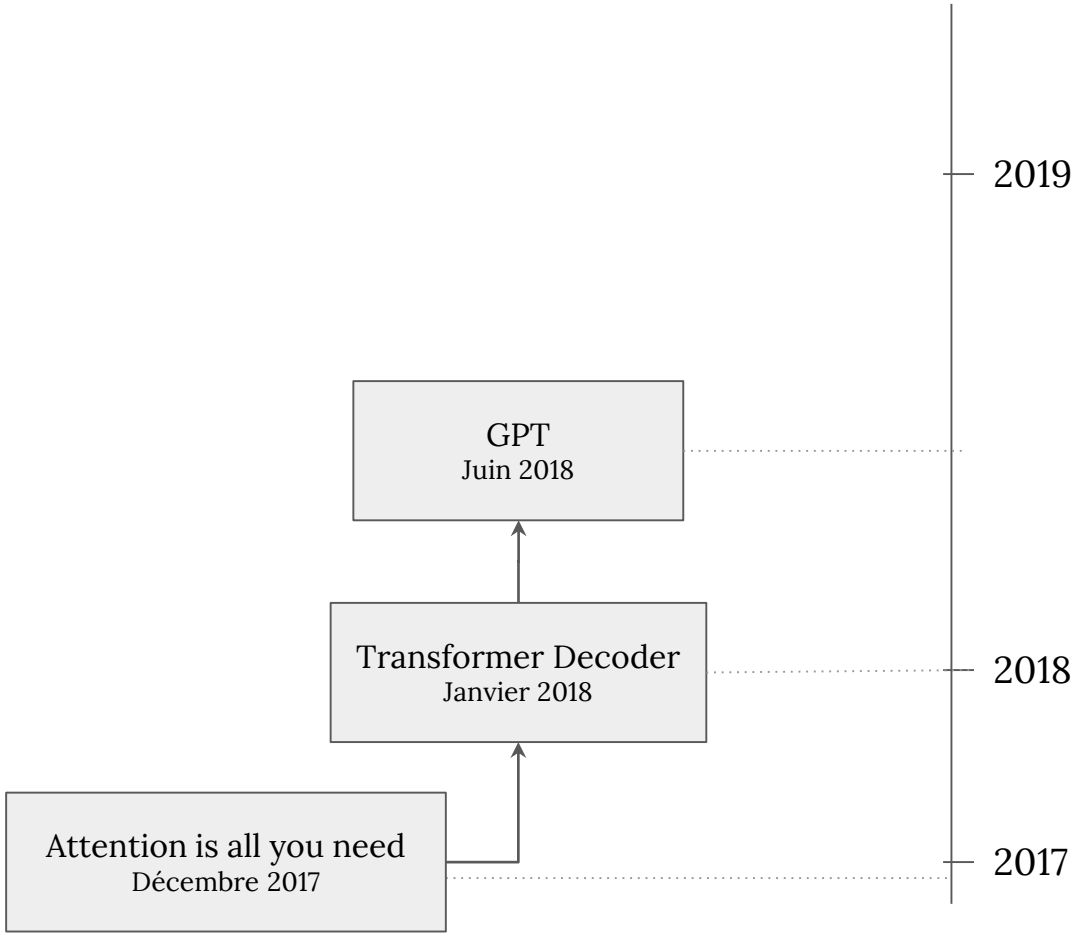
GPT

Improving Language Understanding by **G**enerative **P**re-**T**raining

Radford et al.


2019
GPT
Juin 2018
2018
Transformer Decoder
Janvier 2018
2017
Attention is all you need
Décembre 2017

Outline

- *Generative pre-training* of a language model on a diverse corpus of unlabeled text
- Followed by *discrimitative fine-tuning* on each specific task
- The rise of ImageNet and transfer learning for text!

Dataset

- Large corpus of unlabeled text (BookCorpus)
  - 7000 unique unpublished books
  - Train a Language Model
- Several dataset with manually annotated exemples (translation, Q&A)
  - Specific architecture

Unsupervised pre-training

- The goal is to find a good initialization point
- Acts as a regularization scheme and enables better generalization in deep neural networks

Unsupervised pre-training

- They use the T-D (same as Transformer-Decoder from Peter J. Liu et al.)

