Foundation Models and Prompts

Danna Gurari
University of Colorado Boulder
Spring 2024
Review

• Last lecture:
  • Explosion of transformers
  • GPT
  • BERT
  • ViT
  • Programming tutorial

• Assignments (Canvas):
  • Problem set 4 due Wednesday
  • Project outline due after Spring break

• Questions?
Today’s Topics

• Motivation

• Foundation Models

• NLP: Textual Prompting

• CV: Visual Prompting (In-context Few-shot Learning)

• Parameter-Efficient Tuning Methods

• Latex Tutorial
Today’s Topics

• Motivation

• Foundation Models

• NLP: Textual Prompting

• CV: Visual Prompting (In-context Few-shot Learning)

• Parameter-Efficient Tuning Methods

• Latex Tutorial
What We Learned Works Over Past Decade

Can achieve strong performance with lots of labeled data for target task (aka closed world problems) when training from scratch or fine-tuning

https://dev.mrdbourne.com/tensorflow-deep-learning/04_transfer_learning_in_tensorflow_part_1_feature_extraction/
Open Problems: Beyond Big Data

Learning with Limited Labeled Training Data (e.g., Few-Shot Learning)
Open Problems: Beyond Closed-World Setting

**Closed-world/Closed-set**: assumes all target classes are available at training

**Vocabulary / Language**

**Out-of-domain/Robustness Testing**: same content observed differently

**Open vocabulary and Zero-shot**: generalize to task with no labeled training data for the target task (e.g., novel categories), where the former problem permits annotations with novel category (for a different task)

Open Problems: Beyond Closed-World Setting

Open vocabulary and Zero-shot: generalize to task with no labeled training data for the target task (e.g., novel categories), where the former problem permits annotations with novel category (for a different task).

Closed-world/Closed-set: assumes all target classes are available at training.

Training Domain

Out-of-domain/Robustness Testing: same content observed differently.

Open set classification/Out-of-distribution Detection: predict whether a sample is drawn from the distribution observed at training time.

New Paradigm:

Current Findings Suggest Foundation Models Generalize Well With Limited Training Data and Beyond Closed World Tasks
Today’s Topics

• Motivation

• Foundation Models

• NLP: Textual Prompting

• CV: Visual Prompting (In-context Few-shot Learning)

• Parameter-Efficient Tuning Methods

• Latex Tutorial
Foundation Model: Key Idea

Emergent behavior discovered around 2018 (in NLP) that a foundation model can be used \textit{as is} for many downstream tasks with \textit{prompting}!

Definition of “Foundation Model”

Coined in 2021, it references the recent paradigm shift to develop a single model that can implicitly support many downstream tasks.

Evaluate with modern benchmark datasets for many:

1. **Different tasks** (e.g., object recognition, scene classification)

2. **Different distributions of the same task** (e.g., ImageNet versus data from blind people)

Foundation Models: Why Now?

Key ingredients identified:

1. Transformer model architecture
2. Lots more training data by using Internet data
3. Sufficient hardware with modern GPUs

Caution: Risks of Using Foundation Models

• e.g.,
  • Model biases/limitations can affect all downstream models
  • Computationally expensive models (and so bad for environment)
Today’s Topics

• Motivation
• Foundation Models
• NLP: Textual Prompting
• CV: Visual Prompting (In-context Few-shot Learning)
• Parameter-Efficient Tuning Methods
• Latex Tutorial
Language Models are Few-Shot Learners

<table>
<thead>
<tr>
<th>Tom B. Brown*</th>
<th>Benjamin Mann*</th>
<th>Nick Ryder*</th>
<th>Melanie Subbiah*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jared Kaplan†</td>
<td>Prafulla Dhariwal</td>
<td>Arvind Neelakantan</td>
<td>Pranav Shyam</td>
</tr>
<tr>
<td>Amanda Askell</td>
<td>Sandhini Agarwal</td>
<td>Ariel Herbert-Voss</td>
<td>Gretchen Krueger</td>
</tr>
<tr>
<td>Rewon Child</td>
<td>Aditya Ramesh</td>
<td>Daniel M. Ziegler</td>
<td>Jeffrey Wu</td>
</tr>
<tr>
<td>Christopher Hesse</td>
<td>Mark Chen</td>
<td>Eric Sigler</td>
<td>Mateusz Litwin</td>
</tr>
<tr>
<td>Benjamin Chess</td>
<td>Jack Clark</td>
<td>Christopher Berner</td>
<td></td>
</tr>
<tr>
<td>Sam McCandlish</td>
<td>Alec Radford</td>
<td>Ilya Sutskever</td>
<td>Dario Amodei</td>
</tr>
</tbody>
</table>

OpenAI
Key Idea: Increase Model Size

GPT-3: Model Design

Slightly modified version of GPT with many more layers!

