Overview of Natural Language Processing Part II & ACS L390

Lecture 11: Language Models

Yulong Chen and Weiwei Sun

Department of Computer Science and Technology
University of Cambridge

Michaelmas 2024/25

The Shifts of NLP Paradigms

The Shifts of NLP Paradigms

Paradigms in NLP Research *before* 2017

- Rule and Symbolics
  - Focus on how to better design rules.
- Statistic and Machine Learning
  - Focus on how to design model architectures, e.g., RNN vs Transformer.

The Shifts of NLP Paradigms

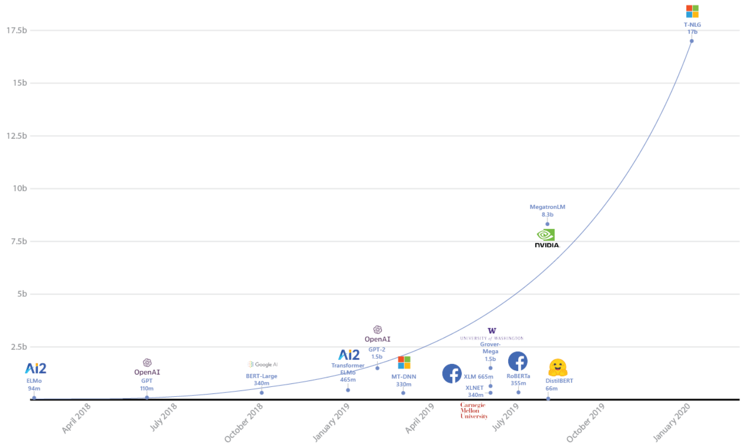
Paradigms in NLP Research *before* 2017

- Rule and Symbolics
  - Focus on how to better design rules.
- Statistic and Machine Learning
  - Focus on how to design model architectures, e.g., RNN vs Transformer.
- Pre-training and Fine-tuning
  - Focus on how to obtain generalizable knowledge (pre-training), and how to efficiently adapt to downstream tasks (fine-tuning).

The Shifts of NLP Paradigms

Paradigms in NLP Research *before* 2017

- Rule and Symbolics
  - Focus on how to better design rules.
- Statistic and Machine Learning
  - Focus on how to design model architectures, e.g., RNN vs Transformer.
- Pre-training and Fine-tuning
  - Focus on how to obtain generalizable knowledge (pre-training), and how to efficiently adapt to downstream tasks (fine-tuning).
- Prompting and Large Language Models
  - Focus on solving real-world problems (compared with traditional NLP tasks).



Lecture 11: Language Models

1. Pre-trained based NLP
2. Prompt learning and LLMs

Pre-trained based NLP

Pre-train and Fine-tune

Suppose a model is a human ...

- Pre-training can be the process of a person learning through early education stages — from infancy to high school. They learn foundational knowledge, and common sense.
- Fine-tuning can be the process of a person entering their high-level education — they will be specialised in a programme. They learn professional knowledge.

Pre-train and Fine-tune

Suppose a model is a human ...

- Pre-training can be the process of a person learning through early education stages — from infancy to high school. They learn foundational knowledge, and common sense.
- Fine-tuning can be the process of a person entering their high-level education — they will be specialised in a programme. They learn professional knowledge.
- Tuning a model from scratch (the second paradigm) can be asking a baby directly to study for their doctorate.

Pre-train and Fine-tune

Suppose a model is a human ...

- Pre-training can be the process of a person learning through early education stages — from infancy to high school. They learn foundational knowledge, and common sense.
- Fine-tuning can be the process of a person entering their high-level education — they will be specialised in a programme. They learn professional knowledge.
- Tuning a model from scratch (the second paradigm) can be asking a baby directly to study for their doctorate.
- **Given the same limited time and knowledge (a doctoral course), an undergrad can learn faster and better than a baby.**

Pre-train and Fine-tune

Suppose a model is a human ...

- Pre-training can be the process of a person learning through early education stages — from infancy to high school. They learn foundational knowledge, and common sense.
- Fine-tuning can be the process of a person entering their high-level education — they will be specialised in a programme. They learn professional knowledge.
- Tuning a model from scratch (the second paradigm) can be asking a baby directly to study for their doctorate.
- **Given the same limited time and knowledge (a doctoral course), an undergrad can learn faster and better than a baby.**
- But in practice, annotated data is small while raw data is large.

