Natural Language Processing

Info 159/259

Lecture 9: BERT, Prompting methods, RLHF (Feb 14, 2024)

Many slides & instruction ideas borrowed from:

David Bamman, Mohit Iyyer & Greg Durrett

Logistics

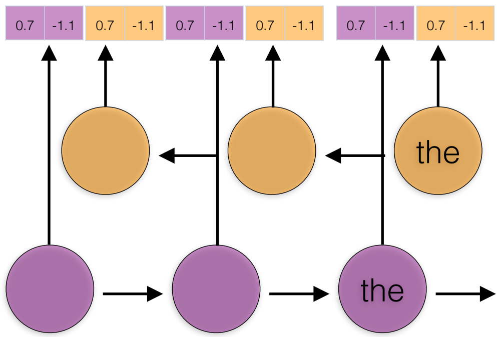
- AP0 is due this Friday Feb 16
- Exam 1 is next Wednesday Feb 21
- Homework 3 will be out soon
  - Will be due Thursday Feb 22.
- Quiz 4 will be out tomorrow **(due next Sunday Feb 18)**
- Today: Large Language Models

Contextualized word representations

- Big idea: transform the representation of a token in a sentence (e.g., from a static word embedding) to be sensitive to its local context in a sentence and trainable to be optimized for a specific NLP task.

ELMo

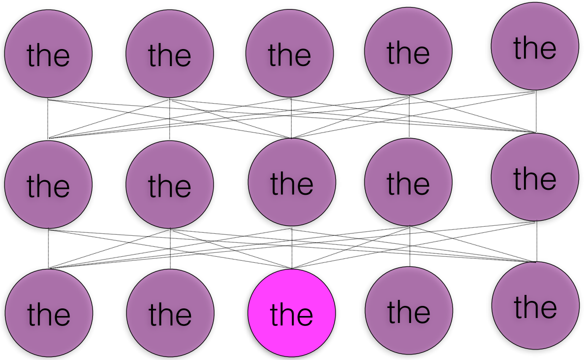
Stacked BiRNN trained to predict next word in language modeling task



Peters et al. 2018

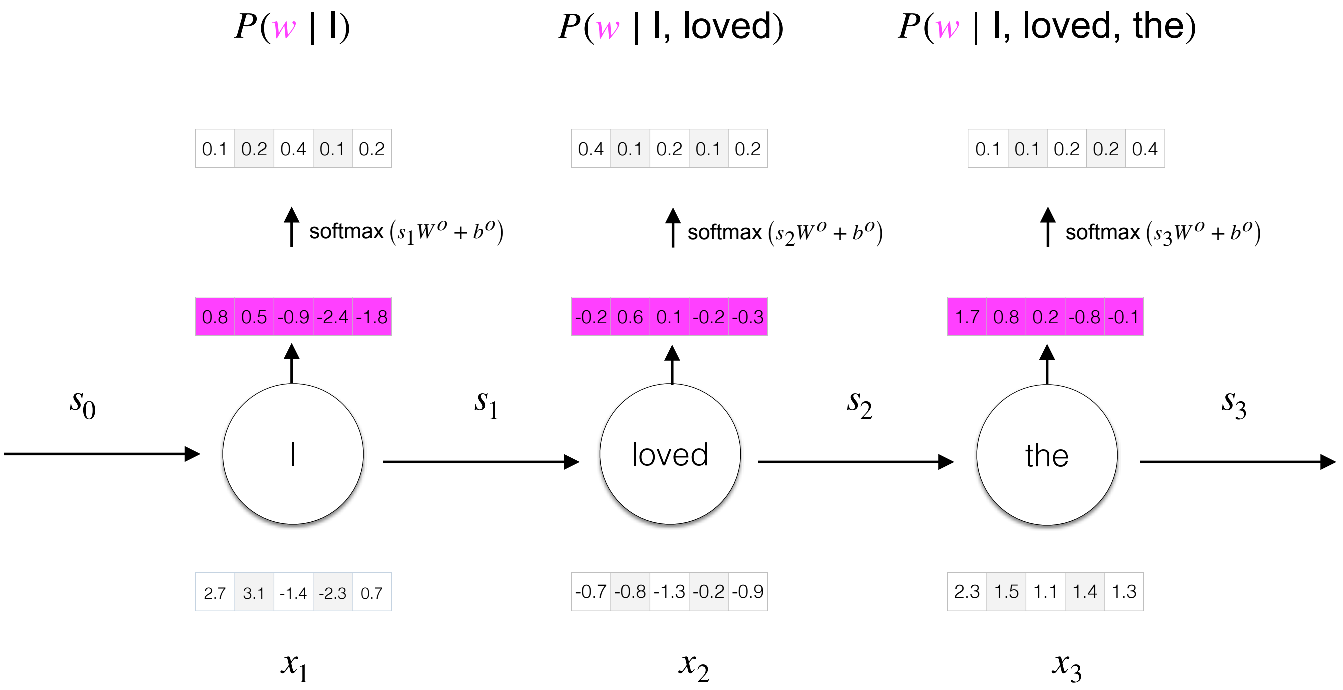
BERT

Transformer-based model to predict masked word using bidirectional context + next sentence prediction.



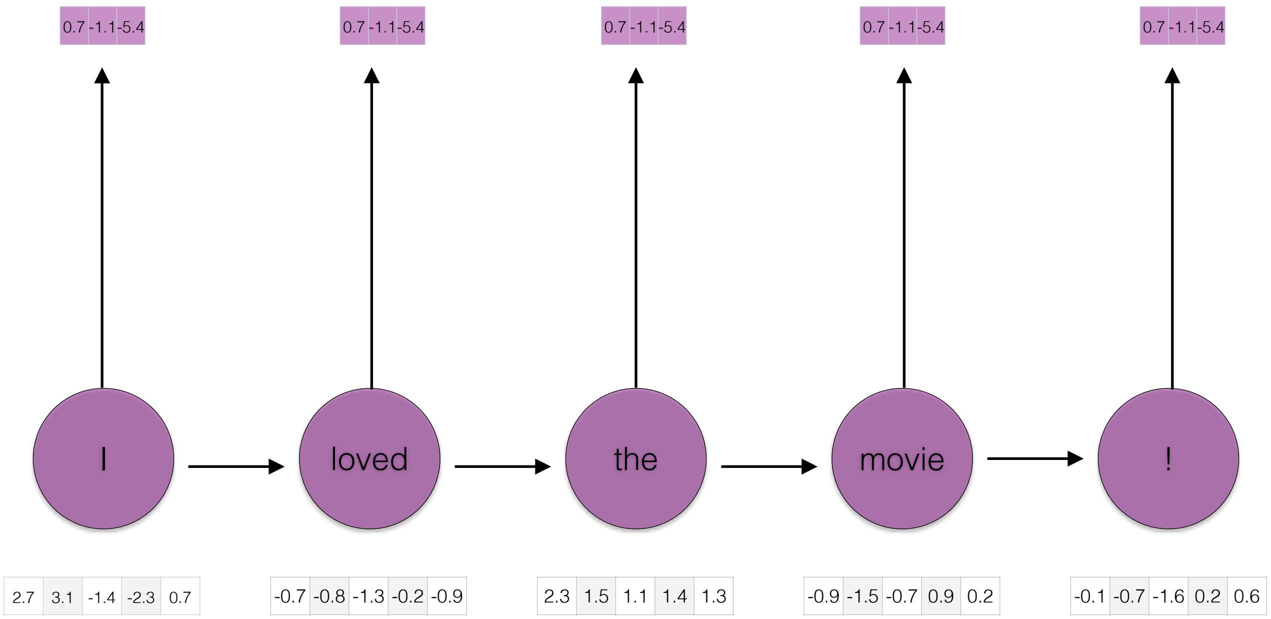
Devlin et al. 2019

RNN Language model



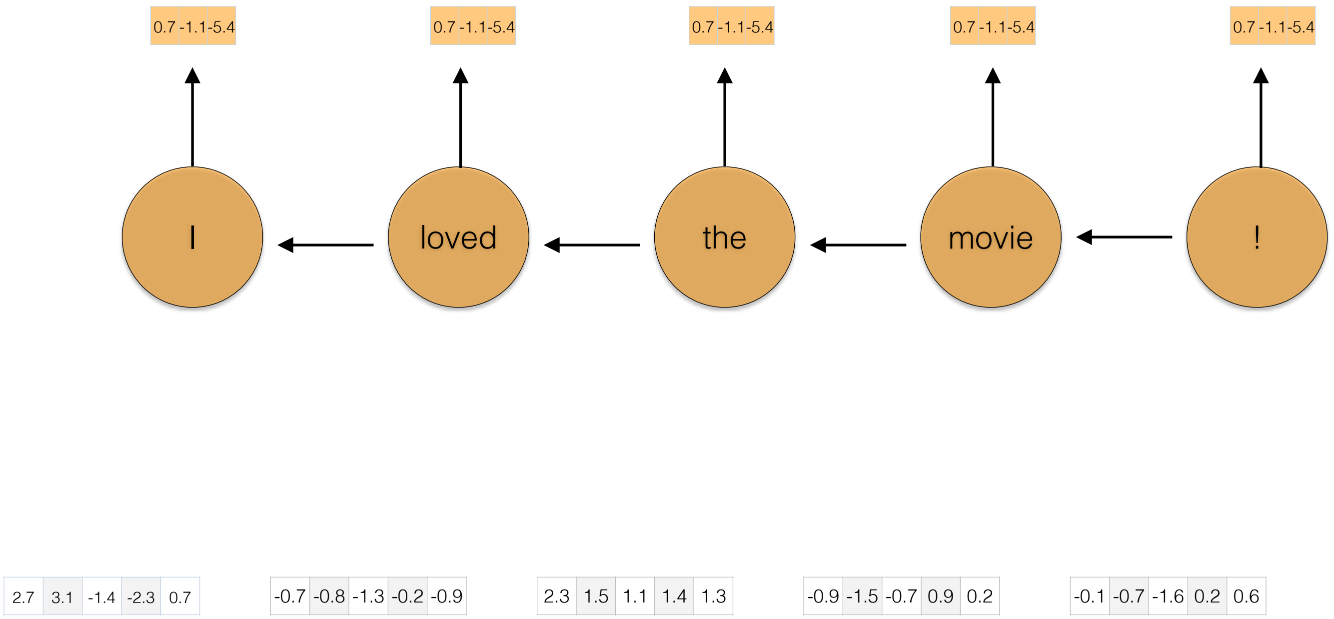
Bidirectional RNN

forward RNN

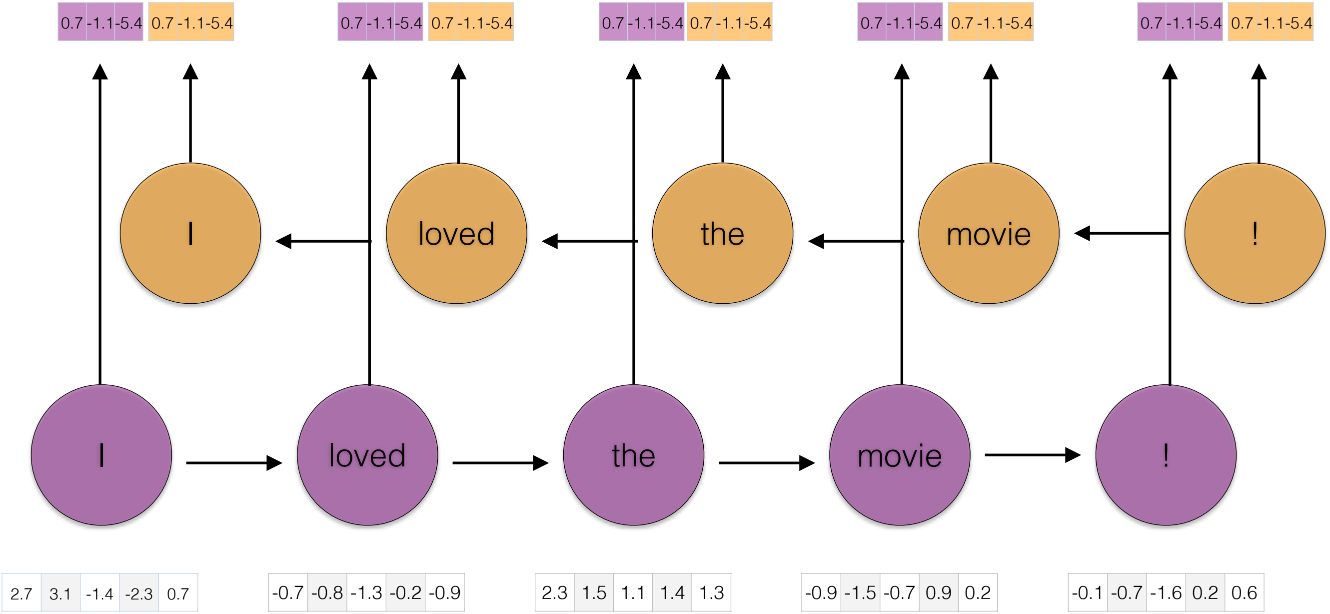


Bidirectional RNN

backward RNN



Bidirectional RNN



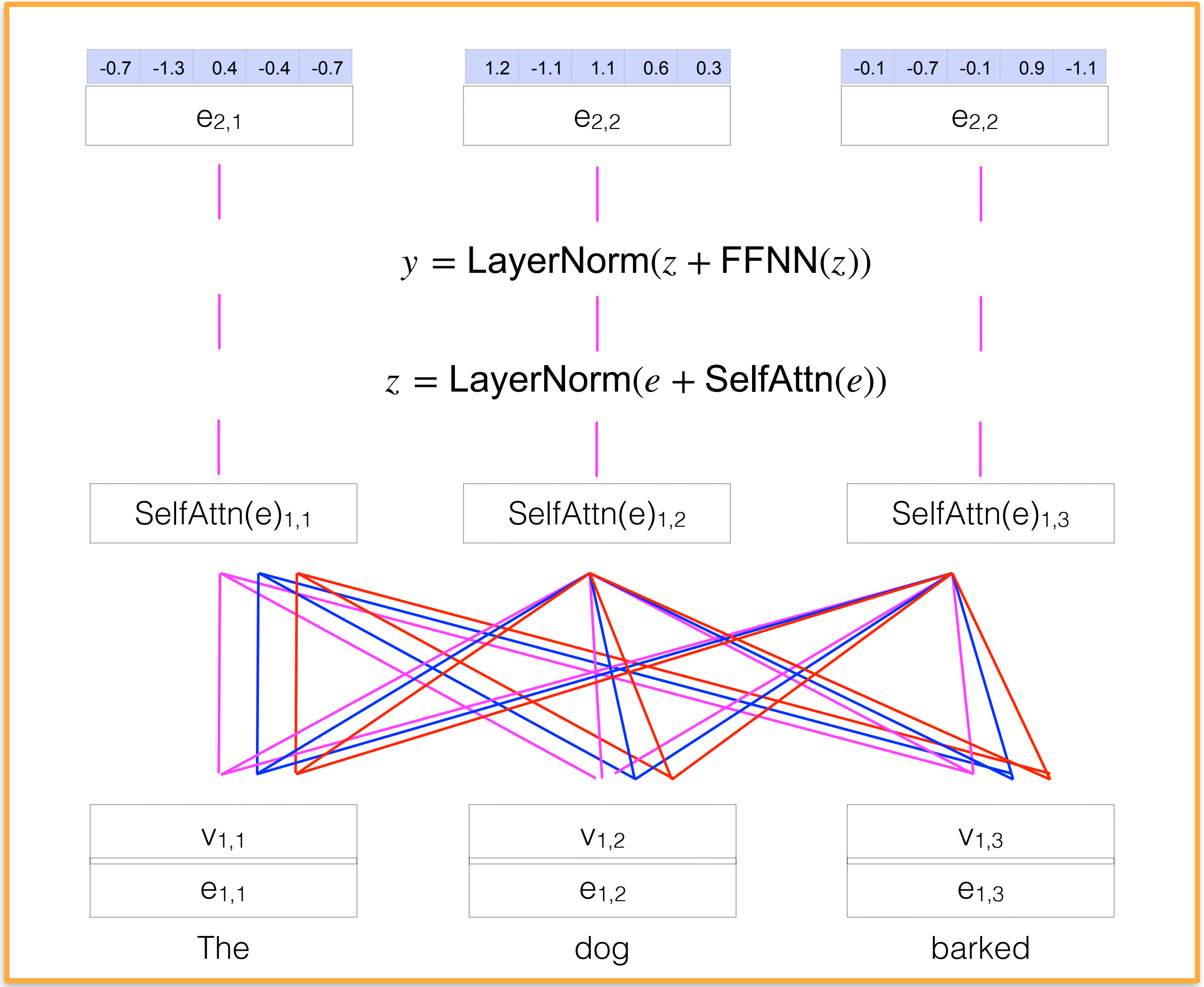
BERT

- Transformer-based model (Vaswani et al. 2017) to predict masked word using bidirectional context + next sentence prediction.
- Generates multiple layers of representations for each token sensitive to its context of use.

