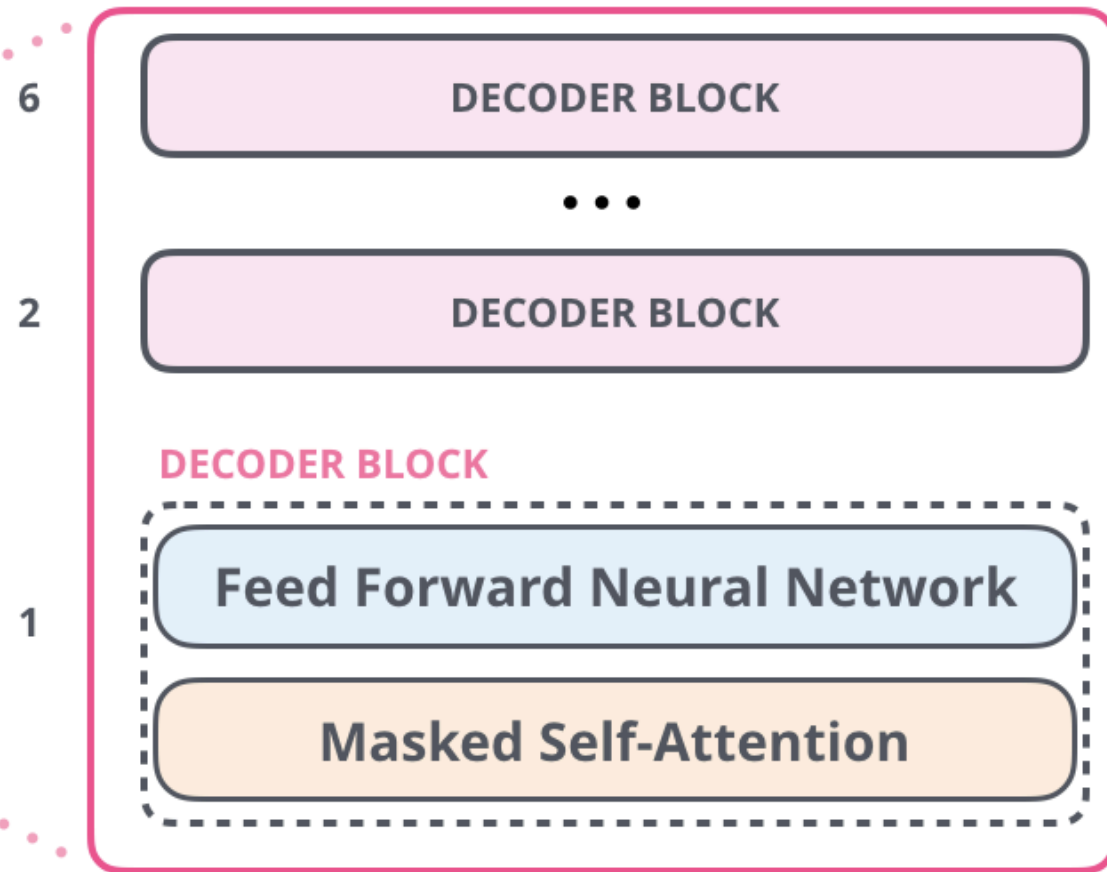
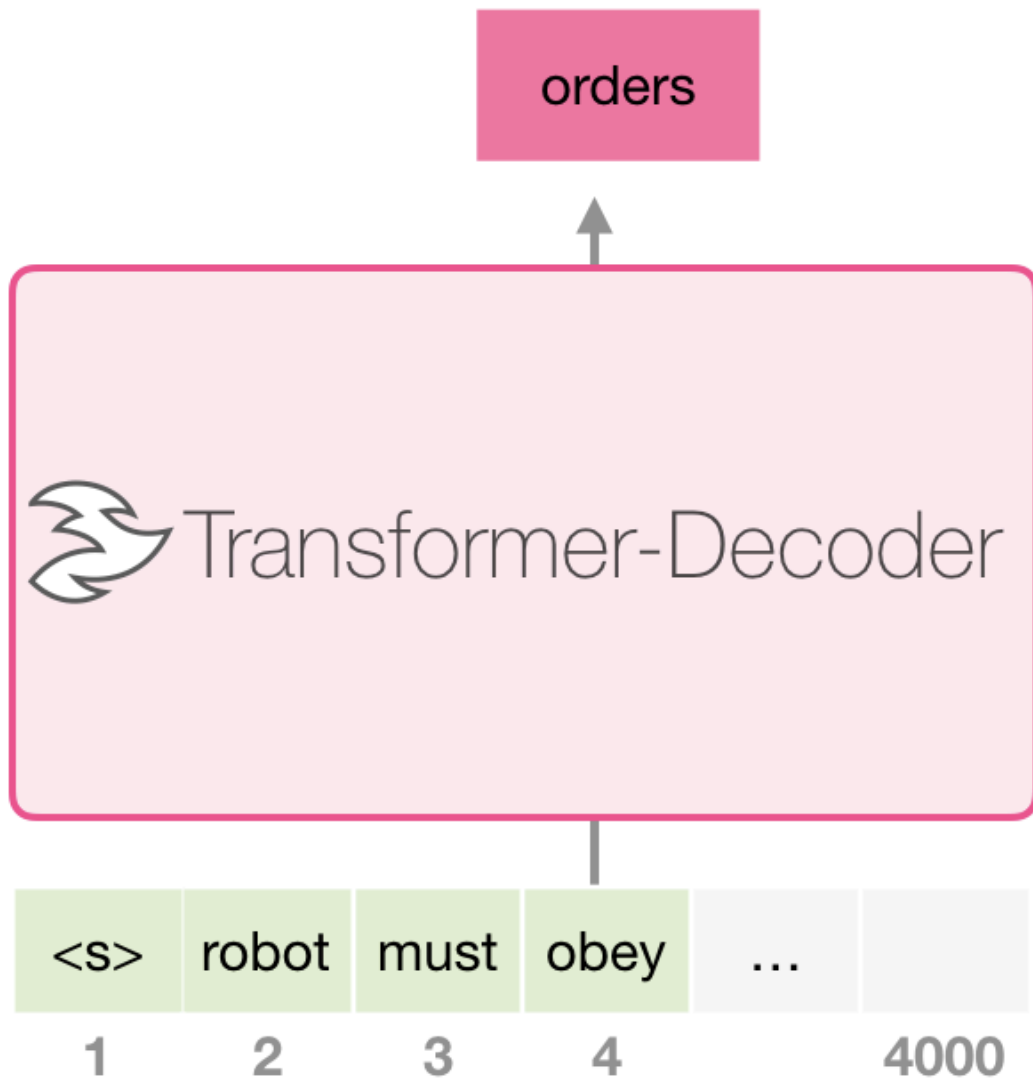
## Self-Attention