Pre-train and Fine-tune

Suppose a model is a human ...

- Pre-training can be the process of a person learning through early education stages — from infancy to high school. They learn foundational knowledge, and common sense.
- Fine-tuning can be the process of a person entering their high-level education — they will be specialised in a programme. They learn professional knowledge.
- Tuning a model from scratch (the second paradigm) can be asking a baby directly to study for their doctorate.
- **Given the same limited time and knowledge (a doctoral course), an undergrad can learn faster and better than a baby.**
- But in practice, annotated data is small while raw data is large.

Can models learn general knowledge from raw text?

Self-supervised Learning

The main idea of self-supervised learning:

- Use the internal signal of a text as the supervising signal.

Self-supervised Learning

The main idea of self-supervised learning:

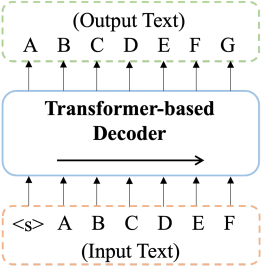
- Use the internal signal of a text as the supervising signal.

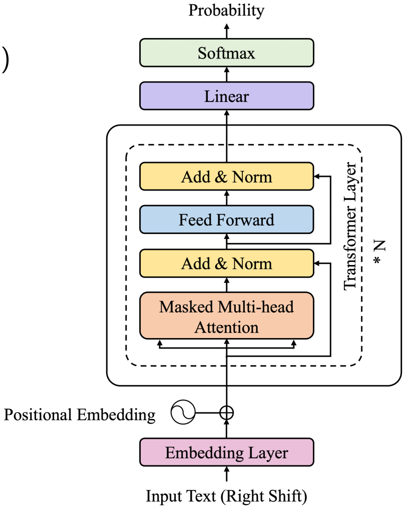
	Supervised	**Self-supervised**
Task	a specific task	re-construct the input
Label	human annotation	generate annotation using the data itself
Resource	limited	large

Decoder-only PLM: GPT

Improving Language Understanding by Generative Pre-Training (GPT)

- Architecture: Transformer decoder
- Pre-training Task:
  - Next Token Prediction (NTP)
- Pre-training Data:
  - BookCorpus

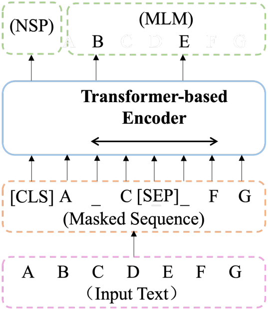


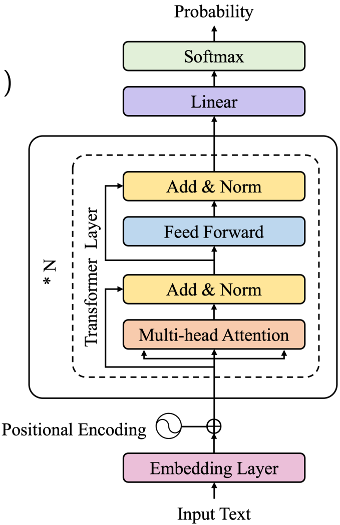


Encoder-only PLM: BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- Architecture: Transformer encoder
- Pre-training Task:
  - Masked Language Modeling (MLM)
  - Next Sentence Prediction (NSP)
- Pre-training Data:
  - BookCorpus and Wikipedia