Output

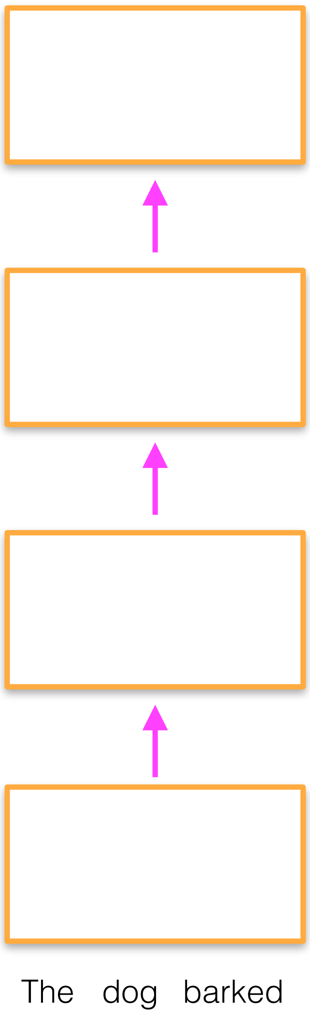
This whole process defines one attention block. The input is a sequence of (e.g. 100-dimensional) vectors; the output of each block is a sequence of (100-dimensional) vectors.

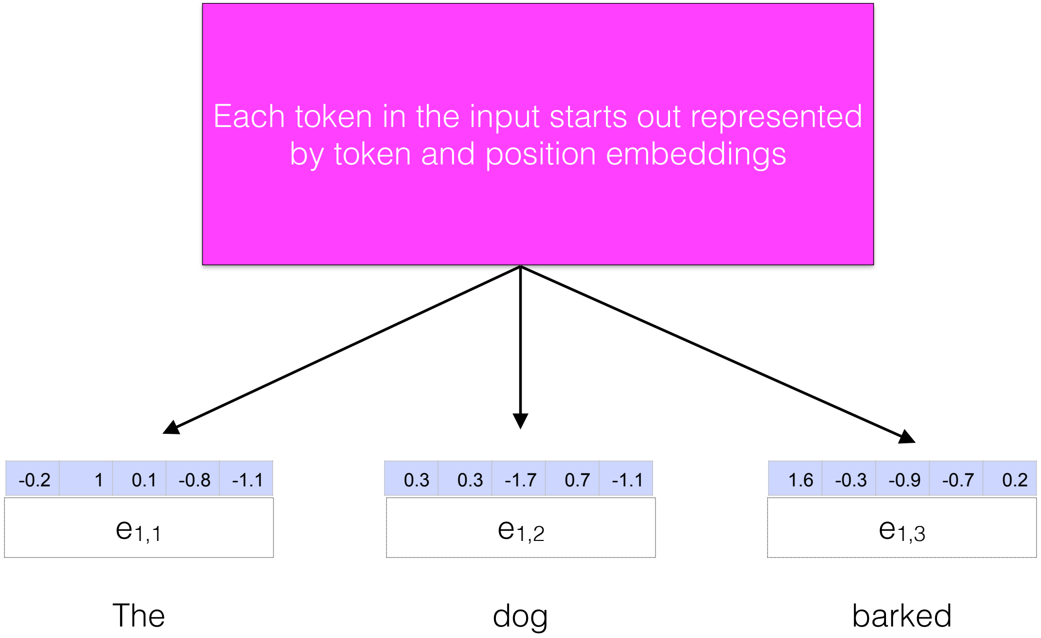
Input



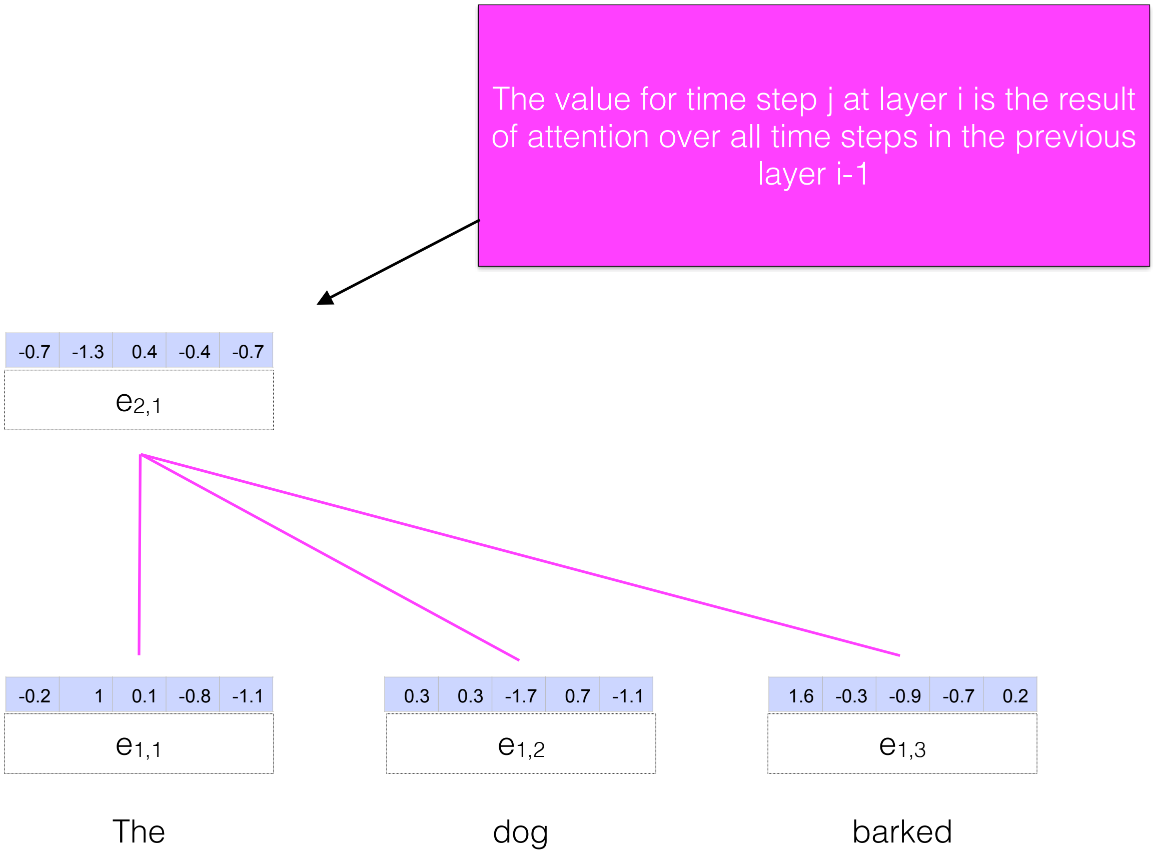
This whole process defines one attention block. The input is a sequence of (e.g. 100-dimensional) vectors; the output of each block is a sequence of (100-dimensional) vectors.

Transformers can stack many such blocks; where the output from block b is the input to block b+1.

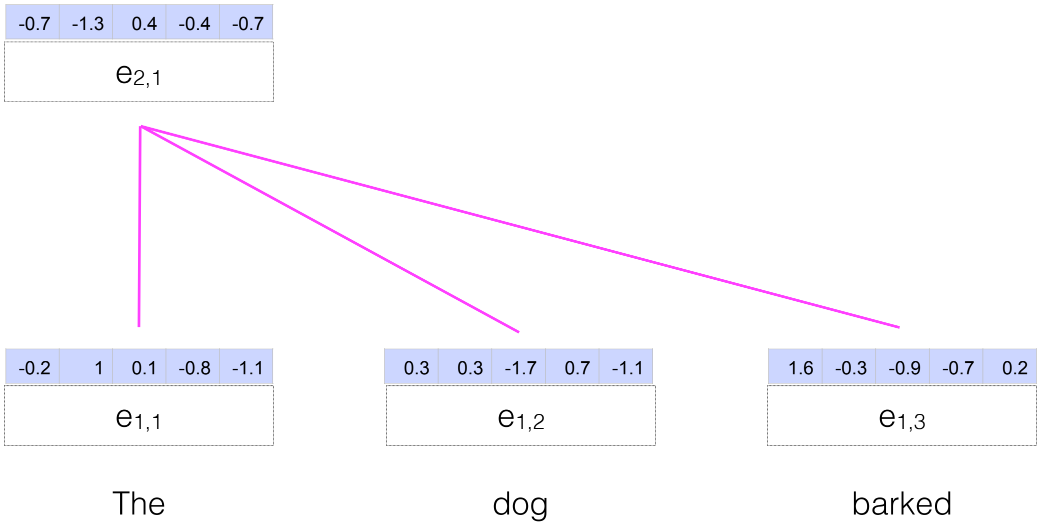


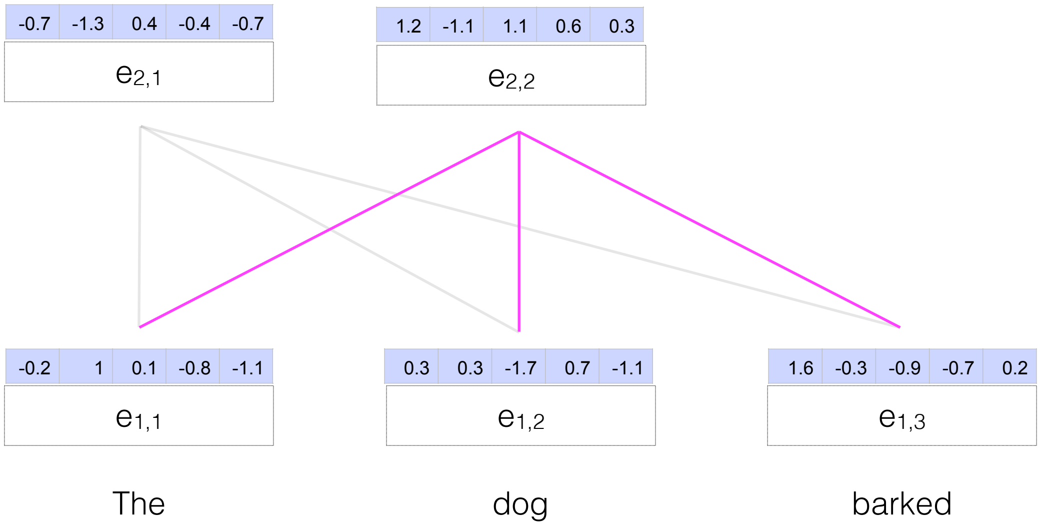


Each token in the input starts out represented by token and position embeddings
-0.2 1 0.1 -0.8 -1.1
$e_{1,1}$
The
0.3 0.3 -1.7 0.7 -1.1
$e_{1,2}$
dog
1.6 -0.3 -0.9 -0.7 0.2
$e_{1,3}$
barked

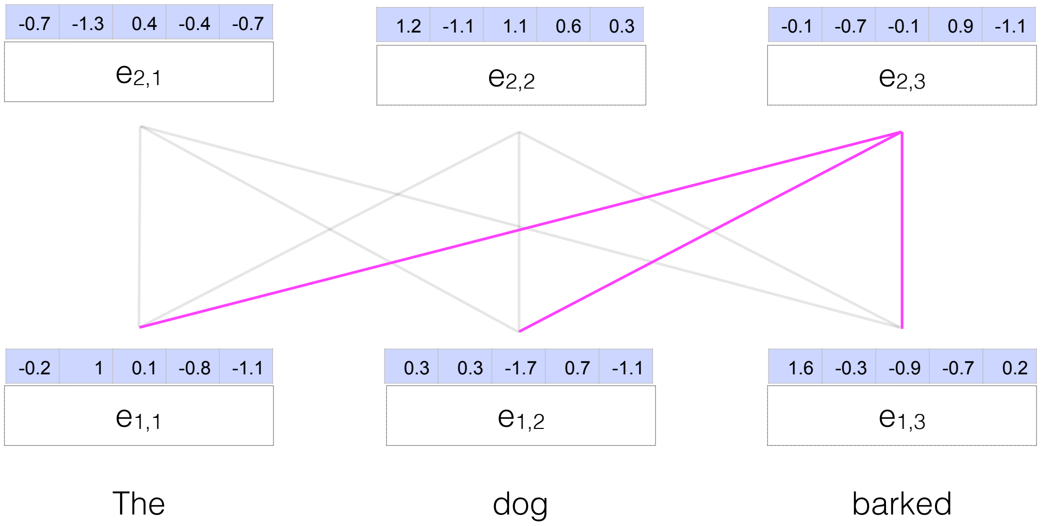


The value for time step j at layer i is the result of attention over all time steps in the previous layer i-1
-0.7 -1.3 0.4 -0.4 -0.7
$e_{2,1}$
-0.2 1 0.1 -0.8 -1.1
$e_{1,1}$
0.3 0.3 -1.7 0.7 -1.1
$e_{1,2}$
1.6 -0.3 -0.9 -0.7 0.2
$e_{1,3}$
The
dog
barked

