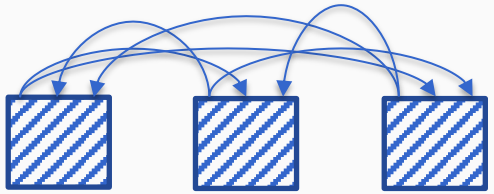
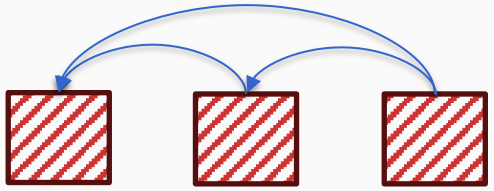
GPT (and decoder models)

Slides credit: Daniel Kashabi, Jay Alammar

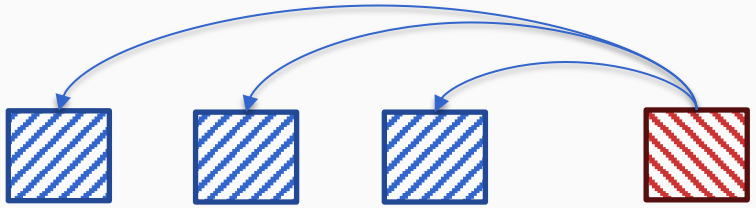
Three different kinds of attention



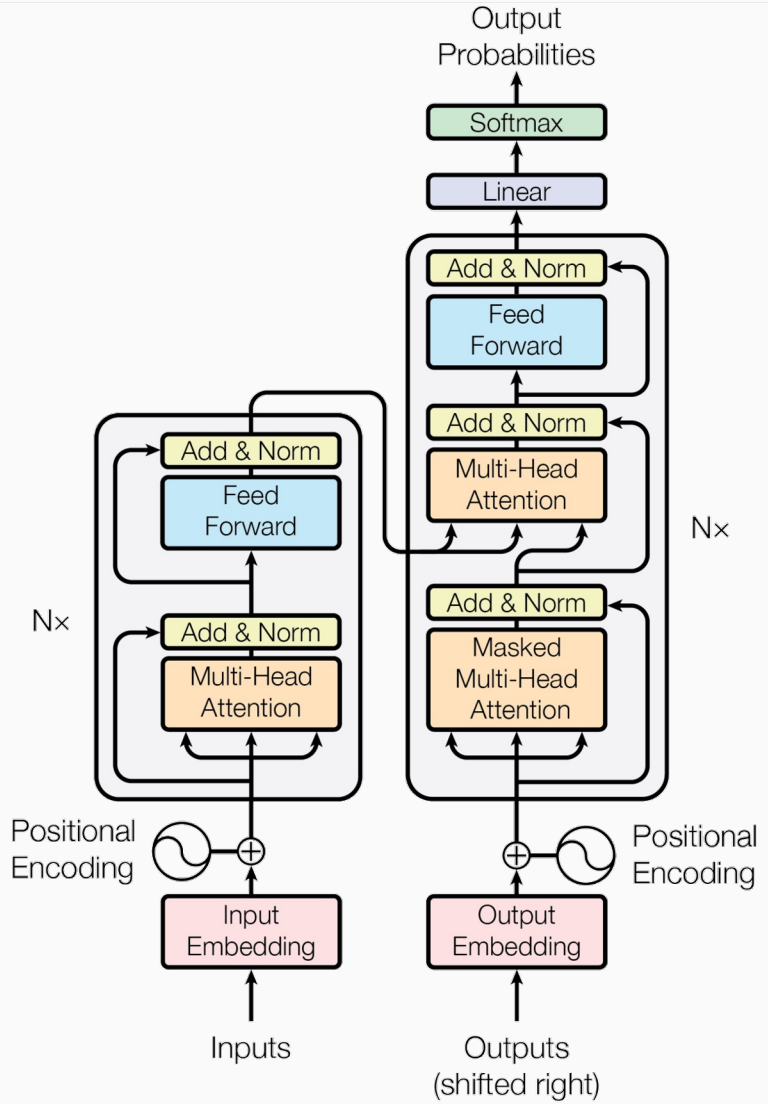
Encoder self-attention



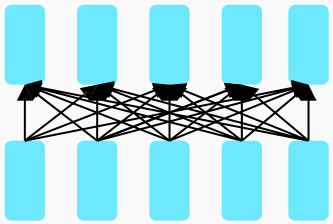
Masked decoder self-attention



Encoder-decoder self-attention



Transformers are the default building blocks for NLP today

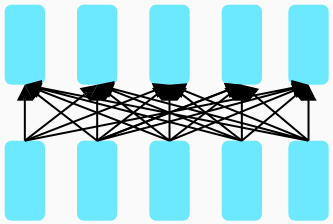


Encoders

Examples: BERT, RoBERTa, SciBERT.

Captures bidirectional context

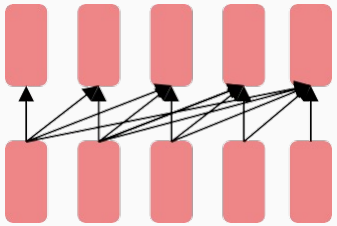
Transformers are the default building blocks for NLP today



Encoders

Examples: BERT, RoBERTa, SciBERT.

Captures bidirectional context



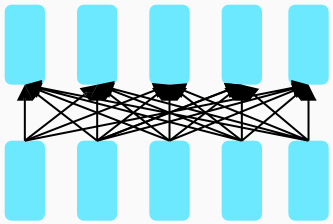
Decoders

Examples: GPT-2, GPT-3, LaMDA

Also known as: causal or auto-regressive language model

Natural if the goal is generation, but can not condition on future words

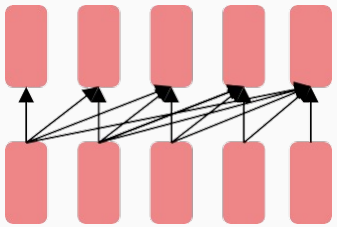
Transformers are the default building blocks for NLP today



Encoders

Examples: BERT, RoBERTa, SciBERT.