## Masked Self-Attention



# Masked Self-Attention



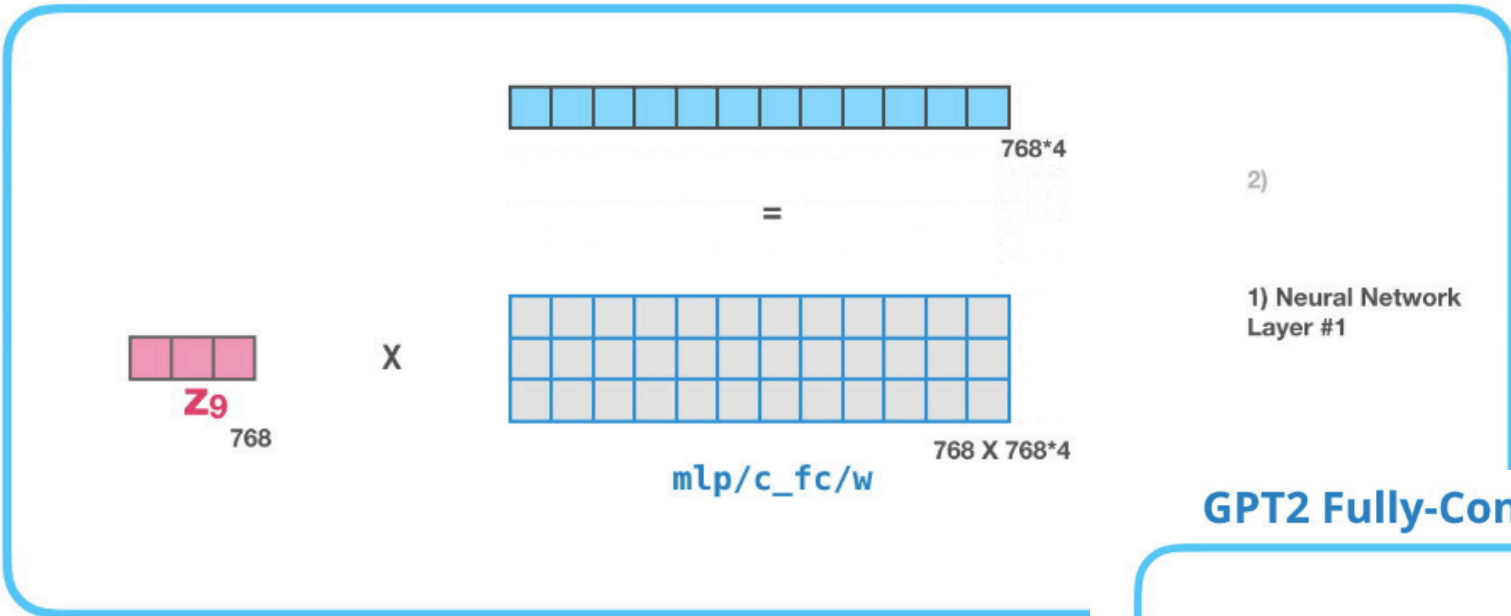
Note: encoder-decoder attention block is gone

## 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.

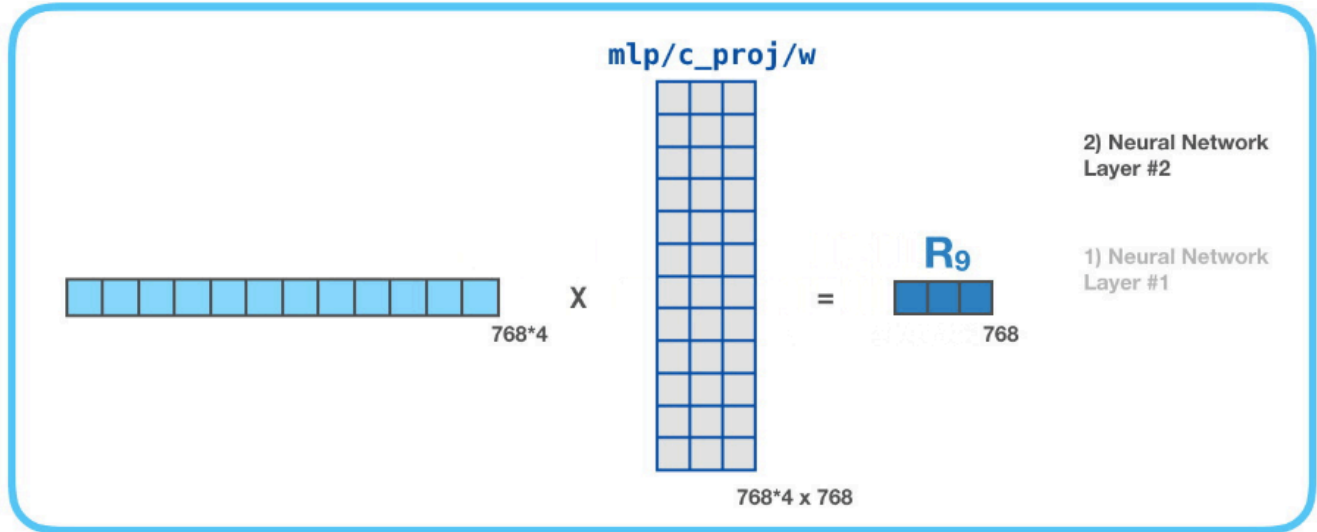
GPT-2 fully connected network has two layers (Example for GPT-2 small)