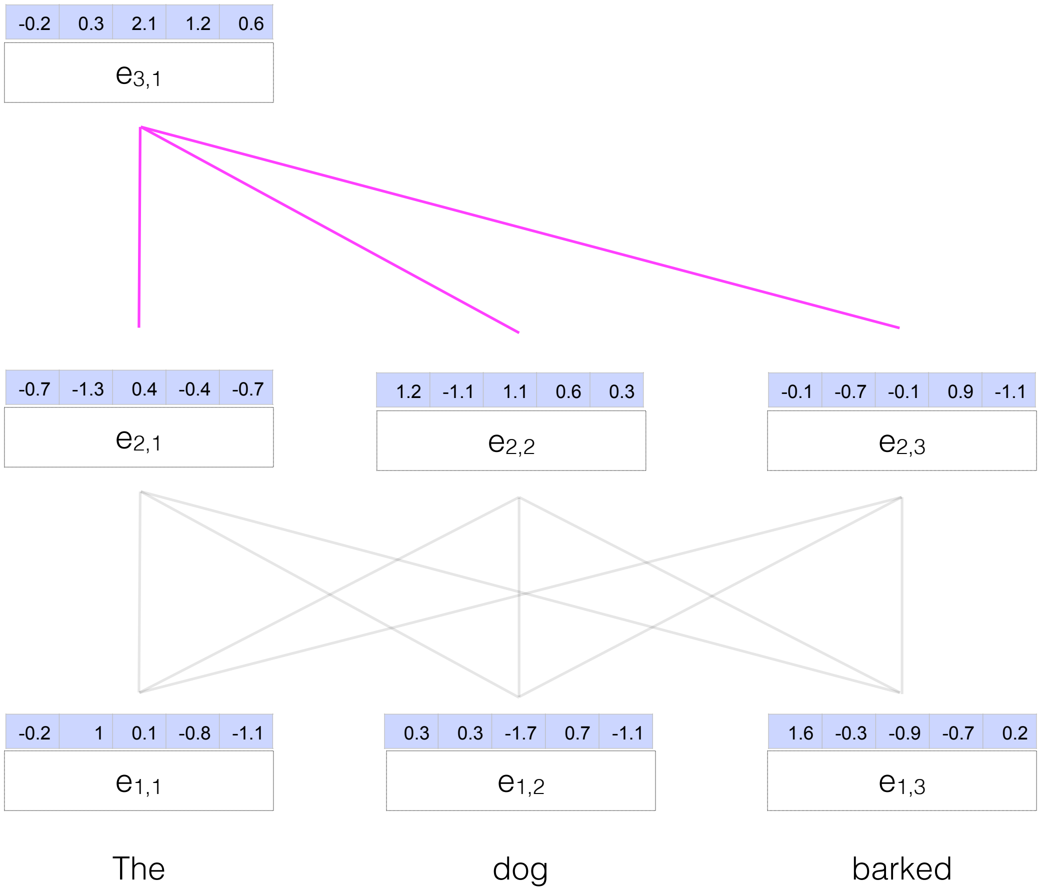

-0.7 -1.3 0.4 -0.4 -0.7
e2,1
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked



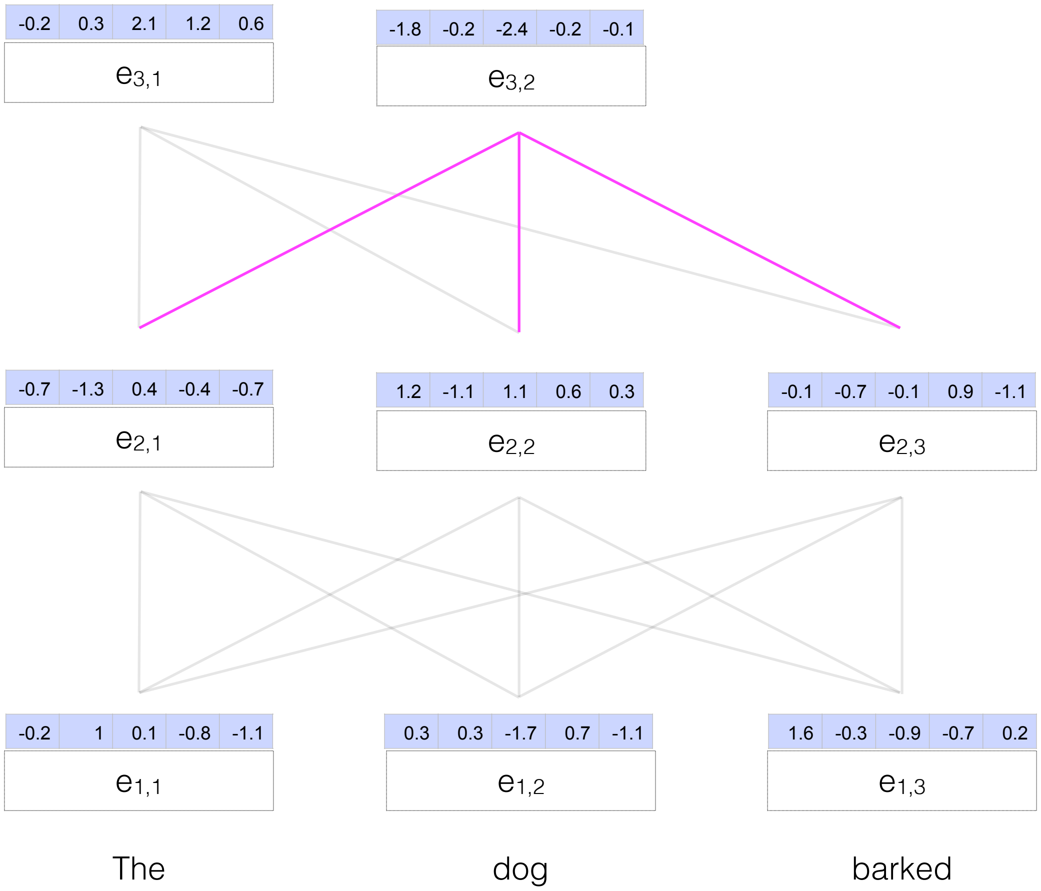
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked



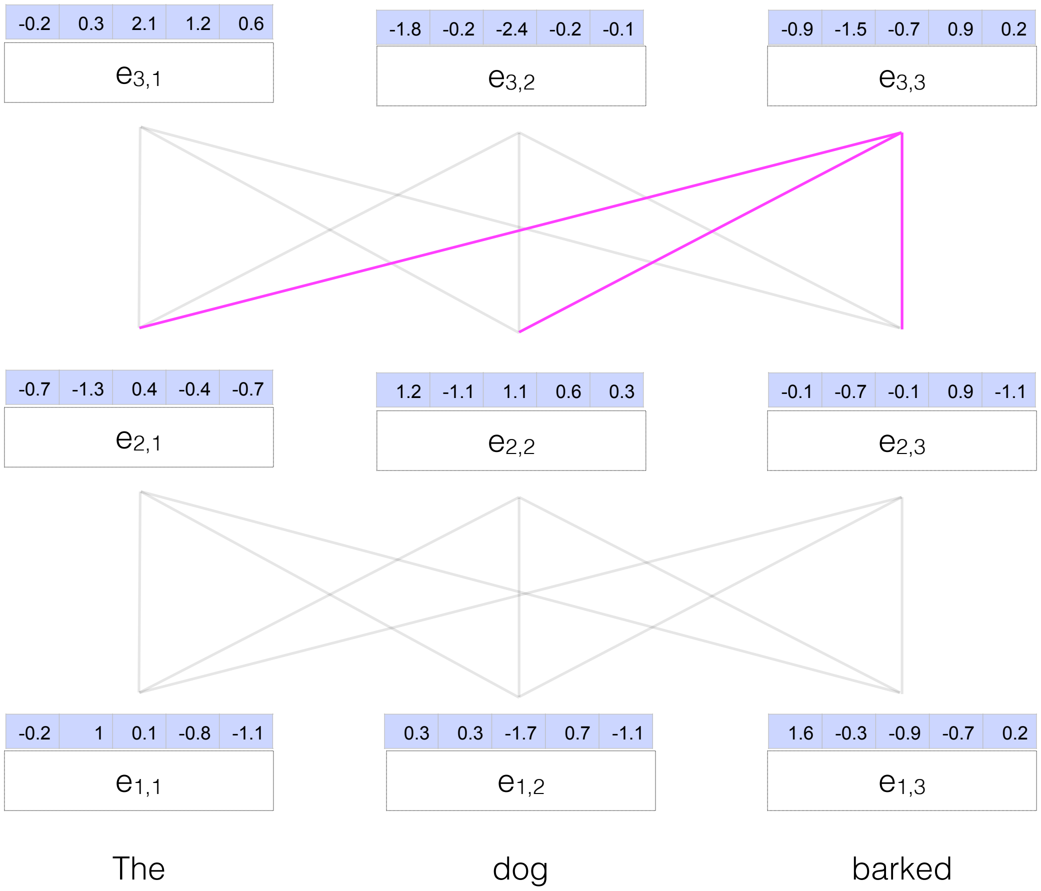
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.1 -0.7 -0.1 0.9 -1.1
e2,3
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked



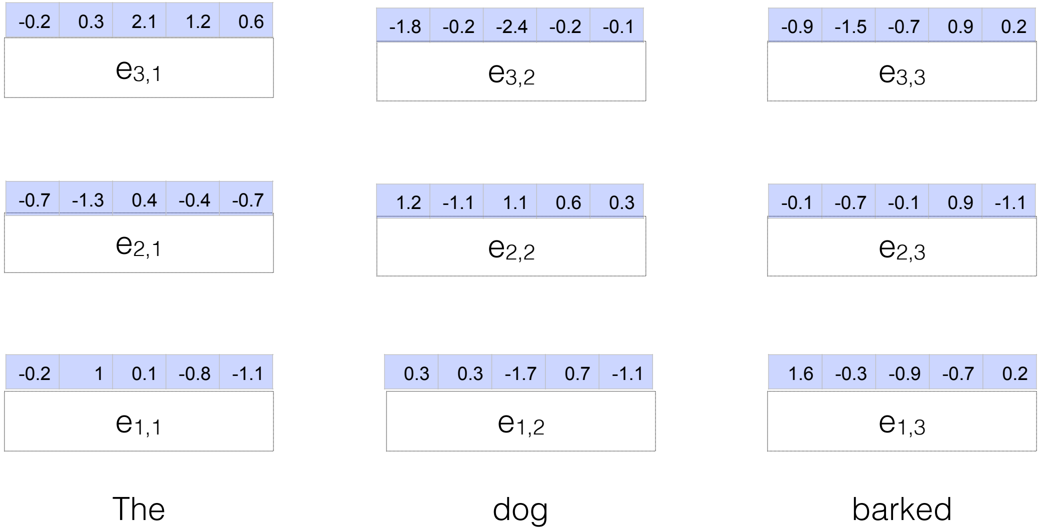
-0.2 0.3 2.1 1.2 0.6
$e_{3,1}$
-0.7 -1.3 0.4 -0.4 -0.7
$e_{2,1}$
1.2 -1.1 1.1 0.6 0.3
$e_{2,2}$
-0.1 -0.7 -0.1 0.9 -1.1
$e_{2,3}$
-0.2 1 0.1 -0.8 -1.1
$e_{1,1}$
0.3 0.3 -1.7 0.7 -1.1
$e_{1,2}$
1.6 -0.3 -0.9 -0.7 0.2
$e_{1,3}$
The
dog
barked



-0.2 0.3 2.1 1.2 0.6
$e_{3,1}$
-1.8 -0.2 -2.4 -0.2 -0.1
$e_{3,2}$
-0.7 -1.3 0.4 -0.4 -0.7
$e_{2,1}$
1.2 -1.1 1.1 0.6 0.3
$e_{2,2}$
-0.1 -0.7 -0.1 0.9 -1.1
$e_{2,3}$
-0.2 1 0.1 -0.8 -1.1
$e_{1,1}$
0.3 0.3 -1.7 0.7 -1.1
$e_{1,2}$
1.6 -0.3 -0.9 -0.7 0.2
$e_{1,3}$
The
dog
barked


-0.2 0.3 2.1 1.2 0.6
e3,1
-1.8 -0.2 -2.4 -0.2 -0.1
e3,2
-0.9 -1.5 -0.7 0.9 0.2
e3,3
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.1 -0.7 -0.1 0.9 -1.1
e2,3
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked

At the end of this process, we have one representation for each layer for each token



WordPiece

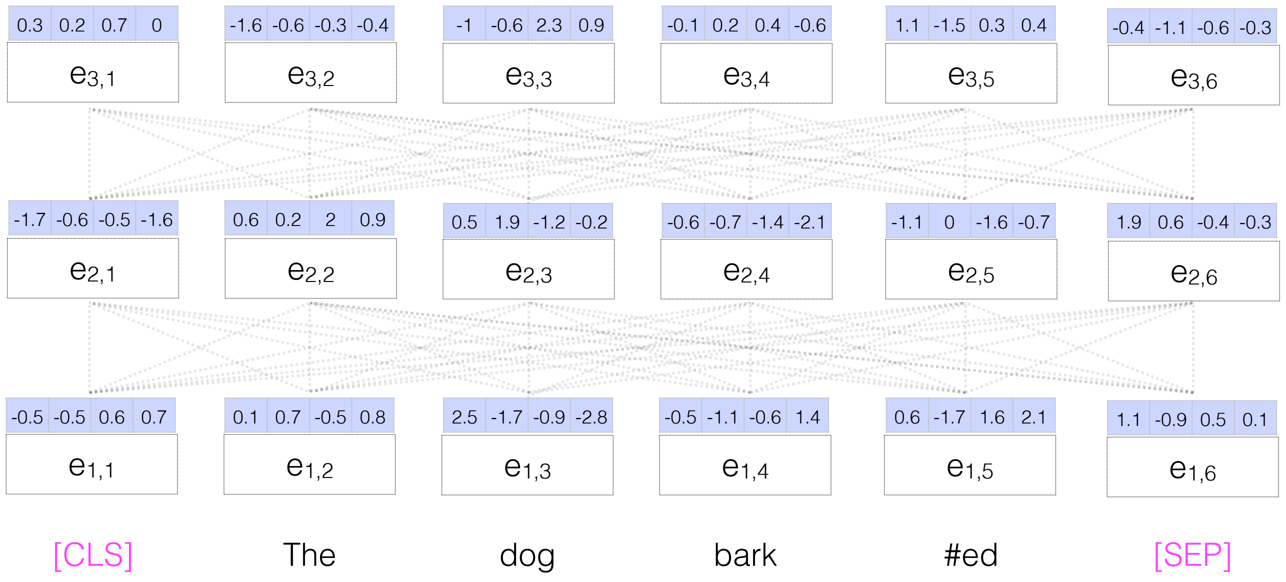
- BERT uses WordPiece tokenization, which segments some morphological structure of tokens

