Deep Learning for Natural Language Processing

Lecture 11 – Text generation 4: Decoder-only Models and GPT

Dr. Martin Tutek

June 27, 2023

Ubiquitous Knowledge Processing
Department of Computer Science
Technical University of Darmstadt

UKP Web

Recap

In the previous lecture we:

- Introduced the **BERT model**
- Introduced the two pretraining tasks for BERT: **MLM** and **NSP**
- Explained the connection between MLM and CBOW-style training
- Explained the purpose of NSP – learning a **sentence embedding**
- Analyzed how to **apply BERT** to various **downstream tasks** such as classification and QA
- Gave an overview of various other pretraining tasks for LLMs

Motivation

Recall: using the **same model** for **multiple tasks** without task-specific decoder heads

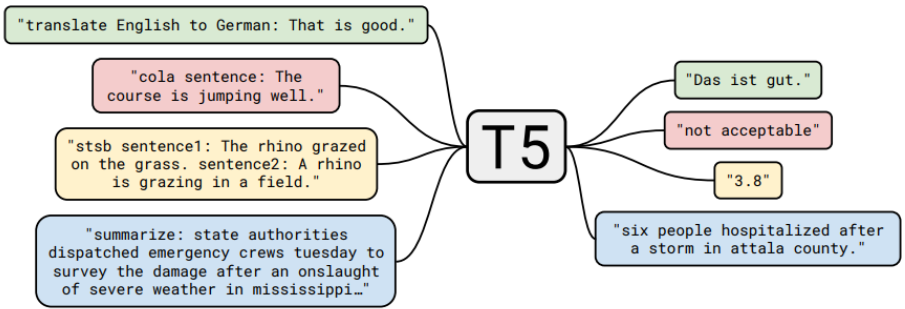


Image from T5 paper

Motivation

Recall: using the **same model** for **multiple tasks** without task-specific decoder heads

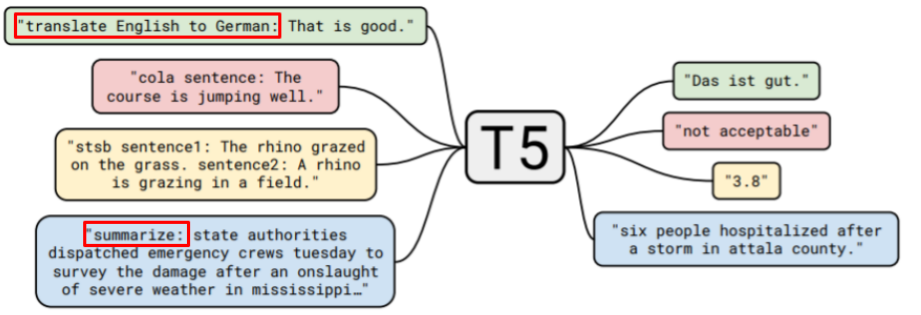
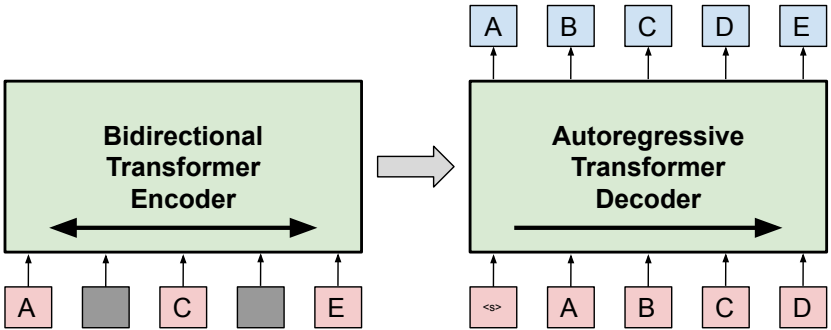