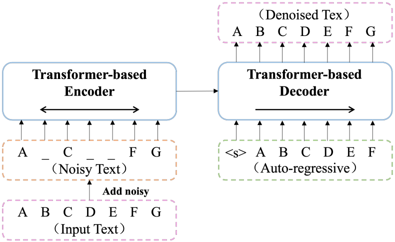


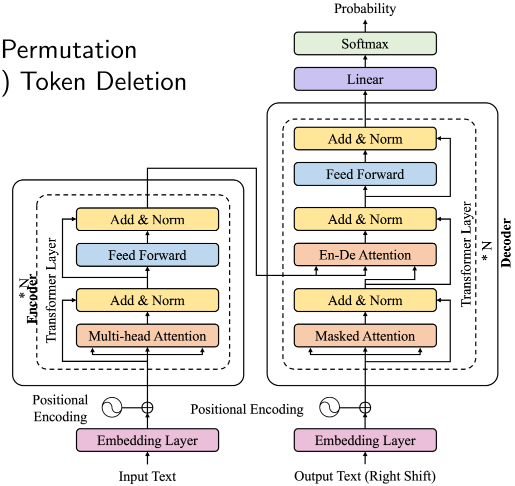


En-De PLM: BART

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

- Architecture: Transformer encoder-decoder
- Pre-training Task:
  - (1) Masking (2) Sentence Permutation
    (3) Document Rotation (4) Token Deletion
    (5) Text Infilling
- Pre-training Data:
  - BookCorpus, Wikipedia
    Webtext, and stories





Adapting PLMs to Downstream Tasks

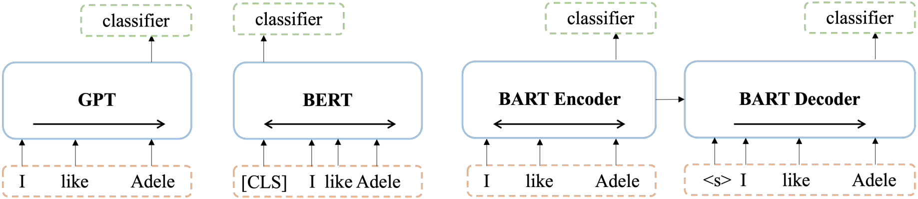
Suppose we want to use PLMs for **a downstream task**, how can we do this?

Adapting PLMs to Downstream Tasks

Suppose we want to use PLMs for **a downstream task**, how can we do this?

Full Fine-tuning

- Update all parameters of a PLM on downstream tasks.
- Case 1: Sentiment Analysis

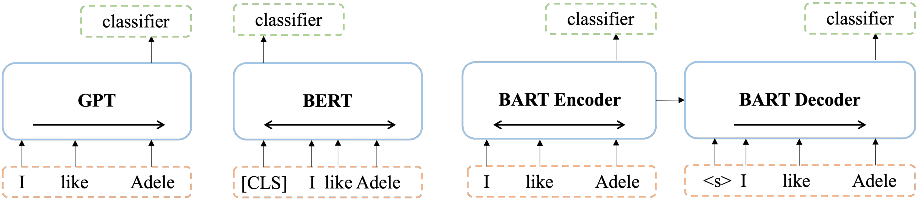


Adapting PLMs to Downstream Tasks

Suppose we want to use PLMs for **a downstream task**, how can we do this?

Full Fine-tuning

- Update all parameters of a PLM on downstream tasks.
- Case 1: Sentiment Analysis



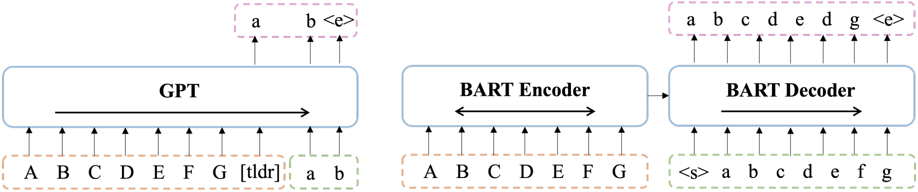
Trick: Continual pre-training (Reading).

Adapting PLMs to Downstream Tasks

After pre-training on large raw data, we fine-tune the PLM to a specific task.

Full Fine-tuning

- Update all parameters of a PLM on downstream tasks.
- Case 2: Summarisation

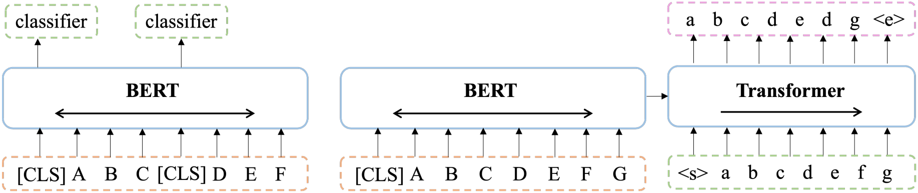


Adapting PLMs to Downstream Tasks

Suppose we want to use PLMs for **a downstream task**, how can we do this?