### GPT2 Fully-Connected Neural Network



2)  
1) Neural Network Layer #1

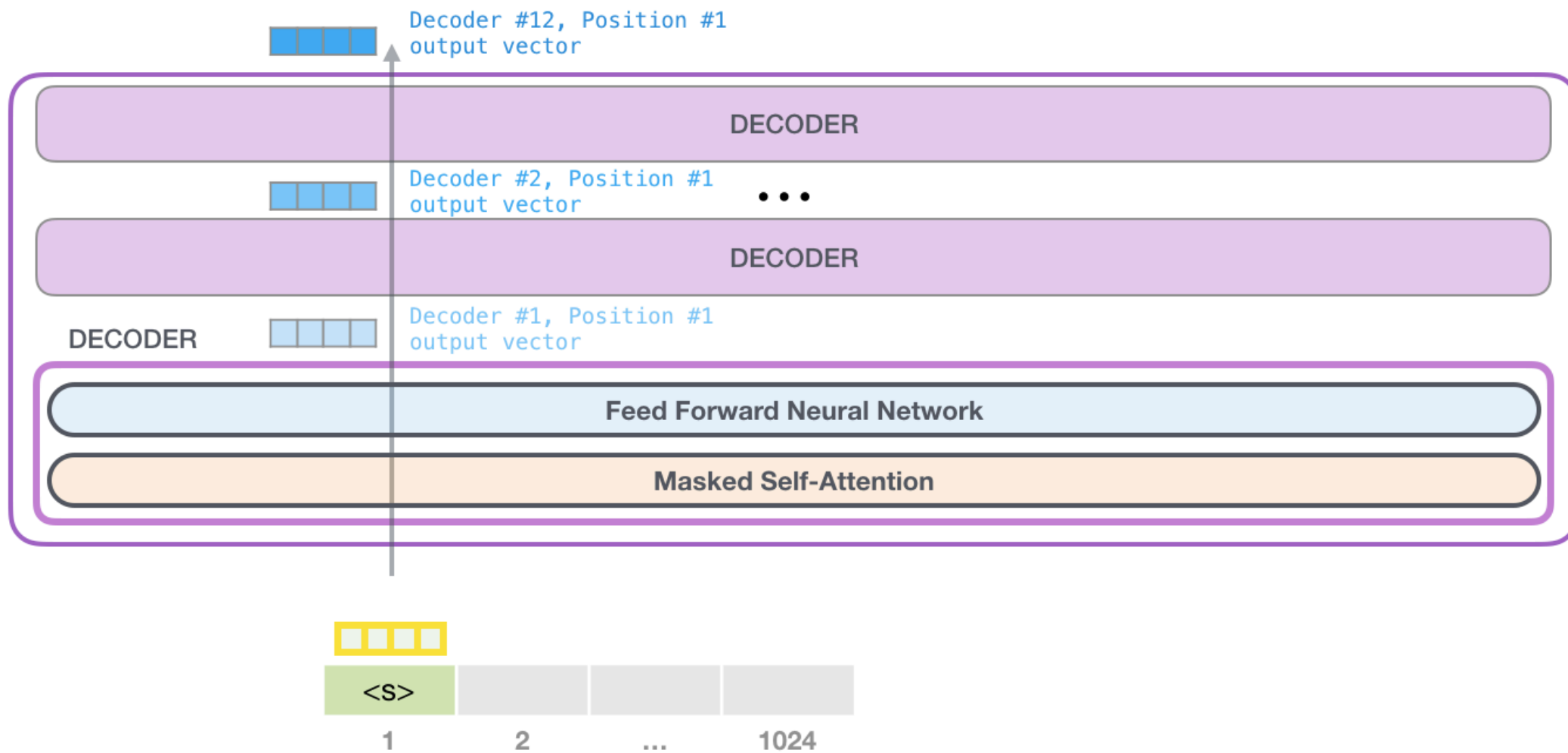
### GPT2 Fully-Connected Neural Network



2) Neural Network Layer #2  
1) Neural Network Layer #1



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







## 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.

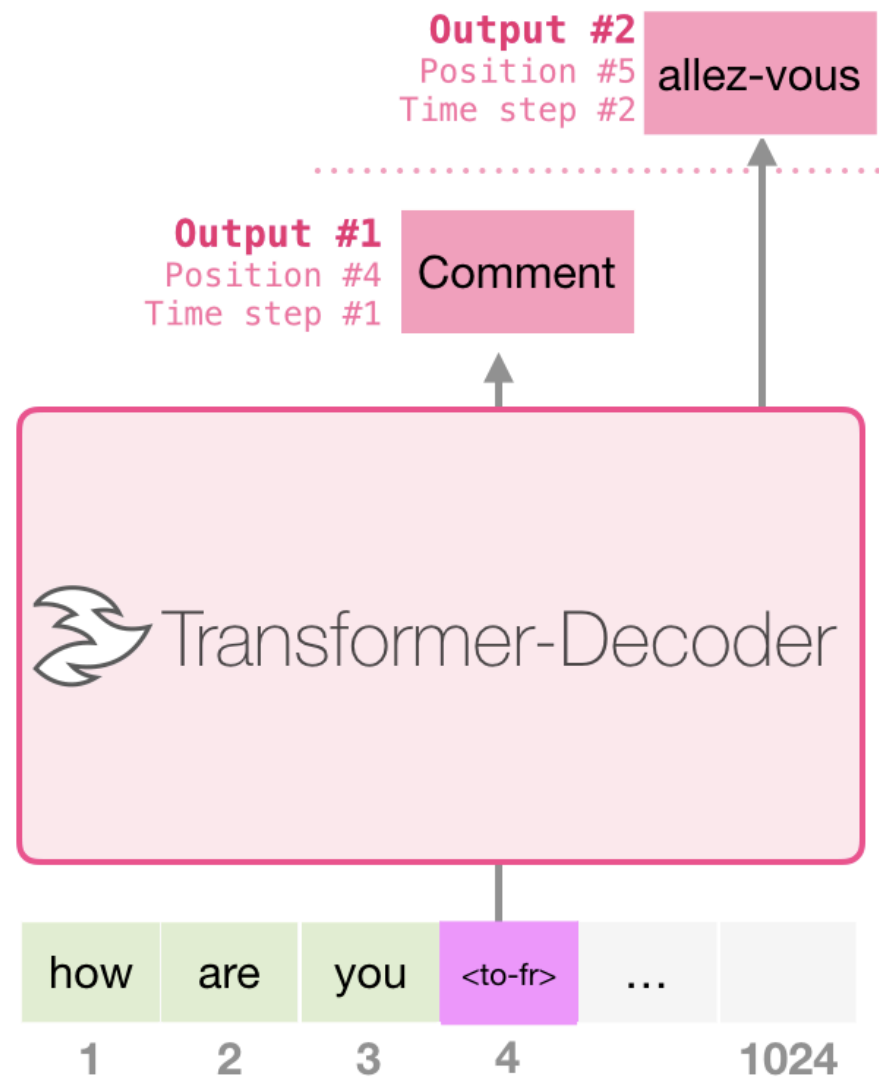
Pérez and the others then ventured further into the valley. ‘By the time we reached the top of one peak, the water looked blue, with some crystals on top,’ said Pérez.

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

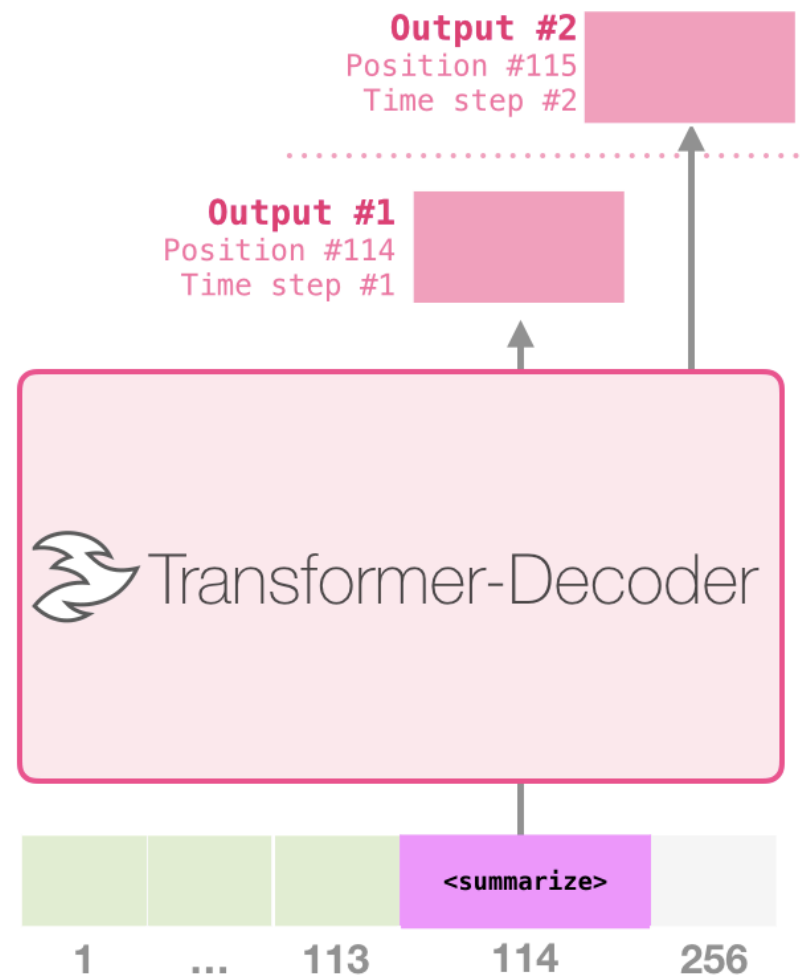
I	am	a	student	<to-fr>	je	suis	étudiant
let	them	eat	cake	<to-fr>	Qu'ils	mangent	de
good	morning	<to-fr>	Bonjour				



# GPT-2 Application: Summarization

## Training Dataset

Article #1 tokens	<summarize>	Article #1 Summary	
Article #2 tokens	<summarize>	Article #2 Summary	padding
Article #3 tokens	<summarize>	Article #3 Summary	



# Using wikipedia data

WIKIPEDIA The Free Encyclopedia

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk

Read | Edit | View history | Search Wikipedia

## Positronic brain

From Wikipedia, the free encyclopedia  
(redirected from *Positronic robot*)

This article is about a fictional technological device. For the manufacturing company based in Springfield, Missouri, see *Positronic Company*.

**This article needs additional citations for verification. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.**  
Find sources: "Positronic brain" – news – newspapers – books – scholar – JSTOR (July 2008) (Learn how and when to remove this template message)

A **positronic brain** is a fictional technological device, originally conceived by science fiction writer Isaac Asimov.<sup>[citation needed]</sup> It functions as a central processing unit (CPU) for robots, and, in some unspecified way, provides them with a form of consciousness recognizable to humans. When Asimov wrote his first robot stories in 1939 and 1942, the position was a newly discovered particle, and so the buzz word *positronic* added a contemporary gloss of popular science to the concept. The short story "Runaround", by Asimov, elaborates on the concept, in the context of his fictional Three Laws of Robotics.