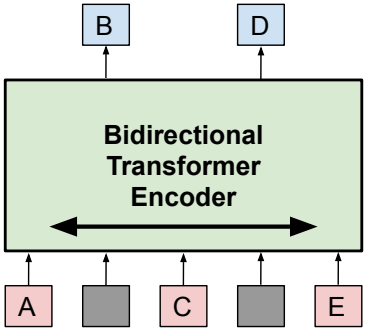
Image from T5 paper

Types of Transformer Architectures

Encoder-Decoder Transformer

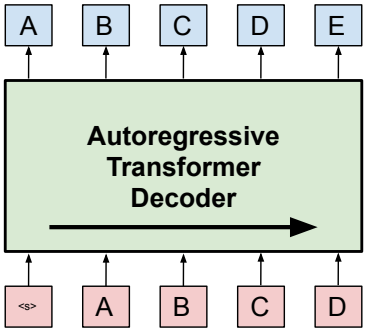


Bidirectional Encoder-only Transformer



- Efficient encoding ✔️
- Versatile base for downstream tasks ✔️
- Can't **really** generate text ❌

Autoregressive Decoder-only Transformer



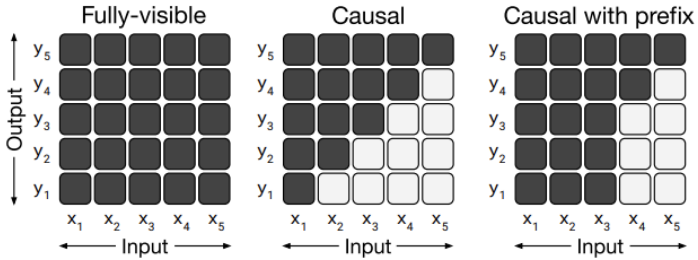
An **autoregressive** (causal) language model uses **past** values of a time series to predict future values.

- Didn't we decide not to use these because they were inefficient?

  (RNNs)

- Yes, but...
  1. Hardware has improved
  2. Autoregressive models are *really* good at generating text

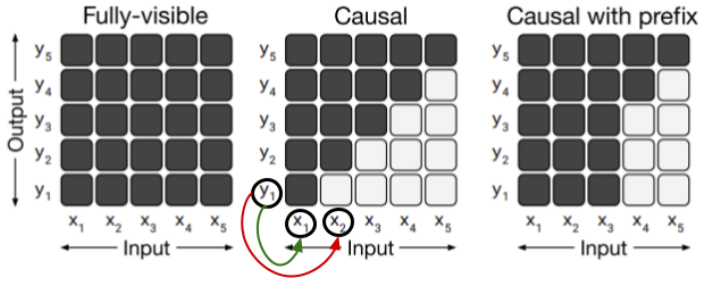
Differences between attention masks



Read: y axis → tokens attending, x axis → tokens attended to.

Black cell → token visible, white cell → token **masked**

Differences between attention masks



Read: y axis → tokens attending, x axis → tokens attended to.

Black cell → token visible, white cell → token **masked**

Attention masks

Recall: the attention mechanism

$$a = \sum_{i}^{n} \alpha_i v_i \qquad \hat{\alpha}_i = \frac{q^T \cdot k_i}{\sqrt{d_{\text{model}}}}$$

How do we do **masking**?