Captures bidirectional context

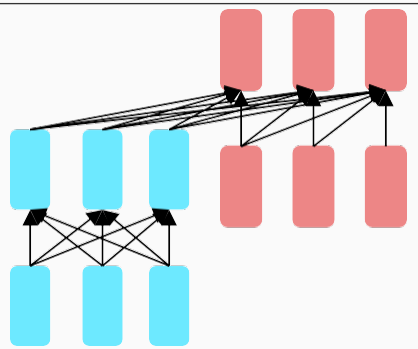


Decoders

Examples: GPT-2, GPT-3, LaMDA

Also known as: causal or auto-regressive language model

Natural if the goal is generation, but can not condition on future words

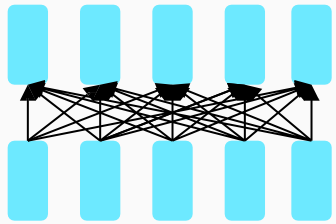


Encoder-Decoders

Examples: BART, T5, Meena

Conditional generation based on an encoded input

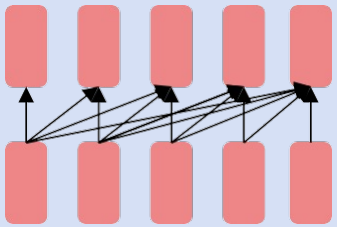
Transformers are the default building blocks for NLP today



Encoders

Examples: BERT, RoBERTa, SciBERT.

Captures bidirectional context

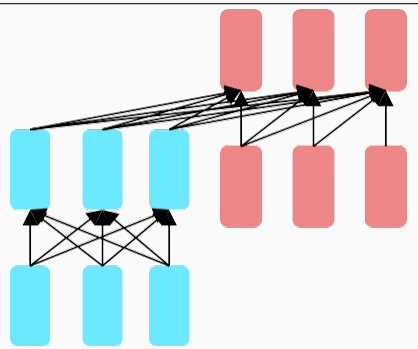


Decoders

Examples: GPT-2, GPT-3, LaMDA

Also known as: causal or auto-regressive language model

Natural if the goal is generation, but can not condition on future words

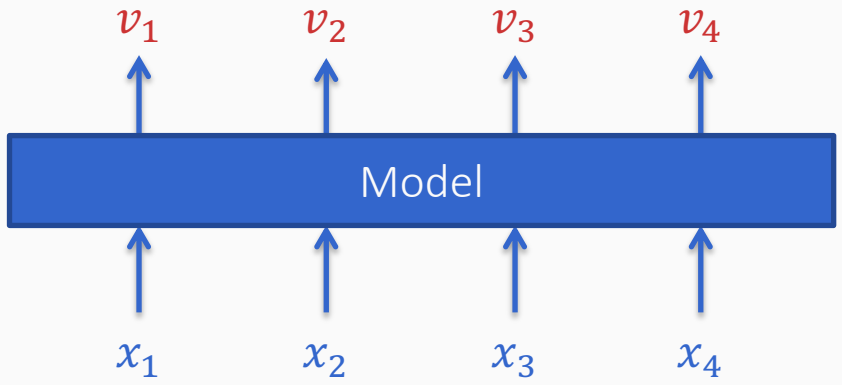


Encoder-Decoders

Examples: BART, T5, Meena

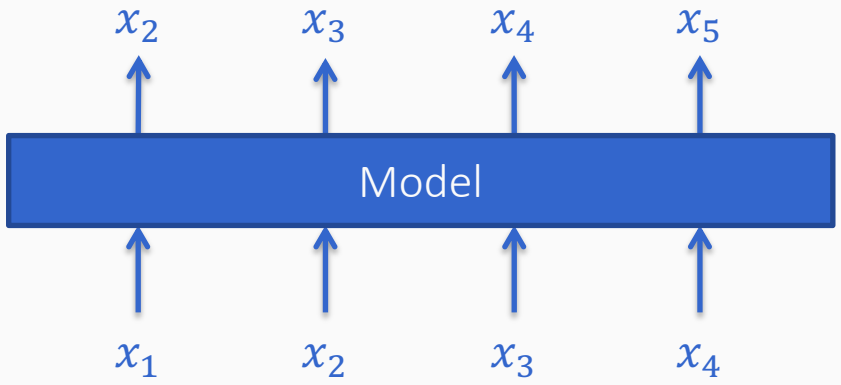
Conditional generation based on an encoded input

Causal or Auto-regressive models



A non-auto-regressive model: Inputs and outputs are different

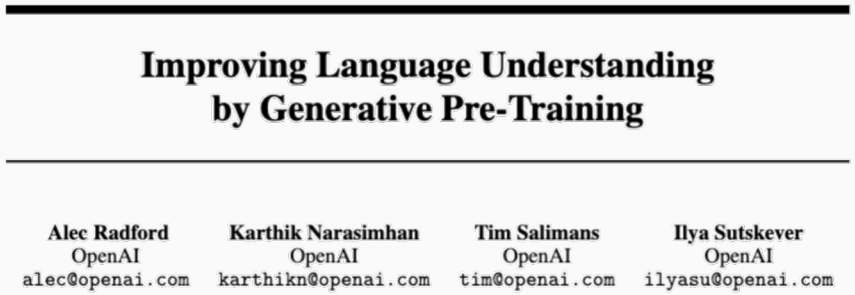
Use case: When we want to assign labels for each word (e.g. part-of-speech tagging)



A causal or an auto-regressive model: Each output is the next input in the sequence

Use case: When we want to generate tokens (e.g. language modeling)

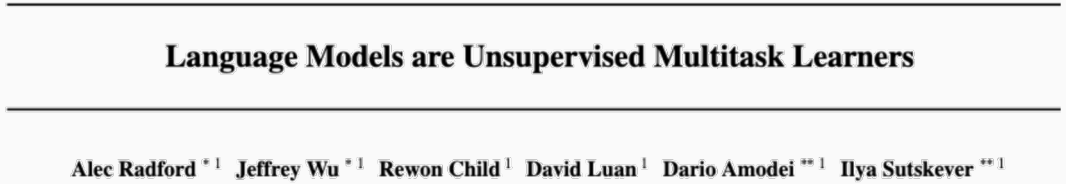
The GPT family



Improving Language Understanding by Generative Pre-Training

Alec Radford, OpenAI, alec@openai.com
Karthik Narasimhan, OpenAI, karthikn@openai.com
Tim Salimans, OpenAI, tim@openai.com
Ilya Sutskever, OpenAI, ilyasu@openai.com

