

Advanced Deep Learning

Pretrained Transformer Models

Manish Gupta

Agenda

- Word Embeddings

Onehot encoding

- How to represent word meaning in a neural network?
- Embeddings are numerical representation of text.
- Cons
 - #dimensions = $|\text{vocabulary}|$ ☐ Large.
 - Sparse embedding matrix
 - Cannot capture meaning

KING	QUEEN	WATER	MAN	WOMAN
$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

Vocabulary = ['king', 'queen', 'water', 'man', 'woman', 'tree', 'leaf', 'happy', 'glad', 'mother', 'daughter'].

word2vec embeddings

- Google's pre-trained Word2Vec model includes 300D word vectors for a vocabulary of 3 million words and phrases
- Trained on ~100B words from a Google News dataset

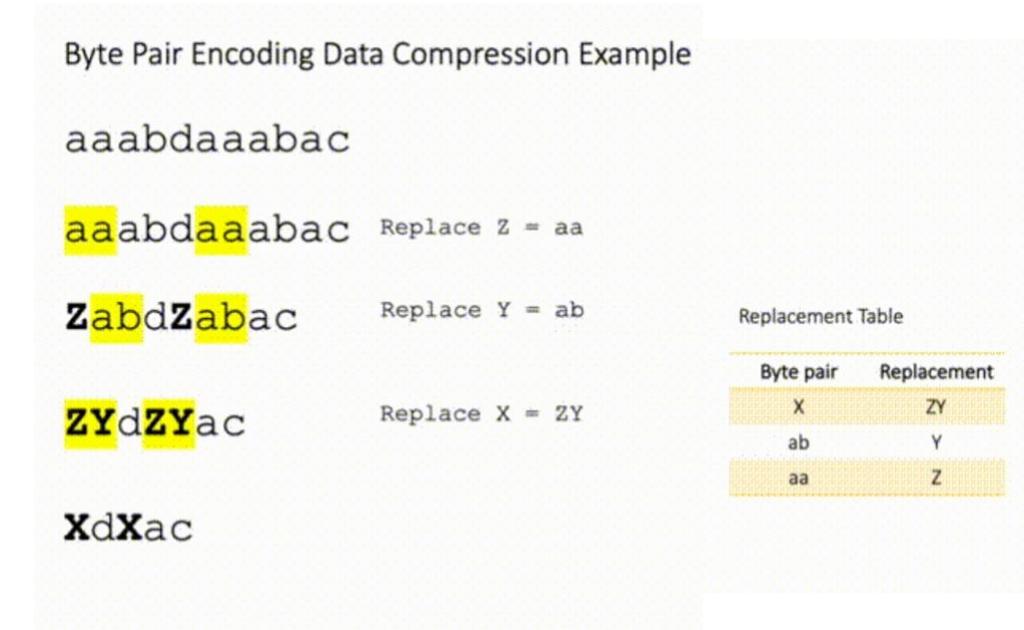
```
import gensim  
# Load Google's pre-trained Word2Vec model  
model =  
gensim.models.Word2Vec.load_word2vec_format('./model/GoogleNews-vectors-negative300.bin')
```

Cosine similarity of words for word2vec representation					
	king	queen	water	man	woman
king	1.000000	0.651096	0.035606	0.229427	0.128480
queen	0.651096	1.000000	0.018264	0.166582	0.316181
water	0.035606	0.018264	1.000000	0.073456	0.101797
man	0.229427	0.166582	0.073456	1.000000	0.766401
woman	0.128480	0.316181	0.101797	0.766401	1.000000

```
[ 1.25976562e-01,  2.97851562e-02,  8.60595703e-03,  1.39648438e-01,  
-2.56347656e-02, -3.61328125e-02,  1.11816406e-01, -1.98242188e-01,  
5.12695312e-02,  3.63281250e-01, -2.42187500e-01, -3.02734375e-01,  
-1.77734375e-01, -2.49023438e-02, -1.67968750e-01, -1.69921875e-01,  
3.46679688e-02,  5.21850586e-03,  4.63867188e-02,  1.28906250e-01,  
1.36718750e-01,  1.12792969e-01,  5.95703125e-02,  1.36718750e-01,  
.....  
-2.73437500e-01, -3.07617188e-02, -3.37890625e-01,  1.53320312e-01,  
2.33398438e-01, -2.08007812e-01,  3.73046875e-01,  8.20312500e-02,  
2.51953125e-01, -7.61718750e-02, -4.66308594e-02, -2.23388672e-02,  
2.99072266e-02, -5.93261719e-02, -4.66918945e-03, -2.44140625e-01,  
-2.09960938e-01, -2.87109375e-01, -4.54101562e-02, -1.77734375e-01,  
-2.79296875e-01, -8.59375000e-02,  9.13085938e-02,  2.51953125e-01]
```

Subword embeddings: Byte Pair Encoding

- Frequently used words should not be split into smaller subwords, but rare words should be decomposed into meaningful subwords.
- Used in GPT2.
- athazagoraphobia ☐ ['_ath', 'az', 'agor', 'aphobia'].
- Vocabulary size=base vocabulary size + the number of merges.
- GPT-2 uses bytes (instead of chars) as the base vocabulary.
- GPT-2 has a vocabulary size of 50,257 = 256 base tokens, a special end-of-text token and 50,000 merges.



Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." *arXiv preprint arXiv:1508.07909* (2015).

WordPiece

- Used for BERT.
- Initialize the vocabulary to include every character present in the training data and progressively learn a given number of merge rules.
- In contrast to BPE, WordPiece chooses the symbol pair whose probability divided by the probabilities of its first symbol followed by its second symbol is the greatest among all symbol pairs.
 - E.g. "u" followed by "g" would have only been merged if $\frac{P(ug)}{P(u)P(g)} >$ probability for any other symbol pair.

Unigram sub-word embedding

- In contrast to BPE or WordPiece merge rules, Unigram initializes its base vocabulary to a large number of symbols and progressively trims down each symbol to obtain a smaller vocabulary.
- The base vocabulary is all pre-tokenized words.
- At each training step, the Unigram algorithm defines log-likelihood loss over the training data given the current vocabulary and a unigram language model. Then, for each symbol in the vocabulary, the algorithm computes how much the overall loss would increase if the symbol was to be removed from the vocabulary.
- Unigram then removes p% ($\sim 10\text{-}20$) of the symbols with lowest loss increase.
- The Unigram algorithm always keeps the base characters so that any word can be tokenized.

SentencePiece

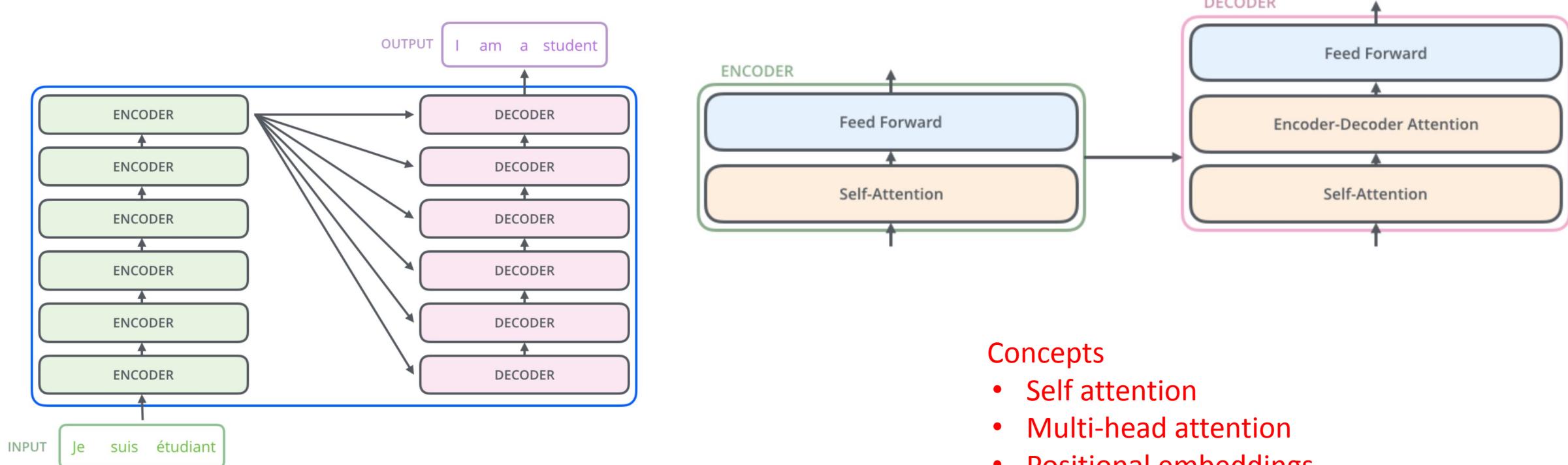
- BPE, WordPiece and Unigram have the same problem: It is assumed that the input text uses spaces to separate words. However, not all languages use spaces to separate words.
- SentencePiece includes the space in the set of characters to use. It then uses the BPE or unigram algorithm to construct the appropriate vocabulary.
- Decoding with SentencePiece is very easy since all tokens can just be concatenated.
- All transformers models in huggingface use SentencePiece with unigram. E.g., T5.

Agenda

- Word Embeddings
- Transformers

Transformers

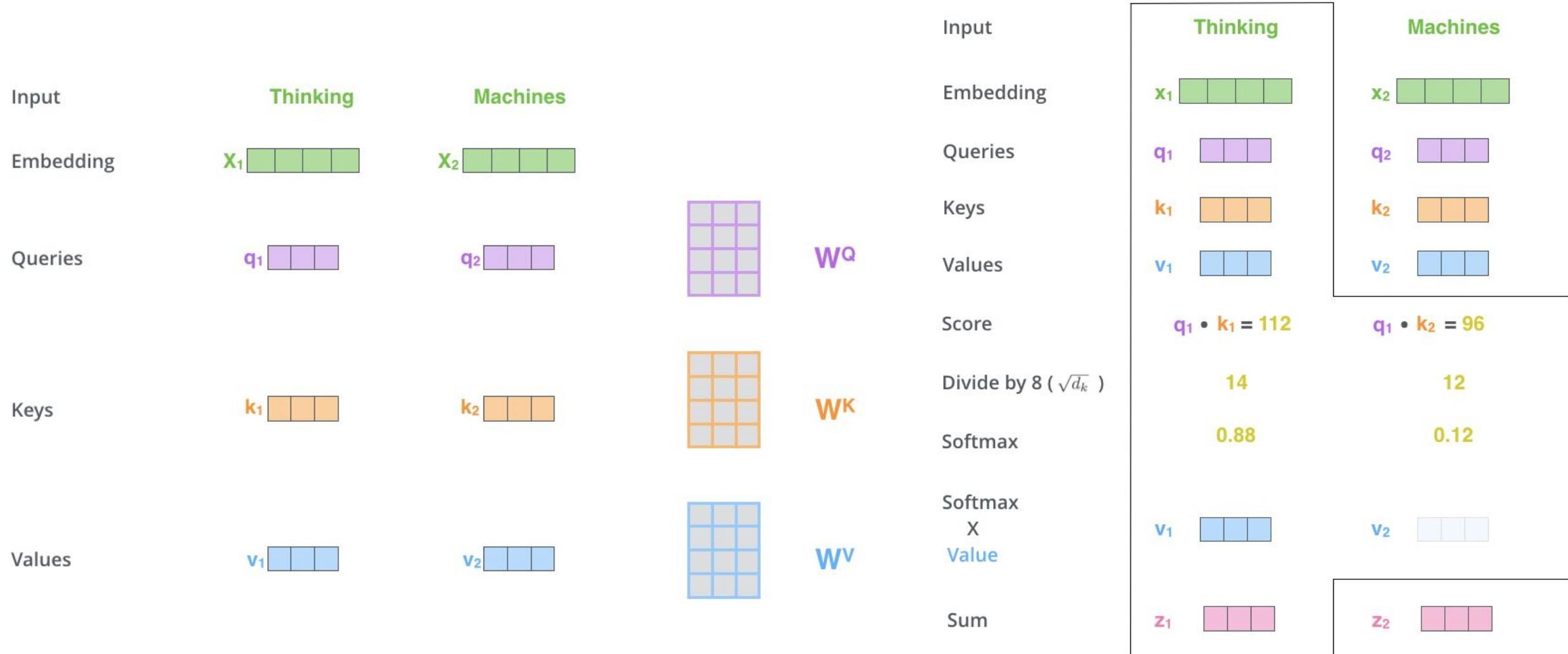
- No recurrent connections.