**Contents** [hide]

- Conceptual overview
- In Allen's Trilogy
- References in other fiction and films
  - Alfred and Cosette Go To Mars
  - The Avengers
  - Doctor Who
  - Star Trek
  - Perry Rhodan
  - I, Robot, 2004 Film
  - Quantum Leap
  - Buck Rogers in the 25th Century
  - Mystery Science Theater 3000
  - Spectreman
  - Ocellaris
- References
- External links

**Conceptual overview** [edit]

Asimov remained vague about the technical details of positronic brains except to assert that their substructure was formed from an alloy of platinum and indium. They were said to be vulnerable to radiation and apparently involve a type of volatile memory (since robots in storage required a power source keeping their brains "alive"). The focus of Asimov's stories was directed more towards the software of robots—such as the Three Laws of Robotics—than the hardware in which it was implemented, although it is stated in his stories that to create a positronic brain without the Three Laws, it would have been necessary to spend years redesigning the fundamental approach towards the brain itself.

Within his stories of robotics on Earth and their development by U.S. Robots, Asimov's positronic brain is less of a plot device and more of a technological item worthy of study.

A positronic brain cannot ordinarily be built without incorporating the Three Laws; any modification thereof would drastically modify robot behavior. Behavioral dilemmas resulting from conflicting potentials set by inexperienced and/or malicious users of the robot for the Three Laws make up the bulk of Asimov's stories concerning robots. They are resolved by applying the science of logic and psychology together with mathematics, the supreme solution finder being Dr. Susan Calvin, Chief Psychopsychologist of U.S. Robots.

The Three Laws are also a bottleneck in brain sophistication. Very complex brains designed to handle world economy interpret the First Law in expanded sense to include humanity as opposed to a single human; in Asimov's later works like *Robots and Empire* this is related to as the "Zeroth Law". At least one brain constructed as a calculating machine, as opposed to being a robot control circuit, was designed to have a flexible, childlike personality so that it was able to pursue difficult problems without the Three Laws inhibiting it completely. Specialized brains created for overseeing world economics were stated to have no personality at all.

Under specific conditions, the Three Laws can be obviated, with the modification of the actual robotic design.

- Robots that are of low enough value can have the **Third Law** deleted; they do not have to protect themselves from harm, and the brain size can be reduced by half.
- Robots that do not require orders from a human being may have the **Second Law** deleted, and therefore require smaller brains again, providing they do not require the Third Law.
- Robots that are disposable, cannot receive orders from a human being and are not able to harm a human, will not require even the **First Law**. The sophistication of positronic circuitry renders a brain so small that it could comfortably fit within the skull of an insect.

Robots of the latter type directly parallel contemporary industrial robotics practice, though real-life robots do contain safety sensors and systems, in a concern for human safety (a weak form of the First Law; the robot is a safe tool to use, but has no "judgment", which is implicit in Asimov's own stories).

**In Allen's trilogy** [edit]

Several robot stories have been written by other authors following Asimov's death. For example, in Roger MacBride Allen's *Caliban* trilogy, a Spacer roboticist called Gubler Anshaw invents the **gravitronic brain**. It offers speed and capacity improvements over traditional positronic designs, but the strong influence of tradition make robotics labs reject Anshaw's work. Only one roboticist, Freda Leving, chooses to adopt gravitronics, because it offers her a blank slate on which she could explore alternatives to the Three Laws. Because they are not dependent upon centuries of earlier research, gravitronic brains can be programmed with the standard Laws, variations of the Laws, or even empty pathways which specify no Laws at all.

WIKIPEDIA The Free Encyclopedia

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk

Read | Edit | View history | Search Wikipedia

## Positronic brain

From Wikipedia, the free encyclopedia  
(redirected from *Positronic robot*)

This article is about a fictional technological device. For the manufacturing company based in Springfield, Missouri, see *Positronic Company*.

**This article needs additional citations for verification. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.**  
Find sources: "Positronic brain" – news – newspapers – books – scholar – JSTOR (July 2008) (Learn how and when to remove this template message)

A **positronic brain** is a fictional technological device, originally conceived by science fiction writer Isaac Asimov.<sup>[citation needed]</sup> It functions as a central processing unit (CPU) for robots, and, in some unspecified way, provides them with a form of consciousness recognizable to humans. When Asimov wrote his first robot stories in 1939 and 1942, the position was a newly discovered particle, and so the buzz word *positronic* added a contemporary gloss of popular science to the concept. The short story "Runaround", by Asimov, elaborates on the concept, in the context of his fictional Three Laws of Robotics.

**SUMMARY**

**Contents** [hide]

- Conceptual overview
- In Allen's Trilogy
- References in other fiction and films
  - Alfred and Cosette Go To Mars
  - The Avengers
  - Doctor Who
  - Star Trek
  - Perry Rhodan
  - I, Robot, 2004 Film
  - Quantum Leap
  - Buck Rogers in the 25th Century
  - Mystery Science Theater 3000
  - Spectreman
  - Ocellaris
- References
- External links

**Conceptual overview** [edit]

Asimov remained vague about the technical details of positronic brains except to assert that their substructure was formed from an alloy of platinum and indium. They were said to be vulnerable to radiation and apparently involve a type of volatile memory (since robots in storage required a power source keeping their brains "alive"). The focus of Asimov's stories was directed more towards the software of robots—such as the Three Laws of Robotics—than the hardware in which it was implemented, although it is stated in his stories that to create a positronic brain without the Three Laws, it would have been necessary to spend years redesigning the fundamental approach towards the brain itself.

Within his stories of robotics on Earth and their development by U.S. Robots, Asimov's positronic brain is less of a plot device and more of a technological item worthy of study.

A positronic brain cannot ordinarily be built without incorporating the Three Laws; any modification thereof would drastically modify robot behavior. Behavioral dilemmas resulting from conflicting potentials set by inexperienced and/or malicious users of the robot for the Three Laws make up the bulk of Asimov's stories concerning robots. They are resolved by applying the science of logic and psychology together with mathematics, the supreme solution finder being Dr. Susan Calvin, Chief Psychopsychologist of U.S. Robots.

The Three Laws are also a bottleneck in brain sophistication. Very complex brains designed to handle world economy interpret the First Law in expanded sense to include humanity as opposed to a single human; in Asimov's later works like *Robots and Empire* this is related to as the "Zeroth Law". At least one brain constructed as a calculating machine, as opposed to being a robot control circuit, was designed to have a flexible, childlike personality so that it was able to pursue difficult problems without the Three Laws inhibiting it completely. Specialized brains created for overseeing world economics were stated to have no personality at all.

Under specific conditions, the Three Laws can be obviated, with the modification of the actual robotic design.

- Robots that are of low enough value can have the **Third Law** deleted; they do not have to protect themselves from harm, and the brain size can be reduced by half.
- Robots that do not require orders from a human being may have the **Second Law** deleted, and therefore require smaller brains again, providing they do not require the Third Law.
- Robots that are disposable, cannot receive orders from a human being and are not able to harm a human, will not require even the **First Law**. The sophistication of positronic circuitry renders a brain so small that it could comfortably fit within the skull of an insect.