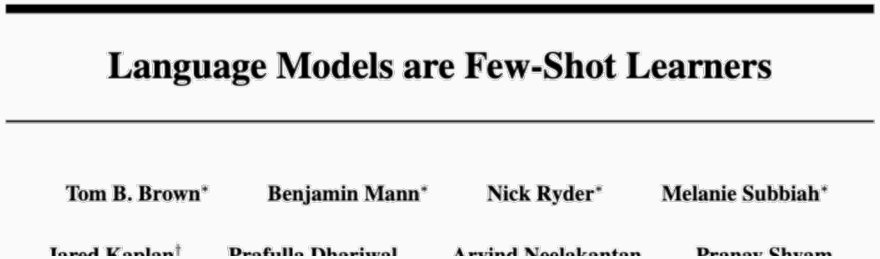
GPT (2018), 117 million parameters



Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

GPT-2 (2019), 1.5 billion parameters

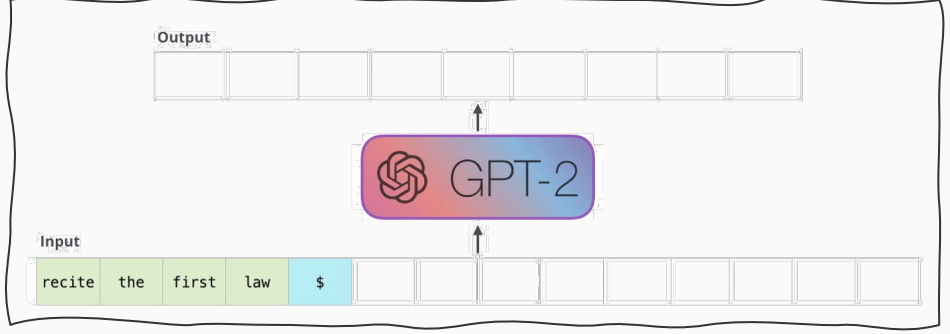


Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

GPT-3 (2020), 175 billion parameters
NeurIPS 2020 best paper

The anatomy of a GPT model



An autoregressive model that predicts the next token given tokens so far (either predicted or given as part of input)

Image from http://jalammar.github.io/illustrated-gpt2/

The anatomy of a GPT model

An autoregressive model that predicts the next token given tokens so far (either predicted or given as part of input)

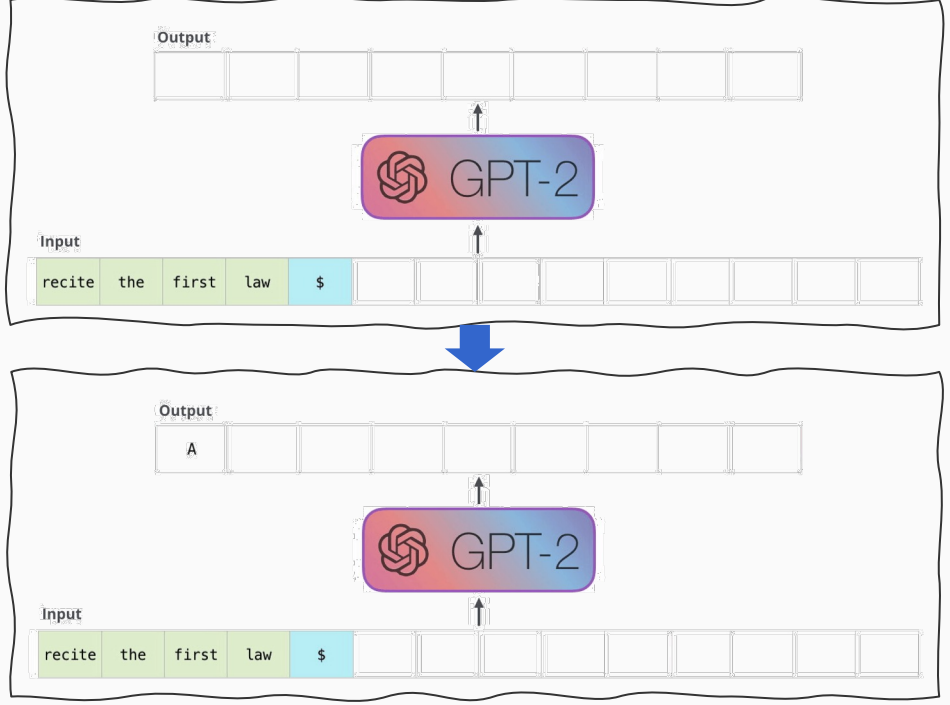


Image from http://jalammar.github.io/illustrated-gpt2/

The anatomy of a GPT model

An autoregressive model that predicts the next token given tokens so far (either predicted or given as part of input)

