GPT-3 & Clip for Image Classification

Gregory Presser, Jacob Khalili *

November 15, 2022

1 Introduction

Humans [and maybe (other) animals too] identify objects using features. For example, a person may identify a dog, because it has 4 legs, fur, and a tail. In Visual Classification via Description from Large Language Models [MV22] the authors attempt to get descriptors of categories from GPT-3 to increase classification accuracy of a model with the CLIP style architecture. We attempt to reproduce these results.

2 Methodology

2.1 Descriptor Generation via GPT-3

We utilized a Open AI's GPT-3 Davinci Model [BMR+20] for generating descriptors based on model categories, via API. We prompted the model in a similar manner to the paper, as shown in Figure 1.

GPT-3 was prompted with category-name replaced for each one of the categories in the Image-Net-1000 classification data-set. [RDS+15]

Q: What are useful visual features for distinguishing a lemur in a photo?
A: There are several useful visual features to tell there is a lemur in a photo:
- four-limbed primate
- black, grey, white, brown, or red-brown
- wet and hairless nose with curved nostrils
- long tail
- large eyes
- furry bodies
- clawed hands and feet

Q: What are useful visual features for distinguishing a {category-name} in a photo?
A: There are several useful visual features to tell there is a {category-name} in a photo:

Figure 1: GPT-3 Prompt to obtain descriptors

2.2 Image Classification via Model with CLIP Architecture

We then passed the descriptors to a pre-trained model with a CLIP architecture. [RKH+21] We used the ViT-B/32 pre-trained model, which was available via Open-AI's Clip library. We passed in the text in form: "a photo of a [class-name]" for the Vanilla Clip Model, and for our model "a photo of a [class-name], which (has/is) [descriptor]"

For the Vanilla clip model, to obtain the a prediction, we use the class with the greatest log-probabilities. However, our model sums the weights of each of the descriptors for a category. We were able to optimize this process via a clever matrix multiplication (similar technique to binary-coding in Comm-Theory).

*Students at the Cooper Union for the Advancement of Science & Art. This work was done as part of ECE-472, a graduate course in Deep Learning

3 Results

Figure 4 shows some sample results of both models test, and the ground truth data. Figure 2 shows the results for different architetures

Model	Ours	CLIP Baseline	Paper Results	Paper Baseline
ViT-B/32	60.52%	59.52%	62.97%	58.46%

Figure 2: Results when comparing a baseline CLIP model VS

4 Complaints/Future Work/Conclusion

4.1 Complaints

The paper did not give enough detail to accurately reproduce their work. There were several assumptions that were made, which may have differed from the authors. This includes: exactly where the data-set was obtained from, and pre-processing of the image (including cropping etc).

Image-Net classes are also confusing to me as a human (this paper was not written by GPT-3), and the classes are sometimes wrong. A notable example that we discovered is shown in Figure 3, which is clearly not a cape.

Figure 3: This image was listed as a cape in the Image-Net dataset

4.2 Future Work

In order to properly replicate this paper, we would have needed more compute available to us. Running these model on the entire Image-Net data-set was not feasible given our compute constraints. This likely explains some of the variation from the the results shown in the paper.

4.3 Conclusion

We were able to reproduce the results of the paper to a lesser effect. The original paper yielded a classification accuracy of 62.97% with a baseline clip model resulting in 58.46% which is approximately a 4% improvement. In our experiments we were able to produce a classification accuracy of 60.52% with

a baseline clip model resulting in 59.52% while classifying 5000 random images which is approximately a 0.5% improvement.

References

[BMR+20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[MV22] Sachit Menon and Carl Vondrick. Visual classification via description from large language models, 2022.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and

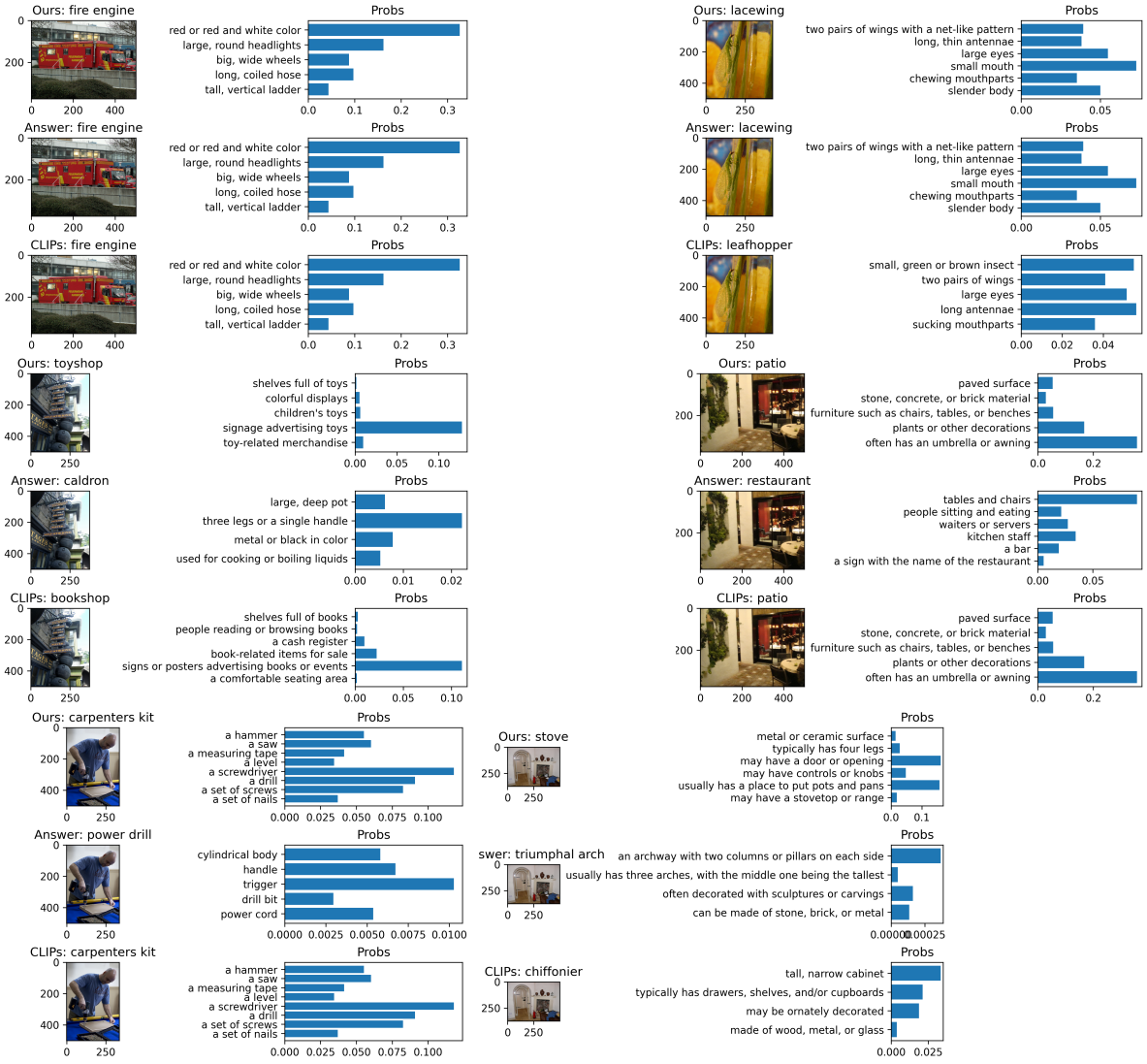

Figure 4: Sample Results from both models, and the correct answer, with probabilities of ground truth descriptors

Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.