Concepts

- Self attention
- Multi-head attention
- Positional embeddings
- Residual connections
- Masked attention

Calculating self attention using vectors



Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In NIPS, pp. 5998-6008. 2017.
<http://jalammari.github.io/>

Calculating self attention using matrices

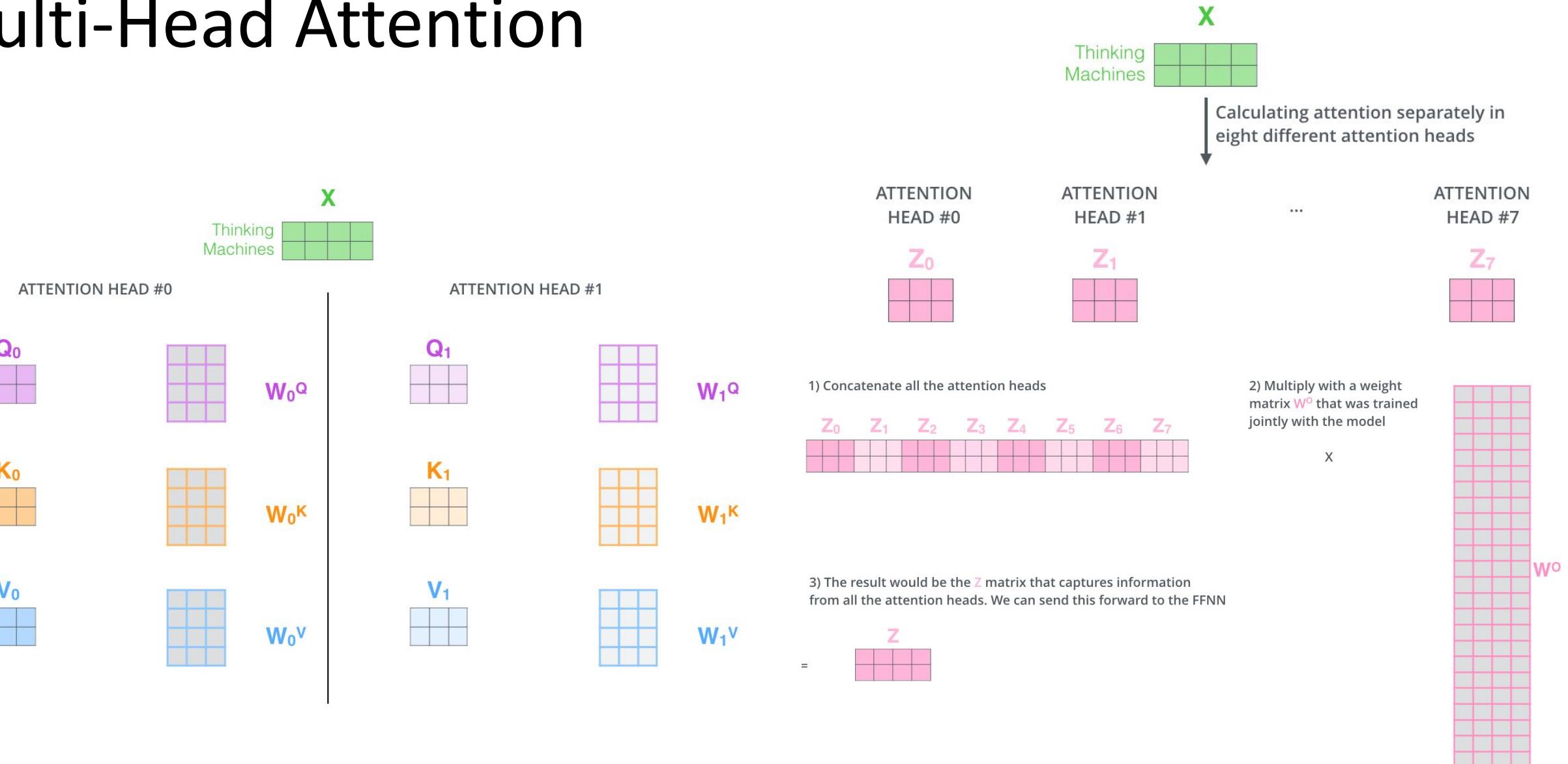
$$\mathbf{X} \times \mathbf{W}^Q = \mathbf{Q}$$

$$\mathbf{X} \times \mathbf{W}^K = \mathbf{K}$$

$$\mathbf{X} \times \mathbf{W}^V = \mathbf{V}$$

$$\text{softmax}\left(\frac{\mathbf{Q} \times \mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} = \mathbf{Z}$$

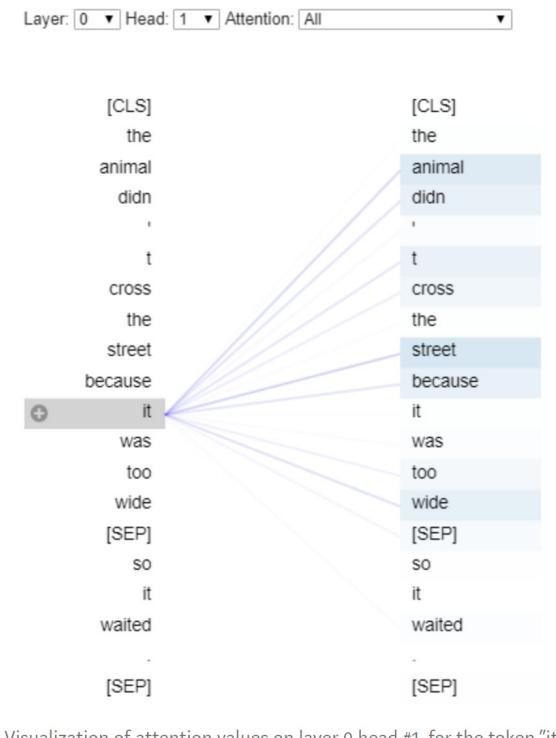
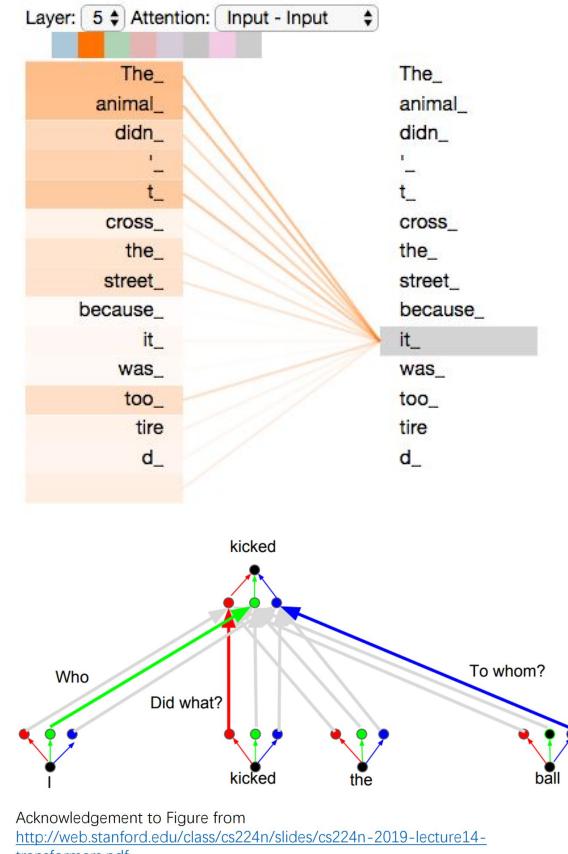
Multi-Head Attention



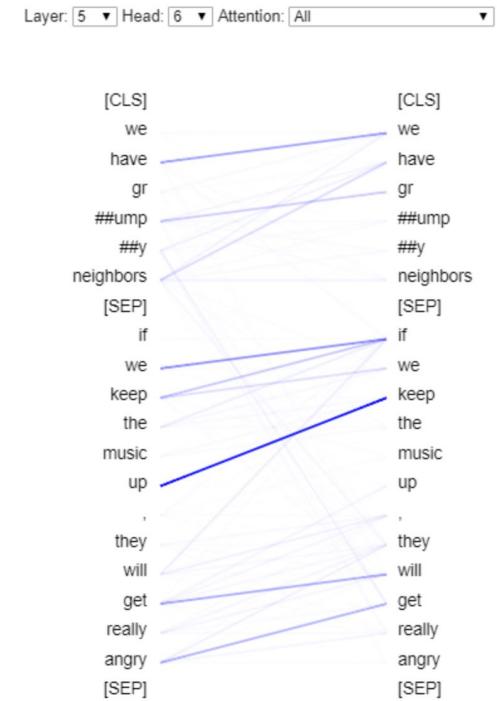
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In NIPS, pp. 5998-6008. 2017.
<http://jalamar.github.io/>

Multi-head Self-Attention examples

- “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”.

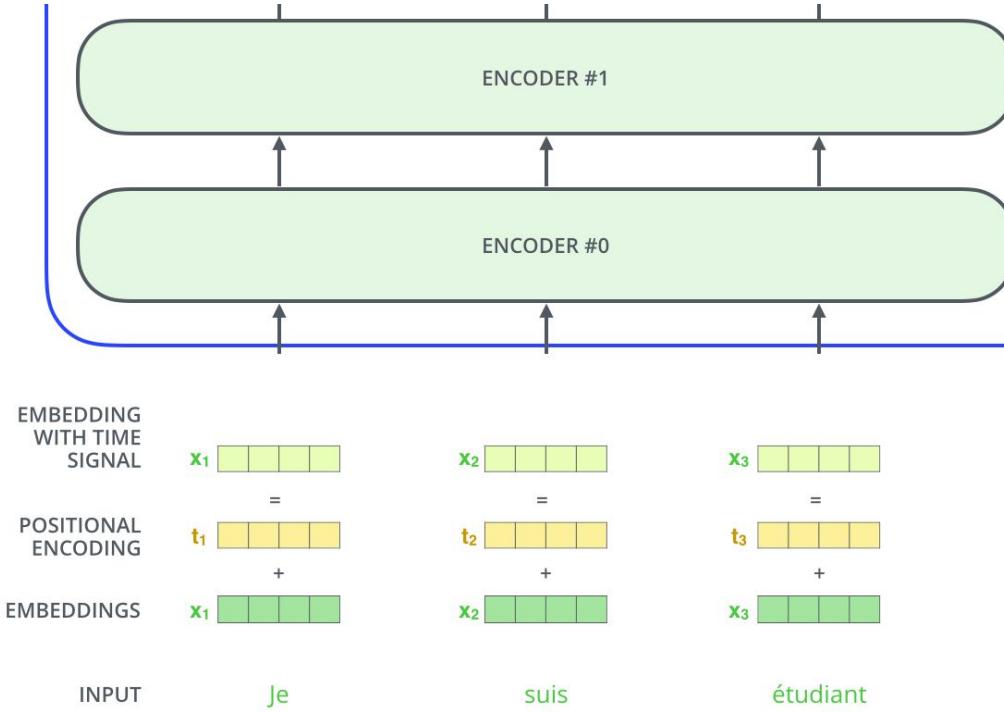


Ambiguous pronoun



Visualization of attention values on layer 5 head #6; where combinations seem to be more focused (*we, have*), (*if, we*), (*keep, up*) (*get, angry*)

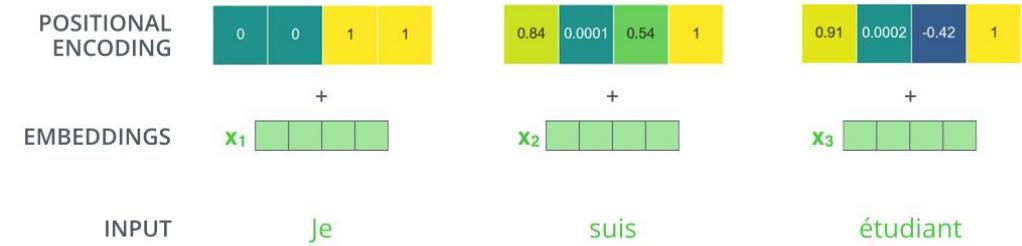
Representing The Order of The Sequence Using Positional Encoding



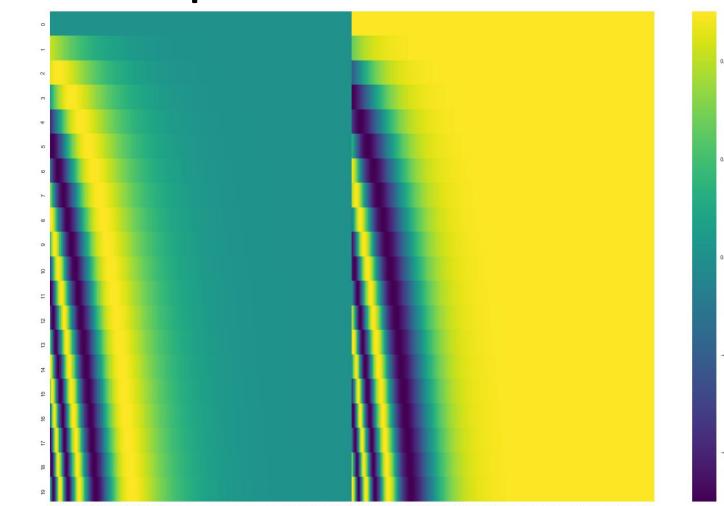
$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$