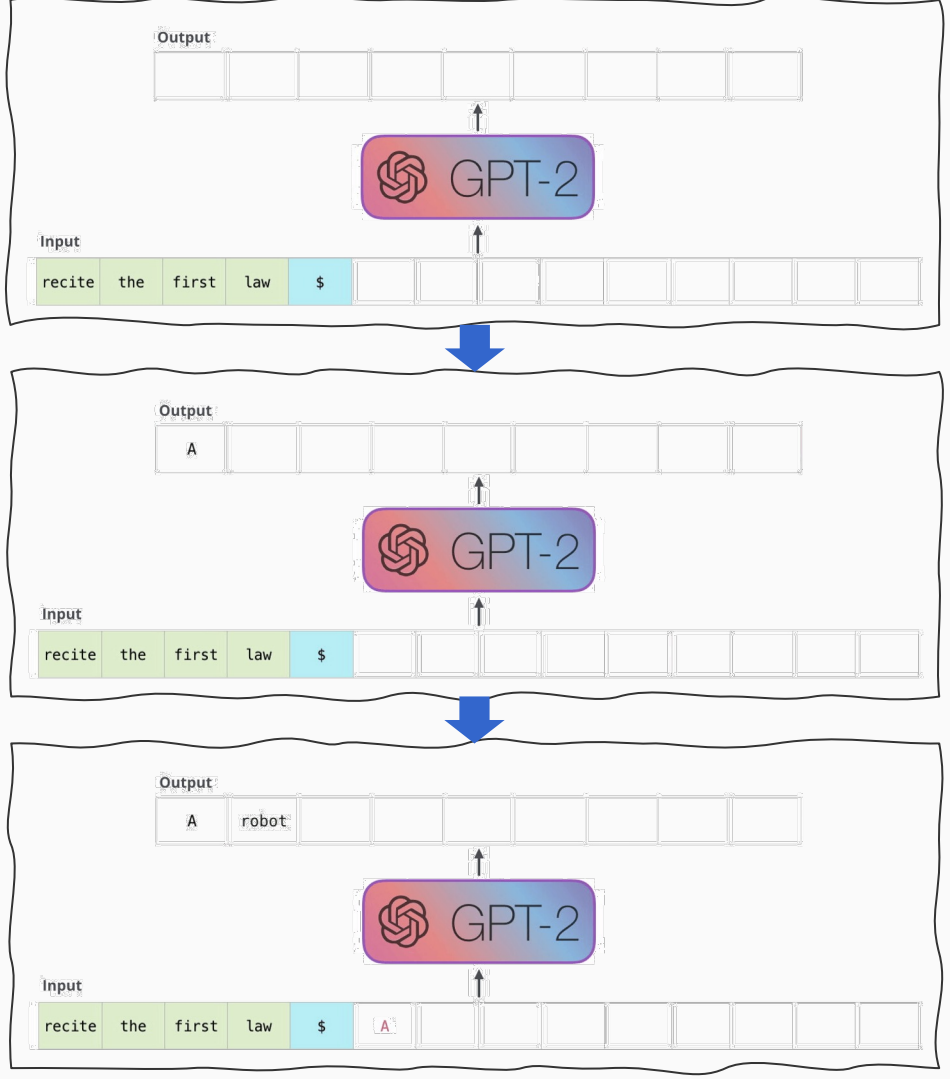
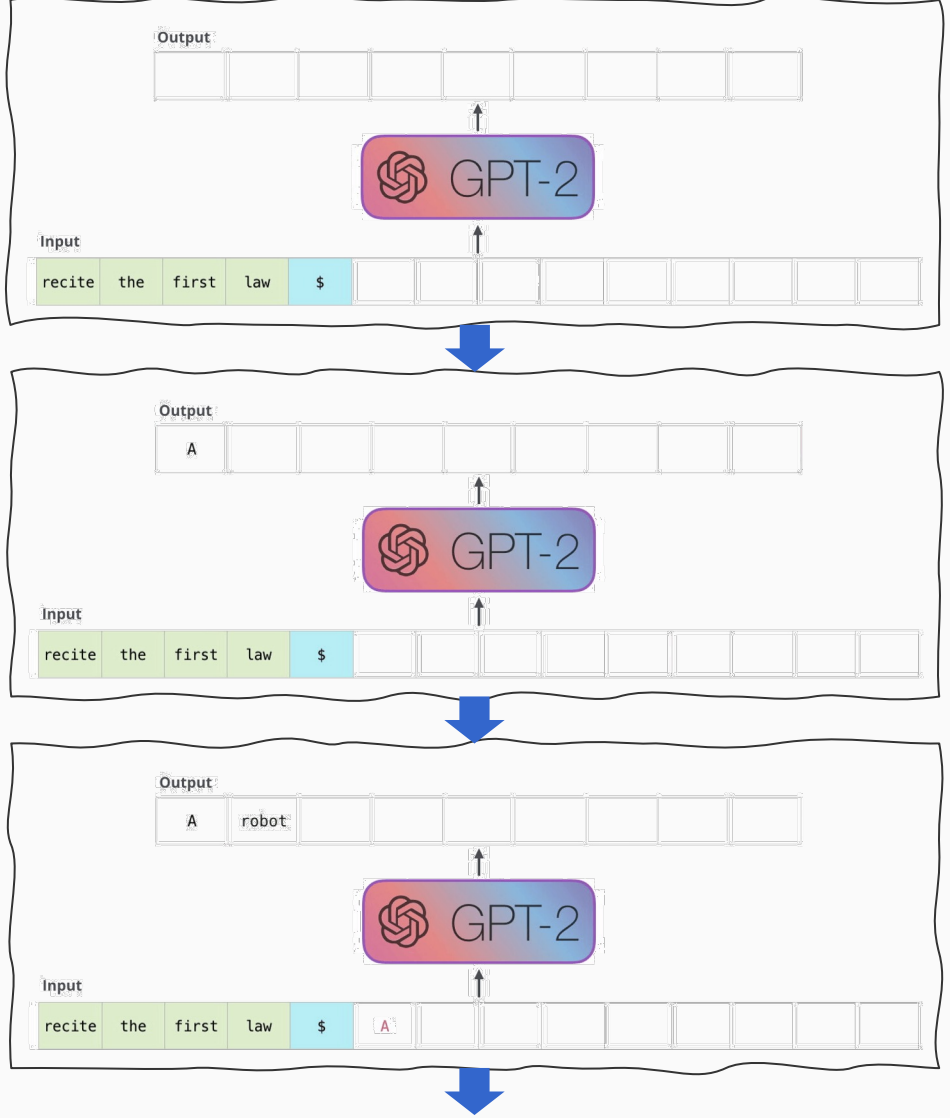


Image from http://jalammar.github.io/illustrated-gpt2/

The anatomy of a GPT model

An autoregressive model that predicts the next token given tokens so far (either predicted or given as part of input)



And so on

Image from http://jalammar.github.io/illustrated-gpt2/

The anatomy of a GPT model

An autoregressive model that predicts the next token given tokens so far (either predicted or given as part of input)