Supervised fine-tuning

- Inputs passed through the pre-trained network
- Simple classification layer on top of it (think ResNet)
- They include *language modeling as an auxiliary objective*

Supervised fine-tuning

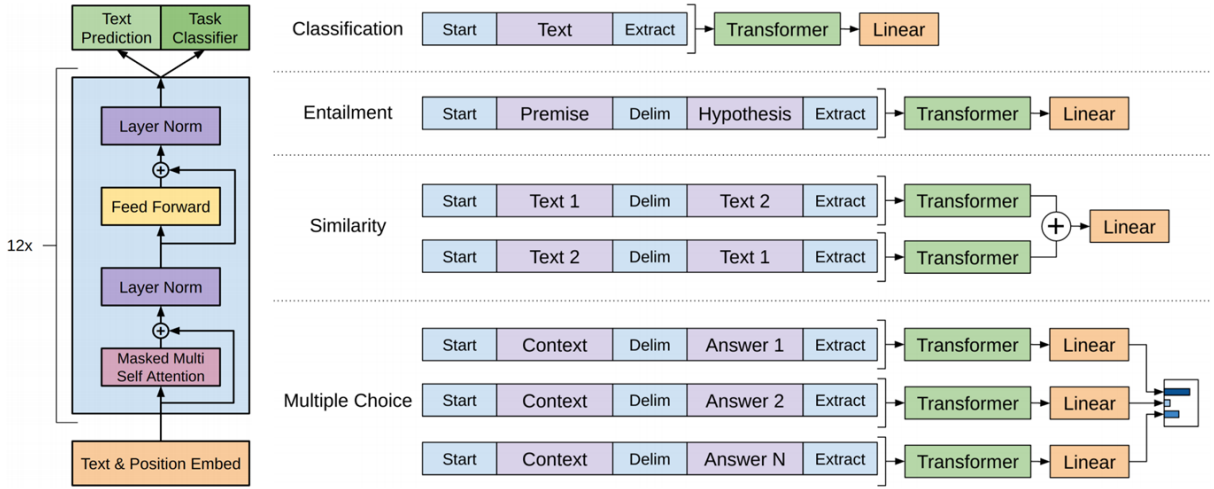


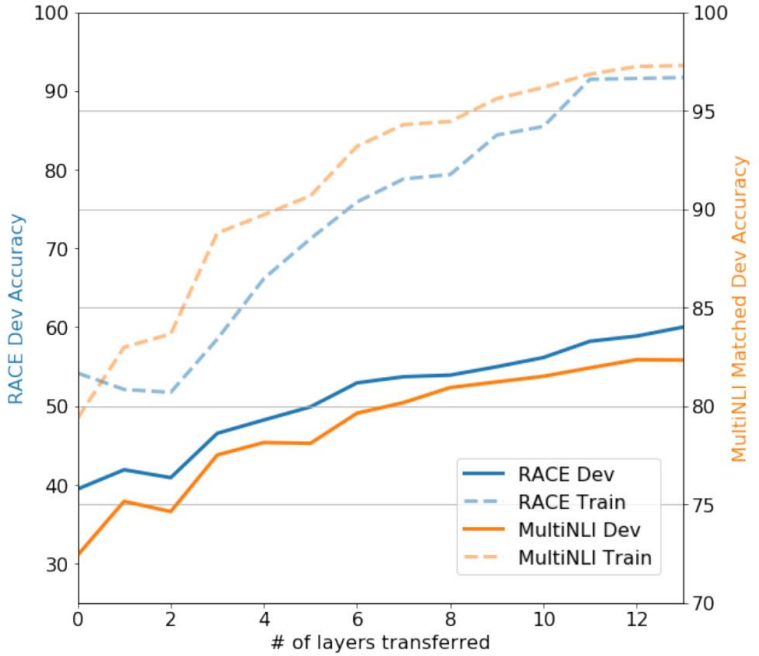
Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Results

Too many, see the paper!

Analysis

Impact of the number of layers transferred



Ablation study

Table 5: Analysis of various model ablations on different tasks. Avg. score is a unweighted average of all the results. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

Method	Avg. Score	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	MNLI (acc)	QNLI (acc)	RTE (acc)
Transformer w/ aux LM (full)	74.7	45.4	91.3	82.3	82.0	**70.3**	**81.8**	**88.1**	**56.0**
Transformer w/o pre-training	59.9	18.9	84.0	79.4	30.9	65.5	75.7	71.2	53.8
Transformer w/o aux LM	**75.0**	**47.9**	**92.0**	**84.9**	**83.2**	69.8	81.1	86.9	54.4
LSTM w/ aux LM	69.1	30.3	90.5	83.2	71.8	68.1	73.7	81.1	54.6

Conclusion

Differences from the Transformer-Decoder

- Doubled the layers (from 6 to 12)
- Specific input depending on the target task
- Applied model to 4 different tasks, 12 datasets in total