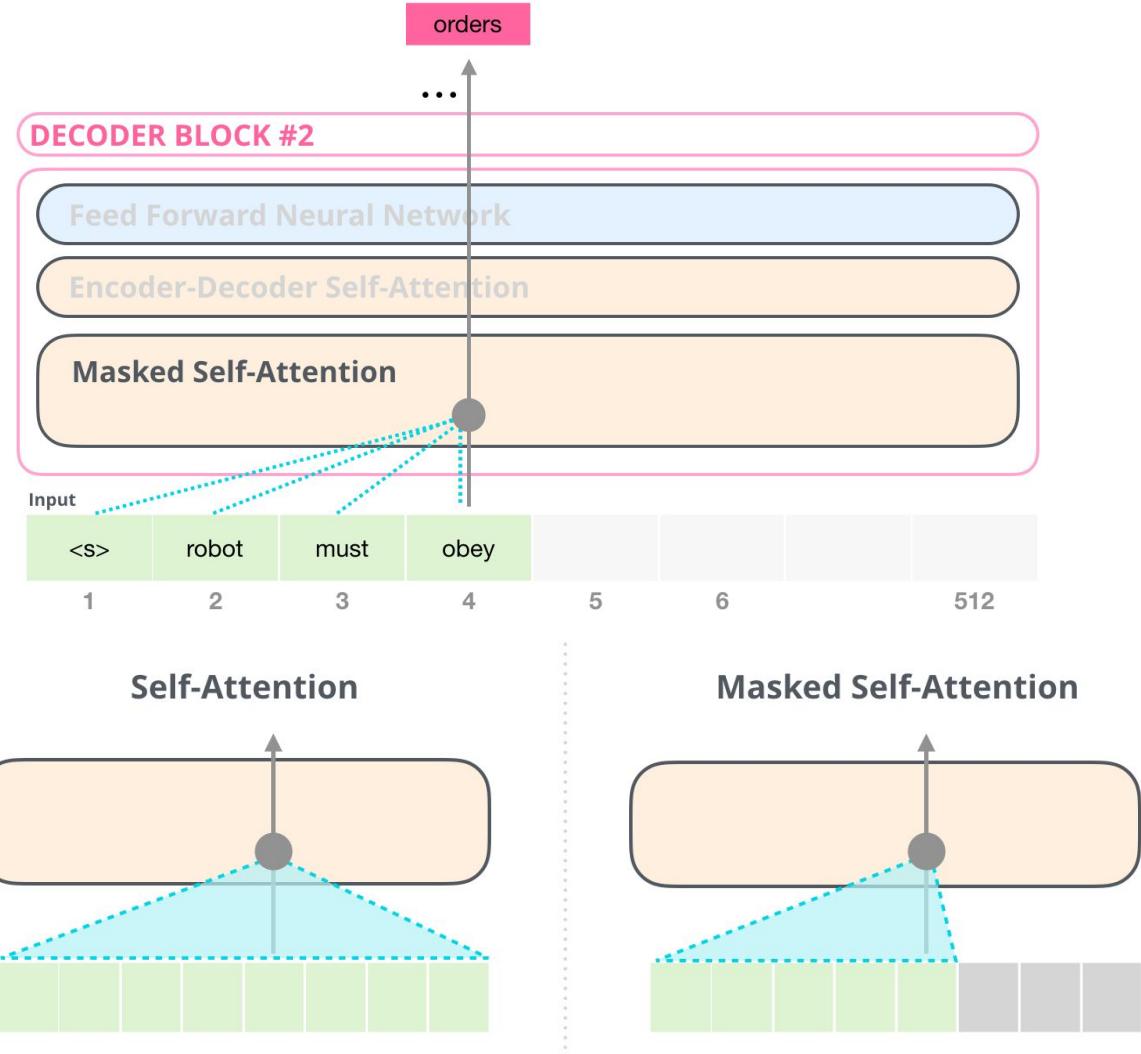
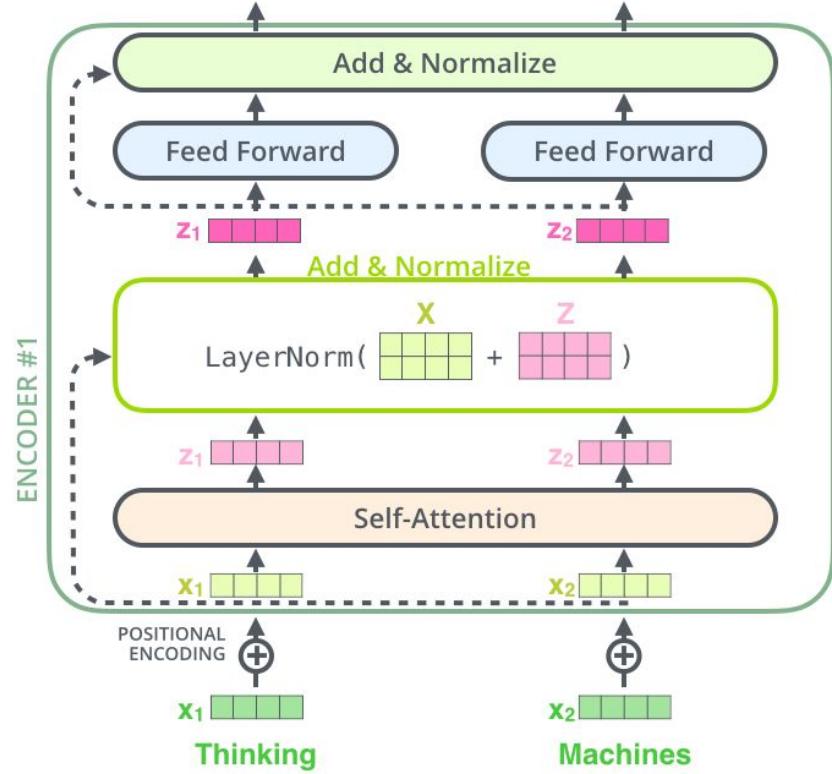
- Example of positional embeddings



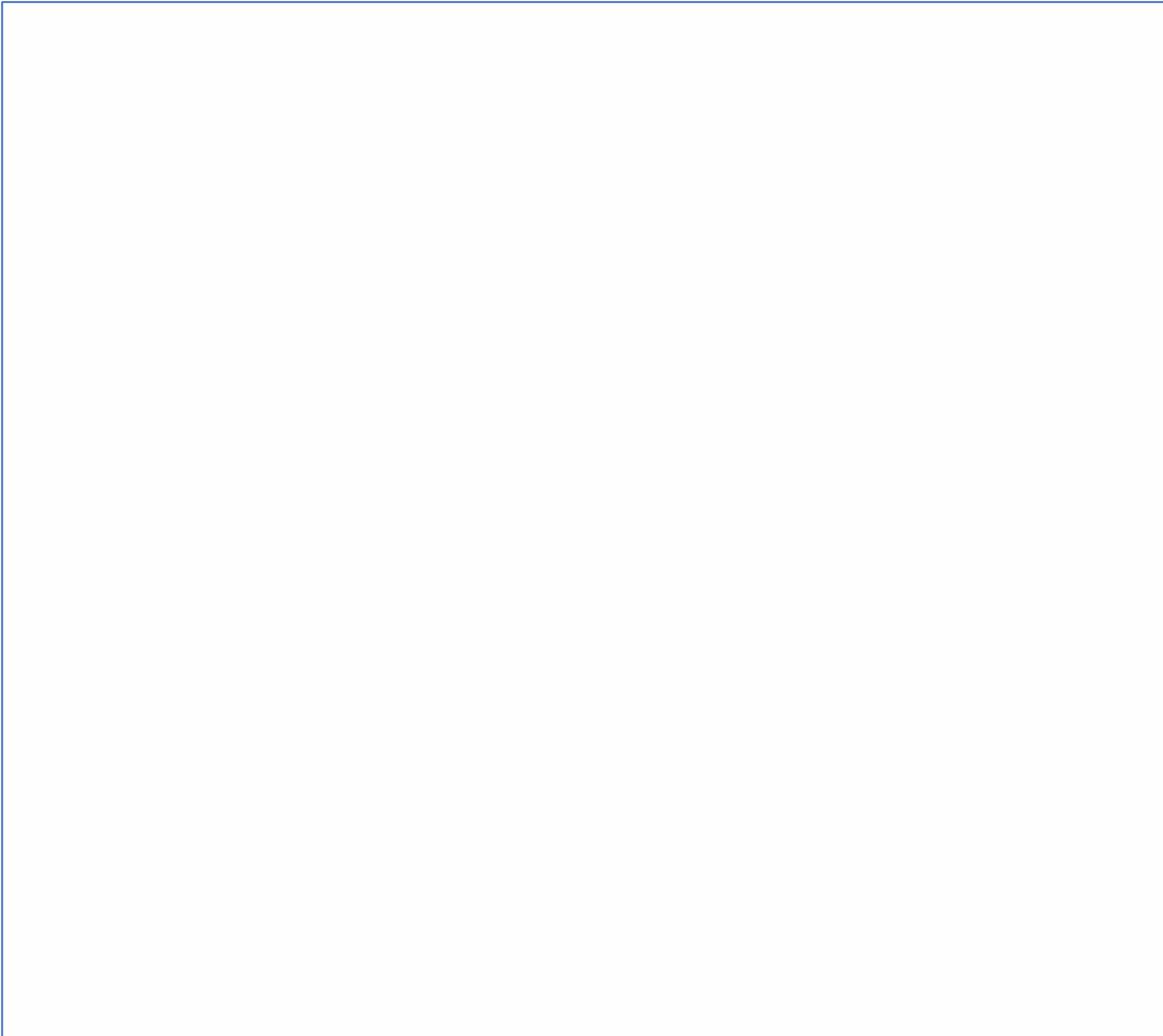
- Another example



Residual Connections and Masked Attention



Model animation from Google Blog



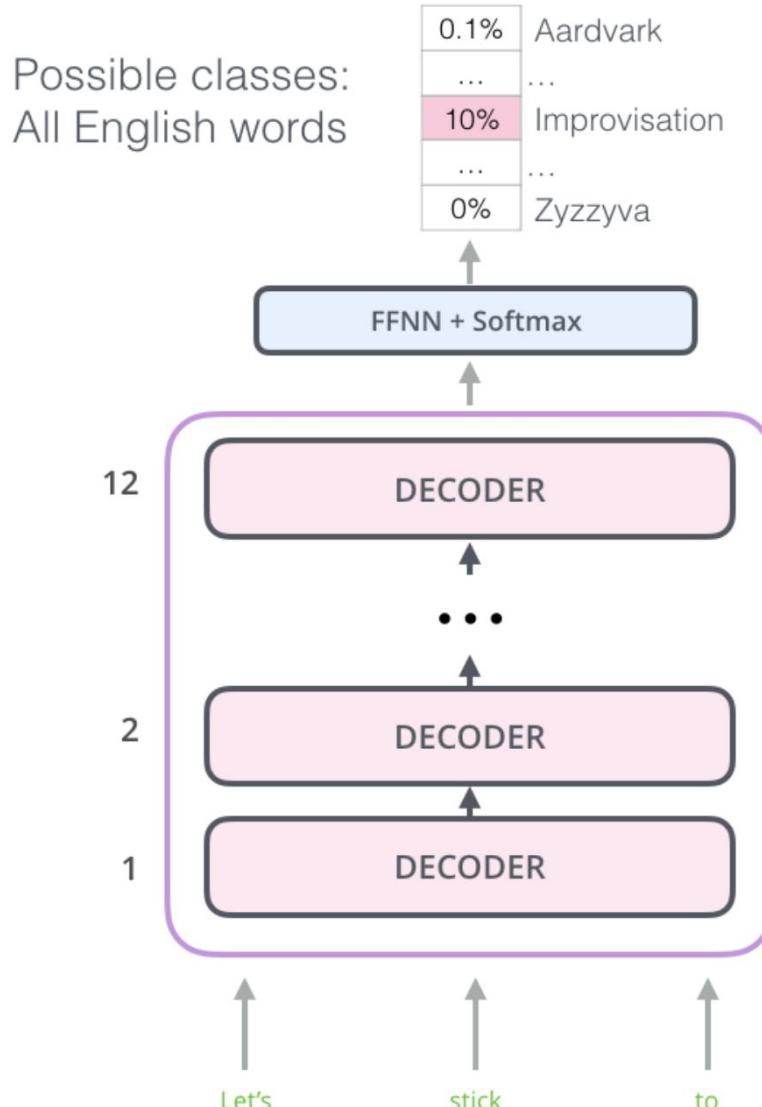
<https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>

Agenda

- Word Embeddings
- Transformers
- GPT

OpenAI GPT

- The model stacked 12 decoder layers.
- No encoder
 - Hence, no encoder-decoder attention sublayer in decoder.
- It would still have the self-attention layer.
- Pretraining with language modeling task: predict the next word using massive (unlabeled) datasets.
- Just, throw the text of 7,000 books at it and have it learn!



Framework=Pretraining

Including language modeling as an auxiliary objective to the fine-tuning helped learning by (a) improving generalization of the supervised model, and (b) accelerating convergence.

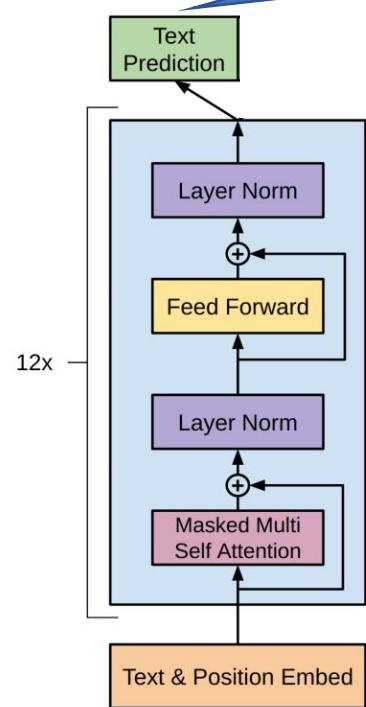
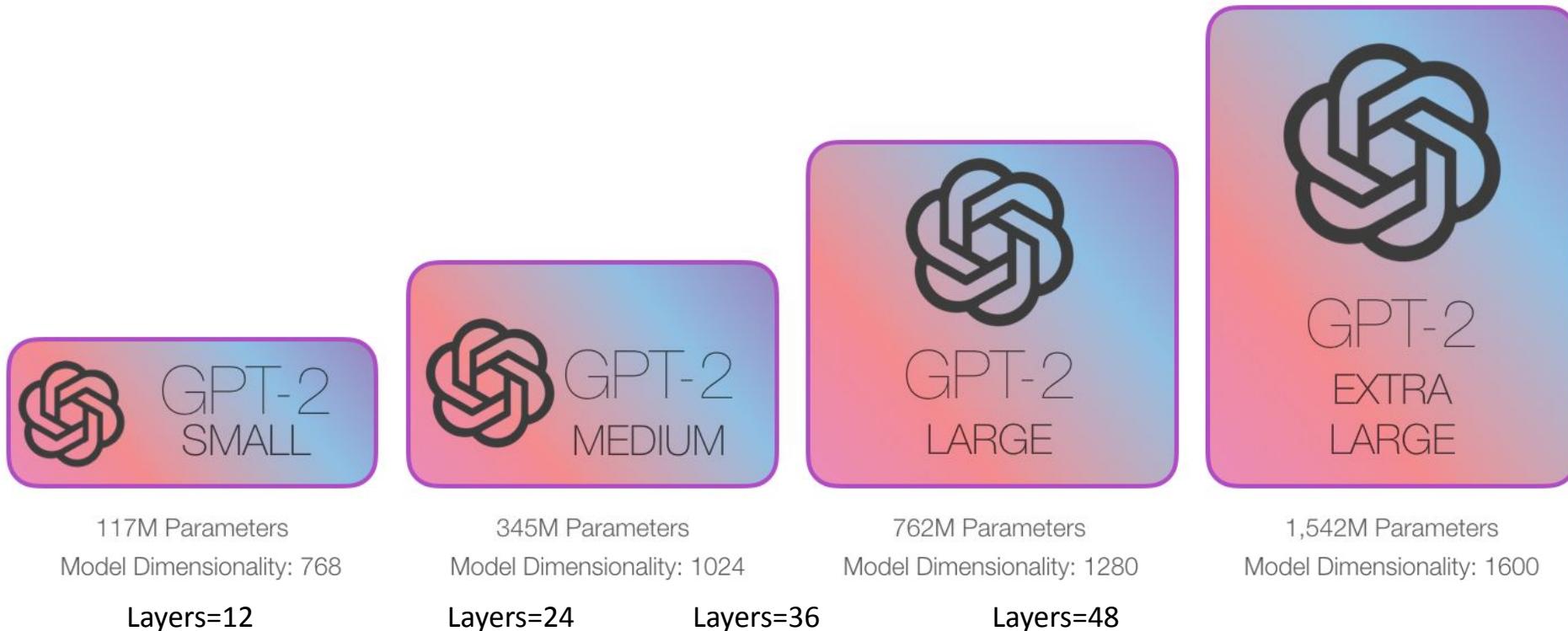


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." (2018).