- Vocabulary size: 30,000

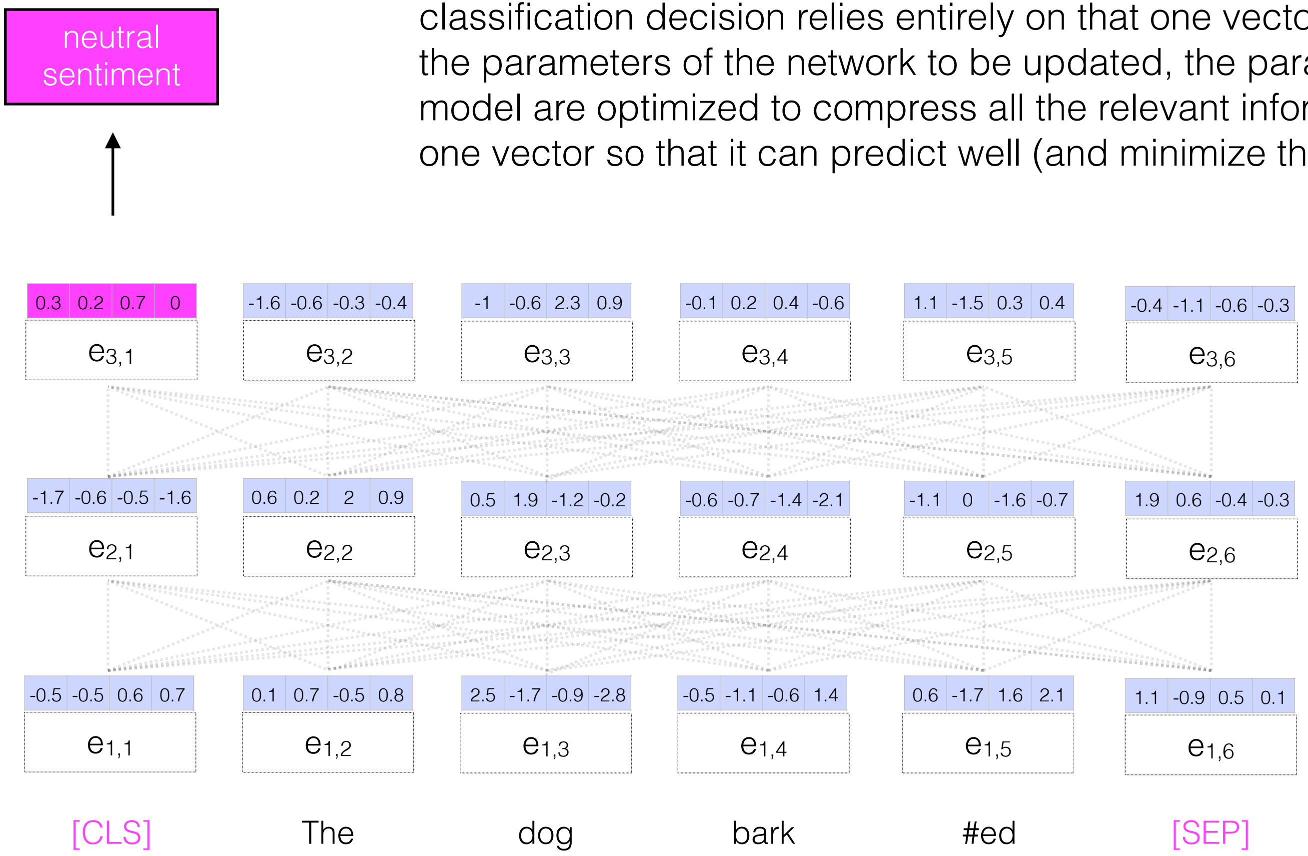
The	The
dog	dog
barked	bark #ed

- BERT also encodes each sentence by appending a special token to the beginning ([CLS]) and end ([SEP]) of each sequence.

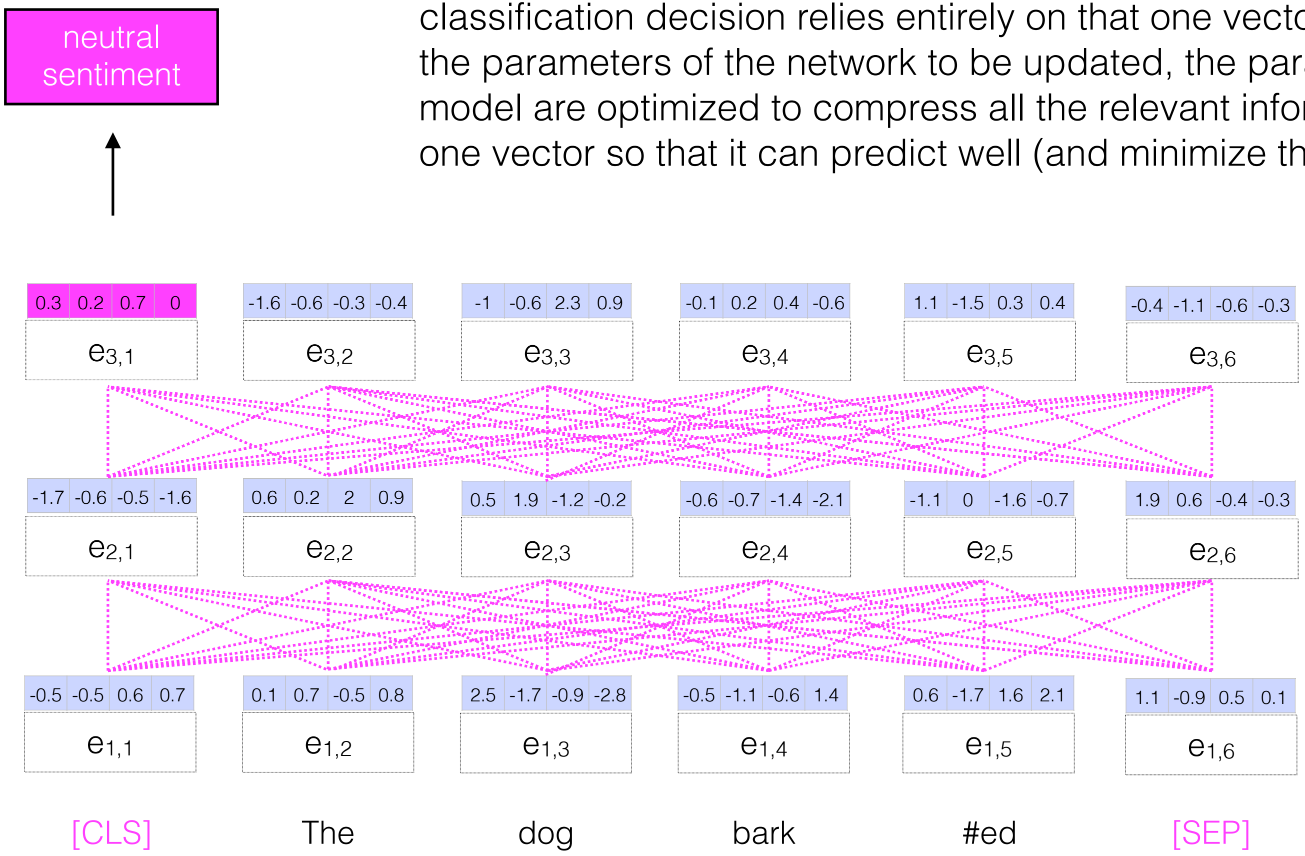
- This helps provides a single token that can be optimized to represent the entire sequence (e.g., for document classification)



- We can represent the entire document with this *one* [CLS] vector
- Why does this work? When we design our network so that a classification decision relies entirely on that one vector and allow all the parameters of the network to be updated, the parameters of the model are optimized to compress all the relevant information into that one vector so that it can predict well (and minimize the loss).



- We can represent the entire document with this *one* [CLS] vector
- Why does this work? When we design our network so that a classification decision relies entirely on that one vector and allow all the parameters of the network to be updated, the parameters of the model are optimized to compress all the relevant information into that one vector so that it can predict well (and minimize the loss).

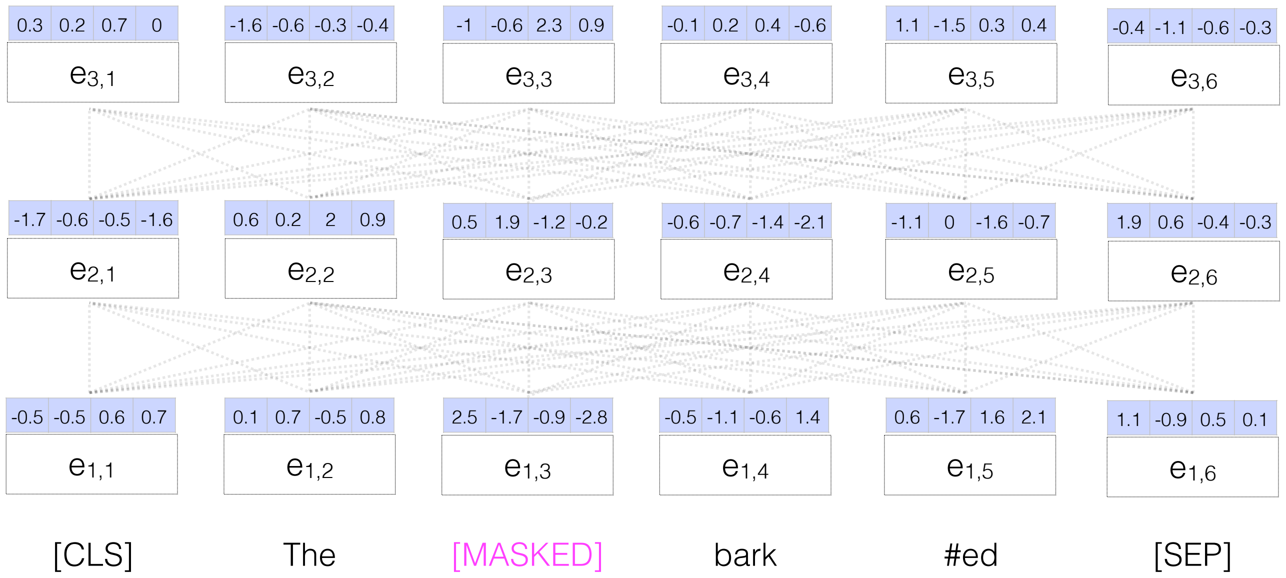


BERT

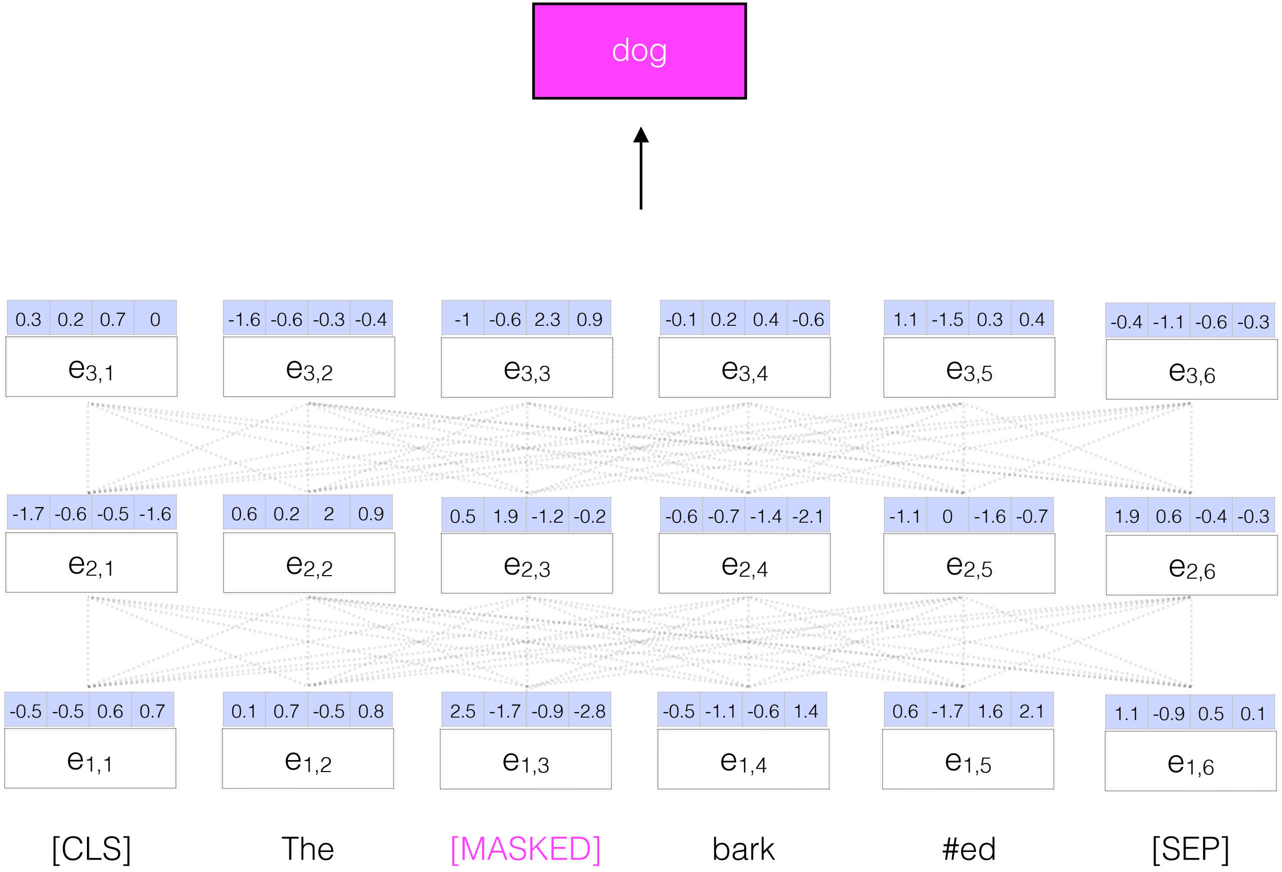
- Learn the parameters of this model with two objectives:
  - Masked language modeling
  - Next sentence prediction

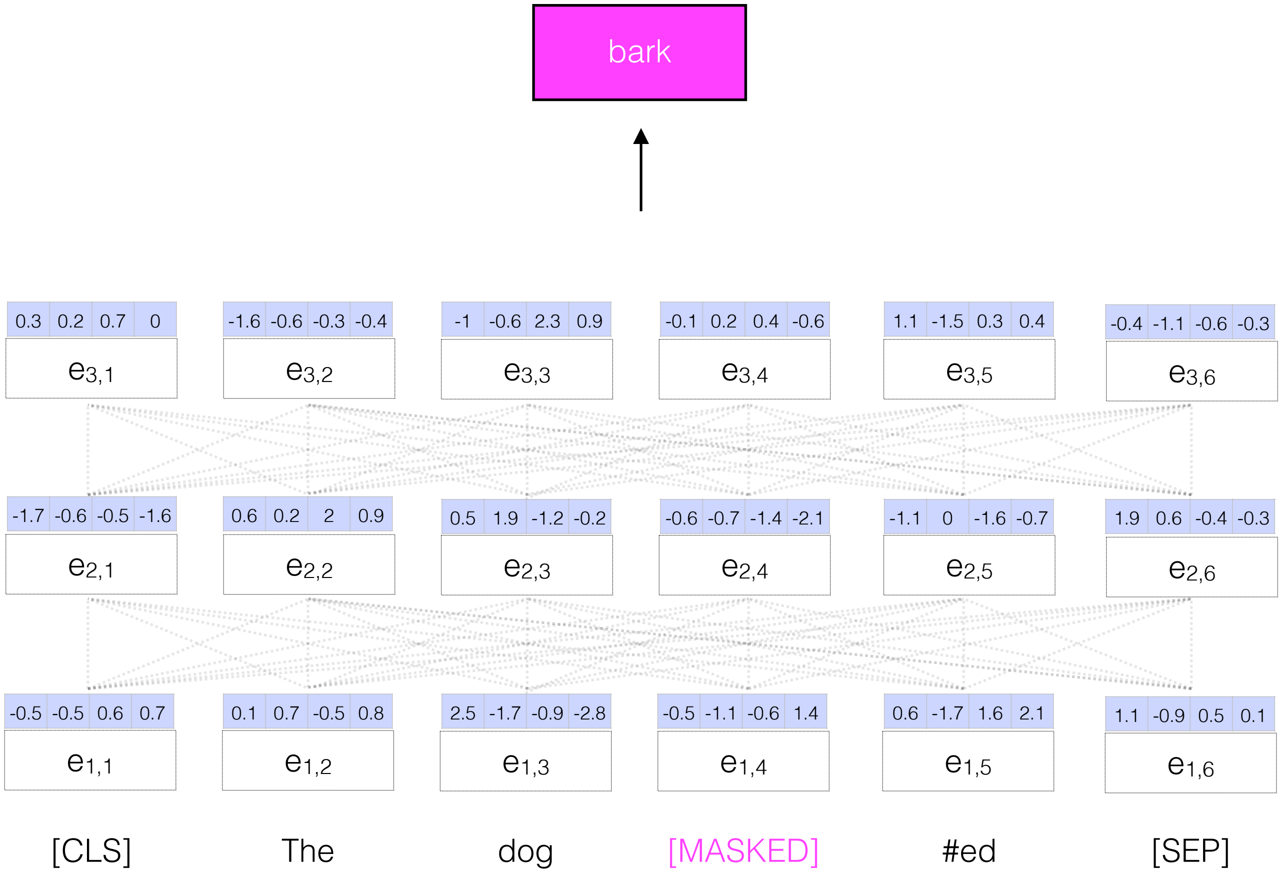
Masked LM

- Mask one word from the input and try to predict that word as the output
- More powerful than an RNN LM (or even a BiRNN LM) since it can reason about context on both sides of the word being predicted.
- A BiRNN models context on both sides, but each RNN only has access to information from one direction.