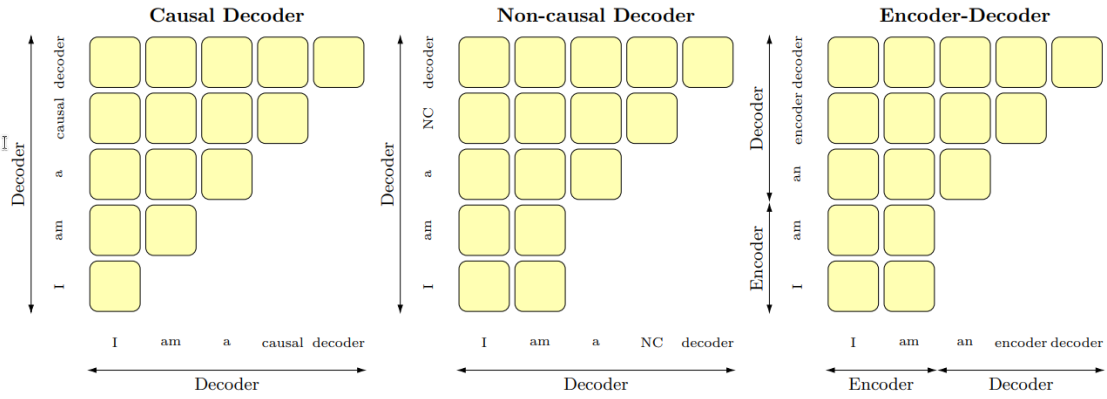
In the **causal** scenario (each token can only attend to **past** tokens);

For a $q = W_q(s_j)$ query computed based on the hidden state s_j at position j

$$\alpha_i = \begin{cases} \alpha_i, & \text{if } j \geq i \\ 0, & \text{otherwise} \end{cases}$$

NB: actually, we set $\hat{\alpha}_i$ to $-\inf$ (before softmax)

Differences between attention masks



Autoregressive decoder-only Models

Types of Transformer Architectures

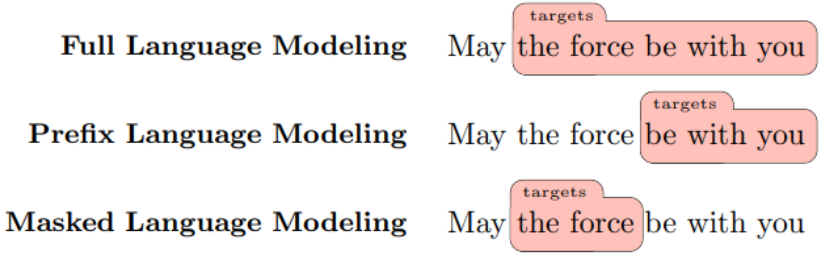
Autoregressive decoder-only Models

- Zero-shot, one-shot and few-shot learning
- Prompting

Prompt-tuning MLMs

A step back

Variants of language modeling



- (Full) language modeling → given previous tokens, predict next token, for **every token** in sequence
- Prefix language modeling → (1) feed a prefix (where mask **does not have to be causal**), (2) full LM starting after prefix
- Masked language modeling → **reconstruct masked** tokens/spans

Autoregressive decoder-only models

Language Models are Unsupervised Multitask Learners

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

Introduction of **GPT-2**, an autoregressive Transformer decoder-only model trained on full language modeling.

GPT-3 is *"just"* a **larger** version of GPT-2

Autoregressive decoder-only models

Language Models are Unsupervised Multitask Learners

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

Introduction of **GPT-2**, an autoregressive Transformer decoder-only model trained on full language modeling.

What does ”unsupervised multitask learners” mean 🤔?

Autoregressive decoder-only Models

Zero-shot, one-shot and few-shot learning

Zero-shot, one-shot and few-shot learning

Recall: T5 was able to perform **multiple tasks** at the same time

… but it was trained on them & on keywords which indicate the task.

For a model that has **not been trained on the downstream task**:

- **Few-shot** learning: tune pretrained model on a **small number** of target task instances, **then perform task (!)**
- **One-shot** learning: tune pretrained model on **one instance (!)** *per class*, then perform task
- **Zero-shot** learning: **don't tune pretrained model(!!!)**, then perform task