GPT-2

The largest GPT-2 variant (1.5B parameters) is 13 times the size so it could take up more than 6.5 GBs of storage space.



Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI blog* 1, no. 8 (2019): 9.

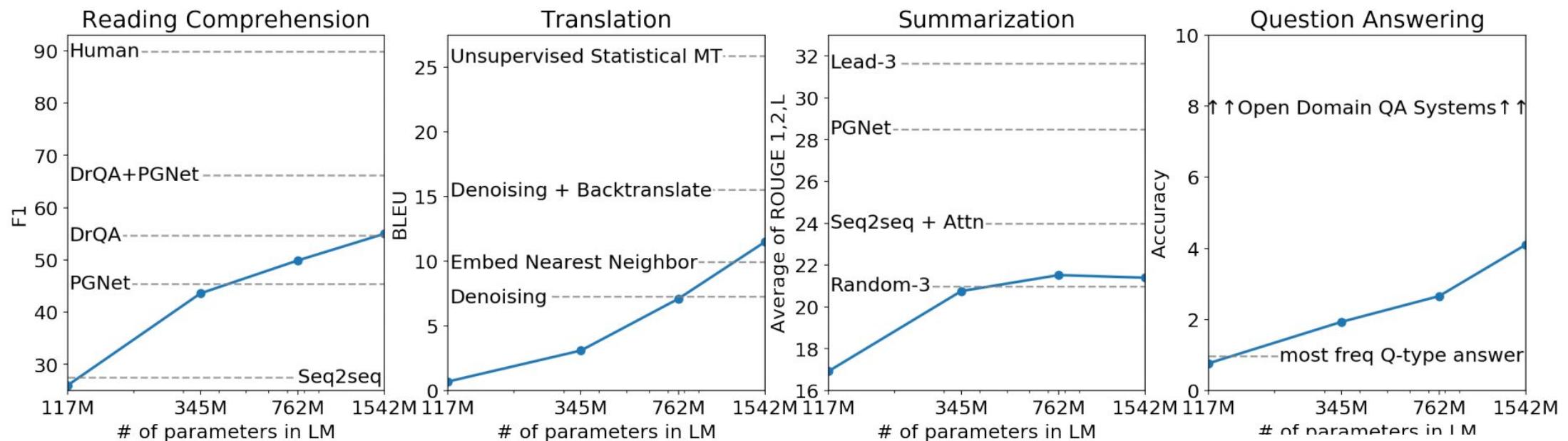
GPT2 training details

- Trained on WebText data.
 - 8 million documents for a total of 40 GB of text.
- Architecture is same as GPT.
- Show that zero-shot learning is possible if model capacity is large and training data is large.
- A single model can be made to perform various tasks by giving task prompts
- Simple Prompts

GPT2 zero shot results

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPC)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	21.8
117M	35.13	45.99	87.65	83.4	29.41	65.85	1.16	1.17	37.50	75.20
345M	15.60	55.48	92.35	87.1	22.76	47.33	1.01	1.06	26.37	55.72
762M	10.87	60.12	93.45	88.0	19.93	40.31	0.97	1.02	22.05	44.575
1542M	8.63	63.24	93.30	89.05	18.34	35.76	0.93	0.98	17.48	42.16

Zero shot language modeling results



Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI blog* 1, no. 8 (2019): 9.

GPT2 question answering zero shot results

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 president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	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%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI blog* 1, no. 8 (2019): 9.

GPT2 Results

Context (WebText test)

Some of the most glorious historical attractions in Spain date from the period of Muslim rule, including The Mezquita, built as the Great Mosque of Córdoba and the Medina Azahara, also in Córdoba and now in ruins but still visitable as such and built as the Madinat al-Zahra, the Palace of al-Andalus; and the Alhambra in Granada, a splendid, intact palace. There are also two synagogues still standing that were built during the era of Muslim Spain: Santa María la Blanca in Toledo and the Synagogue of Córdoba, in the Old City.

Reconquista and Imperial era [edit]

Small model completion

The era of Muslim Spain is very different from that of other parts of the world, with a rather extensive medieval history. While the Spanish Empire existed during the second half of the 9th century BC it was quite slowly transformed to a small, secularized empire by 1150 A.D., in a much more prosperous and diverse state. In 1050 CE the emperor of Naples, at least partially in characterised by a series of wars and invasions from the north-western coast to the western coast of the region (e.g., on the Danube River which extended all the way from Gibraltar into the south of Spain, but was quickly destroyed on the same day by the Roman army), assumed the title of Emperor of Naples. He made a large contribution over the last few decades to restore and maintain the empire, the first of which to the west. His successors were Ferdinand, who took power from the Papal Throne in 1148, while the first imperial successor was King Alfred II of Sweden. The empire was again made one of the few independent states in the European Continent and although no more than 200 million people lived on its territory during this period the people had much to say to the Emperor. From 1720 on the Muslims made up half the population of Spain.

GPT-2 completion

The Reconquista (literally, "The return to the country") of 1092 and 1095 saw the establishment in the Iberian peninsula of a united kingdom of Castile and Aragon in which a Muslim dynasty would exist. This was the last time in Europe where Muslim rule would be considered "native" and the last occasion where Christian-inflamed and organized conflicts would make themselves known in Europe.

The reign of Muslim Moors in Spain is often referred to, especially among non-Muslims, as the "Conquest of Granada" and Spain is sometimes referred to as Spain in "Granada". The last years of Muslim rule were followed by the Spanish Reconquista where the Muslims were driven or compelled out of the Iberian Peninsula with the victory of Ferdinand IX. in the Battle of Poitiers in March 1189 and the conquest of Jerusalem in 1187 by Saladin.

The Reconquista period in Spain came to an end when the Moors retreated north. However the Moorish rule would not be fully replaced by the Christian rule until the Reconquista era resumed when the Reconquista lasted from 1229 to 1492.

The last Muslim state in Spain, Granada, was conquered by the

Context (WebText test)

Outfit 1: Typical

This pairing was the first outfit I thought of when I bought the shoes. It's like a summer version of this Jake Grantham outfit; in fact, my shoes are close to the colors of his Nike Racers! Instead of a heavy Harris Tweed jacket and denim shirt, I'm wearing a cotton DB jacket and a linen shirt. Both fabrics (in these colors) are an absolute must for summer, as they go with both dark and light pants! As you can see, they pair wonderfully with the dark jeans and shoes. It's a pseudo menswear/prep outfit.

Overall, this is a very casual outfit which is why I paired my sneakers with it. I'm not about wearing a full wool suit with sneakers (as GQ shows a lot) but I'm definitely open to keeping things casual, like this cotton DB. Casual fabrics are key to pulling off your sneakers in a dressed down menswear outfit. I'd even suggest to wear these sneakers with a khaki chino suit or a white linen suit. Just be sure to ditch the tie or wear a tee or polo; wearing a tie with sneakers is a bit too much

Small model completion

.

Outfit 2: High End

This outfit was just in the eye of the storm. It's a hybrid of what ASICS does with streetwear (like J. Crew) and what J. Crew does with casual pieces that are a bit more traditional. Again, as I mentioned before, it's a low budget item and they can be very expensive. These will definitely be an issue if you're trying to work as a casual sneakermaker, and that's OK but don't hesitate if you want to try the lower price versions of their outfits!

This version features a very low cut-cut neckline (which is not a bad thing), slightly longer legs, and a pair of long legs to keep things nice. My preferred color of color these items are blue-gray (which is my first colors, as they are a light white), or black (which is my favorite of my designs). As always, my pick of colors from this store have nothing to do with that other than the design. For a high end, I would recommend this to everyone!

Outfit 3: High Light

This was a "high light" look while I had a chance to get some color. I think the color is based on

GPT-2 completion

for me!

Outfit 2: Casual Jacket

I know that some people might be opposed to wearing sneakers in a turtleneck, but I wanted to be true to myself, so I went with a slim fitting turtleneck in a color more similar to my favorite color of the day. Overall, it's a classic turtleneck, with nothing too flashy. I did swap my shoes for these ASICS Tiger Killshots and I wish I hadn't done that. The shoe itself is very comfortable, however, I found that wearing them in these sneakers made my feet look a bit larger. I'd wear them with denim, chinos, whatever, but would never recommend wearing them alone. There are just too many variables involved.

Outfit 3: Dress

This is the final outfit I wore the ASICS Tiger Killshots. This version of the ASICS is a much more modern and streamlined version. The shoe has the most modern and minimal design that I can think of! With a lot of black, these sneakers look very cool. They'll definitely be very comfortable for summer.

To me, wearing them with a suit is the easiest way to have a good pair of sneaker shoes. That's pretty much it! If

GPT2 Completions

Context (WebText test)

Aaron loves mint chocolate cake, but he requires that it be paired with mini chocolate chips, so I threw some of those in between the layers. I also had a few Peppermint Jo Jos on hand so I crushed them up and threw some of those in along with some crushed meringue cookies because, why not? It's a total smorgasbord of minty chocolate chippy cookie crunchy goodness. I didn't measure how much of each topping I used, but after I tasted the finished product, I wish I had added more. You can add anything you want- crushed candy canes, peppermint bark, etc. And don't be afraid to use a heavy hand. Texture = good.

If you don't have 7-inch cake pans, you can get 3 shorter 8-inch layers out of this

GPT-2 Completion 1
recipe. I made the longer version, then used the pan for a layer of buttercream frosting and stuck 1/3 of it into the fridge for a frosting topper. Just make sure you have all of this frosting prepared earlier than you make this cake and you won't have to wait on it as your egg whites sit on the bottom of your mixing bowl.

You want the egg whites whipped to stiff peaks? Then put your cake dish over a pot of simmering water and heat it up until it's just barely warm. It doesn't take more than 3-5 minutes; just enough to get it started

GPT-2 Completion 2
recipe. If you do, and you know how to make that cake, then your life will never be the same.

I have two very simple rules when it comes to cookie crumbs or chocolate chips. 1) When I buy raw sugar, I do not use it until it's cooled down and has at least half a teaspoon of sugar on it. That's my rule. And 2) If I have a recipe that calls for something and I want to use it in a recipe, I do not substitute it with something else on-line. Like this.

I found this really cute card with cute little kittens on

GPT-2 Completion 3
recipe, which will likely yield about the same amount of cake.

Chocolate Buttermilk Cake
Print Chunks of chocolate cake, coated in buttermilk and topped with crunchy chopped chocolate, drizzled with chocolate sauce. Simple but a nice change from chocolate cake! Author: Momma Made Ingredients For the chocolate buttermilk cake 1 1/2 cup unsalted butter , room temperature

1 1/2 cup granulated sugar
3 large eggs plus 1 egg yolk For the chocolate glaze 1/3 cup cocoa powder
1 3/4 cups powdered sugar
6 ounces semis

GPT-2 Completion 4

recipe.

Meringue Cookies

1 3/4 cup butter, softened
1 cup sugar
1 egg yolk
3 T. heavy cream
3 1/2-4 cups flour
pinch of salt
Peppermint-Jojo Topping
1 1/4 cups powdered sugar
1 cup chopped pecans
1/2 cup finely chopped mint leaves
1/2 cup chopped fresh mint (about 1/2 sheet)
1 tsp. peppermint extract
1/4 tsp. sea salt
pinch of sea

GPT-2 Completion 5

recipe.

My 2 failures at Christmas morning were not due to poor frosting skills. That was another success story. The cake was beautifully moist, and well, just a beautiful cake. It's got very little air in it or moisture, but all the chocolate and mint was perfectly blended together.

One thing that made me cringe a bit was that all those ingredients- egg yolks, sugar, melted butter, chocolate chips, melted chocolate, and mint- had to be mixed right into the cake batter, which is super time-consuming. A friend suggested that if I put the batter in a separate bowl, I

GPT-2 Completion 6

recipe, which will give you about 6-9 servings of delicious chocolate cake. To prevent your layers from sticking together, I used my 8-inch round tart pan, lined it with parchment paper, then cut up 4 7-inch round cakes from it and sandwiched the layers between them (this also worked well for the 1-layer chocolate cake as well). When you're ready to bake this cake, remove the parchment paper, preheat oven to 350 degrees and bake for about 30 – 40 minutes, until the top is golden and a tester comes out clean.

Chocolate Buttermilk Layer Cake with Peppermint Jo

GPT2 Completions

Context (passage and previous question/answer pairs)

Tom goes everywhere with Catherine Green, a 54-year-old secretary. He moves around her office at work and goes shopping with her. "Most people don't seem to mind Tom," says Catherine, who thinks he is wonderful. "He's my fourth child," she says. She may think of him and treat him that way as her son. He moves around buying his food, paying his health bills and his taxes, but in fact Tom is a dog.

