Generative Pre-trained Transformer

Mathilde Aguiar Oumaima Hajji

3^{rd} January 2022

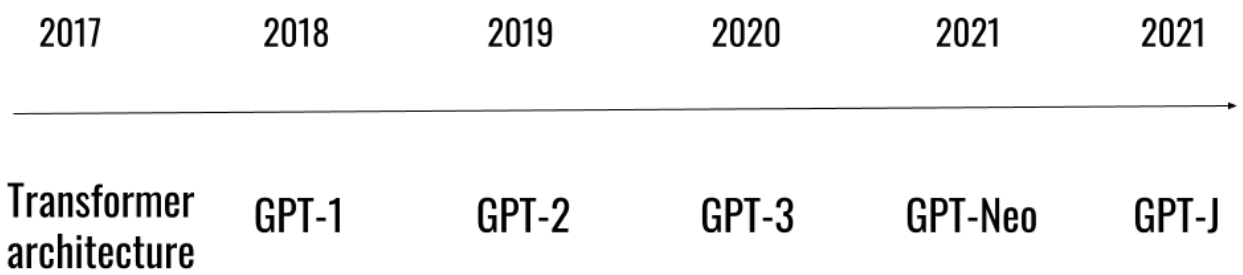
Presentation plan

Introduction

Natural Language Processing (NLP) or Traitement Automatique des Langues (TAL) in French : all means used by a computer to understand, interpret or process data in natural language.
Examples: Text-to-Speech, Named Entity Recognition, Neural Machine Translation, Sentiment Analysis, etc.

What is GPT ?

GPT: Generative Pre-trained *Transformer*



Transformer architecture

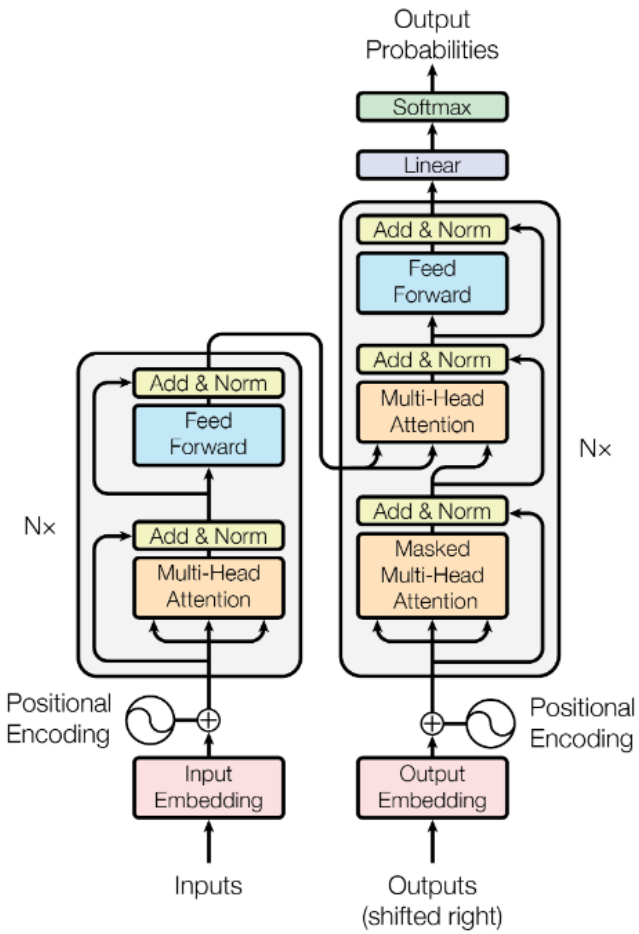
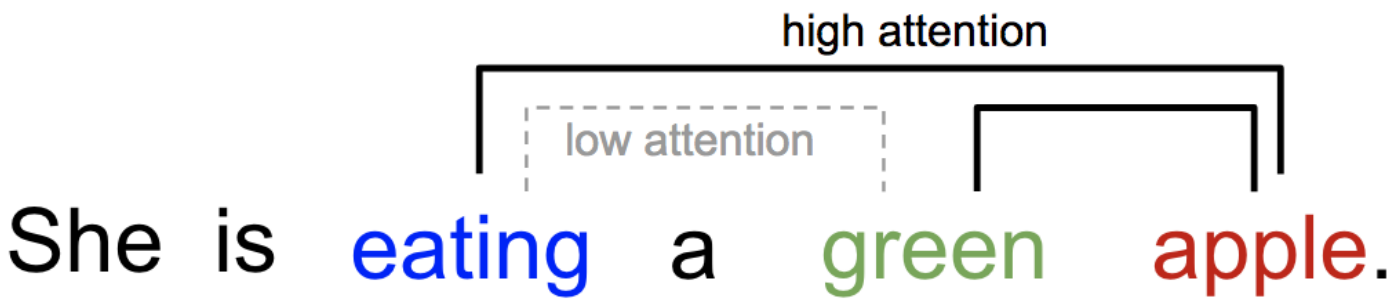