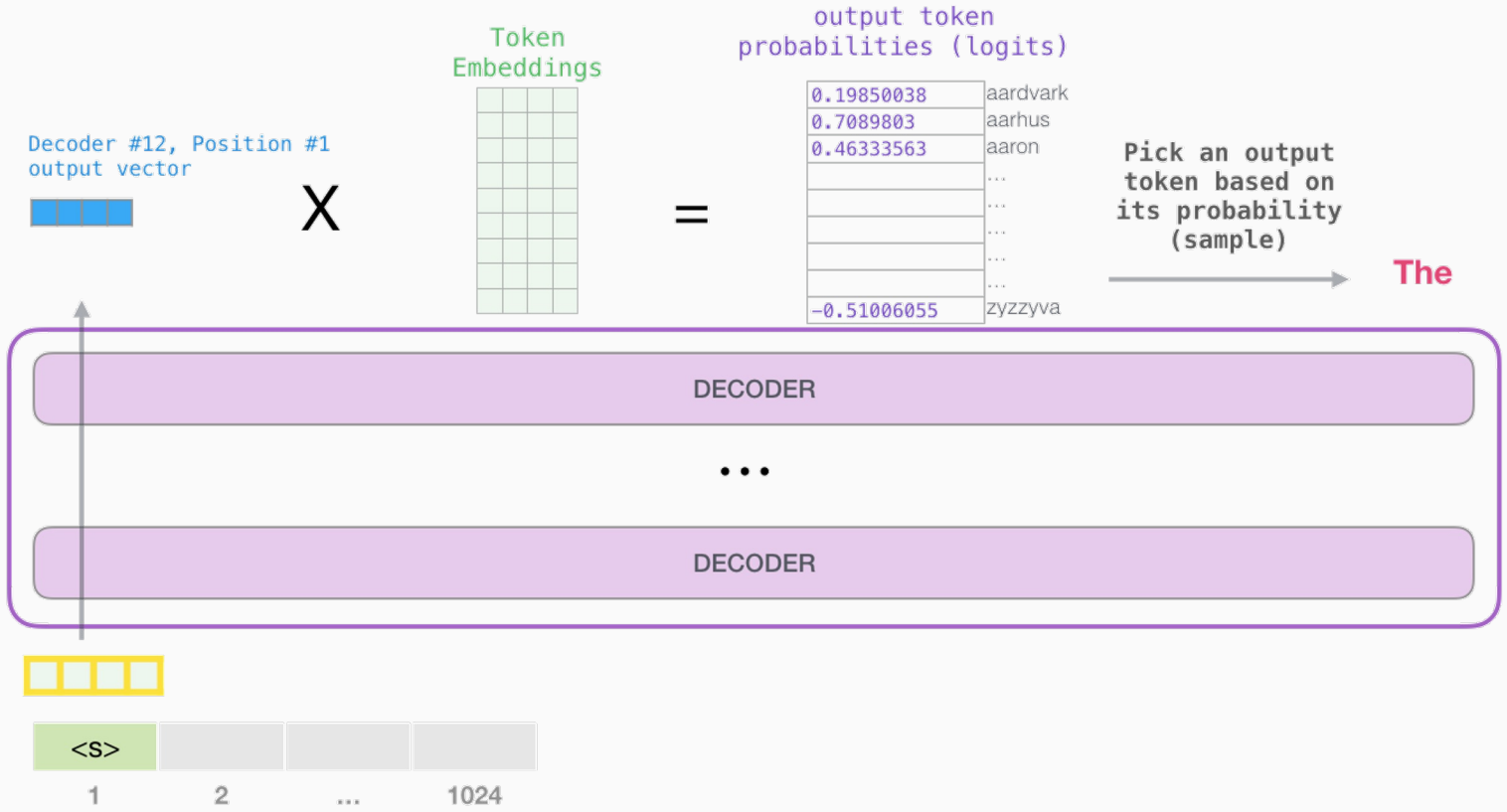


Image from http://jalammar.github.io/illustrated-gpt2/

The anatomy of a GPT model

An autoregressive model that predicts the next token given tokens so far (either predicted or given as part of input)

As it processes each subword, it masks the “future” words and conditions on (i.e. attends to) the previous words

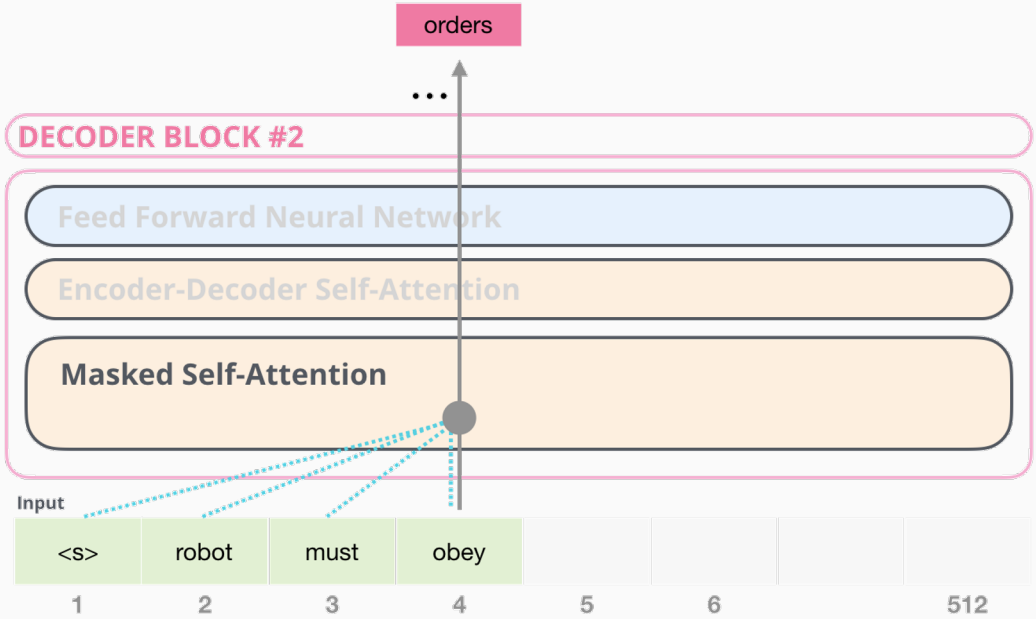
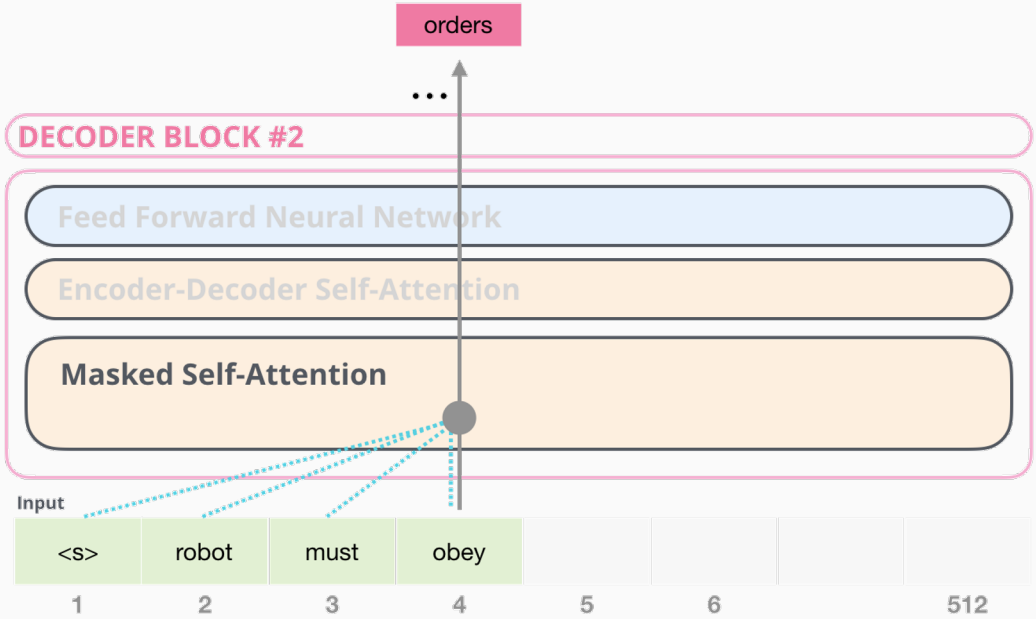