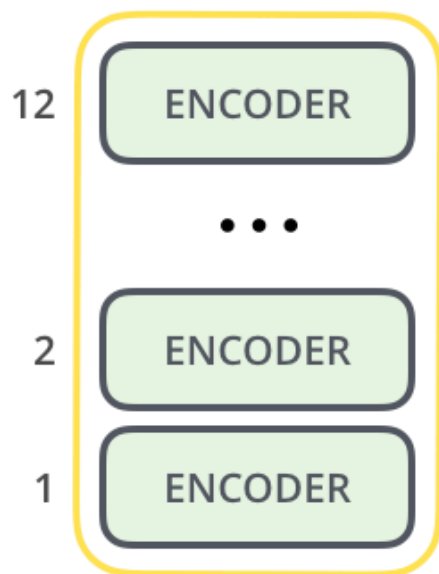
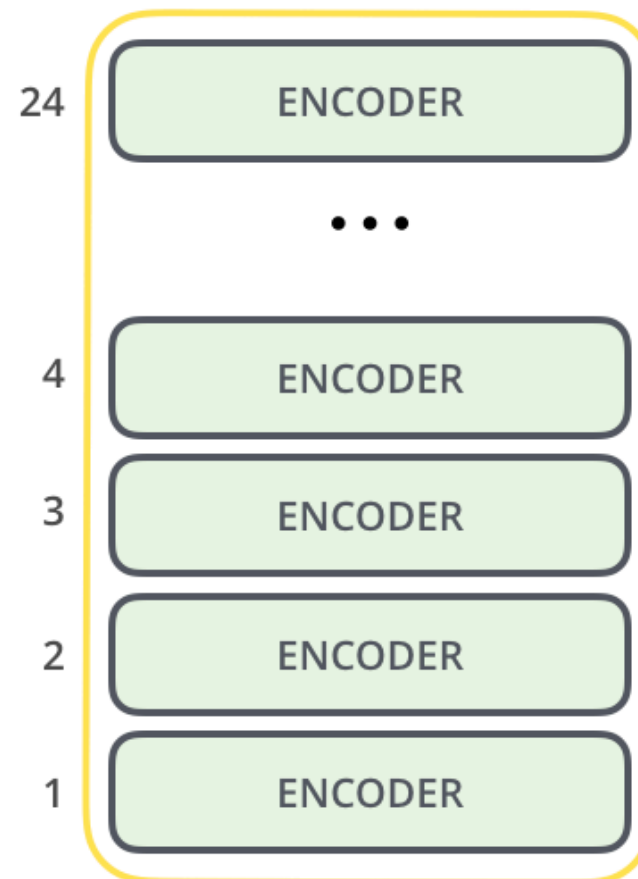
Robots of the latter type directly parallel contemporary industrial robotics practice, though real-life robots do contain safety sensors and systems, in a concern for human safety (a weak form of the First Law; the robot is a safe tool to use, but has no "judgment", which is implicit in Asimov's own stories).

**ARTICLE**

**In Allen's trilogy** [edit]

Several robot stories have been written by other authors following Asimov's death. For example, in Roger MacBride Allen's *Caliban* trilogy, a Spacer roboticist called Gubler Anshaw invents the **gravitronic brain**. It offers speed and capacity improvements over traditional positronic designs, but the strong influence of tradition make robotics labs reject Anshaw's work. Only one roboticist, Freda Leving, chooses to adopt gravitronics, because it offers her a blank slate on which she could explore alternatives to the Three Laws. Because they are not dependent upon centuries of earlier research, gravitronic brains can be programmed with the standard Laws, variations of the Laws, or even empty pathways which specify no Laws at all.

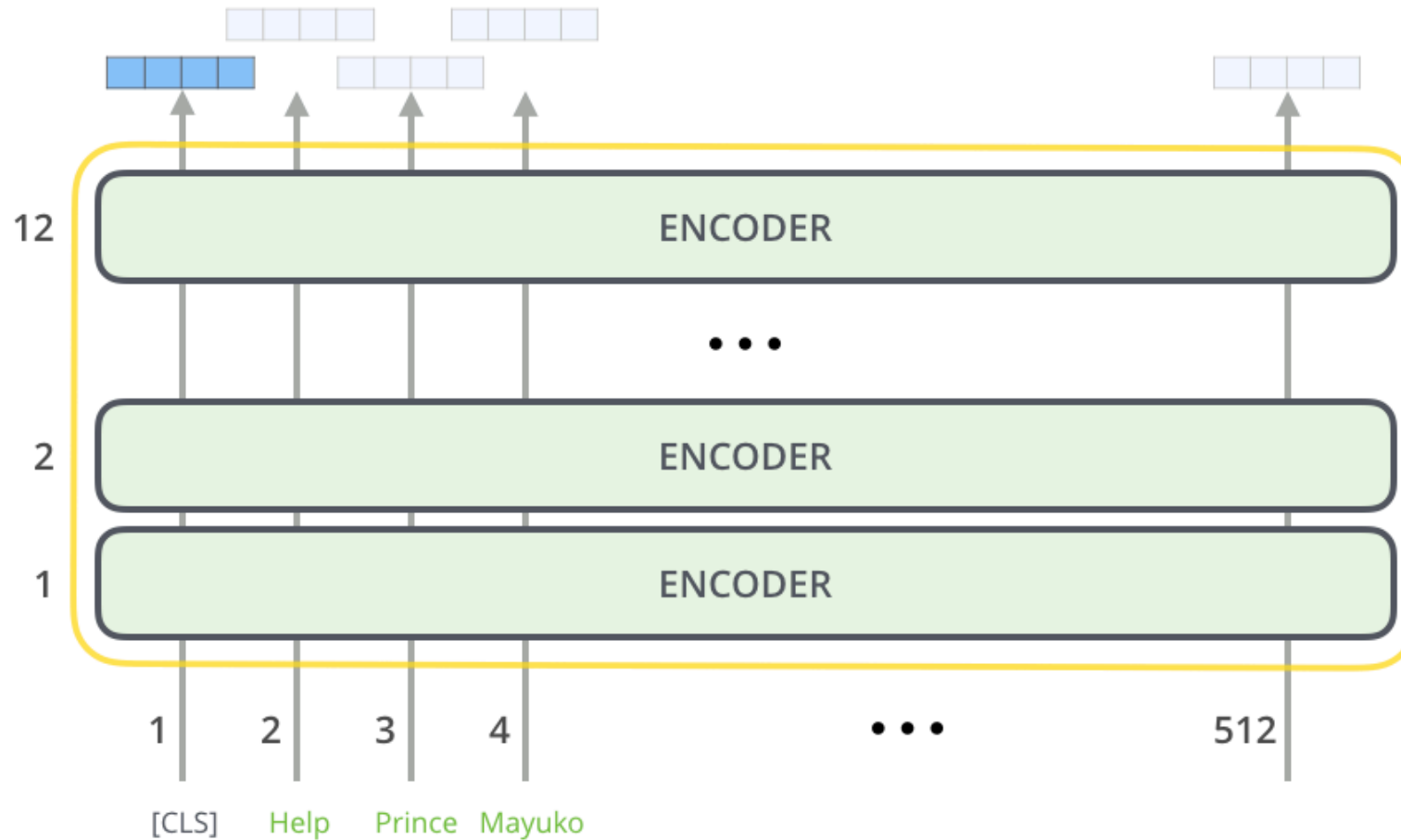
# BERT (Bidirectional Encoder Representation from Transformers)