Extends GPT-2 (Radford et al. OpenAI blog 2019)
## GPT-3: Model Design (8 Tested Variants)

<table>
<thead>
<tr>
<th>Model Name</th>
<th>$n_{\text{params}}$</th>
<th>$n_{\text{layers}}$</th>
<th>$d_{\text{model}}$</th>
<th>$n_{\text{heads}}$</th>
<th>$d_{\text{head}}$</th>
<th>Batch Size</th>
<th>Learning Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3 Small</td>
<td>125M</td>
<td>12</td>
<td>768</td>
<td>12</td>
<td>64</td>
<td>0.5M</td>
<td>$6.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 Medium</td>
<td>350M</td>
<td>24</td>
<td>1024</td>
<td>16</td>
<td>64</td>
<td>0.5M</td>
<td>$3.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 Large</td>
<td>760M</td>
<td>24</td>
<td>1536</td>
<td>16</td>
<td>96</td>
<td>0.5M</td>
<td>$2.5 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 XL</td>
<td>1.3B</td>
<td>24</td>
<td>2048</td>
<td>24</td>
<td>128</td>
<td>1M</td>
<td>$2.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 2.7B</td>
<td>2.7B</td>
<td>32</td>
<td>2560</td>
<td>32</td>
<td>80</td>
<td>1M</td>
<td>$1.6 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 6.7B</td>
<td>6.7B</td>
<td>32</td>
<td>4096</td>
<td>32</td>
<td>128</td>
<td>2M</td>
<td>$1.2 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 13B</td>
<td>13.0B</td>
<td>40</td>
<td>5140</td>
<td>40</td>
<td>128</td>
<td>2M</td>
<td>$1.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>GPT-3 175B or “GPT-3”</td>
<td>175.0B</td>
<td>96</td>
<td>12288</td>
<td>96</td>
<td>128</td>
<td>3.2M</td>
<td>$0.6 \times 10^{-4}$</td>
</tr>
</tbody>
</table>
# GPT-3: Training Data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Quantity (tokens)</th>
<th>Weight in training mix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Crawl (filtered)</td>
<td>410 billion</td>
<td>60%</td>
</tr>
<tr>
<td>WebText2</td>
<td>19 billion</td>
<td>22%</td>
</tr>
<tr>
<td>Books1</td>
<td>12 billion</td>
<td>8%</td>
</tr>
<tr>
<td>Books2</td>
<td>55 billion</td>
<td>8%</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>3 billion</td>
<td>3%</td>
</tr>
</tbody>
</table>

(Additional 4 sources are known high-quality datasets to supplement the noisy, lower quality Common Crawl data)

Language composition: by word count, 93% English

**Common Crawl** (popular NSP source)  
(web archives without HTML markup and non-text content https://commoncrawl.org/)

Common Crawl is a 501(c)(3) non-profit founded in 2007.

We make wholesale extraction, transformation and analysis of open web data accessible to researchers.

- **Over 250 billion pages spanning 17 years.**
- **Free and open corpus since 2007.**
- **Cited in over 10,000 research papers.**
- **3–5 billion new pages added each month.**
GPT-3: Prompts Include Instructions &, Optionally, Examples (Latter Called “In-Context Learning”)

Zero-shot

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

<table>
<thead>
<tr>
<th></th>
<th>task description</th>
<th>prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Translate English to French:</td>
<td>cheese =&gt;</td>
</tr>
</tbody>
</table>

Few-shot

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

<table>
<thead>
<tr>
<th></th>
<th>task description</th>
<th>examples</th>
<th>prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Translate English to French:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>sea otter =&gt; loutre de mer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>peppermint =&gt; menthe poivrée</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>plush giraffe =&gt; girafe peluche</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>cheese =&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One-shot

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

<table>
<thead>
<tr>
<th></th>
<th>task description</th>
<th>example</th>
<th>prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Translate English to French:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>sea otter =&gt; loutre de mer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>cheese =&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prompt Designed Per Dataset; e.g.,

<table>
<thead>
<tr>
<th>Context →</th>
<th>Q: What school did burne hogarth establish?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:</td>
<td></td>
</tr>
<tr>
<td>Target Completion →</td>
<td>School of Visual Arts</td>
</tr>
</tbody>
</table>

**Figure G.35:** Formatted dataset example for WebQA

<table>
<thead>
<tr>
<th>Context →</th>
<th>Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden.</th>
</tr>
</thead>
<tbody>
<tr>
<td>=</td>
<td></td>
</tr>
<tr>
<td>Target Completion →</td>
<td>In no case may they be used for commercial purposes.</td>
</tr>
</tbody>
</table>

**Figure G.36:** Formatted dataset example for De→En. This is the format for one- and few-shot learning, for this and other language tasks, the format for zero-shot learning is “Q: What is the {language} translation of {sentence} A: {translation}.”

Example Result: Fake News Generation

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split – one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church’s annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination. The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church’s history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.
Experimental Findings

Tested on 10s of NLP datasets, showing strong performance overall and occasionally state-of-the-art performance!

Successors: GPT 3.5 and GPT 4

https://openai.com/blog/chatgpt

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

Try ChatGPT ▶️ Read about ChatGPT Plus

GPT-4 is OpenAI's most advanced system, producing safer and more useful responses

https://openai.com/gpt-4
Challenge: What Prompts to Use?

- **Chain-of-thought** prompting can help by guiding model to show its intermediate reasoning steps!