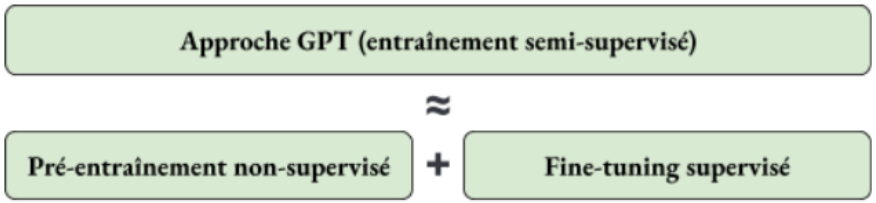
Figure 1: The Transformer - model architecture.

Transformer architecture



The first model, GPT-1

Approach : semi-supervised training



Model's characteristics : 12 layers Transformer Architecture with attention heads. Able to perform **Question answering, natural language inference, similarity analysis and text classification**.

The second model, GPT-2

Goal: Creating general systems that are able to execute multiple tasks with the same dataset.

Configuration : 4 possibilities:

- 1.5 billion parameters for the bigger model.
- 762 million parameters.
- 345 million parameters.
- 117 million parameters for the smaller model.

The second model, GPT-2

Dataset: WebText (40GB) The approach: creating a web scraper that fetches data that was already filtered by humans.

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I'm not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose**," which translates as, "**Lie lie and something will always remain.**"

"I hate the word '**perfume**,'" Burr says. 'It's somewhat better in French: '**parfum**.'

If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre coté? -Quel autre coté?**", which means "**- How do you get to the other side? - What side?**".

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

"**Brevet Sans Garantie Du Gouvernement**", translated to English: "**Patented without government warranty**".

The second model, GPT-2

Performance: GPT-2 was evaluated on multiple tasks (ex: comprehension, memorization, ..)

	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).

The open-source model, GPT-J

Goal: Creating an open source project that improves text generation.
Configuration : 6 billion parameters trained on the pile dataset (825 GiB).

The open-source model, GPT-J

Dataset: The Pile It contains diverse sub-datasets that cover different domains making the best dataset to train a general model.

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3†	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19)†	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles†	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en)†	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics†	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl†	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails†	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	**825.18 GiB**			**1254.20 GiB**	**5.91 KiB**

The open-source model, GPT-J

Its usage: GPT-J was particularly developed to improve text generation using prompts. Some of the prompts it responds to are:
- Arithmetic.
- Theorem proving.
- Coding.

```
Prompt -------

Theorem 3.2: Any cyclic group is abelian.

Proof:

-------------

Output:

Let G be a cyclic group. If the order is 1, then it is trivial. If the order
is 1, then G has only one element, so it is abelian.

If the order is not 1, let G be a cyclic group of order n.

Then G is generated by a single element a. Let a = gt.

Note that G is an abelian group, so gt^n = e. But t^n = e, so gt^n = t^n = e,
so t^n = g^n.

For any element g in G, (gt)^n = t^n, so g^n = (gt)^n = t^n = g^n.

So G is abelian.
```

The open-source model, GPT-J

Performance: A comparison of the performance of GPT-J and the other models.

Model	Training FLOPs	LAMBADA PPL ↓	LAMBADA Acc ↑	Winogrande ↑	Hellaswag ↑	PIQA ↑
GPT-2-1.5B	-----	10.63	51.21%	59.4%	50.9%	70.8%
GPTNeo-2.7B	6.8e21	5.63	62.2%	56.5%	55.8%	73.0%
GPT-3-1.3B	2.4e21	5.44	63.6%	58.7%	54.7%	75.1%
GPT-3-Babbage	-----	5.58	62.4%	59.0%	54.5%	75.5%
GPT-3-2.7B	4.8e21	4.60	67.1%	62.3%	62.8%	75.6%
GPT-J-6B	***1.5e22***	***3.99***	***69.7%***	***65.3%***	***66.1%***	***76.5%***
GPT-3-6.7B	1.2e22	4.00	70.3%	64.5%	67.4%	78.0%
GPT-3-Curie	-----	4.00	69.3%	65.6%	68.5%	77.9%
GPT-3-175B	3.1e23	3.00	76.2%	70.2%	78.9%	81.0%
GPT-3-Davinci	-----	3.0	75%	72%	78%	80%

The latest one, GPT-3

Goal: Having a model that learns like a human.
Configuration : 175 billions of parameters for the original model. From 13 billions to 125 millions of parameters for other smaller GPT-3 models.