BERT<sub>BASE</sub>BERT<sub>LARGE</sub>



Model input dimension 512

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



BERT



## BERT pretraining

**ULM-FiT (2018)**: Pre-training ideas, transfer learning in NLP.

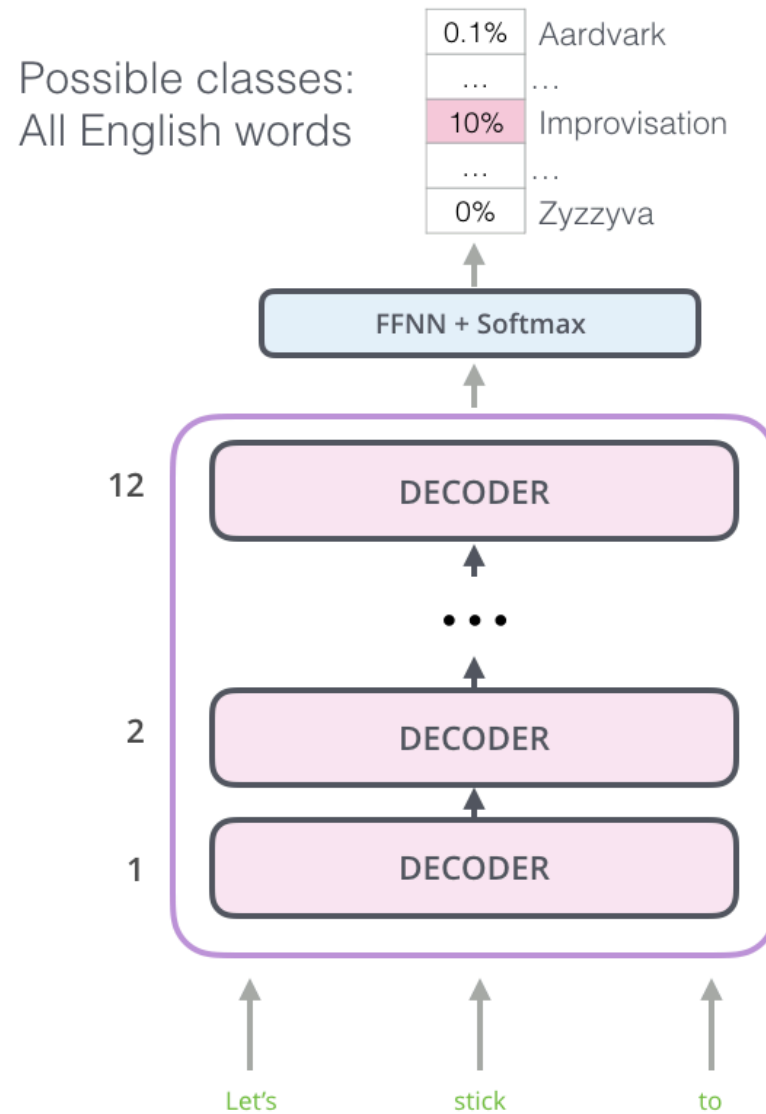
**ELMo**: Bidirectional training (LSTM)

**Transformer**: Although used things from left, but still missing from the right.

**GPT**: Use Transformer Decoder half.

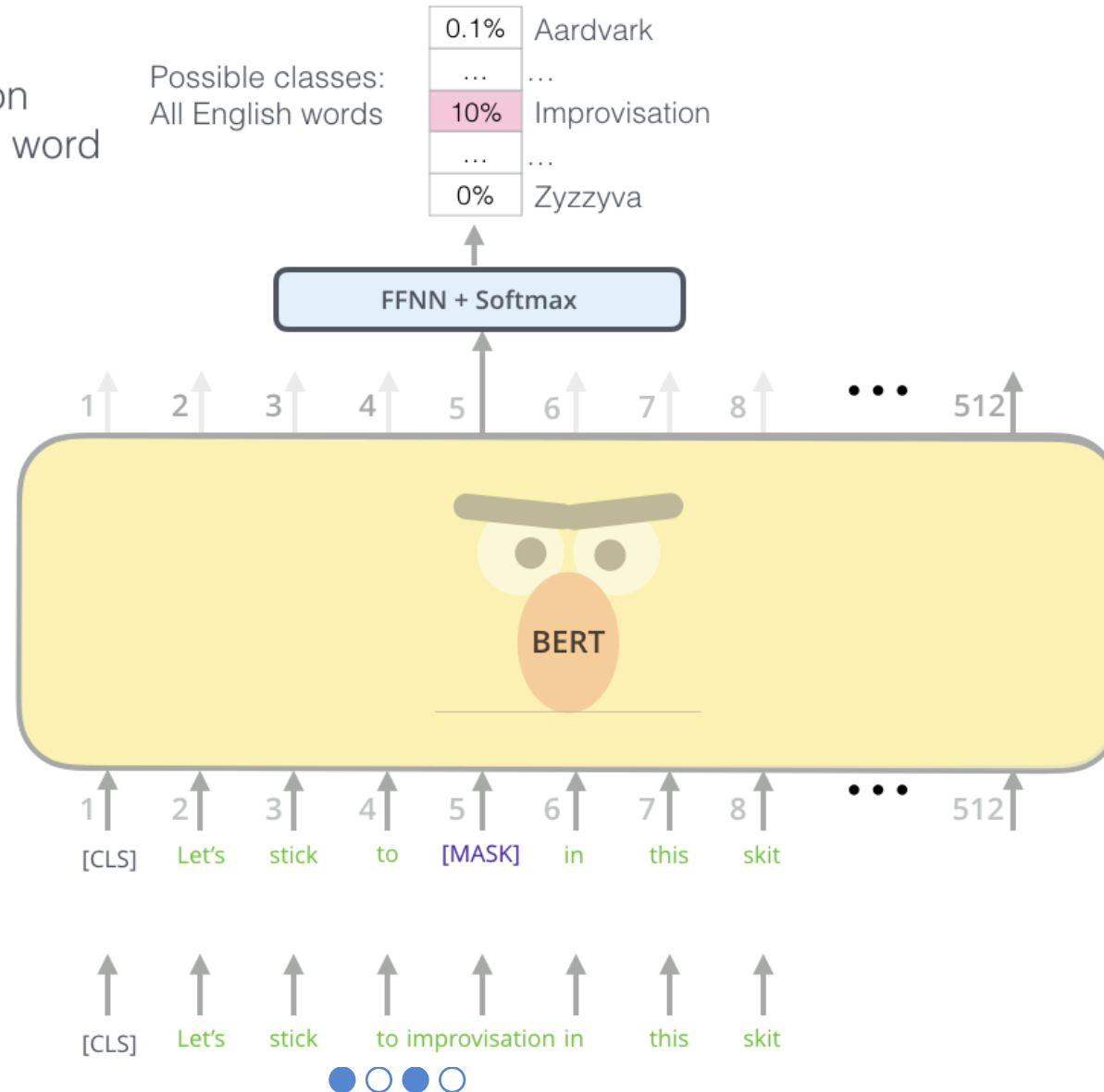
**BERT**: Switches from Decoder to Encoder, so that it can use both sides in training and invented corresponding training tasks: masked language model