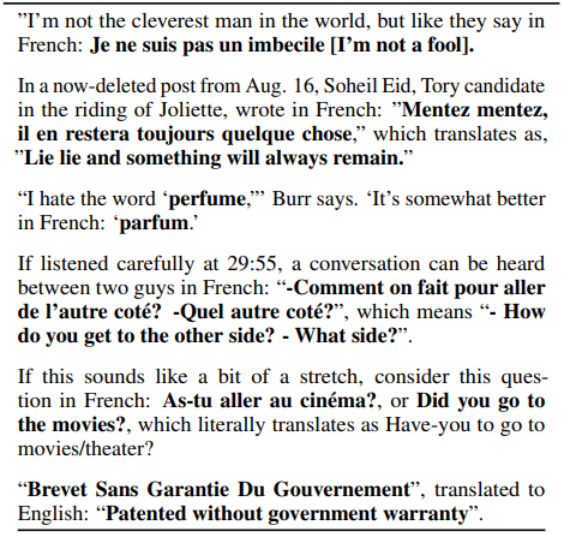
Zero-shot learning

Zero shot learning $\approx$ unsupervised learning

Why $\approx$?

Assumption: when trained on a **massive** corpus of text, the language model is likely to **see some tasks naturally** occur (e.g. question answering).

- We want to **transform** our task into a **generative one** by providing a **prompt** to the model which will make the label of the input instance the **most likely generated sequence**.



”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**”.

Image from GPT2 paper

The internet **does** contain samples of various NLP tasks

- … and a large language model (LLM) **can** remember them;
- … and when **prompted** to perform a task, without seeing the prompt before, **recall it**;
- … and **perform them accurately**.

GPT-2: Zero-shot question answering

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%

Image from GPT2 paper

GPT-2: Prompted one-shot question answering

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

Image from GPT2 paper

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

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

demonstration

Q: where does she live?
A:

prompt

Model answer: Stockholm
Turker answers: Sweden, Sweden, in Sweden, Sweden

Autoregressive decoder-only Models

Prompting

Prompting

A prompt is a piece of text inserted in the input examples, so that the original task **can be formulated as** a (masked) **language modeling** problem.

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
```

Image from GPT3 paper

Prompting

A prompt is a piece of text inserted in the input examples, so that the original task **can be formulated as** a (masked) **language modeling** problem.

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
```

Image from GPT3 paper

Prompting

A prompt is a piece of text inserted in the input examples, so that the original task **can be formulated as** a (masked) **language modeling** problem.

Few-shot

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

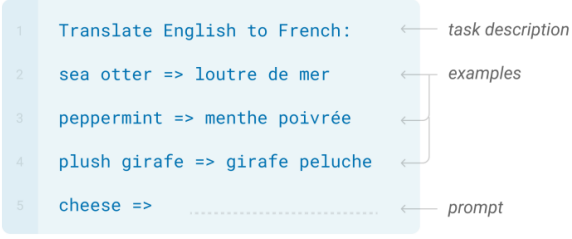


Image from GPT3 paper

Prompting works well

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

Image from GPT3 paper

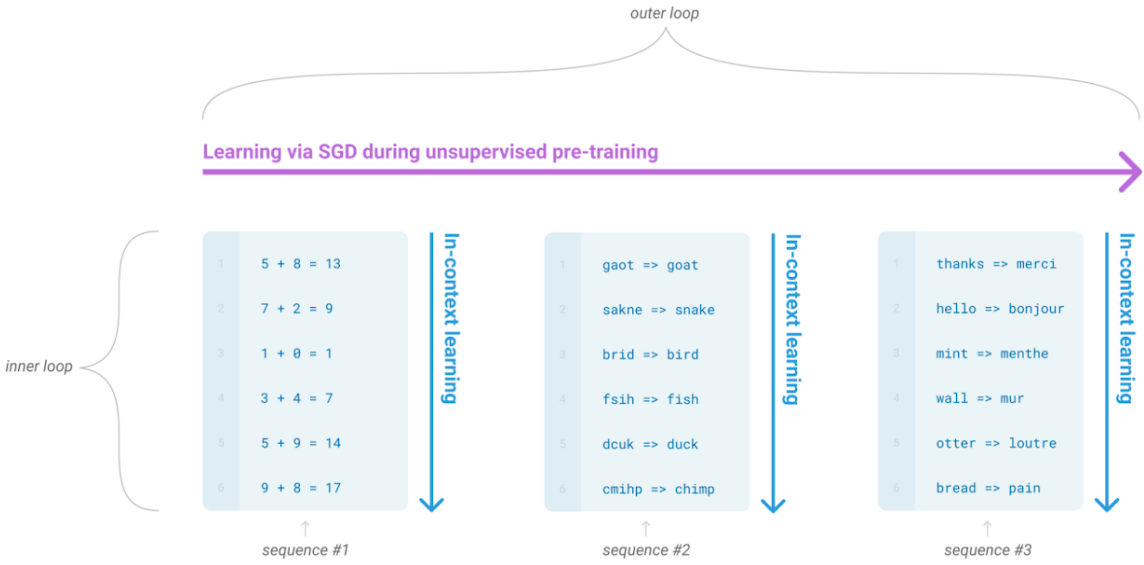
GPT3 **without fine-tuning** performs better than **unsupervised** alternatives, and sometimes even **better** than supervised state-of-the-art!