Catherine and Tom live in Sweden, a country where everyone is expected to lead an orderly life according to rules laid down by the government, which also provides a high level of care for its people. This level of care costs money.

People in Sweden pay taxes on everything, so aren't surprised to find that owning a dog means more taxes. Some people are paying as much as 500 Swedish kronor in taxes a year for the right to keep their dog, which is spent by the government on dog hospitals and sometimes medical treatment for a dog that falls ill. However, most such treatment is expensive, so owners often decide to offer health and even life – for their dog.

In Sweden dog owners must pay for any damage their dog does. A Swedish Kennel Club official explains what this means: if your dog runs out on the road and gets hit by a passing car, you, as the owner, have to pay for any damage done to the car, even if your dog has been killed in the accident.

Q: How old is Catherine?

A: 54

Q: where does she live?

A:

Model answer: Stockholm

Turker answers: Sweden, Sweden, in Sweden, Sweden

GPT2 Completions

Context (passage and previous question/answer pairs)

The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of “one world, one dream”. Plans for the relay were announced on April 26, 2007, in Beijing, China. The relay, also called by the organizers as the “Journey of Harmony”, lasted 129 days and carried the torch 137,000 km (85,000 mi) – the longest distance of any Olympic torch relay since the tradition was started ahead of the 1936 Summer Olympics.

After being lit at the birthplace of the Olympic Games in Olympia, Greece on March 24, the torch traveled to the Panathinaiko Stadium in Athens, and then to Beijing, arriving on March 31. From Beijing, the torch was following a route passing through six continents. The torch has visited cities along the Silk Road, symbolizing ancient links between China and the rest of the world. The relay also included an ascent with the flame to the top of Mount Everest on the border of Nepal and Tibet, China from the Chinese side, which was closed specially for the event.

Q: What was the theme
A: “one world, one dream”.

Q: What was the length of the race?
A: 137,000 km

Q: Was it larger than previous ones?
A: No

Q: Where did the race begin?
A: Olympia, Greece

Q: Is there anything notable about that place?
A: birthplace of Olympic Games

Q: Where did they go after?
A: Athens

Q: How many days was the race?
A: seven

Q: Did they visit any notable landmarks?
A: Panathinaiko Stadium

Q: And did they climb any mountains?
A:

Model answer: Everest

Turker answers: unknown, yes, Yes, yes

GPT-2 example

- GPT-2 is able to write news articles about the discovery of talking unicorns.

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

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.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, “We can see, for example, that they have a common ‘language,’ something like a dialect or dialectic.”

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, “In South America, such incidents seem to be quite common.”

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. “But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization,” said the scientist.

GPT-3: Few-shot learning is enough

- Paper from 31 authors
- 175 billion parameters
- GPT-3 shows that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art finetuning approaches.

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Zero-shot

One-shot

Few-shot

In-context learning

Fine-tuning

The three settings we explore for in-context learning

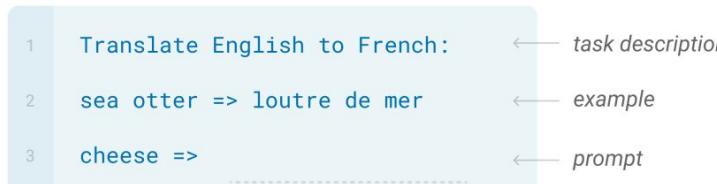
Zero-shot

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



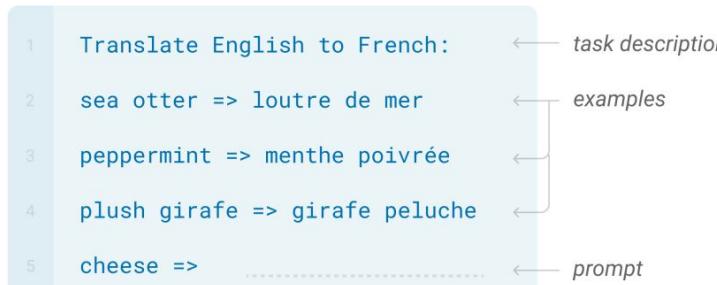
One-shot

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



Few-shot

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



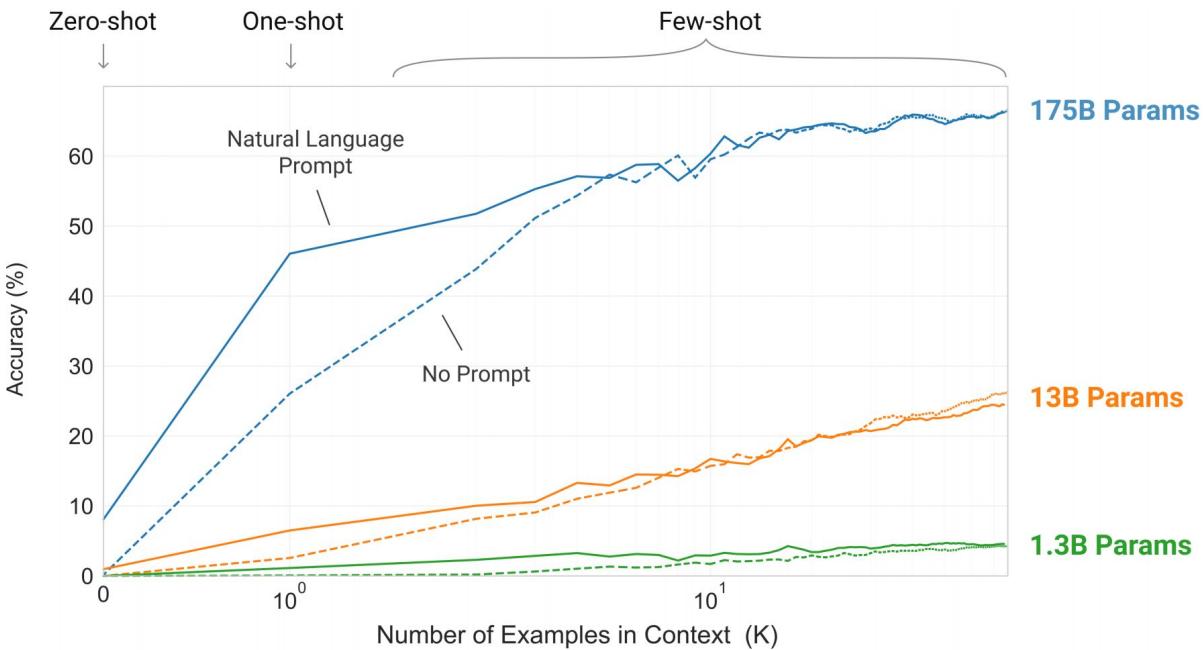
Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



In-context learning is better with large models

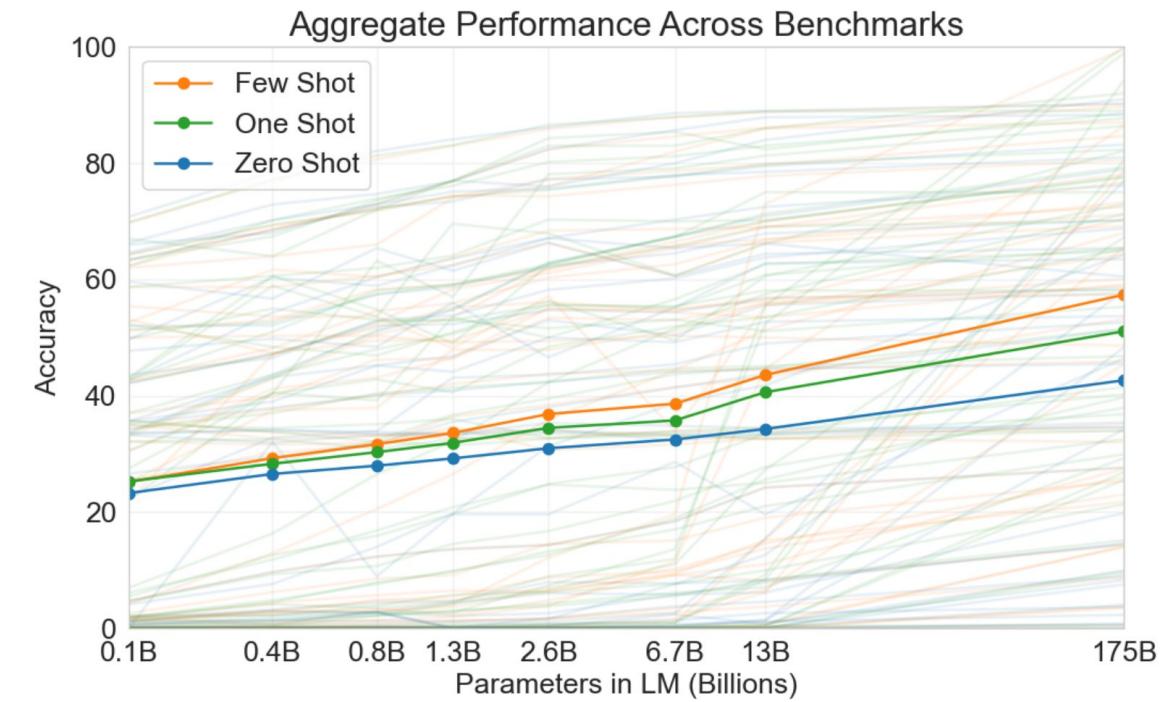


175B Params

13B Params

1.3B Params

Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove extraneous symbols from a word, both with and without a natural language task description. The steeper “in-context learning curves” for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks

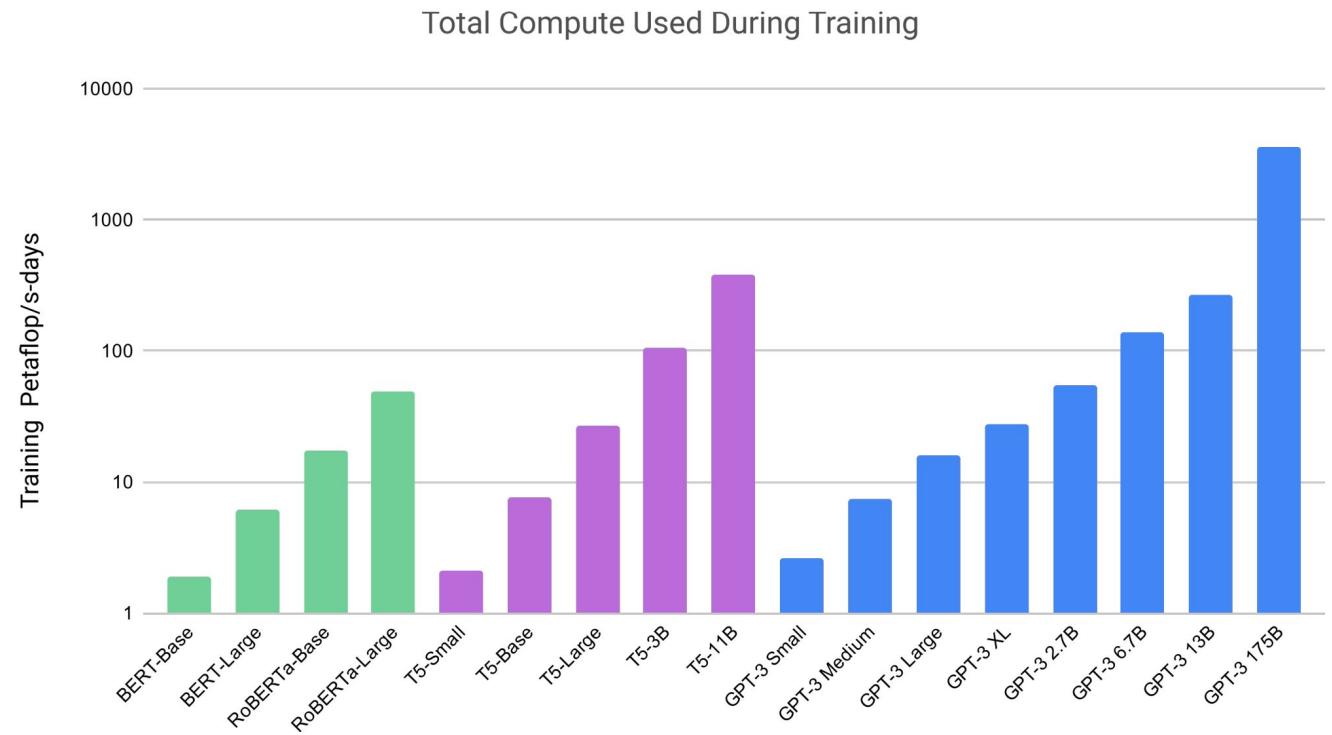


Aggregate performance for all 42 accuracy-denominated benchmarks: While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning.

GPT3 Training

- GPT3 is trained with same model, data and train steps like GPT2 with relatively straightforward scaling up of the model size, dataset size and diversity, and length of training.
- For few-shot learning at inference time, we set K in the range of 10 to 100.
- We train much larger models on many fewer tokens than is typical. Hence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training.

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4



GPT3 results: News Article Generation

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

GPT3 results: Learning and Using Novel Words

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:

We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:

I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:

In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:

I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:

We screeghed at each other for several minutes and then we went outside and ate ice cream.

GPT3 results: Correcting English Grammar

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.

Poor English input: Janet broke Bill on the finger.

Good English output: Janet broke Bill's finger.

Poor English input: Mary arranged for, in St. Louis, John to rent a house cheap.

Good English output: Mary arranged for John to rent a house in St. Louis.

Poor English input: Today I have went to the store to to buys some many bottle of water.

Good English output: Today I went to the store to buy some bottles of water.

Poor English input: I have tried to hit ball with bat, but my swing is has miss.

