How Much Can GPT-3 Benefit Few-Shot Visual Reasoning?

Zhe Gan
Principal Researcher
Language Model Pre-training

- Large-scale language model pre-training has become a central training paradigm for NLP
- Parameter-counts are frequently measured in billions (e.g., GPT-3) rather than millions (e.g., BERT)

<table>
<thead>
<tr>
<th>Model</th>
<th>Company</th>
<th>Param. Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT</td>
<td>OpenAI</td>
<td>110M</td>
</tr>
<tr>
<td>BERT-Large</td>
<td>Google</td>
<td>340M</td>
</tr>
<tr>
<td>GPT-2</td>
<td>OpenAI</td>
<td>1.5B</td>
</tr>
<tr>
<td>MegatronLM</td>
<td>NVIDIA</td>
<td>8.3B</td>
</tr>
<tr>
<td>T-NLG</td>
<td>Microsoft</td>
<td>17B</td>
</tr>
<tr>
<td>GPT-3</td>
<td>OpenAI</td>
<td>175B</td>
</tr>
<tr>
<td>Switch-C</td>
<td>Google</td>
<td>1.6T</td>
</tr>
</tbody>
</table>
Language Models are Few-Shot Learners

• By providing only a few in-context examples, GPT-3 with 175B parameters has demonstrated strong few-shot performance

Can GPT-3 also Benefit Visual Reasoning Tasks?

GQA

VQA

VCR

Referring Expressions

CLEVR

NLVR2
Knowledge-Based VQA

- **OK-VQA**: A VQA benchmark requiring external knowledge not present in the image to correctly answer the question

Previous Methods vs. Ours

• Previous methods:
  • Separate two steps: knowledge retrieval and reasoning
  • Using explicit and structured KBs
  • The retrieved knowledge might be noisy and irrelevant to the question
  • The re-embedded knowledge features during reasoning might deviate from their original meanings in the knowledge source
Previous Methods vs. Ours

- **Previous methods:**
  - Separate two steps: knowledge retrieval and reasoning

- **Our method:**
  - **PICa:** Prompting GPT-3 via the use of Image Captions
  - Treating GPT-3 as an *implicit* and *unstructured* KB
  - 4 shots outperform supervised SOTA
How Do We Prompt GPT-3?

- Convert images into textual descriptions (captions, tags)
- Produce the answer in an open-ended text generation manner
How to Enhance the Performance?

- In-context example selection
  - "Better" in-context examples based on question and image similarity

How to Enhance the Performance?

- In-context example selection
  - “Better” in-context examples based on question and image similarity
- Multi-query ensemble
  - Merge predictions from multiple queries with different examples
PICa Outperforms Supervised SOTA by +8.6 Points

- **PICa-Base**: w/o in-context selection and multi-query ensemble
- **PICa-Base** (43.3) already surpasses SOTA (39.4)
- **PICa-Full** further boosts performance (48.0)
- Both captions and tags are useful for GPT-3 prompting
How Many Shots are Enough?

- 4 shots outperform supervised state-of-the-art (39.4)
- More shots lead to better performance
- PICa outperforms Frozen by a significant margin

<table>
<thead>
<tr>
<th>Method</th>
<th>Image Repr.</th>
<th>n=0</th>
<th>n=1</th>
<th>n=4</th>
<th>n=8</th>
<th>n=16</th>
<th>Example engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Frozen (Tsimploukelli et al. 2021)</td>
<td>Feature Emb.</td>
<td>5.9</td>
<td>9.7</td>
<td>12.6</td>
<td>-</td>
<td>-</td>
<td>×</td>
</tr>
<tr>
<td>(b) PICa-Base</td>
<td>Caption</td>
<td>17.5</td>
<td>32.4</td>
<td>37.6</td>
<td>39.6</td>
<td>42.0</td>
<td>X</td>
</tr>
<tr>
<td>(c) PICa-Base</td>
<td>Caption+Tags</td>
<td>16.4</td>
<td>34.0</td>
<td>39.7</td>
<td>41.8</td>
<td>43.3</td>
<td>X</td>
</tr>
<tr>
<td>(d) PICa-Full</td>
<td>Caption</td>
<td>17.7</td>
<td>40.3</td>
<td>44.8</td>
<td>46.1</td>
<td>46.9</td>
<td>✓</td>
</tr>
<tr>
<td>(e) PICa-Full</td>
<td>Caption+Tags</td>
<td>17.1</td>
<td>40.8</td>
<td>45.4</td>
<td>46.8</td>
<td>48.0</td>
<td>✓</td>
</tr>
</tbody>
</table>
Why GPT-3 are So Powerful on OK-VQA?

