IMPORTANCE OF SMALL MODELS

OpenAI GPT-4.0 Mini Announcement

We're continuing to make advanced AI accessible to all with the launch of GPT-4.0 mini, now available in the API and rolling out in ChatGPT today.

Introducing GPT-4.0 mini! It's our most intelligent and affordable small model, available today in the API. GPT-4.0 mini is significantly smarter and cheaper than GPT-3.5 Turbo.

openai.com/index/gpt-4.0-m...

Pricing

GPT-4.0 mini is more than 60% cheaper than GPT-3.5 Turbo

Note: This chart shows blended pricing assuming 80% input tokens and 20% output tokens.
IMPORATANCE OF SMALL MODELS

GPT-4o mini evaluations
IMPORTANCE OF SMALL MODELS

MMLU vs. Price, Smaller models

MMLU: General reasoning quality benchmark, Price: USD per 1M Tokens

- Most attractive quadrant
- GPT-4o Mini
- GPT-3.5 Turbo
- Gemini 1.5 Flash
- Llama 3 (70B)
- Llama 3 (8B)
- NeMo
- Mistral 7B
- Claude 3 Haiku
- Command-R
- Reka Edge

ArtificialAnalysis.ai
IMPORTANCE OF SMALL MODELS

Andrej Karpathy

LLM model size competition is intensifying... backwards!

My bet is that we'll see models that "think" very well and reliably that are very very small. There is most likely a setting even of GPT-2 parameters for which most people will consider GPT-2 "smart". The reason current models are so large is because we're still being very wasteful during training - we're asking them to memorize the internet and, remarkably, they do and can e.g. recite SHA hashes of common numbers, or recall really esoteric facts. (Actually LLMs are really good at memorization, qualitatively a lot better than humans, sometimes needing just a single update to remember a lot of detail for a long time). But imagine if you were going to be tested, closed book, on reciting arbitrary passages of the internet given the first few words. This is the standard (pre)training objective for models today. The reason doing better is hard is because demonstrations of thinking are "entangled" with knowledge, in the training data.

Therefore, the models have to first get larger before they can get smaller, because we need their (automated) help to refactor and mold the training data into ideal, synthetic formats.

It's a staircase of improvement - of one model helping to generate the training data for next, until we're left with "perfect training set". When you train GPT-2 on it, it will be a really strong / smart model by today's standards. Maybe the MMLU will be a bit lower because it won't remember all of its chemistry perfectly. Maybe it needs to look something up once in a while to make sure.
SUMMARIZING ANDREJ’S TWEET

• A competition has started, which contrasts with the previous competition: to create smaller models

• Creating large models before small models was inevitable and an essential step.

• Models were large because their training was not efficient. However, they have absorbed the knowledge of the internet.

• Now, it is possible to effectively use this knowledge to create smaller models.
QUANTIZATION

LARGE SET OF POSSIBLE VALUES

SMALL SET OF POSSIBLE VALUES
QUANTIZATION: LARGE LANGUAGE MODELS

- Often means reducing the precision of the weights’ values:
  - Smaller precision results in smaller memory requirement
  - Smaller precision results in faster inference
### QUANTIZATION: LARGE LANGUAGE MODELS

<table>
<thead>
<tr>
<th></th>
<th>FP 32</th>
<th></th>
<th>INT 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.34</td>
<td>3.75</td>
<td>5.64</td>
<td>8</td>
</tr>
<tr>
<td>1.12</td>
<td>2.7</td>
<td>-0.9</td>
<td>84</td>
</tr>
<tr>
<td>-4.7</td>
<td>0.68</td>
<td>1.43</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-106</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32</td>
</tr>
</tbody>
</table>
# QUANTIZATION: LARGE LANGUAGE MODELS

<table>
<thead>
<tr>
<th>Precision</th>
<th>Range</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP 32</td>
<td>$-3.4 \times 10^{38} \rightarrow 3.4 \times 10^{38}$</td>
<td>$4.2 \times 10^{9}$</td>
</tr>
<tr>
<td>INT 8</td>
<td>$-128 \rightarrow 127$</td>
<td>256</td>
</tr>
</tbody>
</table>
QUANTIZATION: LARGE LANGUAGE MODELS

\[ X_{\text{quant}} = \text{round} \left( \frac{127}{\max |X|} \cdot X \right) \]

\[ X_{\text{dequant}} = \frac{\max |X|}{127} \cdot X_{\text{quant}} \]
QUANTIZATION: LARGE LANGUAGE MODELS

Qualitative comparison

[Graph showing the relationship between model size (GiB) and perplexity for different model sizes: 7B, 13B, 30B, 65B, both quantized and f16 versions.]
INTRODUCTION

GPT-4

Teacher

Data

Student

Distillation Loss

Llama-7B
INTRODUCTION

• black-box KD and white-box KD
• black-box KD has shown promising results in fine-tuning small models
• white-box KD approaches are mostly studied for small (<1B parameters) language understanding models
**Problem:** Student doesn’t have the capacity of the teacher

**Solution:** Using the teacher output as a signal to improve student performance
Why not use Large Language Models instead
SELF-IMPROVEMENT (DISTILLATION)

USING THE KNOWLEDGE OF THE MODEL TO IMPROVE THE MODEL

Andrei's talk on language models
THANK YOU!