Full Fine-tuning

- Update all parameters of a PLM on downstream tasks.
- Case 2: Summarisation



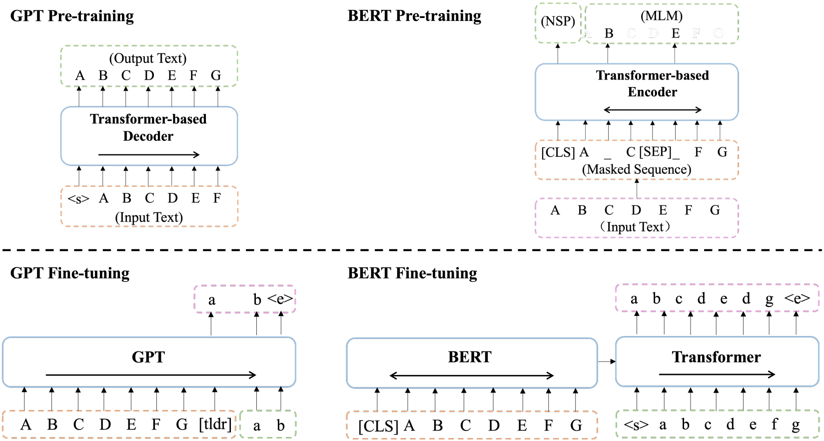
Reading: Text Summarization with Pretrained Encoders

Adapting PLMs to Downstream Tasks

It is empirically considered that the Encoder models are better at NLU tasks, while Decoder and En-De models are better at NLG.

Adapting PLMs to Downstream Tasks

It is empirically considered that the Encoder models are better at NLU tasks, while Decoder and En-De models are better at NLG.

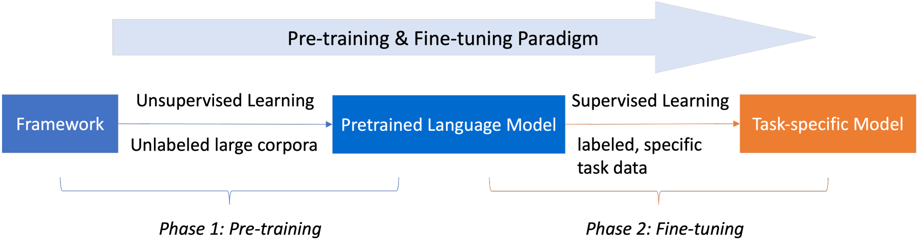


There is a gap between the language modeling task and downstream tasks.

Pre-trained based NLP

Although fine-tuning is less costly than pre-training (**data size**), it still cannot meet the increasing demand:

- Need to update all model parameters (still not cheap).
- One fine-tuned model for one specific task.
- Cannot be used for low-resource settings.



Prompt Learning and Large Language Models

Prompt Learning

Main idea of prompt learning

- Adapting a downstream task into a language modeling format.

Prompt Learning

Main idea of prompt learning

- Adapting a downstream task into a language modeling format.

Take sentiment analysis (SST2) for example:

- Given a movie review $\mathbf{X}$ as the input, e.g.:

  $\mathbf{X}$ = *"it 's about issues most adults have to face in marriage and i think that 's what i liked about it – the real issues tucked between the silly and crude storyline"*

  the task asks a model to generate a binary label (`positive` or `negative`).

Prompt Learning

Main idea of prompt learning

- Adapting a downstream task into a language modeling format.