Image from http://jalammar.github.io/illustrated-gpt2/

The anatomy of a GPT model

An autoregressive model that predicts the next token given tokens so far (either predicted or given as part of input)

As it processes each subword, it masks the “future” words and conditions on (i.e. attends to) the previous words



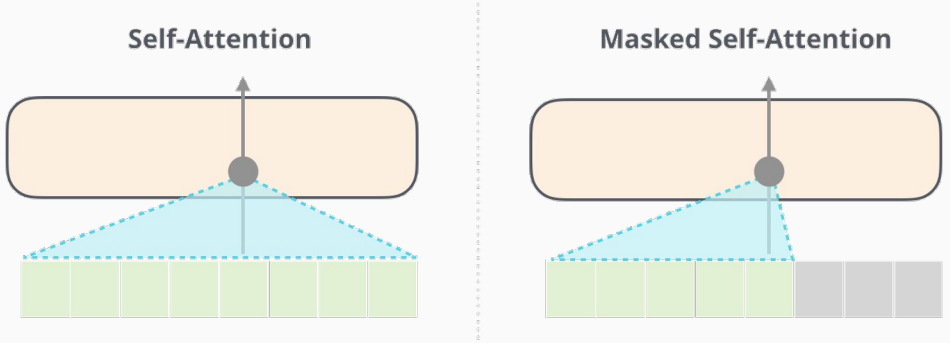


Image from http://jalammar.github.io/illustrated-gpt2/

The anatomy of a GPT model

An autoregressive model that predicts the next token given tokens so far (either predicted or given as part of input)

As it processes each subword, it masks the “future” words and conditions on (i.e. attends to) the previous words

Consists only of decoder transformer blocks (contrast with BERT which consists only of encoders)

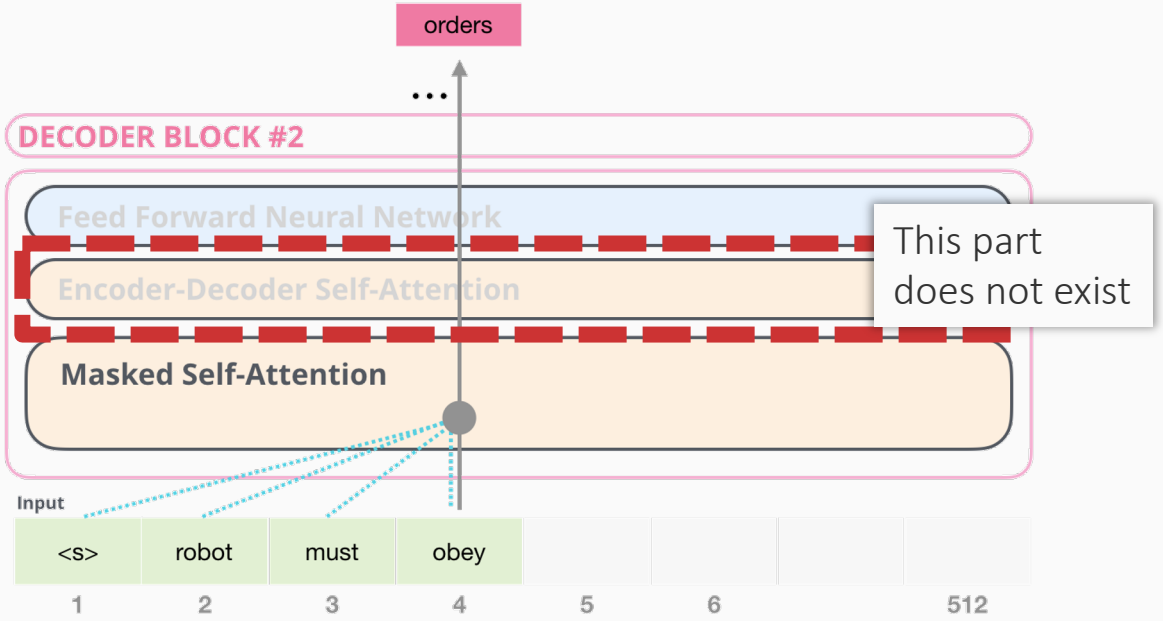


Image from http://jalammar.github.io/illustrated-gpt2/

The anatomy of a GPT model

An autoregressive model that predicts the next token given tokens so far (either predicted or given as part of input)

As it processes each subword, it masks the “future” words and conditions on (i.e. attends to) the previous words

Consists only of decoder transformer blocks (contrast with BERT which consists only of encoders)