0.3 0.2 0.7 0
e3,1
-1.6 -0.6 -0.3 -0.4
e3,2
-1 -0.6 2.3 0.9
e3,3
-0.1 0.2 0.4 -0.6
e3,4
1.1 -1.5 0.3 0.4
e3,5
-0.4 -1.1 -0.6 -0.3
e3,6
-1.7 -0.6 -0.5 -1.6
e2,1
0.6 0.2 2 0.9
e2,2
0.5 1.9 -1.2 -0.2
e2,3
-0.6 -0.7 -1.4 -2.1
e2,4
-1.1 0 -1.6 -0.7
e2,5
1.9 0.6 -0.4 -0.3
e2,6
-0.5 -0.5 0.6 0.7
e1,1
0.1 0.7 -0.5 0.8
e1,2
2.5 -1.7 -0.9 -2.8
e1,3
-0.5 -1.1 -0.6 1.4
e1,4
0.6 -1.7 1.6 2.1
e1,5
1.1 -0.9 0.5 0.1
e1,6
[CLS]
The
[MASKED]
bark
#ed
[SEP]


dog
0.3 0.2 0.7 0
-1.6 -0.6 -0.3 -0.4
-1 -0.6 2.3 0.9
-0.1 0.2 0.4 -0.6
1.1 -1.5 0.3 0.4
-0.4 -1.1 -0.6 -0.3
$e_{3,1}$ $e_{3,2}$ $e_{3,3}$ $e_{3,4}$ $e_{3,5}$ $e_{3,6}$
-1.7 -0.6 -0.5 -1.6
0.6 0.2 2 0.9
0.5 1.9 -1.2 -0.2
-0.6 -0.7 -1.4 -2.1
-1.1 0 -1.6 -0.7
1.9 0.6 -0.4 -0.3
$e_{2,1}$ $e_{2,2}$ $e_{2,3}$ $e_{2,4}$ $e_{2,5}$ $e_{2,6}$
-0.5 -0.5 0.6 0.7
0.1 0.7 -0.5 0.8
2.5 -1.7 -0.9 -2.8
-0.5 -1.1 -0.6 1.4
0.6 -1.7 1.6 2.1
1.1 -0.9 0.5 0.1
$e_{1,1}$ $e_{1,2}$ $e_{1,3}$ $e_{1,4}$ $e_{1,5}$ $e_{1,6}$
[CLS] The [MASKED] bark #ed [SEP]


bark
0.3 0.2 0.7 0
e3,1
-1.6 -0.6 -0.3 -0.4
e3,2
-1 -0.6 2.3 0.9
e3,3
-0.1 0.2 0.4 -0.6
e3,4
1.1 -1.5 0.3 0.4
e3,5
-0.4 -1.1 -0.6 -0.3
e3,6
-1.7 -0.6 -0.5 -1.6
e2,1
0.6 0.2 2 0.9
e2,2
0.5 1.9 -1.2 -0.2
e2,3
-0.6 -0.7 -1.4 -2.1
e2,4
-1.1 0 -1.6 -0.7
e2,5
1.9 0.6 -0.4 -0.3
e2,6
-0.5 -0.5 0.6 0.7
e1,1
0.1 0.7 -0.5 0.8
e1,2
2.5 -1.7 -0.9 -2.8
e1,3
-0.5 -1.1 -0.6 1.4
e1,4
0.6 -1.7 1.6 2.1
e1,5
1.1 -0.9 0.5 0.1
e1,6
[CLS]
The
dog
[MASKED]
#ed
[SEP]

Next sentence prediction

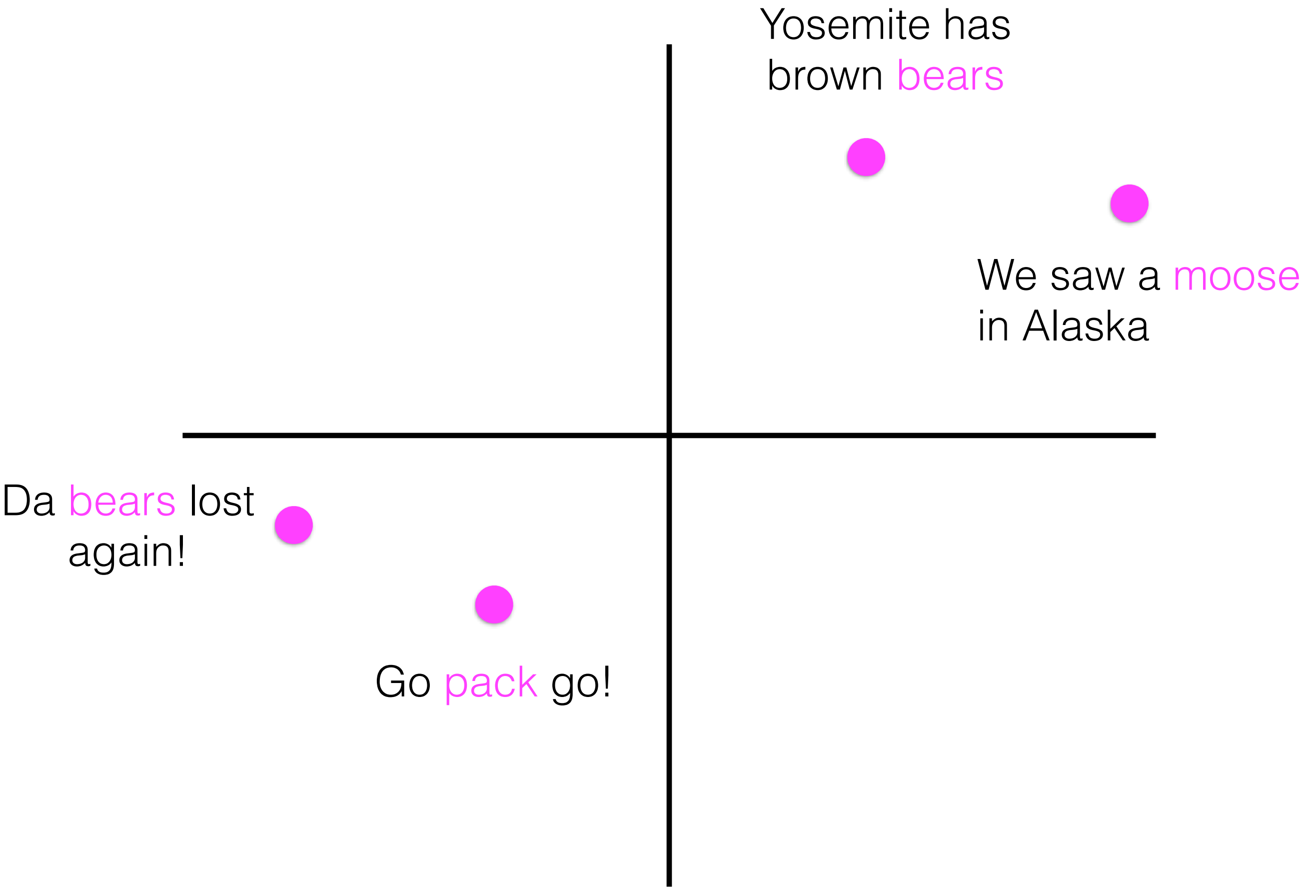
- For a pair of sentences, predict from [CLS] representation whether they appeared sequentially in the training data:

  + [CLS] The dog bark #ed [SEP] He was hungry

  − [CLS] The dog bark #ed [SEP] Paris is in France

BERT

- Deep layers (12 for BERT base, 24 for BERT large)
- Large representation sizes (768 per layer)
- Pretrained on English Wikipedia (2.5B words) and BooksCorpus (800M words)


Yosemite has brown bears
We saw a moose in Alaska
Da bears lost again!
Go pack go!

BERT

	H=128	H=256	H=512	H=768
L=2	**2/128 (BERT-Tiny)**	2/256	2/512	2/768
L=4	4/128	**4/256 (BERT-Mini)**	**4/512 (BERT-Small)**	4/768
L=6	6/128	6/256	6/512	6/768
L=8	8/128	8/256	**8/512 (BERT-Medium)**	8/768
L=10	10/128	10/256	10/512	10/768
L=12	12/128	12/256	12/512	**12/768 (BERT-Base)**

https://github.com/google-research/bert

SIGN IN MODELS FORUM

v4.11.3 ▼

transformers

Star 52,449

Search docs

GET STARTED

Quick tour

Installation

Philosophy

Glossary

USING 🤗 TRANSFORMERS

Summary of the tasks

Summary of the models

Preprocessing data

Docs » Pretrained models

View page source

Pretrained models

Here is a partial list of some of the available pretrained models together with a short presentation of each model.

For the full list, refer to https://huggingface.co/models.

Architecture	Model id	Details of the model
	`bert-base-uncased`	12-layer, 768-hidden, 12-heads, 110M parameters. Trained on lower-cased English text.
	`bert-large-uncased`	24-layer, 1024-hidden, 16-heads, 336M parameters. Trained on lower-cased English text.

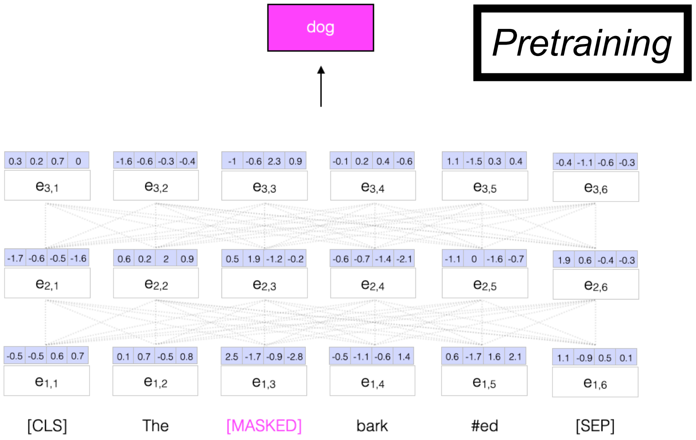
https://huggingface.co/transformers/pretrained_models.html

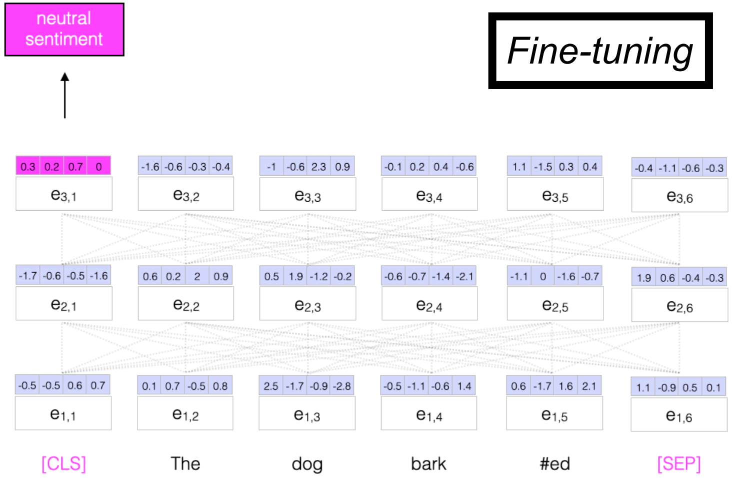
Evolution of the Paradigm

Before 2014	Fully Supervised (feature Engineering)
2014-2019	Architecture Engineering
2019	Pretrain+Finetune: Objective Engineering
….	….

Pretrain + Fine-tune

- The LLM backbone gets trained with its objectives
- The backbone gets fine-tuned for specific task in supervised manner





BERT LANG STREET

Lost in (language-specific) BERT models? We are here to help!

We currently have indexed **31** BERT-based models, **19** Languages and **28** Tasks.

We have a total of **178** entries in this table; we also show **Multilingual Bert (mBERT)** results if available! (see our paper)

Curious which BERT model is the best for named entity recognition in Italian? Just type ***"Italian NER"*** in the search bar!

