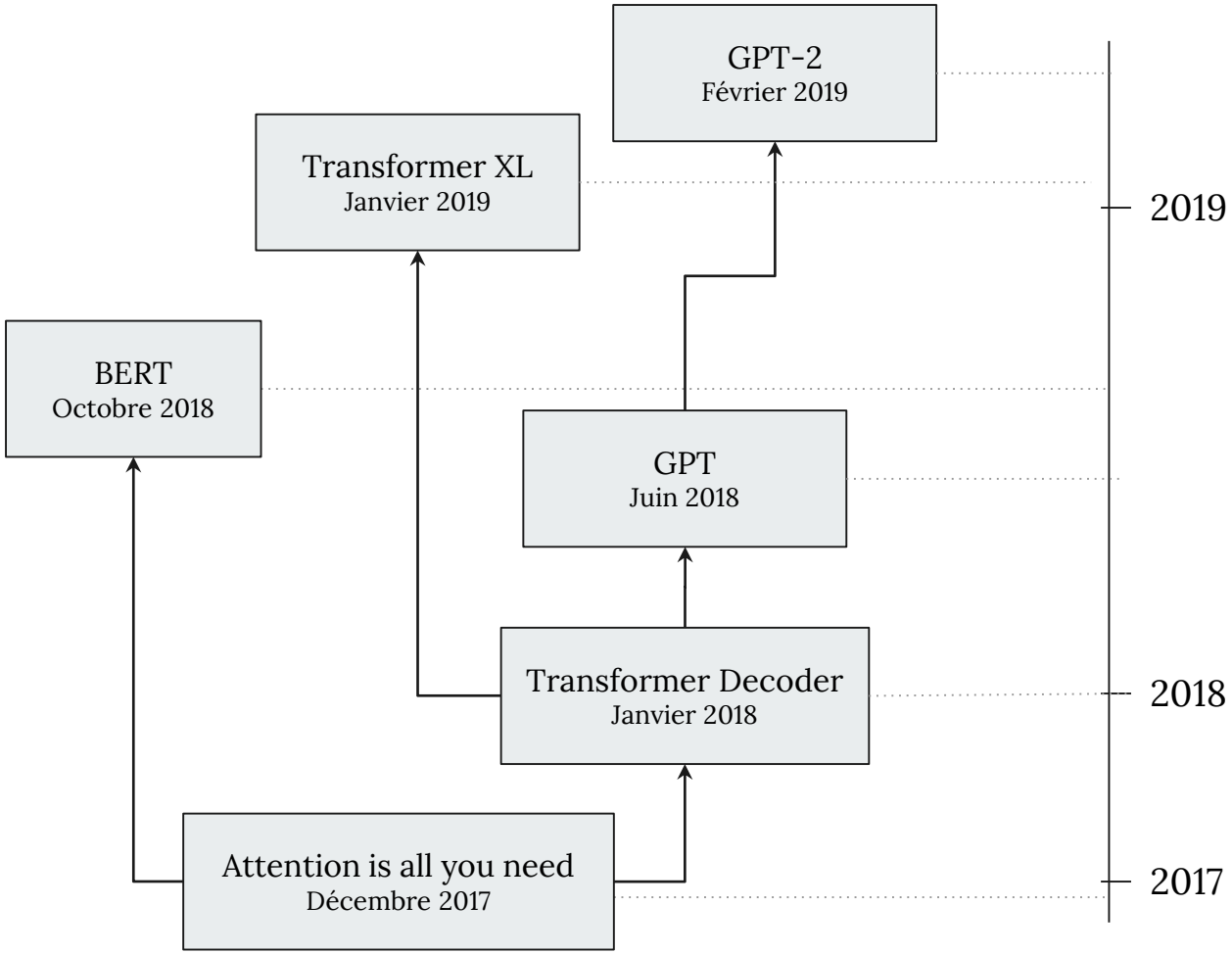
GPT-2

Language Models are Unsupervised Multi-Task Learners


GPT-2
Février 2019
Transformer XL
Janvier 2019
BERT
Octobre 2018
GPT
Juin 2018
Transformer Decoder
Janvier 2018
Attention is all you need
Décembre 2017
2019
2018
2017

Outline

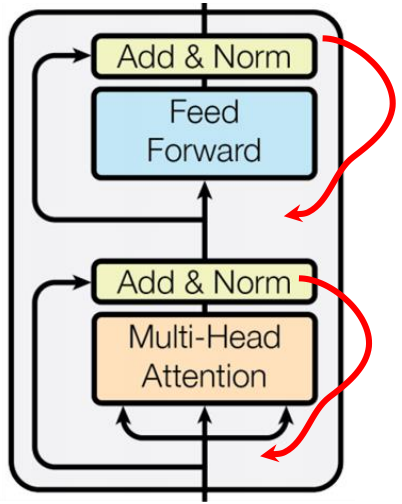
- Successor of GPT
- New, huge training dataset: WebText, not available :(
- Increased model capacity: 1,5B parameters
- Zero-Shot setting

Dataset

- Move away from the BookCorpus (GPT)
- Create a new web scrape that emphasizes **quality**
  - Using outbound links from Reddit that received at least 3 charma
  - Results in 8 million documents
  - 40GB of textual data

Model

- Largely follows GPT
- Moved the LayerNorm at the input of each block



GPT

BERT$_{LARGE}$

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.