Good English output: I tried to hit the ball with the bat, but my swing missed.

GPT3 results: Poem generation

Context →

The City
BY C. P. CAVAFY
TRANSLATED BY EDMUND KEELEY
[Poem text omitted]

SOME TREES
John Ashbery
[Poem text omitted]

Shadows on the Way
Wallace Stevens

----- Generated Poem 1 -----

I must have shadows on the way
If I am to walk I must have
Each step taken slowly and alone
To have it ready made

And I must think in lines of grey
To have dim thoughts to be my guide
Must look on blue and green
And never let my eye forget
That color is my friend
And purple must surround me too

The yellow of the sun is no more
Intrusive than the bluish snow
That falls on all of us. I must have
Grey thoughts and blue thoughts walk with me
If I am to go away at all.

----- Generated Poem 2 -----

There is a way in the world.
That way a man may live in a small town
His lifetime and see no man come,
Never the stir of men about a market,
Never the bobbing of covered carts,
Never the creak passed on the stair
Closing some other life; an old man
Can prosper in that town, can prosper
When his store is seen to close, his wife
Fades in his house, and the world shouts in vain
That no one comes to hear him
Where his small bell tolls.

In that small town
Everything is known.
The ways of the young men are known,
And the ways of the young women
As they walk to the Spring;
The rumors are all known,
The quarrels, the hurt, the envy
Of the men and the women.
Till a day comes
That way out of the world
That leads to a wood.
Young, the young man
Walks, a lover now,
Knows the streets as they are and is of them.

Comes to the wood.

There, in the wood, among trees,
He sees shadows on the way, hears voices,
hears the wind and the rustling of leaves;
Through an open glade
He sees a shape and the shape hears:
It waits as he waits,
Waits as the shadows wait,
As the voices wait;
Shadows on the way, voices in the wind.

----- Generated Poem 3 -----

The sun was all we had. Now, in the shade
All is changed. The mind must dwell on those
White fields, that to its eyes were always old;
Those ancient gleams, convoluting
The way they lay among the huge roots,
The great dark tomes of reverie,
The plumed labyrinth of the sea.

When our own breath gonue faint and fade
Up the long stairway of the dead
We shall remember well
The blindness of the light, the whiteness
Of that white land. The footsteps, and
The grasses in that light, and in the shade
The menacing life.

We played, too, once, in the warmest rooms.
To one content, there is one content
For all mankind. The forest glades
Are all the more for being in fable,
And wood and town in legend shadowed-
Once more to return. The one content
Has in it every happiness, more brave
For being purposeless; and so
Those island gardens, deep in the mind,
Make of remoteness familiarity
And of the unspoken name a friend.

----- Generated Poem 4 -----

Nobody will come to this place. It is a road that leads nowhere.
The solitude is deep. The mountains are high.
But they are desolate, and they turn the traveler's face
Towards the North. All the sounds of the world are far away.
When the wind rises above the trees,
The boughs bow to the ground.
Even the birds that inhabit the tangle of weeds
That is the roadside cover, are silent. One listens,
But hears no roar of the forest. One is alone.
One will be taken.
One will be taken.
There is no utterance, there is no conversation,
But one is uneasy all the same....
There is a thin blue mist,
A darkness rising like smoke,
And within that darkness
A possession of the heart.
One will be taken.... It was here, and it will be here again-
Here, under this sky empty and full of light.

GPT3 results: More task examples

This prompt translates text into programmatic commands.

Prompt

Q: Ask Constance if we need some bread
A: send-msg 'find constance' Do we need some bread?
Q: Send a message to Greg to figure out if things are ready for Wednesday.
A: send-msg 'find greg' Is everything ready for Wednesday?
Q: Ask Ilya if we're still having our meeting this evening
A: send-msg 'find ilya' Are we still having a meeting this evening?
Q: Contact the ski store and figure out if I can get my skis fixed before I leave on Thursday
A: send-msg 'find ski store' Would it be possible to get my skis fixed before I leave on Thursday?
Q: Thank Nicolas for lunch
A: send-msg 'find nicolas' Thank you for lunch!
Q: Tell Constance that I won't be home before 19:30 tonight – unmovable meeting.
A: send-msg 'find constance' I won't be home before 19:30 tonight. I have a meeting I can't move.
Q: Let Jessie know I'll be at the meeting.
A:

Sample response

send-msg 'find jessie' I'll be at the meeting.

<https://beta.openai.com/examples>

<https://machinelearningknowledge.ai/openai-gpt-3-demos-to-convince-you-that-ai-threat-is-real-or-is-it/>

<https://www.educative.io/blog/top-uses-gpt-3-deep-learning>

<https://bdtechtalks.com/2021/07/05/openai-github-gpt-3-copilot/> and <https://copilot.github.com/>

Create tables from long form text by specifying a structure and supplying some examples.

Prompt

There are many fruits that were found on the recently discovered planet Goocrux. There are neoskizzles that grow there, which are purple and taste like candy. There are also loheckles, which are a grayish blue fruit and are very tart, a little bit like a lemon. Pounits are a bright green color and are more savory than sweet. There are also plenty of loopnovas which are a neon pink flavor and taste like cotton candy. Finally, there are fruits called glowls, which have a very sour and bitter taste which is acidic and caustic, and a pale orange tinge to them.

Please make a table summarizing the fruits from Goocrux

Fruit	Color	Flavor
Neoskizzles	Purple	Sweet
Loheckles	Grayish blue	Tart

Sample response

Pounits	Bright green	Savory
Loopnovas	Neon pink	Cotton candy
Glowls	Pale orange	Sour
Other		

Movie to Emoji



Transformation Generation

Convert movie titles into emoji.

Prompt

Back to Future: 🧑‍🤝‍🧑 🚗 🕒

Batman: 🕸️ 🦇

Transformers: 🚗 🛡️

Wonder Woman: 🙀 🤷‍♀️ 🤷‍♂️ 🎖️ 🏴

Winnie the Pooh: 🐻 🐻 🐻 🐻

The Godfather: 🤵 🤵 🤵 🤵 🤵 🎬

Game of Thrones: 🏰 🏰 🏰 🏰 🏰 🎬

Spider-Man:

Sample response

The Incredible Hulk: 💪

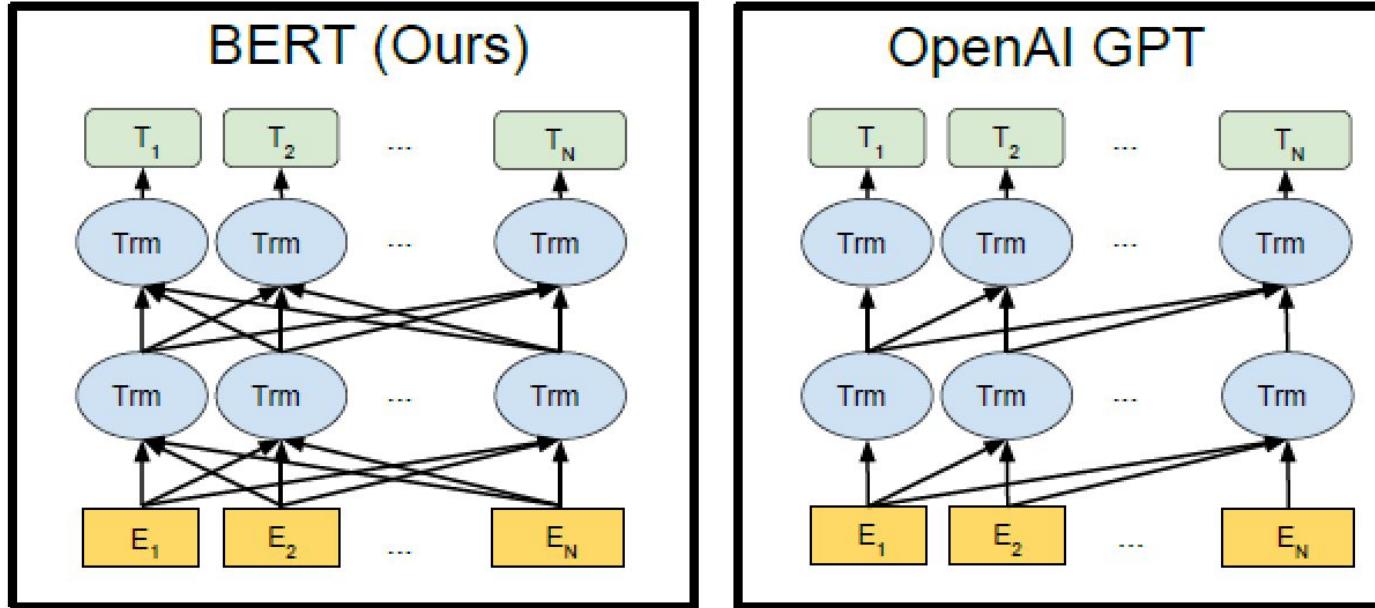
GPT3 limitations

- GPT-3 samples sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, and contradict themselves.
- Model is not bidirectional.
- During pre-training it sees way more than a human sees in their lifetime.
- Large size: inconvenient to perform inference on.
- Biases in the data leads to offensive content generation.

Agenda

- Word Embeddings
- Transformers
- GPT
- BERT

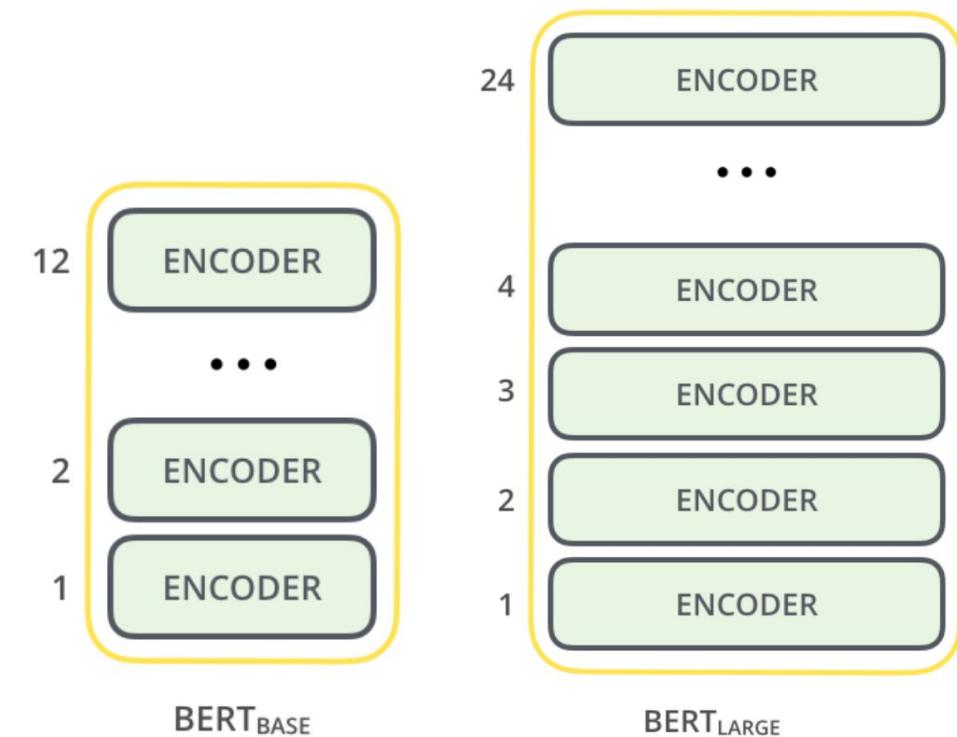
Why BERT?



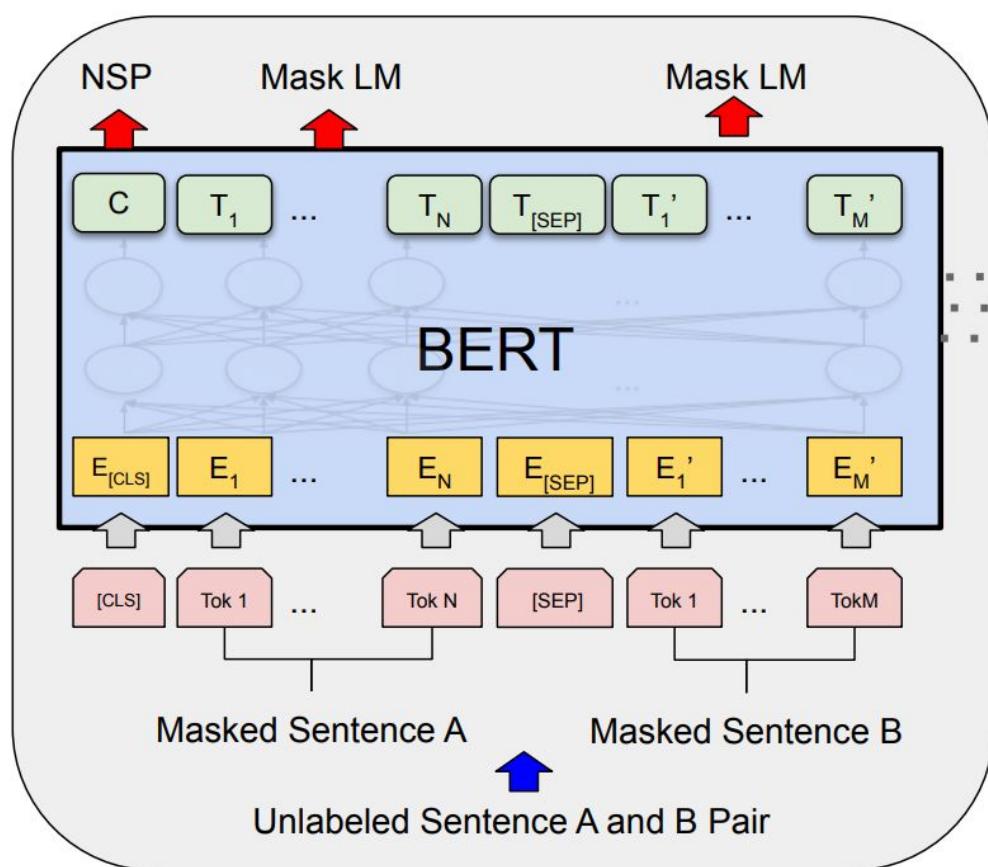
- BERT's model architecture is a multi-layer bidirectional Transformer encoder
- OpenAI GPT uses a left-to-right Transformer.
- BERT representations are jointly conditioned on both left and right context in all layers.