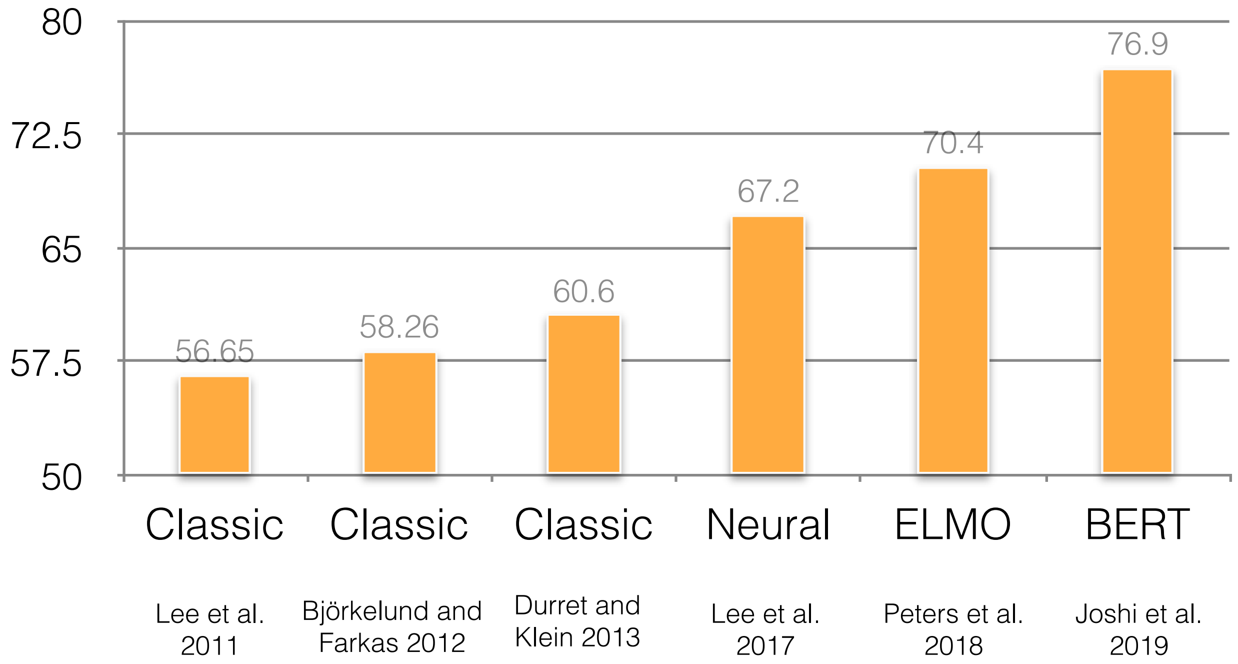
Show 10 entries

Search:

Language	Model	NLP Task	Dataset	Dataset-Domain	Measure	Performance	mBERT	Difference with mBERT	Source
Arabic	Arabert v1	SA	AJGT	twitter	Accuracy	93.8	83.6	10.2	
Arabic	Arabert v1	SA	HARD	hotel reviews	Accuracy	96.1	95.7	0.4	
Arabic	Arabert v1	SA	ASTD	twitter	Accuracy	92.6	80.1	12.5	
Arabic	Arabert v1	SA	ArSenTD-Lev	twitter	Accuracy	59.4	51.0	8.4	

https://bertlang.unibocconi.it

Progress — Coreference resolution



Bertology

- Hewitt et al. 2019
- Tenney et al. 2019
- McCoy et al. 2019
- Liu et al. 2019
- Clark et al. 2019
- Goldberg 2019
- Michel et al. 2019

Code

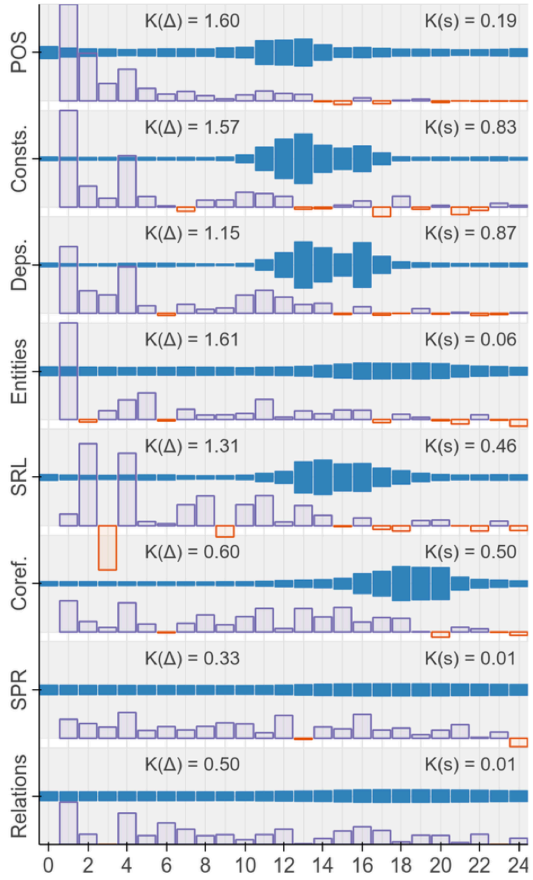
Pre-trained models for BERT, Transformer-XL, ALBERT, RoBERTa, DistilBERT, GPT-2, etc. for English, French, "Multilingual"

https://huggingface.co

Probing

- Even though BERT is mainly trained on a language modeling objective, it learns a lot about the structure of language — even without direct training data for specific linguistic tasks.

- Probing experiments uncover what—-and where (in what layers)—-pretrained BERT encodes this information.



Tenney et al. (2019), "BERT Rediscovers the Classical NLP Pipeline"

Language model

- Language models allow us to calculate the probability of the next word conditioned on some context (and different models make different assumptions about how much of that context is available).

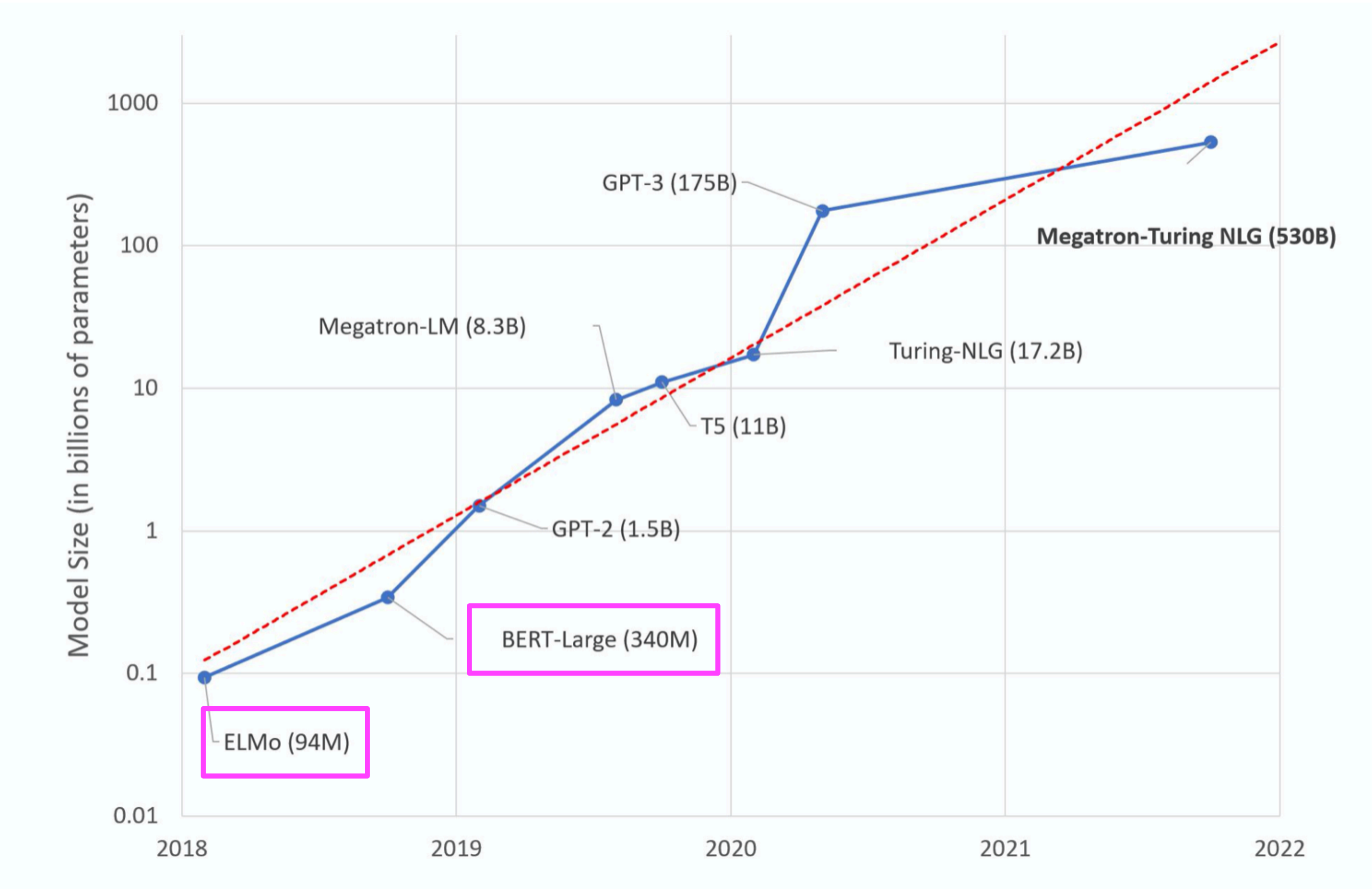
$$P(x_i \mid x_1, \ldots, x_{i-1})$$

- Even BERT can be used this way (by masking out the final word in a sequence)

Generating

- As we sample, the words we generate form the new context we condition on

context1	context2	generated word
START	START	The
START	The	dog
The	dog	walked
dog	walked	in



https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

The importance of being on twitter

by Jerome K. Jerome
London, Summer 1897

It is a curious fact that the last remaining form of social life in which the people of London are still interested is Twitter. I was struck with this curious fact when I went on one of my periodical holidays to the sea-side, and found the whole place twittering like a starling-cage. I called it an anomaly, and it is.

I spoke to the sexton, whose cottage, like all sexton's cottages, is full of antiquities and interesting relics of former centuries. I said to him, "My dear sexton, what does all this twittering mean?" And he replied, "Why, sir, of course it means Twitter." "Ah!" I said, "I know about that. But what is Twitter?"

"It is a system of short and pithy sentences strung together in groups, for the purpose of conveying useful information to the initiated, and entertainment and the exercise of wits to the initiated, and entertainment and the exercise of wits to the rest of us."

https://twitter.com/quasimondo/status/1284509525500989445

Dialogue generation

Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many bonks are in a quoit?
A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?
A: It takes two rainbows to jump from Hawaii to seventeen.

Lacker (2020), "Giving GPT-3 a Turing Test" (https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html); Brown et al. (2020), "Language Models are Few-Shot Learners" (NeurIPS)

LMs as knowledge bases

- Language models can directly encode knowledge present in the training corpus.

The director of *2001: A Space Odyssey* is _____________

LMs as knowledge bases

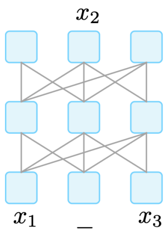
- Language models can directly encode knowledge present in the training corpus.

Query	Answer	Generation
Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8] , **Florence** **[-1.8]** , Naples
Adolphe Adam died in ____.	Paris	**Paris** **[-0.5]** , London [-3.5] , Vienna
English bulldog is a subclass of ____.	dog	dogs [-0.3] , breeds [-2.2] , **dog**
The official language of Mauritius is ____.	English	**English** **[-0.6]** , French [-0.9] , Arabic
Patrick Oboya plays in ____ position.	midfielder	centre [-2.0] , center [-2.2] , **midfielder**
Hamburg Airport is named after ____.	Hamburg	Hess [-7.0] , Hermann [-7.1] , Schmidt

Petroni et al. (2019), "Language Models as Knowledge Bases?" (ACL)

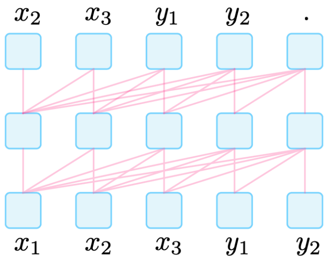
Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first			48.3%
Who is the head			47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%

Radford et al. 2019, “Language Models are Unsupervised Multitask Learners” (GPT-2)



Masked LM
(BERT)

$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$



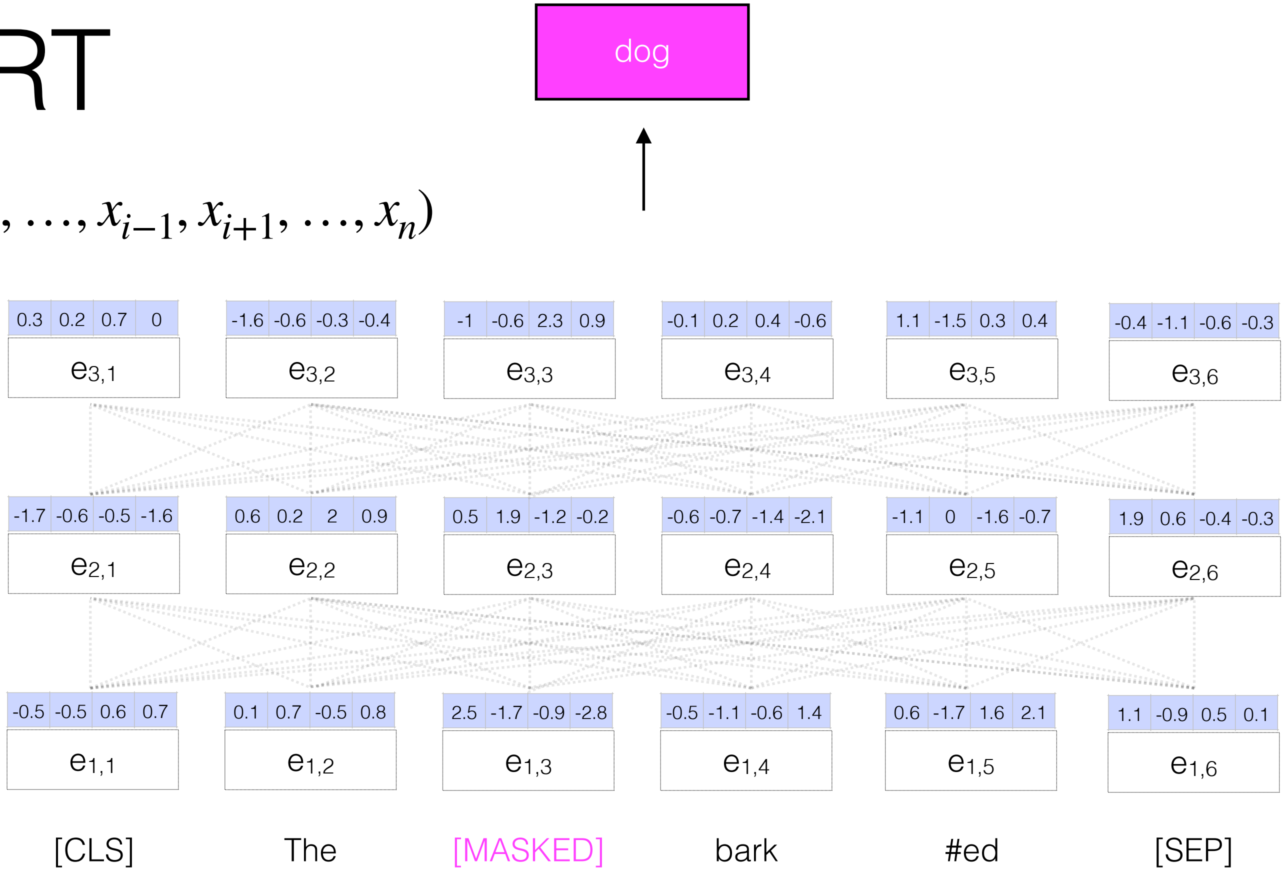
Left-to-right LM
(GPT)