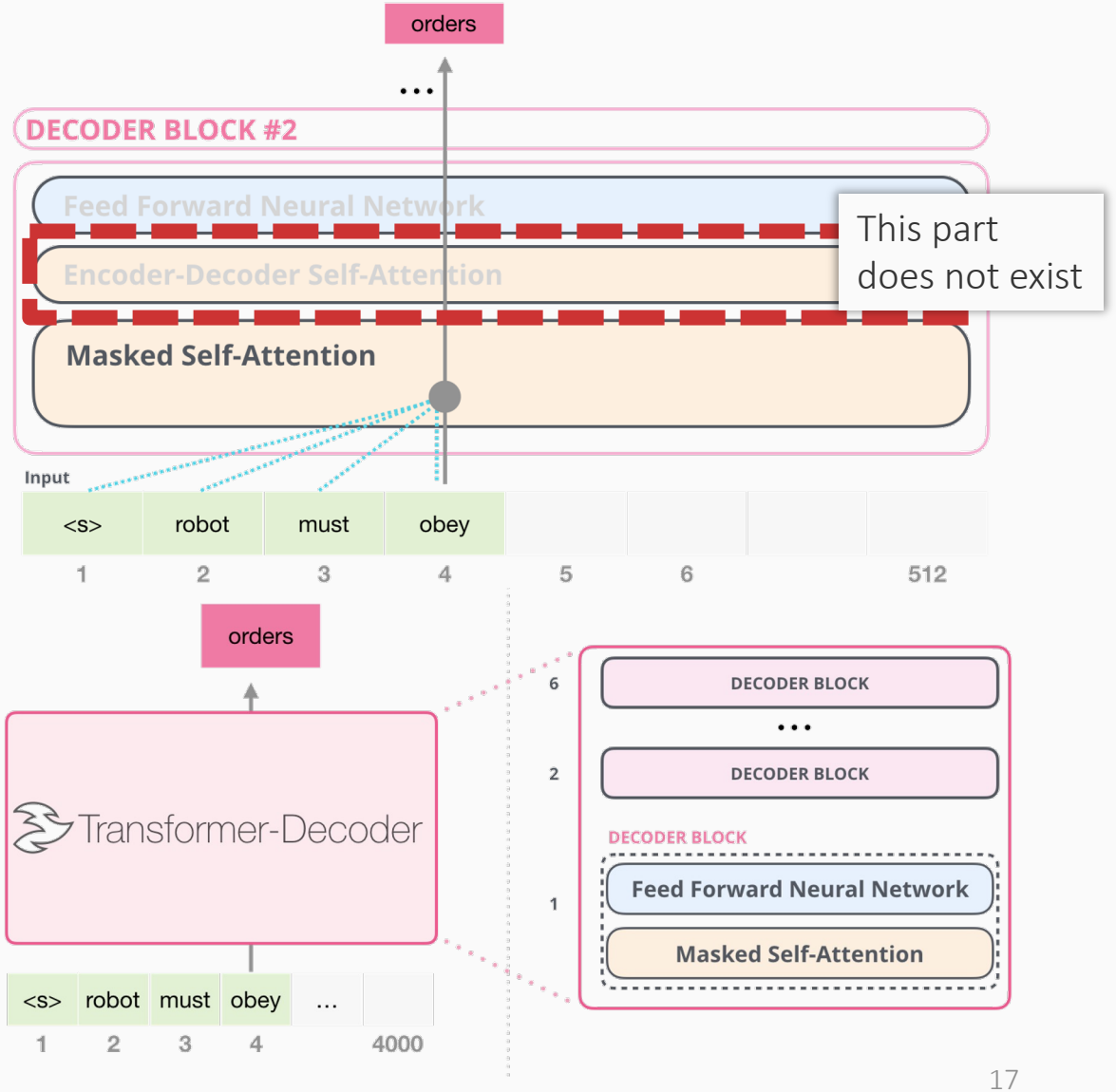
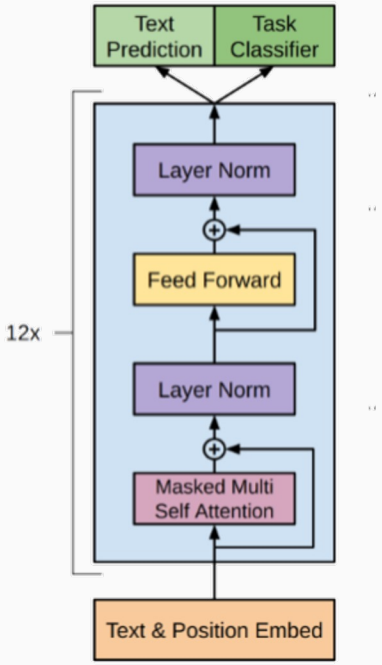


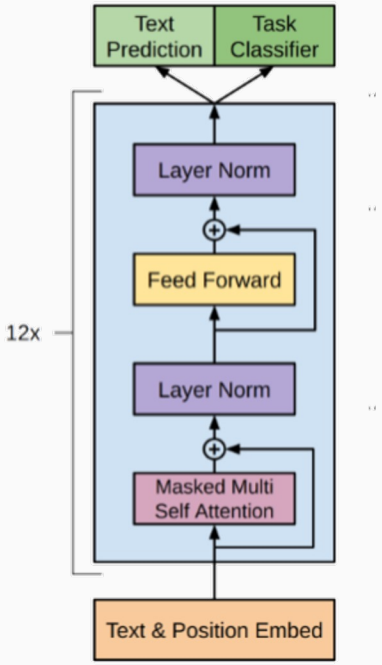
Image from http://jalammar.github.io/illustrated-gpt2/

The first GPT model (sometimes called GPT-1)



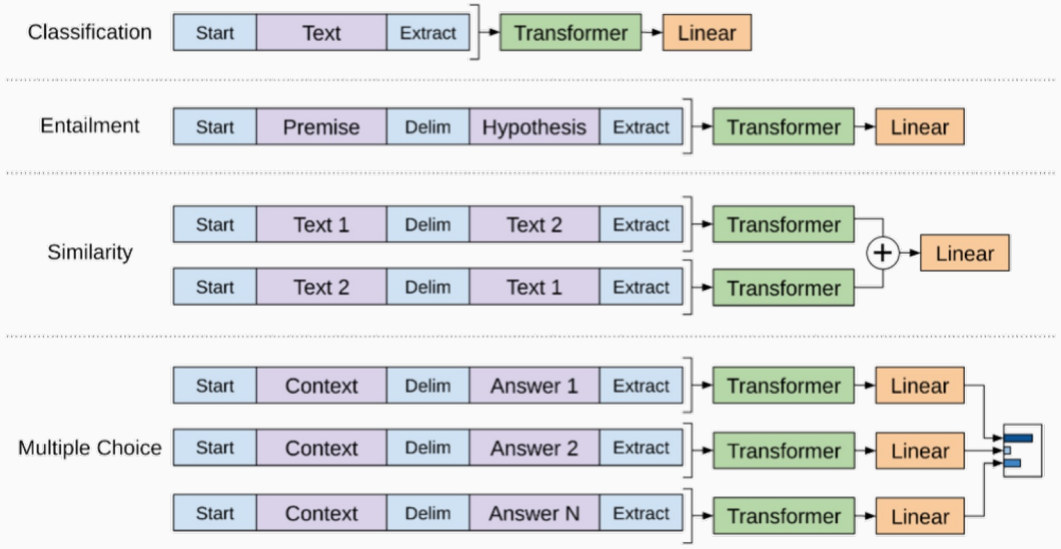
Pretrained on the BooksCorpus

The first GPT model (sometimes called GPT-1)