In-context learning

In-context learning is the paradigm in which a LLM learns to solve a new task at inference time **without any change to its weights**, based only on examples in the **prompt**.

$\approx$ umbrella term for zero-, one- and few-shot learning with task descriptions also contained in prompt.

*"During **unsupervised pre-training**, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. We use the term "in-context learning" to describe the inner loop of this process, which occurs **within the forward-pass** upon each sequence."* – from GPT3 paper


outer loop
Learning via SGD during unsupervised pre-training
inner loop
1 5 + 8 = 13
2 7 + 2 = 9
3 1 + 0 = 1
4 3 + 4 = 7
5 5 + 9 = 14
6 9 + 8 = 17
In-context learning
sequence #1
1 gaot => goat
2 sakne => snake
3 brid => bird
4 fsih => fish
5 dcuk => duck
6 cmihp => chimp
In-context learning
sequence #2
1 thanks => merci
2 hello => bonjour
3 mint => menthe
4 wall => mur
5 otter => loutre
6 bread => pain
In-context learning
sequence #3

Prompt-tuning MLMs

Types of Transformer Architectures
Autoregressive decoder-only Models
    Zero-shot, one-shot and few-shot learning
    Prompting
Prompt-tuning MLMs
A step back

Prompt-tuning MLMs

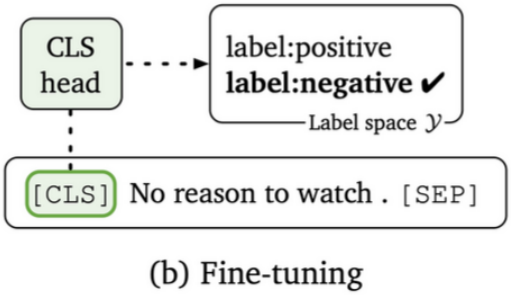
Can we only use prompting with autoregressive models?

- No – we can also use it with bidirectional decoder-only models!
  - … but it is **more difficult** because they have not been trained to generate texts
  - … because the downstream task is **less natural** (further from the pretraining task) to the model

How to overcome this gap between the **pretraining task** and the **prompting-transformed downstream task**?