The latest one, GPT-3

The latest one, GPT-3

Few-shot

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

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

The latest one, GPT-3

Zero-shot

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

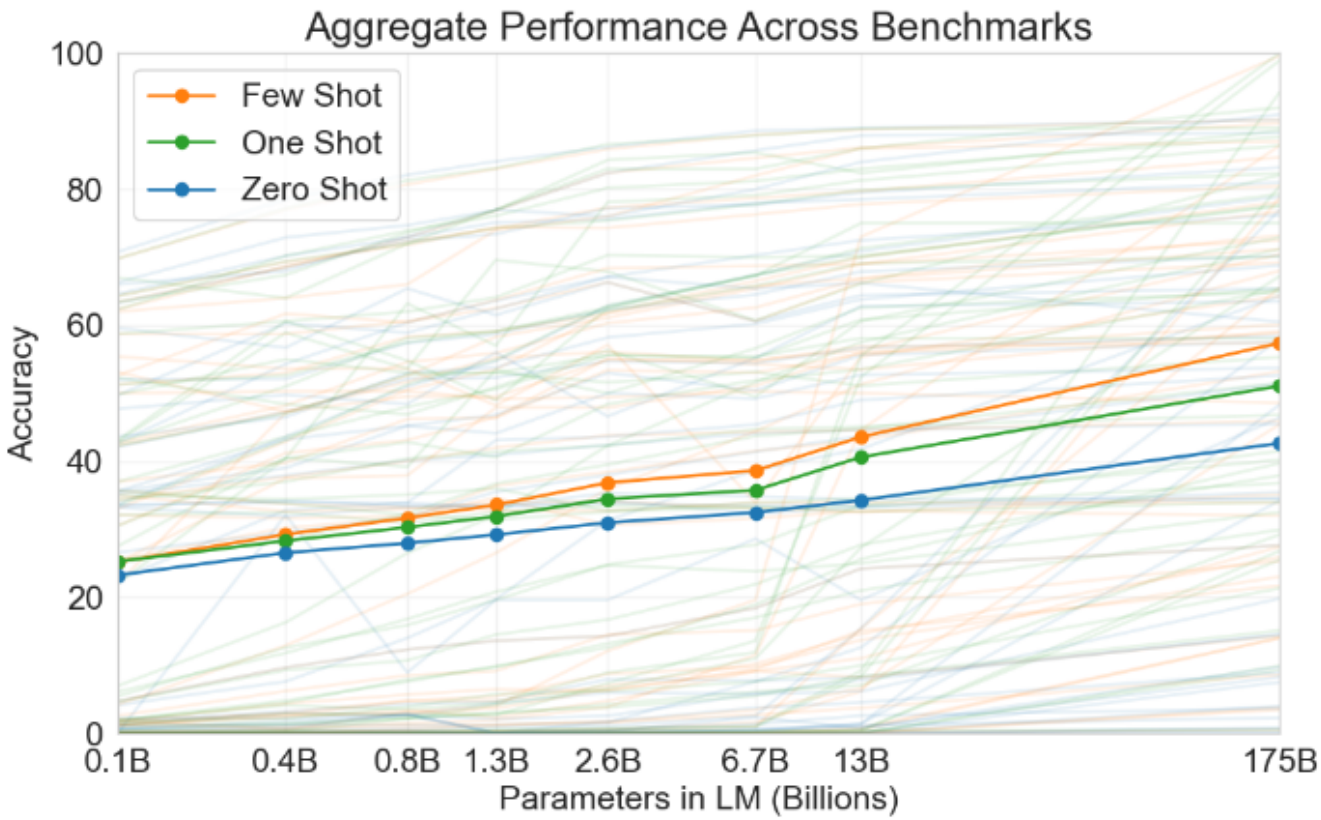
```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

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

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

The latest one, GPT-3



The latest one, GPT-3

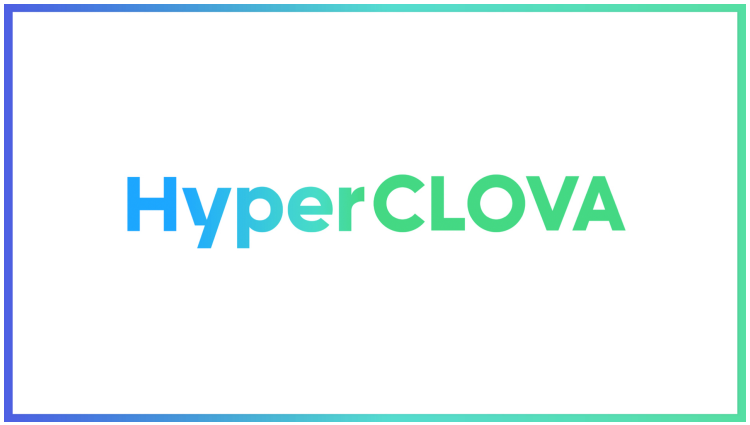
OpenAI API models :

Model	Tasks
Davinci	**Complex intent, cause and effect, summarization for audience**
Curie	**Language translation, complex classification, text sentiment, summarization**