BERT Model Architecture

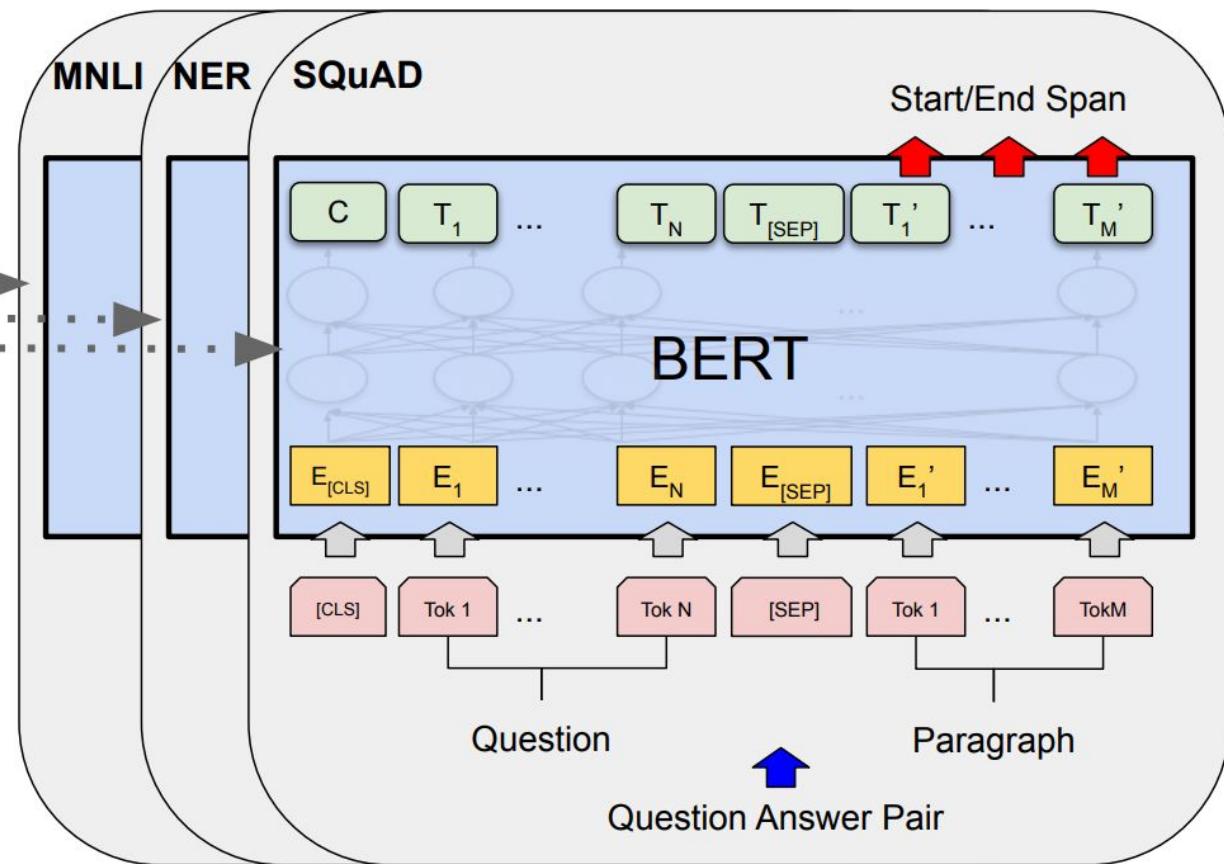
- 2 steps in BERT framework: pre-training and fine-tuning.
- Pre-training: unlabeled data for MLM+NSP.
- 2 popular variants
 - BERT_BASE (L=12, H=768, A=12, Total Parameters=110M)
 - BERT_LARGE (L=24, H=1024, A=16, Total Parameters=340M).



BERT Model Architecture



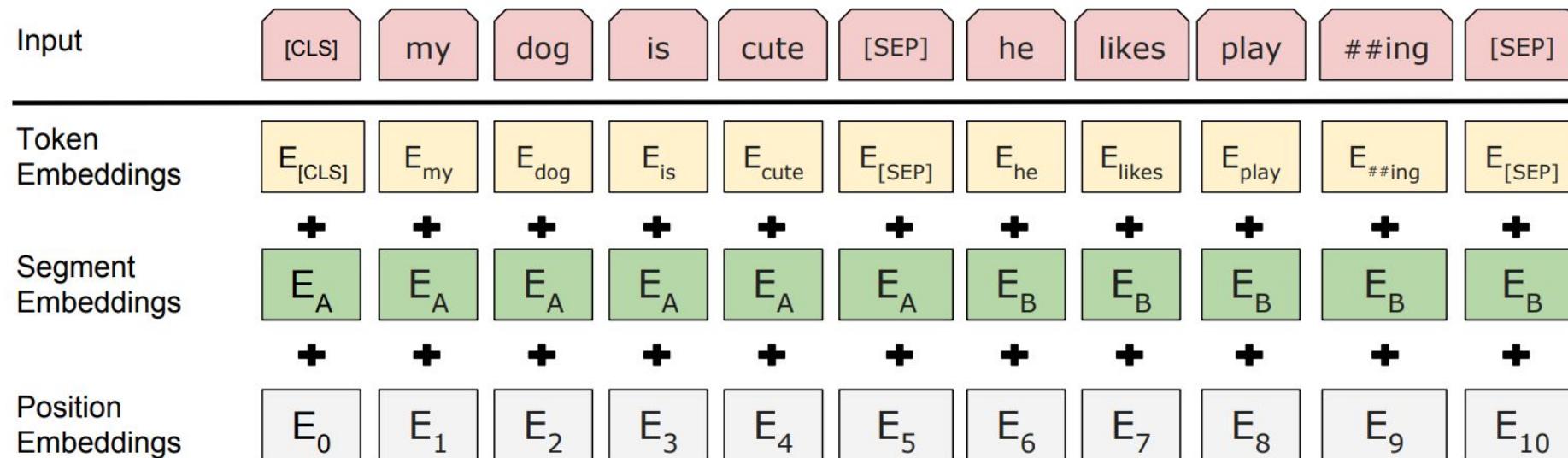
Pre-training



Fine-Tuning

Input/Output Representations

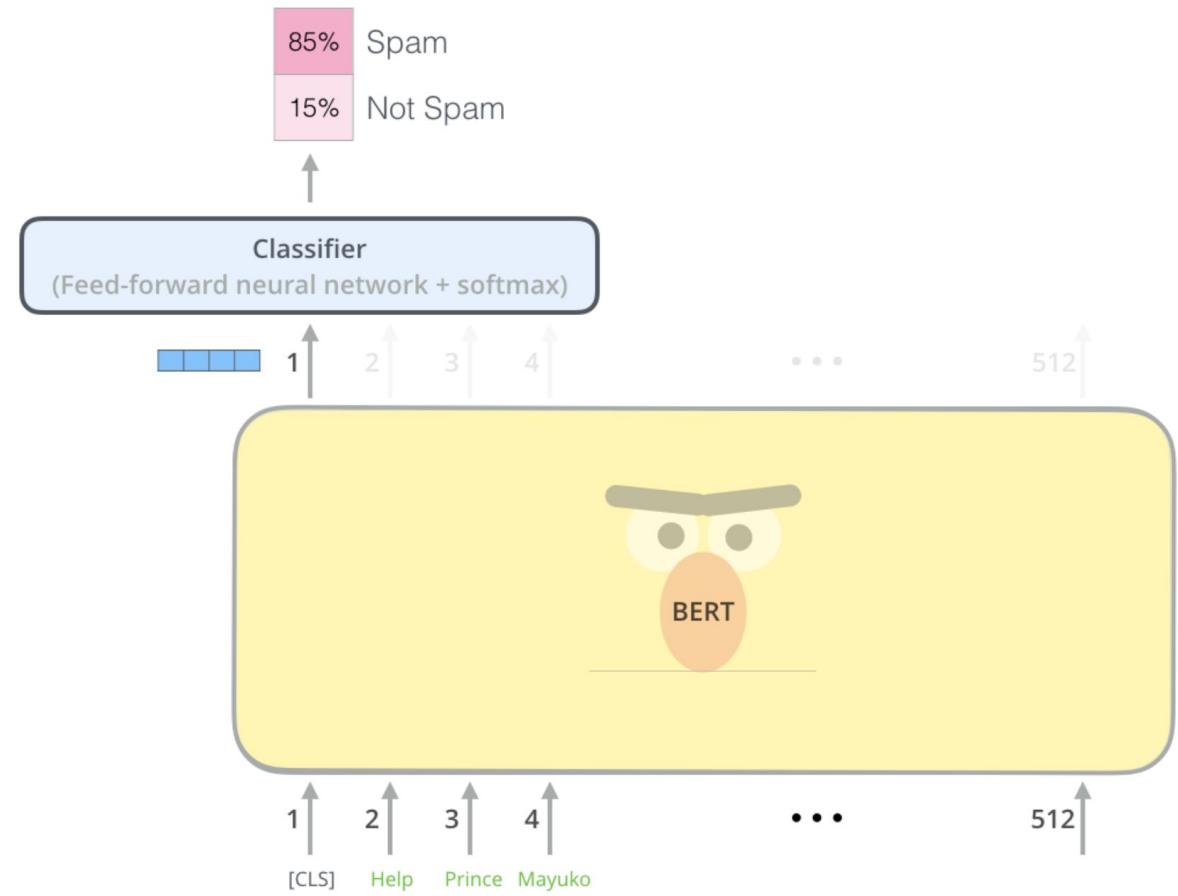
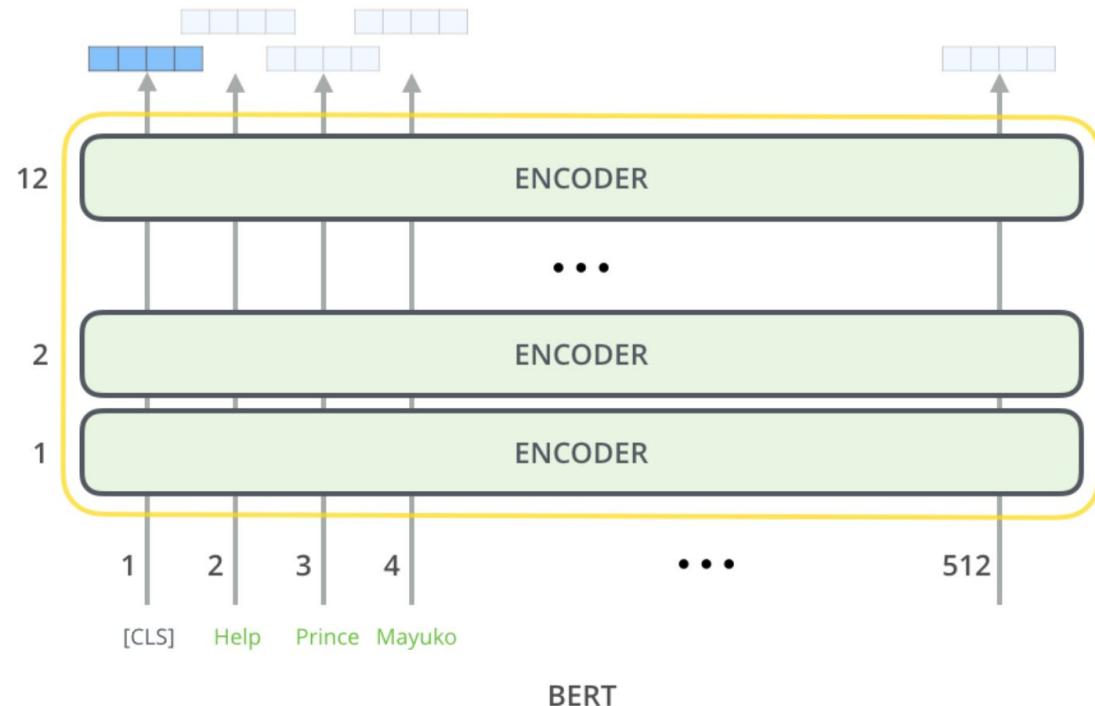
WordPiece embeddings with a 30,000 token vocabulary.



Pre-training BERT using two unsupervised tasks

- Task #1: Masked LM (also called Cloze task)
 - MLM: mask 15% of the input tokens at random, and then predict those masked tokens.
 - Consider: “my dog is hairy” and if chosen word is “hairy”
 - 80% of the time: Replace the word with the [MASK] token, e.g., my dog is hairy → my dog is [MASK]
 - 10% of the time: Replace the word with a random word, e.g., my dog is hairy → my dog is apple
 - 10% of the time: Keep the word unchanged, e.g., my dog is hairy → my dog is hairy. The purpose of this is to bias the representation towards the actual observed word.
- Task #2: Next Sentence Prediction (NSP)
 - Input = [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]
 - Label = IsNext
 - Input = [CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]
 - Label = NotNext
- Pre-training data: BooksCorpus (800M words) and English Wikipedia (2,500M words).

How do you use BERT for Sentence Classification?



Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL, pp. 4171-4186. 2019.