Take sentiment analysis (SST2) for example:

- Given a movie review $\mathbf{X}$ as the input, e.g.:

  $\mathbf{X}$ = *"it 's about issues most adults have to face in marriage and i think that 's what i liked about it – the real issues tucked between the silly and crude storyline"*

  the task asks a model to generate a binary label (`positive` or `negative`).
- We can define a pattern function for NTP PLM (e.g., GPT):

$$P(\mathbf{X}) = \mathbf{X}. \textit{ In summary, this movie is} \tag{1}$$

Prompt Learning

Take sentiment analysis (SST2) for example:

- The input is converted as:
  $P(\mathbf{X})$ = *"it 's about issues most adults have to face in marriage and i think that 's what i liked about it – the real issues tucked between the silly and crude storyline. In summary, this movie is"*

- Then, we define a verbalizer function:

$$V(\textit{"good"}) = \texttt{pos} \tag{2}$$

$$V(\textit{"bad"}) = \texttt{neg} \tag{3}$$

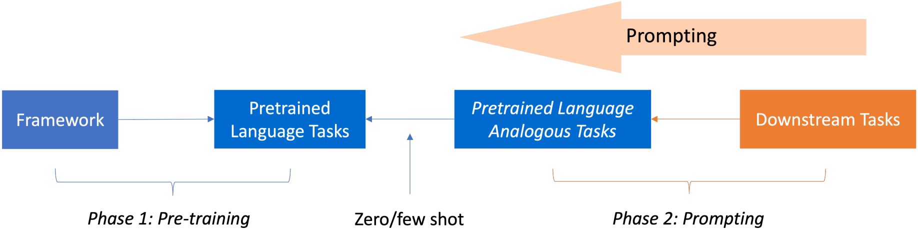
- Finally, we can ask GPT (θ) to directly perform NTP:

$$p(\texttt{pos}|\mathbf{X}) = \theta(y = \textit{"good"}|P(\mathbf{X})) \tag{4}$$

$$p(\texttt{neg}|\mathbf{X}) = \theta(y = \textit{"bad"}|P(\mathbf{X}) \tag{5}$$

Prompt Learning

- Make better use of PLM's pre-training knowledge.
- Zero-shot/few-shot Performance.
- One model for multiple NLP tasks (one PLM with multiple prompts).
- Unify NLP tasks in an NLG manner.



Reading: Language Models are Few-Shot Learners and Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference

Scaling from PLMs to LLMs

Intuitively:

- Model of larger parameters has better performance.
- Model trained on more data has better performance.

Scaling from PLMs to LLMs

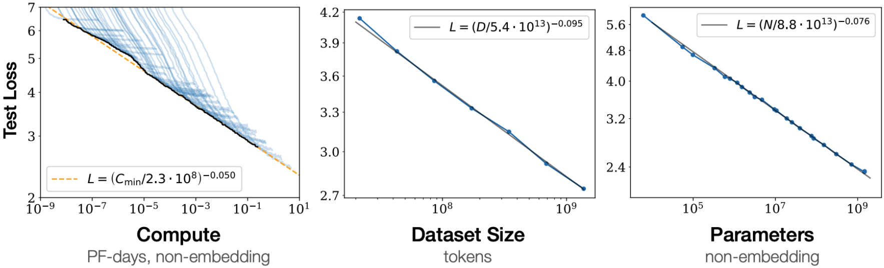
Intuitively:

- Model of larger parameters has better performance.
- Model trained on more data has better performance.

But in practical:

- But training large models on small data can be **overfitting**, and training small models are large data can be **underfitting**. How can we find the balance?
- Training budgets are **limited**. How can we make the best use of training time to maximize performance?

Scaling from PLMs to LLMs



Main findings:

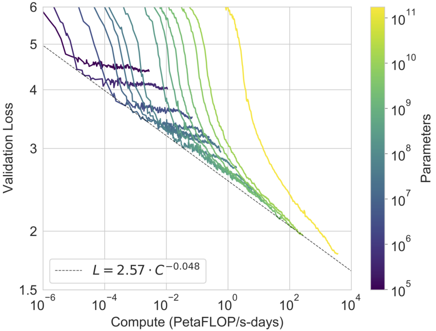
- Model performance L depends the most on amount of compute (C), size of datasets (D) and parameters (N), and each has a power-law relationship with L.
- Efficiency on the ratio of $N^{0.74}/D$.
- When C is fixed, increase large N with small D.