Prompt-tuning MLMs

So far, we have **fine-tuned** masked language models

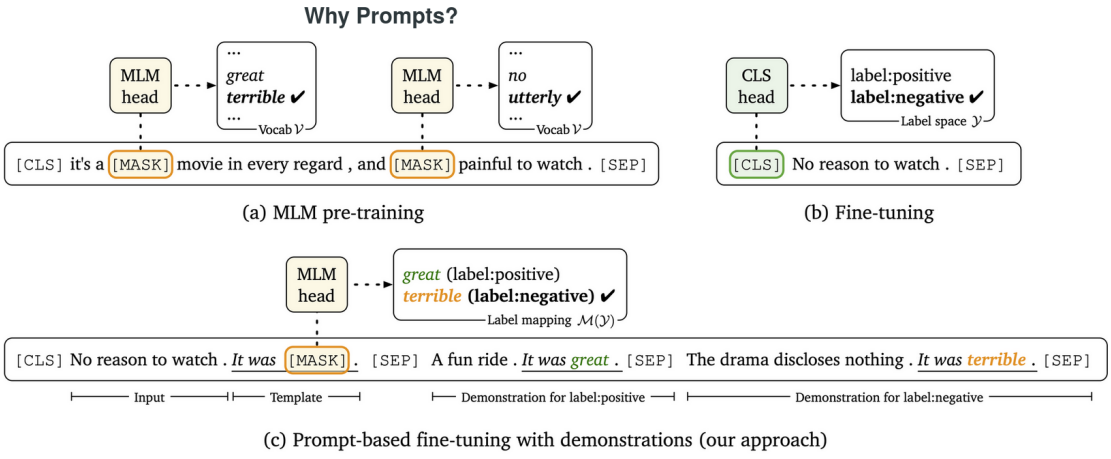


(b) Fine-tuning

Figure from The Gradient

Can we frame our downstream task **as MLM**?

Prompt-tuning MLMs



Prompt-tuning MLMs

We transform the target task (e.g. sentiment analysis) to **masked language modeling**.

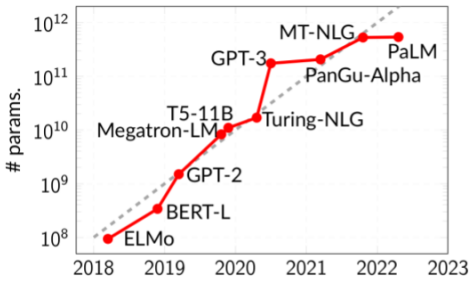
1. Choose the prompt and word/token used for each label
   - Choice of label token **important**
   - Template design also **important**
2. Demonstrate task through a few samples
   - Usually through **fine-tuning**
3. **No new parameters needed** to perform task!

A step back

Incredible Performance of Large Language Models

So... what caused LLMs to be **so good** all of a sudden?

- More available data (more data → better models)
- Training tricks (from experience)
- Hardware advancements (faster training of larger models)



Discrete and continuous prompts

So far, we have shown **discrete prompts**: actual text that we prepend/append to existing data which triggers the LLM to perform our task.

Can we learn **continuous prompts**? (dense vectors which we prepend, e.g. as a token)

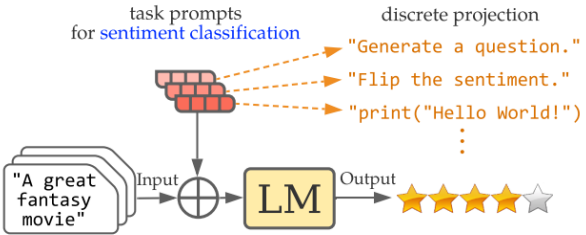


Figure from Prompt Waywardness

Takeaways

- Three types of Transformer-based architectures for LLM pretraining:
  - **Encoder-decoder** (T5)
  - **Bidirectional encoder-only** (BERT)
  - **Autoregressive decoder-only** (GPT-2)
- The **attention masks** of these models differ
- There are three variants of language modeling for pretraining LLMs
- GPT-2 (and 3) are autoregressive decoder-only transformers
- We introduced zero-, one- and few-shot learning
- We introduced prompting and its variants
  - Autoregressive vs MLM prompting
  - Continuous vs discrete prompts
  - In-context learning

Useful resources

- The Gradient: Prompting by Tianyu Gao
- The Gradient: In Context Learning by Daniel Bashir
- Understanding in-context learning by Sang Michael Xie and Sewon Min

License and credits

Credits

Martin Tutek

Content from ACL Anthology papers licensed under CC-BY `https://www.aclweb.org/anthology`