- Encyclopedia and commonsense knowledge
Why GPT-3 Are So Powerful on OK-VQA?

- GPT-3 also generates answer rationales reasonably well

![Images and text explanations for questions (a) to (e) with rationales for selected examples.](image-url)
Key Takeaways

- The first study of using GPT-3 for multimodal tasks
- With 16 in-context examples, GPT-3 surpasses the supervised SOTA by an absolute +8.6 points on the challenging OK-VQA dataset
Limitations of Using GPT-3 for Visual Reasoning

- Converting images into captions could lose important visual info.
- We test PICa on VQAv2, with questions on detailed visual contents
- Human performance:
  - 40.8% with question only
  - 57.5% with question and captions
  - 83.3% with question and images

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Oscar (Li et al. 2020)</td>
<td>Feature Emb.</td>
<td>X</td>
<td>73.8</td>
</tr>
<tr>
<td>Frozen</td>
<td>Feature Emb.</td>
<td>✓</td>
<td>38.2</td>
</tr>
<tr>
<td>PICa–Base</td>
<td>Caption</td>
<td>✓</td>
<td>53.2</td>
</tr>
<tr>
<td>PICa–Base</td>
<td>Caption+Tags</td>
<td>✓</td>
<td>54.3</td>
</tr>
<tr>
<td>PICa–Full</td>
<td>Caption</td>
<td>✓</td>
<td>55.9</td>
</tr>
<tr>
<td>PICa–Full</td>
<td>Caption+Tags</td>
<td>✓</td>
<td><strong>56.1</strong></td>
</tr>
<tr>
<td>PICa–Full†</td>
<td>GT-Caption-5</td>
<td>✓</td>
<td>59.7</td>
</tr>
</tbody>
</table>
Limitations of Using GPT-3 for Visual Reasoning

- **Success**: knowledge, related descriptions, and reasonable guesses
- **Failures**: missing detailed visual information

---

(a) What kind of sign is above the doorway?
**Context**: A computer monitor sitting on top of a desk.
**Answer**: exit.
**GT Answer**: [exit, 'exit', 'exit', 'exit', 'exit', 'exit', 'exit', 'exit', 'exit', 'exit']
**Acc.**: 1.0

(b) Why is the cow laying down?
**Context**: A group of cows in a barn with one standing and one laying down.
**Answer**: tired
**GT Answer**: [tired, 'tired', 'tired', 'tired', 'tired', 'tired', 'tired', 'tired', 'tired', 'tired']
**Acc.**: 1.0

(c) What color is the fridge?
**Context**: A silver refrigerator freezer sitting inside of a kitchen.
**Answer**: silver
**GT Answer**: [silver, 'silver', 'silver', 'silver', 'silver', 'silver', 'silver', 'silver', 'silver', 'silver']
**Acc.**: 1.0

(d) What color are the walls?
**Context**: A bedroom with a desk, chair, computer and pictures on the wall.
**Answer**: white
**GT Answer**: [white, 'white', 'white', 'white', 'white', 'white', 'white', 'white', 'white', 'white']
**Acc.**: 1.0

(e) What color is the man's jacket?
**Context**: A man flying through the air while riding a snowboard.
**Answer**: black
**GT Answer**: [black, 'black', 'black', 'black', 'black', 'black', 'black', 'black', 'black', 'black']
**Acc.**: 0.0

(f) How many giraffes are there?
**Context**: A herd of giraffes standing next to a wooden fence.
**Answer**: 3
**GT Answer**: [3, '3', '3', '3', '3', '3', '3', '3', '3', '3']
**Acc.**: 0.0
Future Direction

- **Looking back**, Microsoft has been an important player in the vision-language pre-training (VLP) space.
- **Looking forward**, when can we have the GPT-3 moment for VLP?

![Timeline of VLP Models](image)
Future Direction

- Multimodal GPT-3:
  - Instead of converting images into captions, *learn a vision encoder* to align with the language embedding space in GPT-3

Collaborators

Zhengyuan Yang                     Jianfeng Wang                        Xiaowei Hu
Yumao Lu                             Zicheng Liu                          Lijuan Wang
Thank you!