Reading: Scaling Laws for Neural Language Models

Scaling from PLMs to LLMs

From GPT-1 to GPT-3:

	GPT-1	**GPT-2**	**GPT-3**
Model	Transformer	Transformer	Transformer
Parameter	120M	1.5B	175B
Data Size	1.3B	10B	300B
Emergent	No	No	ICL

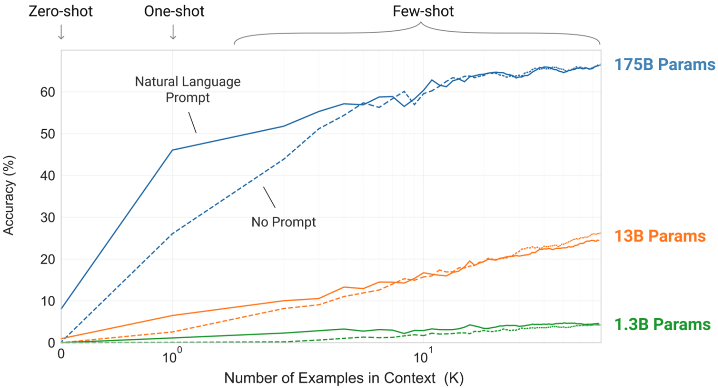


Large Language Models

In-Context Learning (ICL) in GPT-3

The ICL is regarded as an **emergent ability** of GPT-3.

- Different from task prompt (task descriptions) & Can be combined together with prompts.
- New few-shot learning paradigm (pattern recognition at inference time).

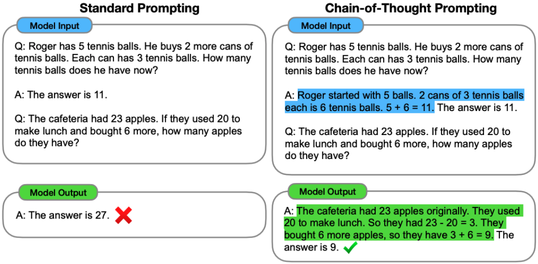


Large Language Models

Although GPT-3 is very strong at standard NLP tasks (e.g., text classification), it shows poor performance on **complex tasks**.

Pre-training on Code.

- Code-trained model shows better performance on other tasks (in particular the mathematical and logical reasoning tasks). Why?
- Many assume that *Step-by-step* reasoning (Chain-of-Thoughts, CoT) is an emergent ability from code training.



Reading: Evaluating Large Language Models Trained on Code and Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

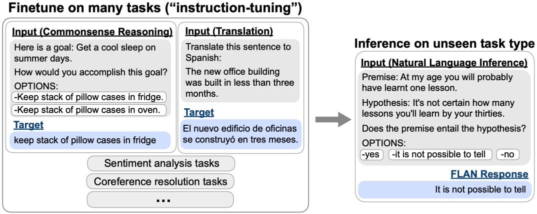
Large Language Models

Although GPT-3 is very strong at standard NLP tasks (e.g., text classification), it shows poor performance on **complex tasks**.

- Standard prompting on GPT-3.
  - GPT-3 may not understand the prompt well.
  - GPT-3 cannot perform complex task.

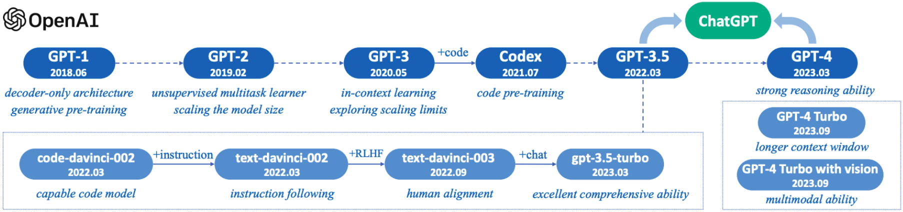
Tuning with Instructions.

- Explicit describe the goal of tasks in natural language:
  - "*Is the sentiment of this movie review positive or negative?*"
  - "*Translate 'how are you' into Chinese.*"



Reading: Finetuned Language Models Are Zero-Shot Learners

Large Language Models



More interesting topics

- Multi-modal LLMs
- LLMs as Agents (with Tools)
- Retrieval-Augmented Generation
- Hallucination in LLMs
- etc

Reading

- Chapter 3: N-gram Language Models. D Jurafsky and J Martin. *Speech and Language Processing*
  Other reading papers are embedded with hyperlinks in the previous slides of this lecture.