# Transformer / GPT prediction



# BERT Pretraining Task 1: masked words

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

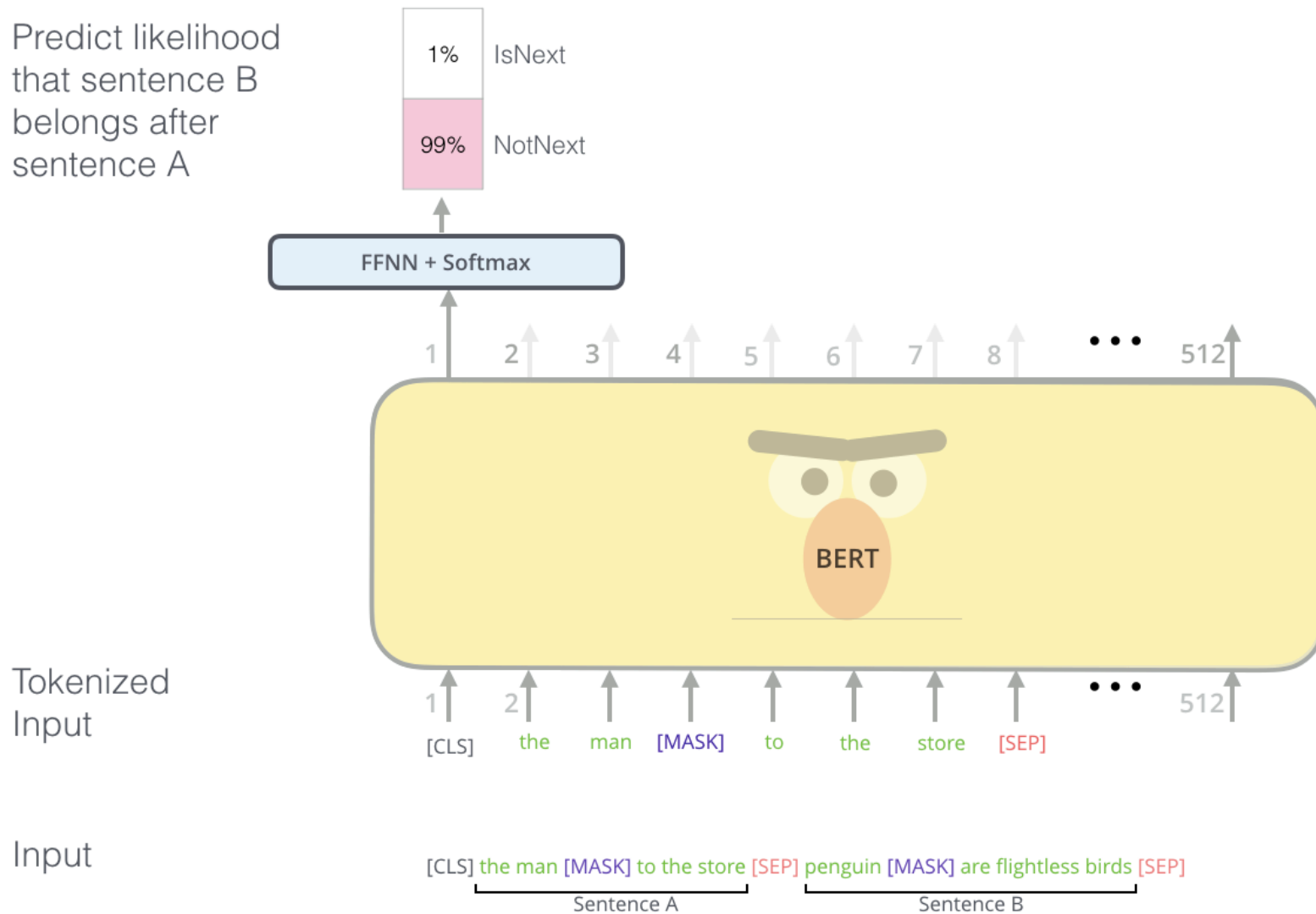


Out of this 15%,  
80% are [Mask],  
10% random words  
10% original words

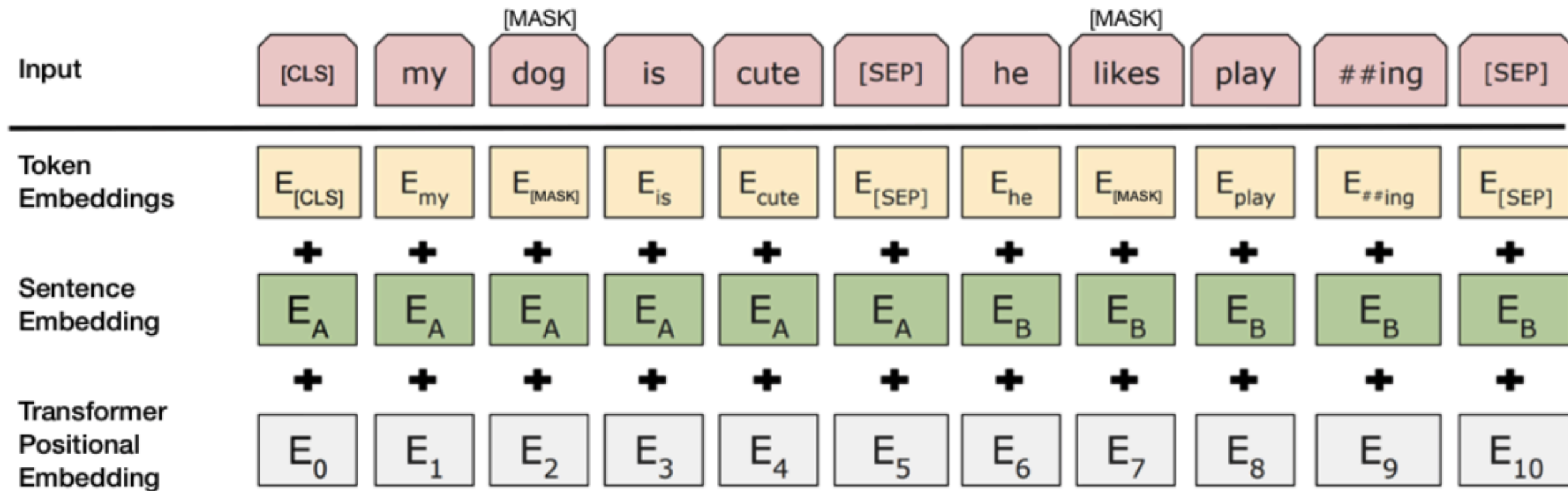
Randomly mask  
15% of tokens

# BERT Pretraining Task 2: two sentences

Predict likelihood that sentence B belongs after sentence A



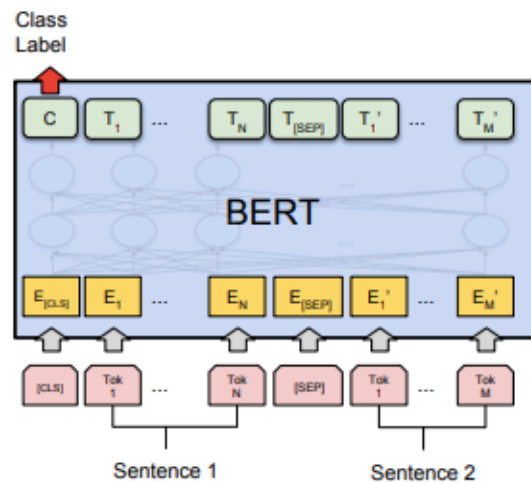
## BERT Pretraining Task 2: two sentences



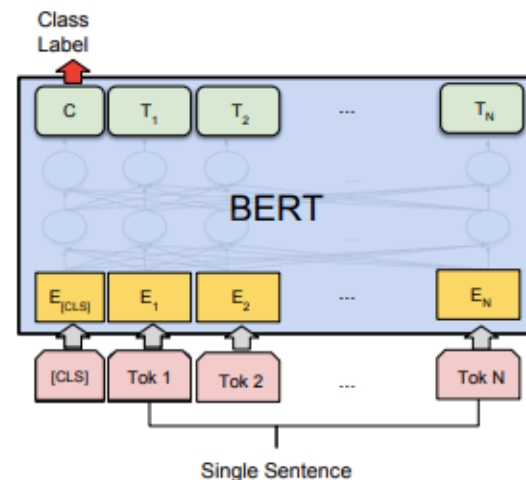
50% true second sentences

50% random second sentences

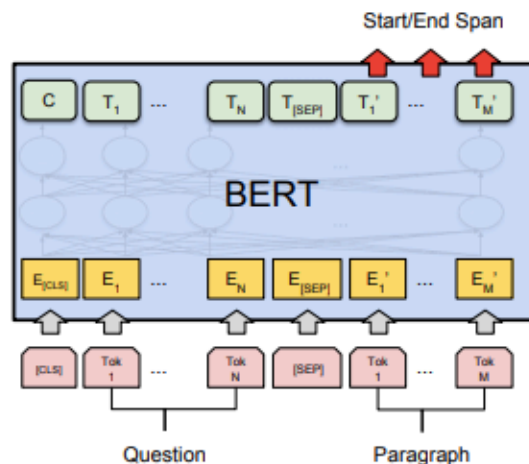
# Fine-tuning BERT for other specific tasks