LM Results in a 0-shot setting

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	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

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). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Generation examples

AMY GOODMAN: This is Democracy Now!, democracynow.org, The War and Peace Report. I'm Amy Goodman. I'm Amy Goodman, with John F. Kennedy. I'm on the phone from Chicago. I want to ask you this question: Why do so many Americans support the president's decision to use executive orders to restrict immigration? And why does the president want to use them? And you say the executive branch has the power to take away the right to travel without visas, the right to work, the right to travel to the United States for work permits, the right to have a place of study and the right to have a job. What is your view?

Generation examples

JOHN F. KENNEDY: Well, I think that the president, if he were to make a decision that he did not think would be consistent with the Constitution, would not be able to take away the right to travel or to have a place of study. And I mean, if that is going on in this country, it's not going to be consistent with my understanding of the constitutional principles that I believe.

Generation examples

We're seeing the president of the United States in the White House using executive and legislative power, which he used to try to limit the free movement of people, and to try to limit the free mobility of people, to try to restrict the right to work, to try to limit the ability to work to help the poor. And I think, for example, he's using these executive executive and legislative powers to try to make sure that, as a consequence of the executive order, the people in this country that were denied visas, who were working and earning $5 an hour and were looking for work, are going to be denied visas, which means that people who were in these situations are going to be denied visas. And we're going to see more and more people denied visas. And I think we're seeing an increase in people that have been denied visas because they're looking for work, or they're looking for a job that is not here now, or they are trying to get a job to come here and get a job that doesn't exist, because that's not who they are at that moment at this point in time.

Generation examples

AMY GOODMAN: What do you mean, by looking at the president's actions?

JOHN F. KENNEDY: Well, what I mean is, as a result of the executive order

Generation examples

We're seeing the president of the United States in the White House using executive and legislative power, which he used to try to limit the free movement of people, and to try to limit the free mobility of people, to try to restrict the right to work, to try to limit the ability to work to help the poor. And I think, for example, he's using these executive executive and legislative powers to try to make sure that, as a consequence of the executive order, the people in this country that were denied visas, who were working and earning $5 an hour and were looking for work, are going to be denied visas, which means that people who were in these situations are going to be denied visas. And we're going to see more and more people denied visas. And I think we're seeing an increase in people that have been denied visas because they're looking for work, or they're looking for a job that is not here now, or they are trying to get a job to come here and get a job that doesn't exist, because that's not who they are at that moment at this point in time.

Generation examples

AMY GOODMAN: What do you mean, by looking at the president's actions?

JOHN F. KENNEDY: Well, what I mean is, as a result of the executive order

The Perfect Toblerone

Ingredients

50g dark chocolate

1 tbsp milk

1 tsp vanilla extract

1 cup butter (melted)

1/2 tsp rum or triple sec (optional)

Instructions

1. Melt the chocolate (in a heatproof bowl or pan over a saucepan of boiling water) until it spreads. If it starts to melt slightly too quickly, pour to the sieve and allow to cool for 5 minutes before pouring straight into the fridge to solidify.

2. Cream the butter and add the rum, then the vanilla, and then gradually mix in your melted chocolate mixture.

3. Once all the chocolate's mixed in, turn up the heat to low and allow to cook for 2-3 minutes, or until the chocolate is all melted and smooth. It is important to stir constantly to prevent the chocolate from splattering.

4. Spread the melted chocolate over the cake, and then top with a dusting of cocoa powder.

5. Place back into the fridge for around an hour to harden down.

Nutritional Info per serving: 541 calories, 36g fat, 35g protein, 38g carbs, 9g fibre, 28g sugar

0-shot in other task

- Key idea is to condition the language model for that task

Summarization

- Introduce TL;DR at the end of the article
- Generate 100 tokens
- Use the first 3 sentences generated

	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 `TL;DR:`	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

Translation

- Condition the LM by aligning;
  - `english sentence = french sentence`
- Prompt `english sentence =`
  - Sample from the model with greedy decoding

Question answering

- Same condition as translation
  - `question = answer`
- Prompt `question =`
  - Sample from the model with greedy decoding

How is the conditioning exactly done?

- *See Appendix page 23

Conclusion

- Demonstrated that a **very large** language model train on a lot of textual data *_can_* generalize to multiple NLP tasks
- Not much change in the architecture here