Pretrained on the BooksCorpus

Also shows results on fine-tuning for end tasks, where inputs and outputs are converted to text



GPT-2 is identical to GPT-1, but:

- Layer norm moved to the input of each sub-block
- Vocabulary extended to 50,257 tokens and context size increased from 512 to 1024
- Trained on 8 million docs from the web (Common Crawl), minus Wikipedia

GPT2: Model Sizes

Play with it here: https://huggingface.co/gpt2

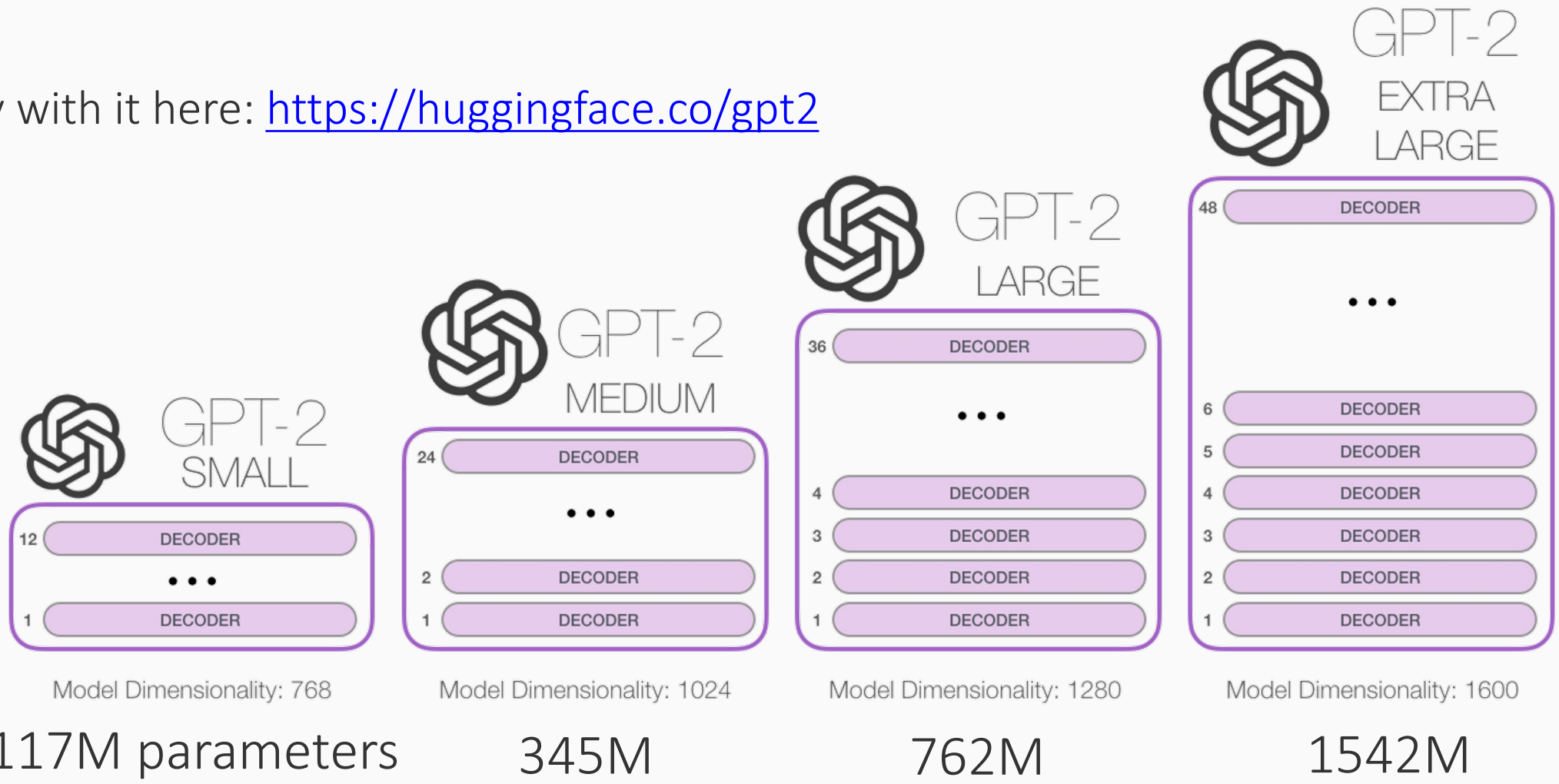


Image from http://jalammar.github.io/illustrated-gpt2/

GPT2: Some Results

Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	56.25	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

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). Other language model results are from (Dai et al., 2019).

GPT-2 completions

Larger models fare better

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 and a linen shirt. Both fabrics (in these colors) are an absolute must for summer, as they go with both dark and 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	**GPT-2 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 de-	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 did 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

GPT-3: More everything

GPT-3: A Very Large Language Model (2020)

- More layers & parameters
- Bigger dataset
- Longer training
- Larger embedding/hidden dimension
- Larger context window

Slide credit: Sbhya Chhabria & Michael Tang

Size Comparisons

- BERT-Base model has 12 transformer blocks, 12 attention heads,
  - 110M parameters
- BERT-Large model has 24 transformer blocks, 16 attention heads,
  - 340M parameters
- GPT-2 is trained on 40GB of text data (8M webpages)!
  - 1.5B parameters
- GPT-3 is an even bigger version of GPT-2, but isn’t open-source
  - 175B parameters

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%).

Summary

- The GPT family: Decoder only models
- General theme: Train the largest language model your resources allow on the largest dataset you can find
- Impressive generation performance
  - Even more impressive: Zero-shot capabilities