  **Standard Prompting**
  
  **Model Input**
  
  Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
  
  A: The answer is 11.
  
  Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
  
  **Model Output**
  
  A: The answer is 27. ✗

  **Chain-of-Thought Prompting**
  
  **Model Input**
  
  Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
  
  A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.
  
  Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
  
  **Model Output**
  
  A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Challenge: What Prompts to Use?

- Why COT prompting works? examples may reveal the target output format as performance still improves with invalid examples; e.g.,

| CoT                  | Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read 24 pages. So she read a total of $12 + 24 = 36$ pages. Now she has $120 - 36 = 84$ pages left. Since she wants to read half of the remaining pages, she should read $84 / 2 = 42$ pages. The answer is 42. ✓
| Originally, Leah had 32 chocolates and her sister had 42. So in total they had $32 + 42 = 74$. After eating 35, they had $74 - 35 = 39$ pieces left in total. The answer is 39. |

| Invalid Reasoning   | Julie is reading a 120-page book. Yesterday, she read 12 pages. Today, she read $12 * 2 = 24$ pages. So she read a total of $12 + 24 = 36$ pages. Now she needs to read $120 - 36 = 84$ more pages. She wants to read half of the remaining pages tomorrow, so she needs to read $84 / 2 = 42$ pages tomorrow. The answer is 42. ✓
| Originally, Leah had 32 chocolates and her sister had 42. So her sister had $42 - 32 = 10$ chocolates more than Leah has. After eating 35, since $10 + 35 = 45$, they had $45 - 6 = 39$ pieces left in total. The answer is 39. |

Challenge: What Prompts to Use?


Uses most consistent answer from multiple predicted answers! (as is, only applicable when answers are from a fixed set)
Challenge: What Prompts to Use?

Models can also perform better when asked to show their reasoning steps without seeing reasoning examples. e.g., for GPT-3:

Today’s Topics

• Motivation

• Foundation Models

• NLP: Textual Prompting

• CV: Visual Prompting (In-context Few-shot Learning)

• Parameter-Efficient Tuning Methods

• Latex Tutorial
Motivating Observation

• Foundation models achieved better performance for NLP tasks when provided “in-context” examples.
  
  • i.e., [Task description, Examples, Prompt]
  • e.g., “Translate English to Spanish. Computer -> Computadora. Vision ->

• Idea: Use in-context few-shot learning for image-based prompts
Novel Idea: Image Inpainting

Designed to adapt to any “image-to-image translation” task by using the model as is (e.g., no fine-tuning required)

Idea

Image inpainting for prompting introduced in 2022 by Bar et al.

Edge detection  Colorization  Inpainting  Segmentation  Style transfer

Idea extended in 2023 by Wang et al. on standard vision benchmark datasets.

Training: Masked Image Modeling

Uses self-supervised learning such that the model predicts values in masked-out patches.

Uses standard vision benchmarks for each evaluated task.

Experimental Results

(Used in prompt the best performing example-per pair per task from all examples in the training dataset)

Model achieves state-of-the-art performance on depth estimation for NYUv2 dataset and outperforms other generalist models on several more tasks.
Qualitative Results: In-Domain Results

Qualitative Results: In-Domain Results

Qualitative Results: In-Domain Results

Qualitative Results: Open-Vocabulary Results (i.e., Categories Not Seen at Training)

Shows in-context examples, prompts, and predictions for keypoint detection, object segmentation, and instance segmentation.

Today’s Topics

• Motivation

• Foundation Models

• NLP: Textual Prompting

• CV: Visual Prompting (In-context Few-shot Learning)

• Parameter-Efficient Tuning Methods

• Latex Tutorial
Motivation: Improve Model Performance Without Humans

Manually engineering prompts is challenging to do well (leading to MANY prompt marketplaces)
Idea: Replace Manually-Authored Prompts with Learnable Parameters

Learned prompts adapt frozen model (e.g., no fine-tuning required) to different target tasks

More Generally, Fix Model and Modify Some Added Parameters
(As Shown, There Are Many Approaches)

Figure 1: Illustration of the transformer architecture and several state-of-the-art parameter-efficient tuning methods. We use blocks with dashed borderlines to represent the added modules by those methods.

Competitive Performance With Full Fine-Tuning While Training 15% or Less Parameters

Figure 2: Performance of different methods on the XSum (Narayan et al., 2018) summarization task. The number of fine-tuned parameters is relative to the tuned parameters in full fine-tuning.

What Are Benefits of Tuning Methods?

• Typically, little training data is needed because only a limited amount of parameters need to be trained

• Few task-specific parameters need to be learned and stored to support a new task, compared to model fine-tuning

• Prevents overfitting generalizable knowledge and overfitting to the task

• Provides a static knowledge-base
Today’s Topics

• Motivation

• Foundation Models

• NLP: Textual Prompting

• CV: Visual Prompting (In-context Few-shot Learning)

• Parameter-Efficient Tuning Methods

• Latex Tutorial
Today’s Topics

• Motivation

• Foundation Models

• NLP: Textual Prompting

• CV: Visual Prompting (In-context Few-shot Learning)

• Parameter-Efficient Tuning Methods

• Latex Tutorial