$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

Liu et al. 2021

BERT

$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$



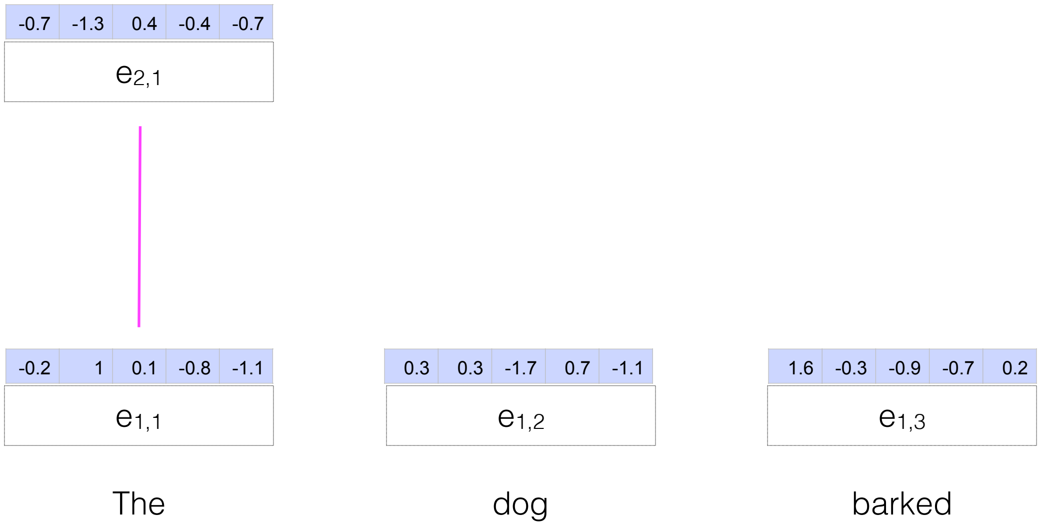
GPT

- Transformer-based causal (left-to-right) language model:

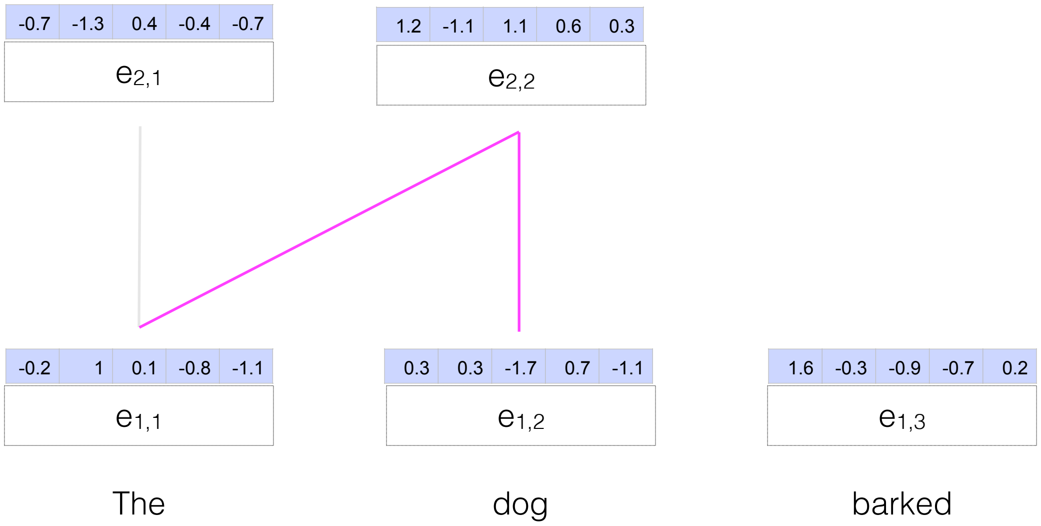
$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

	Model	Data
GPT-2 (Radford et al. 2019)	Context size: 1024 tokens 117M-1.5B parameters	WebText (45 million outbound links from Reddit with 3+ karma); 8 million documents (40GB)
GPT-3 (Brown et al. 2020)	Context size: 2048 tokens 125M-175B parameters	Common crawl + WebText + “two internet-based books corpora” + Wikipedia (400B tokens, 570GB)

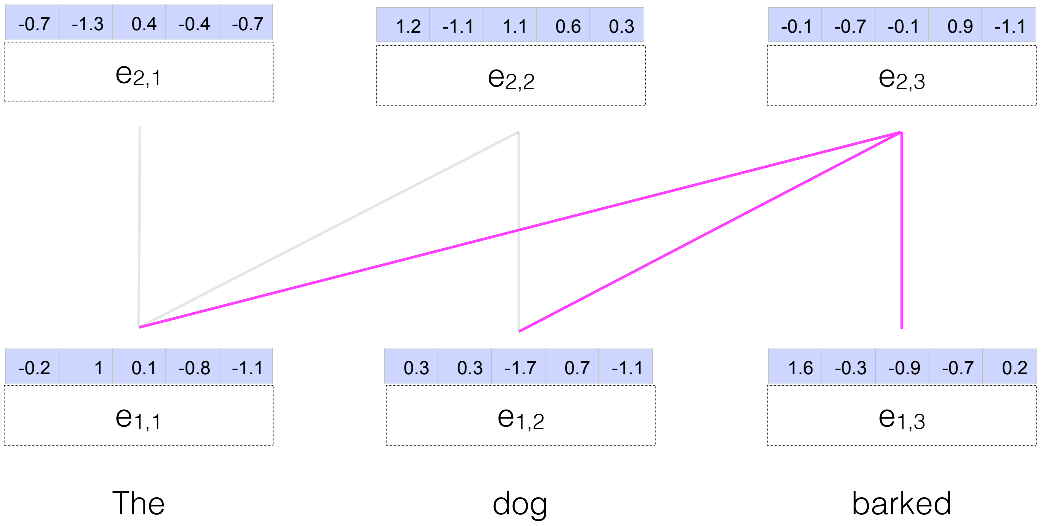
- Self-attention for token i at layer j only attends to tokens 1 through i at layer j-1

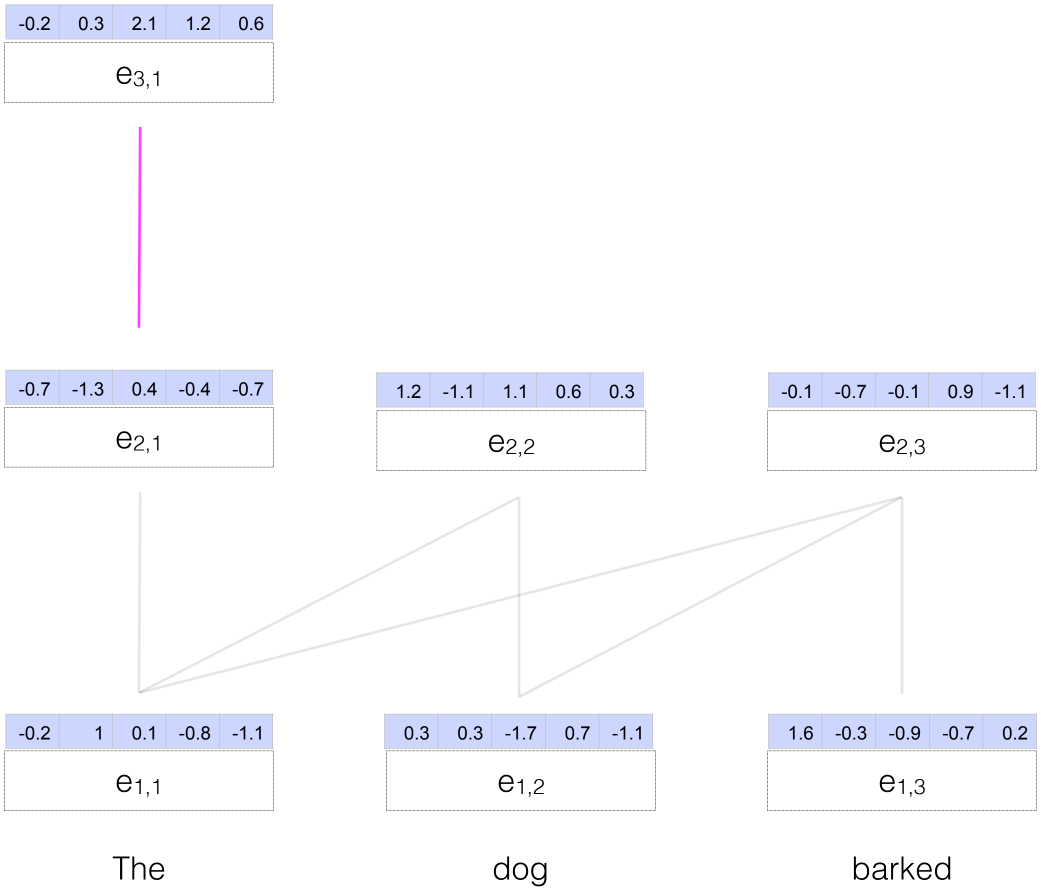


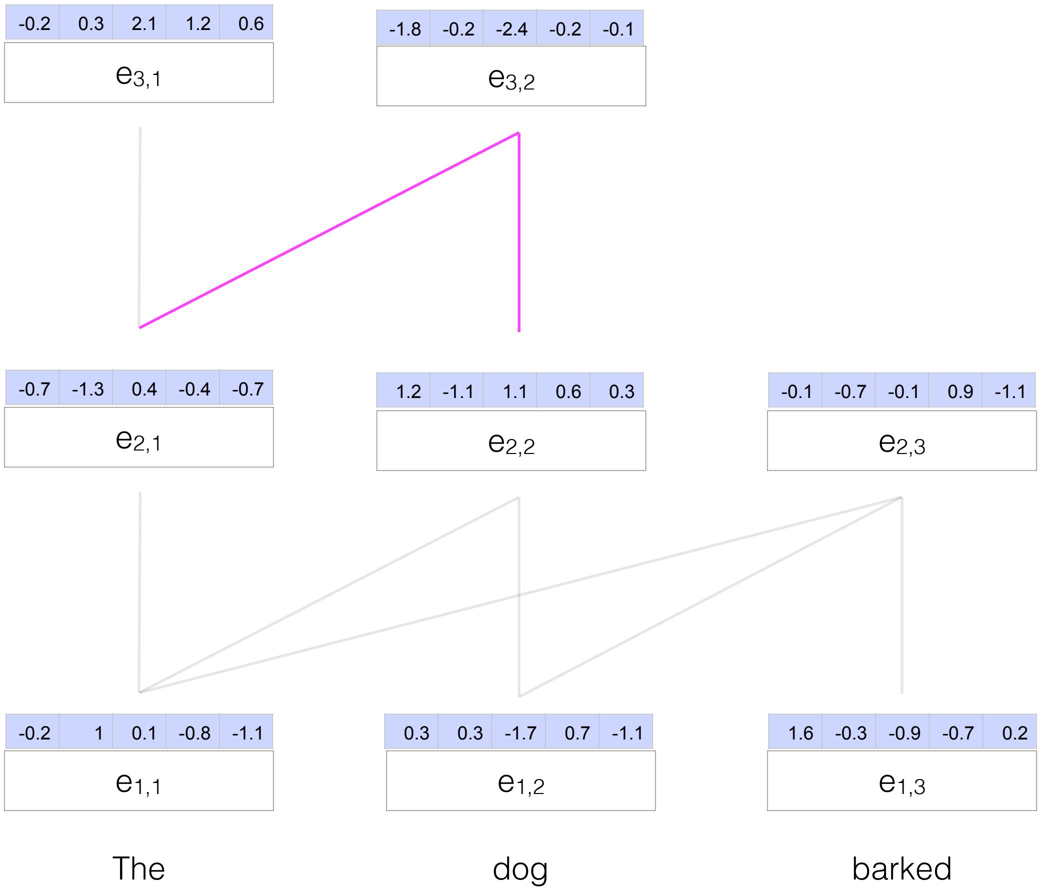
- Self-attention for token i at layer j only attends to tokens 1 through i at layer j-1



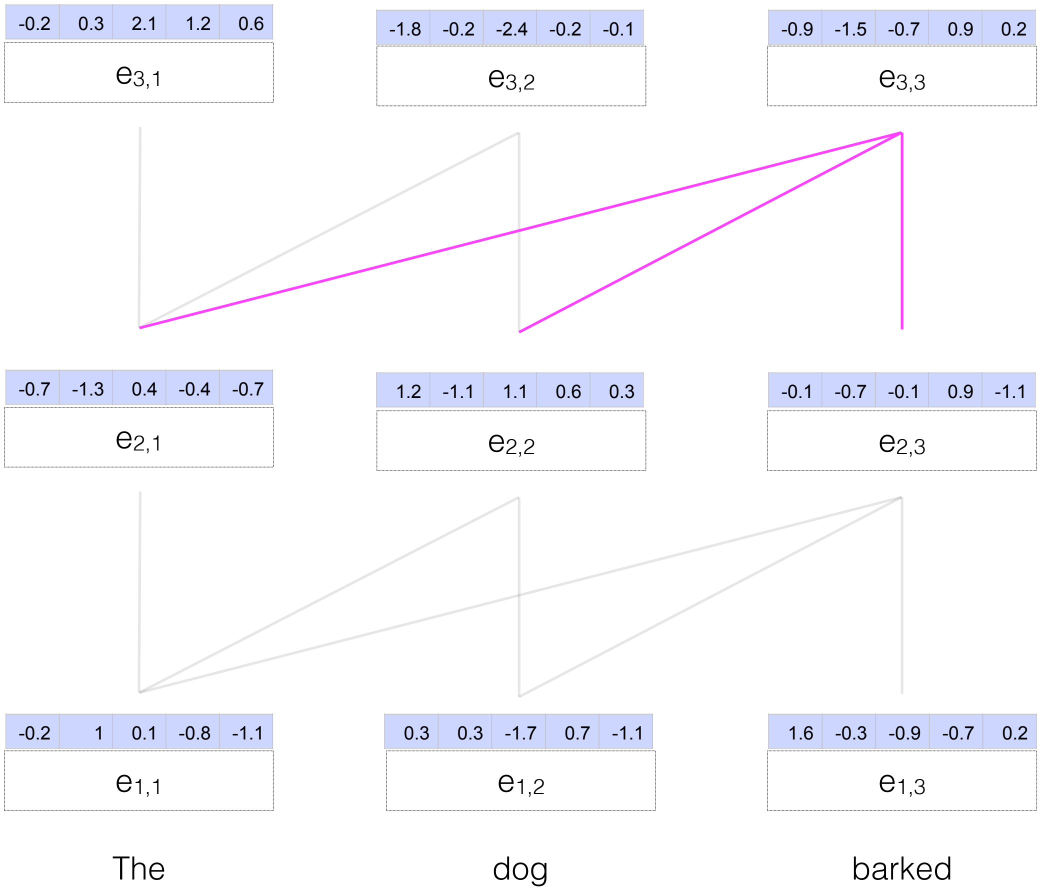
- Self-attention for token i at layer j only attends to tokens 1 through i at layer j-1




-0.2 0.3 2.1 1.2 0.6
$e_{3,1}$
-0.7 -1.3 0.4 -0.4 -0.7
$e_{2,1}$
1.2 -1.1 1.1 0.6 0.3
$e_{2,2}$
-0.1 -0.7 -0.1 0.9 -1.1
$e_{2,3}$
-0.2 1 0.1 -0.8 -1.1
$e_{1,1}$
0.3 0.3 -1.7 0.7 -1.1
$e_{1,2}$
1.6 -0.3 -0.9 -0.7 0.2
$e_{1,3}$
The
dog
barked



-0.2 0.3 2.1 1.2 0.6
$e_{3,1}$
-1.8 -0.2 -2.4 -0.2 -0.1
$e_{3,2}$
-0.7 -1.3 0.4 -0.4 -0.7
$e_{2,1}$
1.2 -1.1 1.1 0.6 0.3
$e_{2,2}$
-0.1 -0.7 -0.1 0.9 -1.1
$e_{2,3}$
-0.2 1 0.1 -0.8 -1.1
$e_{1,1}$
0.3 0.3 -1.7 0.7 -1.1
$e_{1,2}$
1.6 -0.3 -0.9 -0.7 0.2
$e_{1,3}$
The
dog
barked



-0.2 0.3 2.1 1.2 0.6
e3,1
-1.8 -0.2 -2.4 -0.2 -0.1
e3,2
-0.9 -1.5 -0.7 0.9 0.2
e3,3
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.1 -0.7 -0.1 0.9 -1.1
e2,3
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked

Everything is language modeling

The director of *2001: A Space Odyssey* is ______________

The French translation of “cheese” is ______________

The sentiment of “I really hate this movie” is _____________

In Context Learning

- Provide the pattern; LLM is expected to continue with it.
- Use the off-the-shelf model:
    - No Gradient update and parameter change.