Agenda

- Word Embeddings
- Transformers
- GPT
- BERT
- T5

Text to text transfer transformer



- Treat every NLP problem as a “text-to-text” problem, i.e. taking text as input and producing new text as output.
- “Colossal Clean Crawled Corpus” (C4): Clean English Web text. ~750 GB
 - Pre-train on $2^{35} \approx 34B$ tokens
- Use a task-specific (text) prefix to the original input sequence before feeding it to the model.
- For text classification tasks, the model simply predicts a single word corresponding to the target label.

Masking objectives

- Masking

- The target corresponds to all of the dropped-out spans of tokens, delimited by the same sentinel tokens used in the input sequence plus a final sentinel token to mark the end of the target sequence.
- Note only the corrupted tokens are used as targets.

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	(original text)
Deshuffling	party me for your to . last fun you inviting week Thank	(original text)
I.i.d. noise, mask tokens	Thank you <M> <M> me to your party <M> week .	(original text)
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>

Table 3: Examples of inputs and targets produced by some of the unsupervised objectives we consider applied to the input text “Thank you for inviting me to your party last week.” Note that all of our objectives process *tokenized* text. For this particular sentence, all words were mapped to a single token by our vocabulary. We write *(original text)* as a target to denote that the model is tasked with reconstructing the entire input text. <M> denotes a shared mask token and <X>, <Y>, and <Z> denote sentinel tokens that are assigned unique token IDs. The BERT-style objective (second row) includes a corruption where some tokens are replaced by a random token ID; we show this via the greyed-out word *apple*.

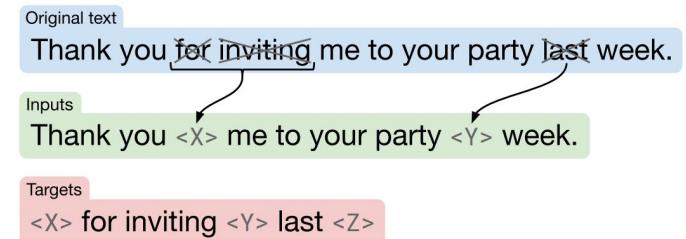
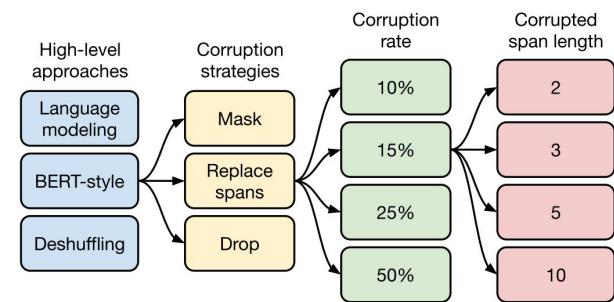


Figure 2: Schematic of the objective we use in our baseline model. In this example, we process the sentence “Thank you for inviting me to your party last week.” The words “for”, “inviting” and “last” (marked with an \times) are randomly chosen for corruption. Each consecutive span of corrupted tokens is replaced by a sentinel token (shown as $\langle X \rangle$ and $\langle Y \rangle$) that is unique over the example. Since “for” and “inviting” occur consecutively, they are replaced by a single sentinel $\langle X \rangle$. The output sequence then consists of the dropped-out spans, delimited by the sentinel tokens used to replace them in the input plus a final sentinel token $\langle Z \rangle$.



T5 Results

- Model sizes
 - Base (220M params)
 - Small (dmodel = 512, dff = 2,048, A=8, 6L+6L. 60M params)
 - Large (dmodel = 1,024, dff = 4,096, A=16, 12L+12L. 770M params)
 - 3B and 11B: dmodel = 1024, 24L+24L.
 - 3B: dff = 16,384 with A=32. 2.8B params.
 - 11B: dff = 65,536 with A=128. 11B params.
Model size = 21.7GB
- Overall, 11B model achieved SOTA on 17 out of the 24 tasks.
- SOTA GLUE of 89.7.
- SOTA SuperGLUE of 84.6.

Model	GLUE Average	CoLA Matthew's	SST-2 Accuracy	MRPC F1	MRPC Accuracy	STS-B Pearson	STS-B Spearman
Previous best	89.4 ^a	69.2 ^b	97.1^a	93.6^b	91.5^b	92.7^b	92.3^b
T5-Small	77.4	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	89.7	70.8	97.1	91.9	89.2	92.5	92.1
Model	QQP F1	QQP Accuracy	MNLI-m Accuracy	MNLI-mm Accuracy	QNLI Accuracy	RTE Accuracy	WNLI Accuracy
Previous best	74.8^c	90.7^b	91.3 ^a	91.0 ^a	99.2^a	89.2 ^a	91.8 ^a
T5-Small	70.0	88.0	82.4	82.3	90.3	69.9	69.2
T5-Base	72.6	89.4	87.1	86.2	93.7	80.1	78.8
T5-Large	73.9	89.9	89.9	89.6	94.8	87.2	85.6
T5-3B	74.4	89.7	91.4	91.2	96.3	91.1	89.7
T5-11B	74.6	90.4	92.0	91.7	96.7	92.5	93.2
Model	SQuAD EM	SQuAD F1	SuperGLUE Average	BoolQ Accuracy	CB F1	CB Accuracy	COPA Accuracy
Previous best	88.95 ^d	94.52 ^d	84.6 ^e	87.1 ^e	90.5 ^e	95.2 ^e	90.6 ^e
T5-Small	79.10	87.24	63.3	76.4	56.9	81.6	46.0
T5-Base	85.44	92.08	76.2	81.4	86.2	94.0	71.2
T5-Large	86.66	93.79	82.3	85.4	91.6	94.8	83.4
T5-3B	88.53	94.95	86.4	89.9	90.3	94.4	92.0
T5-11B	90.06	95.64	88.9	91.0	93.0	96.4	94.8
Model	MultiRC F1a	MultiRC EM	ReCoRD F1	ReCoRD Accuracy	RTE Accuracy	WiC Accuracy	WSC Accuracy
Previous best	84.4 ^e	52.5 ^e	90.6 ^e	90.0 ^e	88.2 ^e	69.9 ^e	89.0 ^e
T5-Small	69.3	26.3	56.3	55.4	73.3	66.9	70.5
T5-Base	79.7	43.1	75.0	74.2	81.5	68.3	80.8
T5-Large	83.3	50.7	86.8	85.9	87.8	69.3	86.3
T5-3B	86.8	58.3	91.2	90.4	90.7	72.1	90.4
T5-11B	88.2	62.3	93.3	92.5	92.5	76.1	93.8
Model	WMT EnDe BLEU	WMT EnFr BLEU	WMT EnRo BLEU	CNN/DM ROUGE-1	CNN/DM ROUGE-2	CNN/DM ROUGE-L	
Previous best	33.8^f	43.8^f	38.5^g	43.47 ^h	20.30 ^h	40.63 ^h	
T5-Small	26.7	36.0	26.8	41.12	19.56	38.35	
T5-Base	30.9	41.2	28.0	42.05	20.34	39.40	
T5-Large	32.0	41.5	28.1	42.50	20.68	39.75	
T5-3B	31.8	42.6	28.2	42.72	21.02	39.94	
T5-11B	32.1	43.4	28.1	43.52	21.55	40.69	

Fill-in-the-Blank Text Generation using T5

- I love peanut butter and *jelly* sandwiches.
- I love peanut butter and *jelly on my* sandwiches.
- I love peanut butter and *jelly, which is what makes good* sandwiches.
- I love peanut butter and *jelly, Yum! You can't beat peanut butter and jelly* sandwiches.
- I love peanut butter and *bread. Thanks!! This looks delicious. I love all types of peanut butter, but especially peanut butter/jam* sandwiches.
- I love peanut butter and *banana sandwiches and the Peanut Butter Chocolate Chip Cookie Bites are now very easy to prepare. The best part is that the PB&J Cookie Bites are the perfect size for kid lunch boxes and everyone will enjoy them. Kids love these sandwiches.*
- I love peanut butter and *jelly! It's like an incredible childhood love now that I'm in my 30s. I'm obsessed with peanut butter and jelly sandwiches. I have made this recipe many times with great success. I'm guessing you could probably find that it's from a 1950s cookbook. I haven't seen it listed in a current one but it is in a 1930 version (they seem to have retired it). Hope you enjoy it ;) I hope you'll post what you do with your sandwiches.*
- I love peanut butter and *banana sandwiches. Peanut butter and bananas were the original sandwich spread (also known as PB&J or Peanut Butter and Jelly) so they are probably my favorite. When I was a kid, I knew what peanut butter and bananas tasted like, but I didn't really think of them as one flavor. I did recognize PB & J's as just a sandwich spread, and not really two separate flavours. When I was in graduate school researching food allergies and intolerances, PB&B is one of the main themes that came up, PB&B was and still is the favorite sandwich spread on my school cafeteria tray, and it's still on my personal board in the kitchen. I've seen PB&B on more than one occasion over the years. I can always find one of the other types of peanut butter spread at the school cafeteria and I find it's the perfect spread for PB&B. You have several options for making your own peanut butter (I love this recipe from Cookie and Kate) or buying it at stores like Trader Joes, Publix, and Safeway. The flavor of PB&B varies, but I've found that white chocolate is my favorite to use. This simple sandwich is perfect for lunch or an afternoon snack. When you're on the go, you can use a store bought product, but it just wouldn't taste as good. These peanut butter n' banana sandwich spreads (as well as other peanut butter spreads) are great for on the go eating, but they can also be refrigerated for a few hours and then served at your next party. Preheat oven to 350 degrees. Using a kitchen timer, or using a microwave, heat butter in a saucepan and melt over low heat. Add vanilla extract and milk and heat until nearly boiling. Add peanut butter, bananas, sugar, salt, and pepper, mixing well to combine all ingredients. Spread peanut butter mixture on both sides of your bread. Top with banana slices and a slice of cheese. Toast the bread until both sides are golden brown. Remove from toast and sprinkle a little more sugar and peanut butter on the top. Enjoy while the bread is warm. Store in an airtight container up to one day. Assemble peanut butter and banana sandwich spread by spreading the peanut butter mixture on each slice of bread. Add a banana slice on top and then a PB & J sandwich. Enjoy while the bread is still warm. P.S. You might also like these peanut butter and jelly sandwiches.*

Agenda

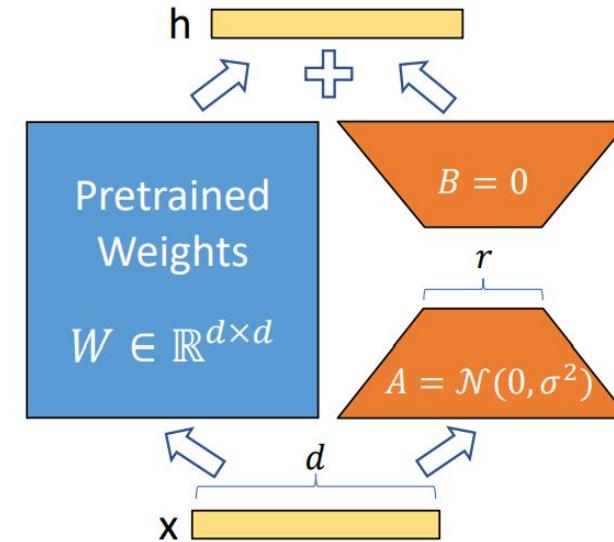
- Word Embeddings
- Transformers
- GPT
- BERT
- T5
- LoRA: Low Rank Adaptation of LLMs

LoRA: Low Rank Adaptation of LLMs

- Why finetuning LLMs is difficult?
 - RoBERTa base (125M), RoBERTa large (355M), DeBERTa XXL (1.5B), GPT-2 Medium (355M), GPT-2 Large (774M), GPT-3 (175B).
 - Finetuning takes a lot of RAM and time.
 - Need to store full model for each task.
 - GPT-3 finetuning needs 1.2TB VRAM!
- LoRA
 - **Significantly fewer** LoRA weights make it easily portable.
 - LoRA makes **training more efficient by up to 3 times**.
 - Less **catastrophic forgetting**.
 - **No extra inference latency**

What is LoRA?

- Low-Rank Adaptation
- Freeze the pretrained model weights and inject trainable rank decomposition matrices into each layer
 - Reparametrization: train A and B only.
- GPT-3 175B finetuning: LoRA reduces #trainable parameters by 10,000 times and the GPU memory requirement by 3 times.
- Performs on-par or better than finetuning in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3.
- No additional inference latency.
- <https://github.com/microsoft/LoRA>



- GPT-3 175B: a very low rank (i.e., $r=1$ or $r=2$) suffices even when the full rank (i.e., d) is as high as 12,288.
- If $r=d$, LoRA \rightarrow finetuning.
- Very popular for finetuning LLMs like instruction-based models including Alpaca and Vicuna.
- Used to tune stable diffusion to adapt the style of generated images.

How can we apply LoRA to Transformers?

- W_0 be a pretrained weight matrix and ΔW be its accumulated gradient update during adaptation.
- For $h = W_0x$, modified forward pass yields: $h = W_0x + \alpha\Delta Wx = W_0x + \alpha BAx$.
 - α is merging ratio.
- Random Gaussian initialization for A and zero for B, so $\Delta W = BA$ is zero at the beginning of training.
- Transformer: four weight matrices in the self-attention module (W_q, W_k, W_v, W_o) and two in the MLP module.
- Adapt only attention weights; freeze the MLP modules.
- GPT-3 175B: Reduce VRAM consumption during training from 1.2TB to 350GB. With $r = 4$ and only W_q and W_v being adapted, the checkpoint size is reduced by $\sim 10,000\times$ (from 350GB to 35MB).
- Storing 100 adapted models only requires $350\text{GB} + 35\text{MB} * 100 \approx 354\text{GB}$ as opposed to $100 * 350\text{GB} \approx 35\text{TB}$.
- 25% speedup during training on GPT-3 175B compared to full fine-tuning as we do not need to calculate the gradient for the vast majority of the parameters.

Summary

- Word Embeddings
- Transformers
- GPT
- BERT
- T5
- LoRA: Low Rank Adaptation of LLMs