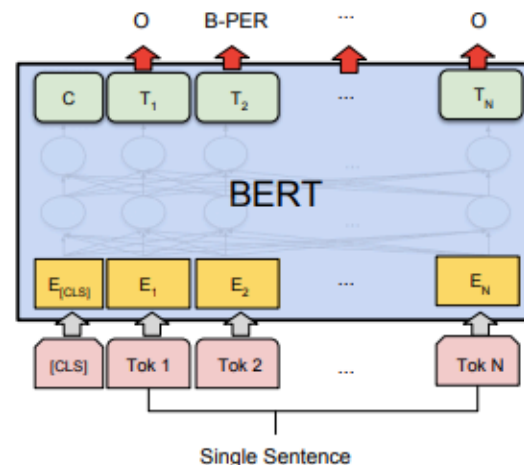
(a) Sentence Pair Classification Tasks:  
MNLi, QQP, QNLI, STS-B, MRPC,  
RTE, SWAG



(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.

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

Met my first girlfriend that way.	FACE-TO-FACE <b>contradiction</b> C C N C	I didn't meet my first girlfriend until later.
8 million in relief in the form of emergency housing.	GOVERNMENT <b>neutral</b> 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 <b>neutral</b> 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 <b>entailment</b> E E E E	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE <b>neutral</b> 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 <b>contradiction</b> 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.

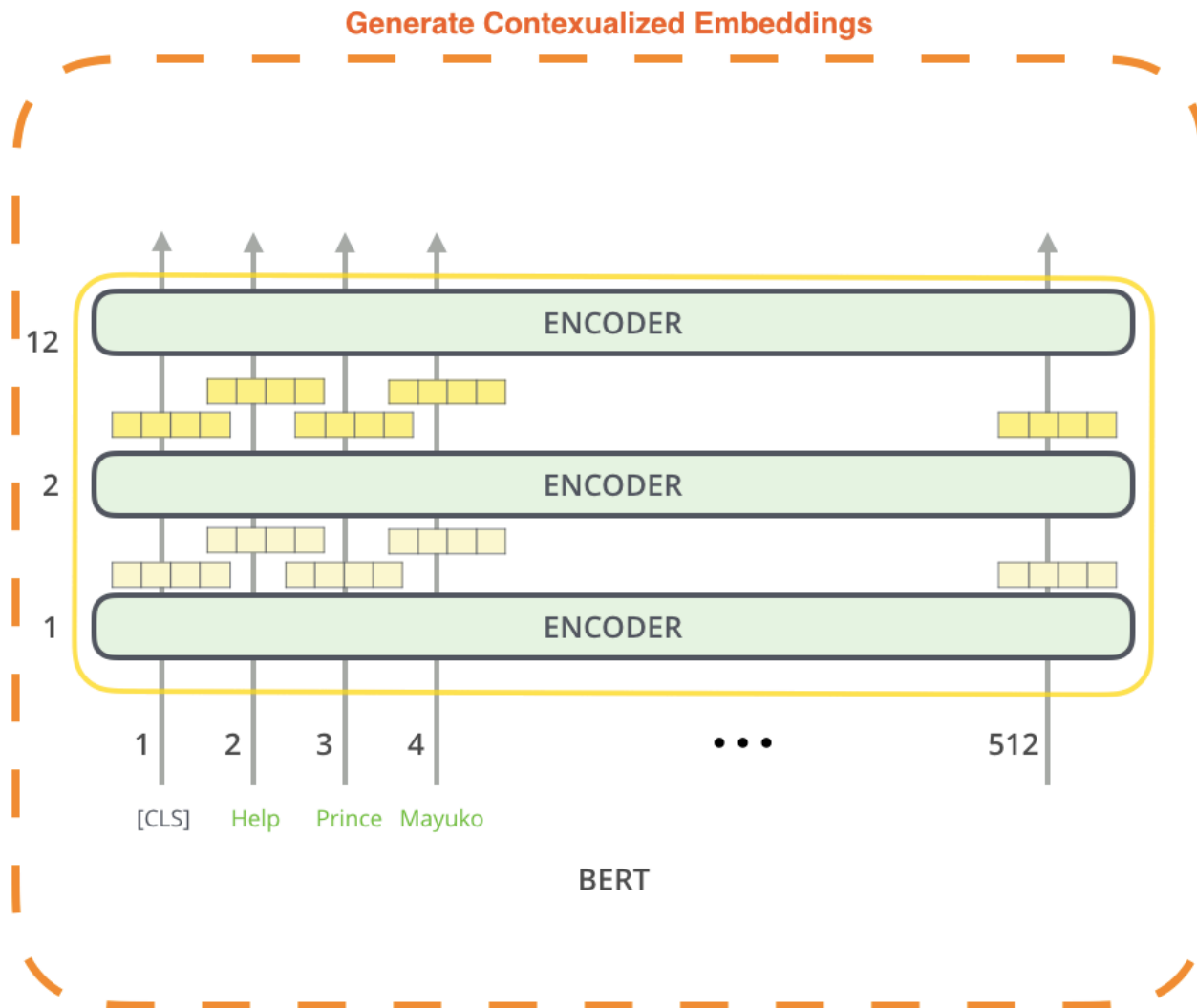
Which NFL team represented the AFC at Super Bowl 50?

Ground Truth Answers: Denver Broncos

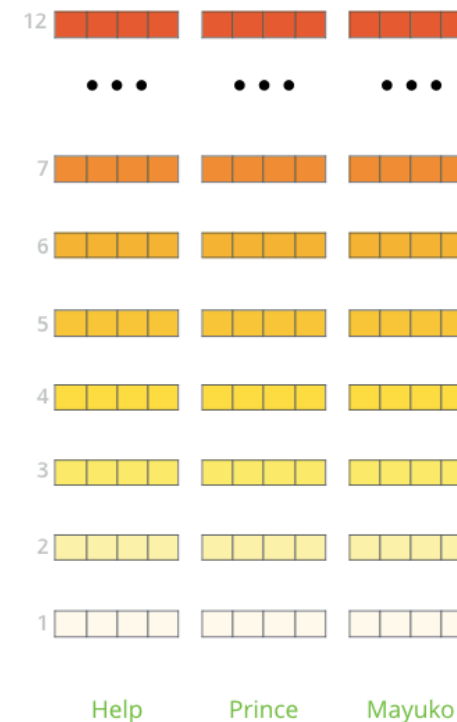
Which NFL team represented the NFC at Super Bowl 50?

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?

What is the best contextualized embedding for “Help” in that context?  
For named-entity recognition task CoNLL-2003 NER

		Dev F1 Score	
12	First Layer Embedding	91.0	
...	Last Hidden Layer	94.9	
7	Sum All 12 Layers	95.5	
6			12
5			+
4			...
3			+
2	2	95.9	
1	1		
	=		
	Second-to-Last Hidden Layer	95.6	
	Sum Last Four Hidden	95.9	
			12
			+
			11
			+
	10	96.1	
	+		
	9		
	=		
	Concat Last Four Hidden		
	9 10 11 12		

Help



## 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. Howear, S. Ruder., 2018

Jay Alamm, The illustrated GPT-2, <https://jalammar.github.io/illustrated-gpt2/>