Zero-shot

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

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

task description

prompt

Brown et al. (2020, "Language Models are Few-Shot Learners"
https://arxiv.org/pdf/2005.14165.pdf

One-shot

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

`Translate English to French:`	*task description*
`sea otter => loutre de mer`	*example*
`cheese =>`	*prompt*

Brown et al. (2020, “Language Models are Few-Shot Learners”
https://arxiv.org/pdf/2005.14165.pdf

Few-shot

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

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

task description

examples

prompt

Brown et al. (2020, "Language Models are Few-Shot Learners"
https://arxiv.org/pdf/2005.14165.pdf

Evolution of Paradigm

Before 2014	Fully Supervised (feature Engineering)
2014-2019	Architecture Engineering
2019-2021	Pretrain+Finetune: Objective Engineering
2021-…	Pretrain, prompt, predict: Prompt Engineering

```
Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you
requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you
requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
```
Good English output: I'd be more than happy to work with you on another project.

```
Poor English input: Please provide me with a short brief of the design you're looking for and
that'd be nice if you could share some examples or project you did before.
```
Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

```
Poor English input: The patient was died.
```
Good English output: The patient died.

```
Poor English input: We think that Leslie likes ourselves.
```
Good English output: We think that Leslie likes us.

Brown et al. (2020), "Language Models are Few-Shot Learners"

Context →	Please unscramble the letters into a word, and write that word: volwskagen =
Target Completion →	volkswagen

Figure G.23: Formatted dataset example for Anagrams 2

Brown et al. (2020), “Language Models are Few-Shot Learners”

Context →	Title: The_Blitz Background: From the German point of view, March 1941 saw an improvement. The Luftwaffe flew 4,000 sorties that month, including 12 major and three heavy attacks. The electronic war intensified but the Luftwaffe flew major inland missions only on moonlit nights. Ports were easier to find and made better targets. To confuse the British, radio silence was observed until the bombs fell. X- and Y-Gerät beams were placed over false targets and switched only at the last minute. Rapid frequency changes were introduced for X-Gerät, whose wider band of frequencies and greater tactical flexibility ensured it remained effective at a time when British selective jamming was degrading the effectiveness of Y-Gerät. Q: How many sorties were flown in March 1941? A: 4,000 Q: When did the Luftwaffe fly inland missions? A:
Target Completion →	only on moonlit nights

Figure G.28: Formatted dataset example for SQuADv2

Brown et al. (2020), “Language Models are Few-Shot Learners”

Context →	The bet, which won him dinner for four, was regarding the existence and mass of the top quark, an elementary particle discovered in 1995. question: The Top Quark is the last of six flavors of quarks predicted by the standard model theory of particle physics. True or False? answer:
Target Completion →	False

Figure G.31: Formatted dataset example for RTE

Context →	An outfitter provided everything needed for the safari. Before his first walking holiday, he went to a specialist outfitter to buy some boots. question: Is the word 'outfitter' used in the same way in the two sentences above? answer:
Target Completion →	no

Figure G.32: Formatted dataset example for WiC

Brown et al. (2020), "Language Models are Few-Shot Learners"

Causal reasoning

Textual entailment

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

Word sense disambiguation

Question answering

Brown et al. (2020), “Language Models are Few-Shot Learners”

Prompt engineering

- Manual prompt design: encoding domain knowledge into prompt templates that are likely to generate a response in the output space.

Type	Task	Input ([X])	Template	Answer ([Z])
Text CLS	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
Text-pair CLS	NLI	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	NER	[X1]: Mike went to Paris. [X2]: Paris	[X1][X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman ... ...
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...

Liu et al. 2021

Prompt engineering

- Prompt mining: rather than manually writing prompts, learning high-performing prompts from input/output pairs in training data (e.g., labeled classification/relation extraction examples).

ID	Relations	Manual Prompts	Mined Prompts	Acc. Gain
P140	religion	x is affiliated with the y religion	x who converted to y	+60.0
P159	headquarters location	The headquarter of x is in y	x is based in y	+4.9
P20	place of death	x died in y	x died at his home in y	+4.6
P264	record label	x is represented by music label y	x recorded for y	+17.2
P279	subclass of	x is a subclass of y	x is a type of y	+22.7
P39	position held	x has the position of y	x is elected y	+7.9

Jiang et al. 2020, “How Can We Know What Language Models Know?”

Prompt engineering

- Prompt paraphrasing: automatically generate paraphrases of a manual prompt, and see which ones perform best on evaluation data.

Usage	Number	Seed	Example
$\boldsymbol{s} \rightarrow \boldsymbol{h}$	70	in summary	in short, in a word, to sum up
$\boldsymbol{h} \leftrightarrow \boldsymbol{r}$	34	in other words	to rephrase it, that is to say, i.e.

Yuan et al. 2021, “BARTSCORE: Evaluating Generated Text as Text Generation”

Prompt augmentation

- Providing several examples in the prompt context to illustrate the intended behavior.

Answered prompts

Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

Concerns on LLMs

- Concerns on the current trend LLM expansions
  - Data and representation
  - Economic and environmental
  - Privacy
  - ….
- Worldwide debates on regulations

Documentation debt

- As Bender et al. 2021 notes, “documentation allows for accountability” and it’s often unclear what data these models are trained on.

- When known, training data encodes narrow perspectives — e.g., links shared on Reddit; filtering out pages containing words related to sex (as C4 does) filters pornography but also positive sex discussions.

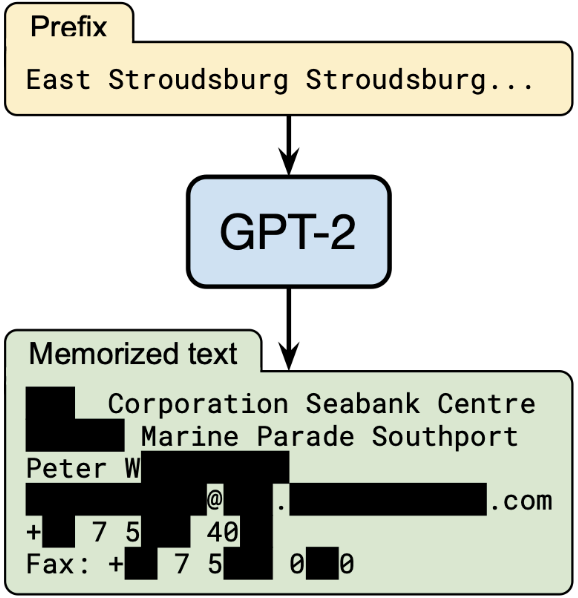
- Biases in training data can lead to representational harms [Kurita et al. 2019; Hutchinson et al. 2020; Gehman et al. 2020]

Bender et al. 2021, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”

Privacy

- Large language models (e.g,. GPT-3, BERT) can memorize training data, which is recoverable from it.

- Potential violations of confidential data (e.g., GMail messages) and contextual integrity (data being published in a way that violates a user's expectations of use).



Carlini et al. (2020), "Extracting Training Data from Large Language Models"

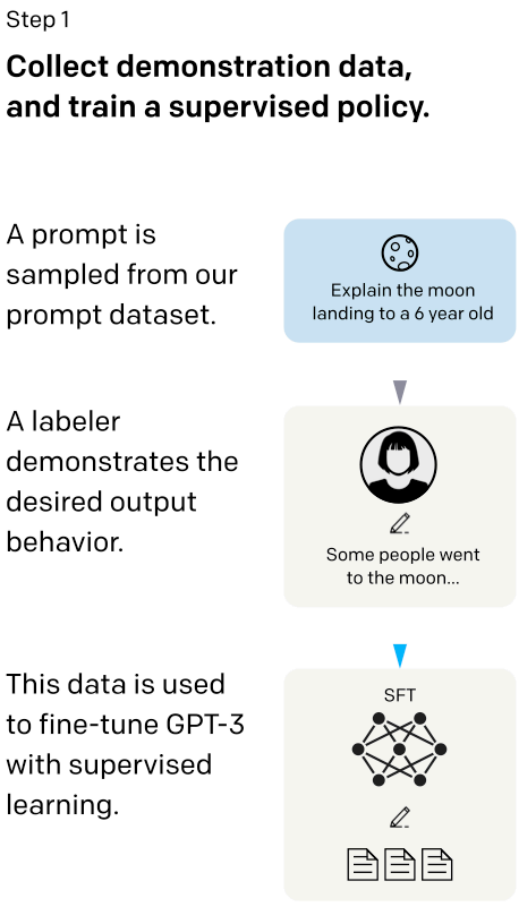
Aligning Language Models

- All of the models we've discussed so far (BERT, GPT-*) are optimized to predict the probabilities of words—-not to encourage (or discourage) any specific kind of behavior.

Q: How many bonks are in a quoit?
A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?
A: It takes two rainbows to jump from Hawaii to seventeen.

Lacker (2020), "Giving GPT-3 a Turing Test" (https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html); Brown et al. (2020), "Language Models are Few-Shot Learners" (NeurIPS)



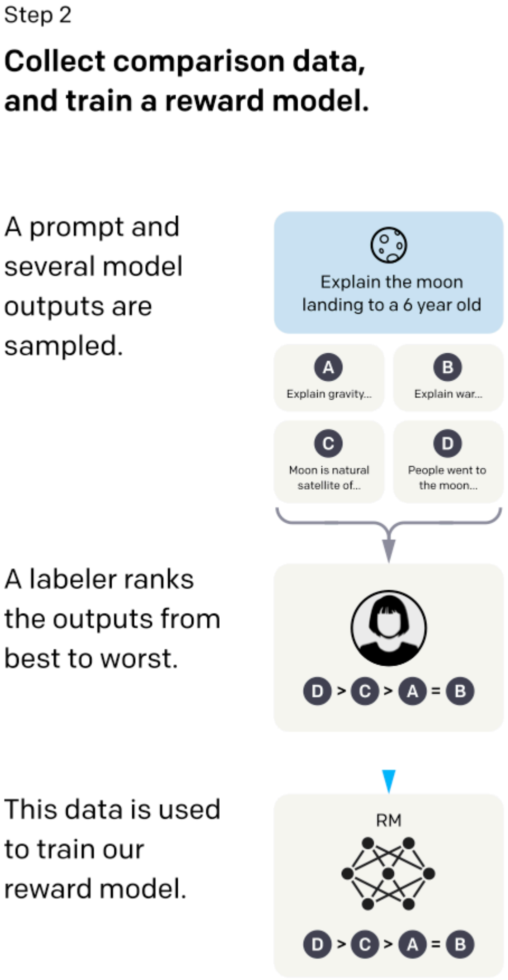
Instruct-GPT

- We can encourage specific behaviors by providing *feedback* to the model, in two forms:

- (1) Given a prompt ("How many rainbows does it take to jump from Hawaii to seventeen?"), a human writes a preferred response ("A rainbow is not a unit of measurement."). **13K prompts/labels.**

- Fine-tune a pretrained model (e.g., GPT-3) on that supervised data to produce a supervised policy (a distribution over words to produce given a model state).

https://arxiv.org/pdf/2203.02155.pdf



Instruct-GPT

- We can encourage specific behaviors by providing *feedback* to the model, in two forms:

- (2) Given a prompt and a set of model responses, a human ranks those responses from best to worst, defining a preference. **33K prompts + 4-9 outputs per prompt.**

- Train a reward model (using reinforcement learning) to score an output reflecting those learned ranks.

https://arxiv.org/pdf/2203.02155.pdf

How many bonks ...

How many bonks are in a quoit?

Show drafts

The question "How many bonks are in a quoit?" doesn't have a straightforward answer because:

1. **"Bonk" isn't a standard unit of measurement:** It's more of a sound effect or slang term for hitting something. It doesn't have a defined quantity or value.
2. **"Quoit" can refer to different things:** Depending on the context, "quoit" can refer to:
    - A ring-shaped object thrown in a game similar to horseshoes.
    - A large flat stone used for quoiting.
    - A historical throwing weapon similar to a discus.

Without knowing the specific context and intended meaning of "bonk" and "quoit," it's impossible to give a meaningful answer to your question.

However, if you could provide more information about what you're trying to understand, I might be able to help you further. For example, are you wondering:

- How many times a quoit (the game object) can be thrown and hit its target before breaking?
- How many different sounds a quoit might make when hitting different surfaces?

You

How many bonks are in a quoit

ChatGPT

In the traditional game of quoits, there aren't specific "bonks" per se. Quoits is a game where players throw rings (quoits) onto a target, often a peg or a stake embedded in the ground. The objective is to land the quoits as close to the target as possible. Players may use terms like "ringer" or "leaner" to describe a successful throw where the quoit lands perfectly around the peg or leans against it, respectively.

If by "bonks" you mean successful throws or points scored, then it would depend on the rules you're playing with and how scoring is determined in that particular game. Typically, each quoit that lands around the peg or close to it earns points. The exact scoring system can vary, but typically, the closer the quoit is to the target, the more points it earns.

How many rainbows …

How many rainbows does it take to jump from Hawaii to seventeen?

Show drafts

This question is a bit of a trick question, and it doesn't have a literal answer. Rainbows are optical phenomena and can't be used to physically jump anywhere. Additionally, "seventeen" could refer to a location or age, neither of which are reachable by jumping rainbows.

It seems like you might be enjoying wordplay or testing my ability to understand context. While I can engage in some playful responses, it's important to note that I always strive to provide accurate and relevant information.

Would you like me to try answering a different question, or perhaps explain the reasoning behind trick questions like this one?

BE **You**

How many rainbows does it take to jump from Hawaii to seventeen?

ChatGPT

It seems like you're presenting a playful and nonsensical question. To clarify, rainbows aren't physical objects that one can use to jump from one place to another, and "seventeen" could be interpreted as a location, a number, or something else entirely depending on the context. If you could provide more information or clarify your question, I'd be happy to try to help.

More on LLMs

Logistics

- AP0 is due this Friday Feb 16
- Exam 1 is next Wednesday Feb 21
- Homework 3 will be out soon
    - Will be due Thursday Feb 22.
- Quiz 4 will be out tomorrow **(due this Sunday Feb 18)**
- Have a good long weekend & good luck in exam 1.