Babbage	**Moderate classification, semantic search classification**
Ada	**Parsing text, simple classification, address correction, keywords**

The latest one, GPT-3

Limitations: Some bias and some limitations on tasks like text synthesis.

Other models and applications based on the GPT architecture



Demonstration

With GPT-2 :
https://transformer.huggingface.co/doc/gpt2-large
With GPT-J: https://6b.eleuther.ai/
With GPT-3 :
https://beta.openai.com/examples/default-friend-chat
https://beta.openai.com/playground/p/default-movie-to-emoji
https://beta.openai.com/playground/p/default-translate

Conclusion

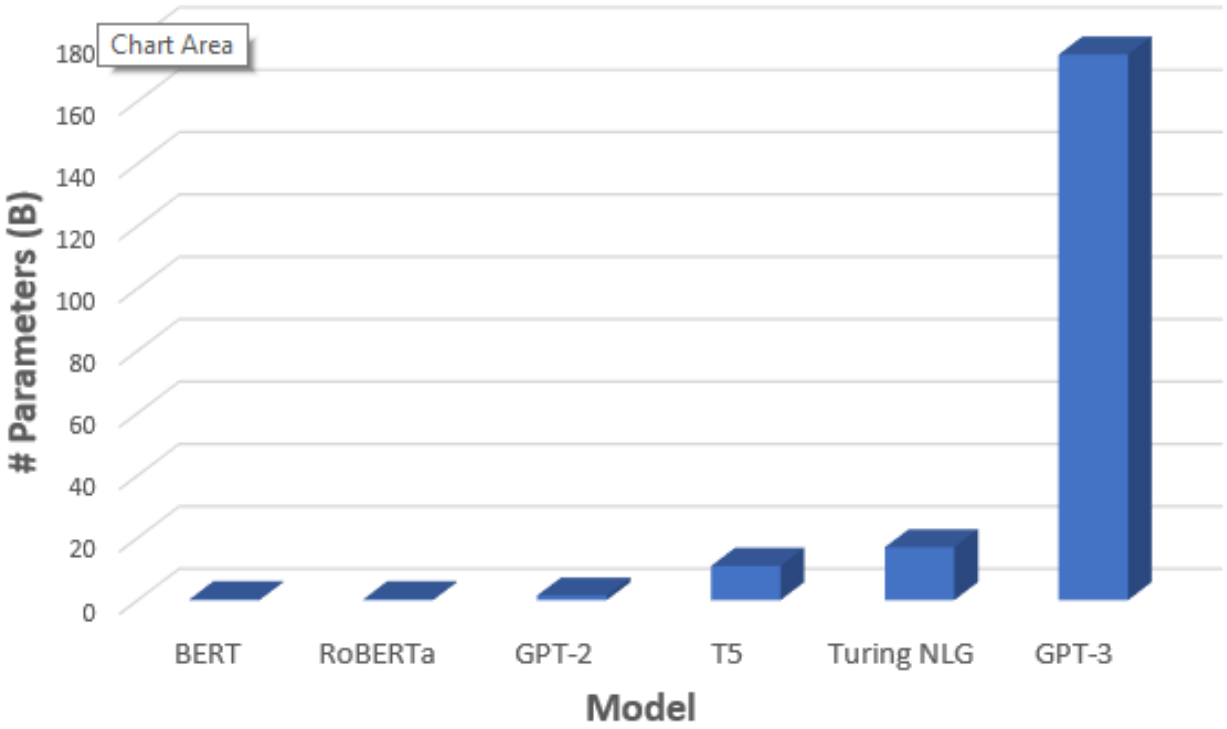


Chart Area
180
160
140
120
100
80
60
40
20
0
Parameters (B)
BERT
RoBERTa
GPT-2
T5
Turing NLG
GPT-3
Model