
ChatGPT-Powered Hierarchical Comparisons for Image Classification Supplementary Material

Zhiyuan Ren, Yiyang Su, and Xiaoming Liu

Department of Computer Science and Engineering, Michigan State University
East Lansing, MI 48824
{renzhiy1, suyiyang1, liuxm}@msu.edu

1 Illustrations

1.1 ChatGPT (2) Responses

We show two examples of our summary-based comparison and direct comparison in Fig. 1, Fig. 2, Fig. 3 and Fig. 4. In each example, we compare the comparative descriptions with the baseline descriptions (6) to show our method’s strength. Compared with the baseline, our method can give set-aware and comparative descriptions, which benefits classification.

Q: summarize the following categories with one sentence:
Salmon, Goldfish, Piranha, Zebra Shark, Whale Shark,
Snapper, Swordfish, Bass, Trout?
A: this is a dataset of various fishes

[more examples ...]

Q: summarize the following categories with one
sentence: cobwebbed, interlaced, meshed, stratified,
zigzagged?
A: this dataset includes different types of patterns or
structures involving interlocking elements

(a) Summary.

Q: What are useful visual features for distinguishing a Clay Colored Sparrow in a photo in a dataset: This dataset consists of various sparrows?

A: There are several useful visual features to tell there is a Clay Colored Sparrow in a photo:

- a distinct pale crown stripe or central crown patch
- a dark eyeline and a pale stripe above the eye
- brownish-gray upperparts
- conical-shaped bill

[more examples ...]

Q: What are useful features for distinguishing a cobwebbed in a photo: this dataset includes different types of patterns or structures involving interlocking elements?

A: There are several useful visual features to tell there is a cobwebbed in a photo:

- Presence of a central hub or anchor point from which the threads radiate
- Concentration of threads in a specific area, creating a dense network
- Delicate and fine thread thickness
- Potential presence of captured prey or debris within the web structure

(b) Summary-based comparison (ours).

Q: What are useful visual features for distinguishing a Clay Colored Sparrow in a photo?

A: There are several useful visual features to tell there is a Clay Colored Sparrow in a photo:

- a distinct pale crown stripe or central crown patch
- a dark eyeline and a pale stripe above the eye
- brownish-gray upperparts
- conical-shaped bill

[more examples ...]

Q: What are useful features for distinguishing a cobwebbed in a photo?

A: There are several useful visual features to tell there is a cobwebbed in a photo:

I'm sorry, but I'm not familiar with the term "cobwebbed" in the context of a specific species or object. Could you please provide more information or clarify your question?

(c) Baseline.

Figure 1: Example 1. A detailed illustration of summary-based comparison compared with the baseline.

Q: summarize the following categories with one sentence:
Salmon, Goldfish, Piranha, Zebra Shark, Whale Shark,
Snapper, Swordfish, Bass, Trout?
A: this is a dataset of various fishes

[more examples ...]

Q: summarize the following categories with one sentence:
club sandwich, grilled cheese sandwich, hamburger, hot
dog, pizza, pulled pork sandwich, tacos?
A: this dataset lists different types of popular food
items

(a) Summary.

Q: What are useful visual features for distinguishing a Clay Colored Sparrow in a photo in a dataset: This dataset consists of various sparrows?

A: There are several useful visual features to tell there is a Clay Colored Sparrow in a photo:

- a distinct pale crown stripe or central crown patch
- a dark eyeline and a pale stripe above the eye
- brownish-gray upperparts
- conical-shaped bill

[more examples ...]

Q: What are useful features for distinguishing a club sandwich in a photo: this dataset lists different types of popular food items?

A: There are several useful visual features to tell there is a club sandwich in a photo:

- Generally cut into quarters or halves, often held together by toothpicks
- Typically includes layers of turkey or chicken, bacon, lettuce, and tomato
- May include cheese or mayonnaise
- Often visible from the side due to the triple-decker structure

(b) Summary-based comparison (ours).

Q: What are useful visual features for distinguishing a Clay Colored Sparrow in a photo?

A: There are several useful visual features to tell there is a Clay Colored Sparrow in a photo:

- a distinct pale crown stripe or central crown patch
- a dark eyeline and a pale stripe above the eye
- brownish-gray upperparts
- conical-shaped bill

[more examples ...]

Q: What are useful features for distinguishing a club sandwich in a photo?

A: There are several useful visual features to tell there is a club sandwich in a photo:

- held together with toothpicks or skewers
- usually made with three slices of bread, creating a triple-decker structure
- a visually appealing appearance

(c) Baseline.

Figure 2: Example 2. A detailed illustration of summary-based comparison compared with the baseline.

Q: What are useful visual features for distinguishing Hooded Oriole from Scott Oriole, Baltimore Oriole in a photo

A: There are several useful visual features to tell there is a Hooded Oriole in a photo:

- distinctive bright orange or yellow and black coloration
- orange or yellow body and underparts
- noticeably curved downwards bill
- a black bib or "hood" that extends up over the head and down the back

[more examples ...]

Q: What are useful features for distinguishing a crosshatched from lined in a photo?

A: There are several useful visual features to tell there is a crosshatched in a photo:

- The presence of intersecting lines that create a grid-like pattern
- The lines are typically evenly spaced and perpendicular to each other
- The areas between the lines are filled with parallel lines in a different direction, creating a crosshatch effect

(a) Direct Comparison (ours).

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

[more examples ...]

Q: What are useful features for distinguishing a crosshatched in a photo?

A: There are several useful visual features to tell there is a crosshatched in a photo:

- The lines may vary in thickness, density, or color
- The areas between the lines may be filled with parallel lines in a different direction

(b) Baseline.

Figure 3: Example 1. A detailed illustration for direct comparison.

Q: What are useful visual features for distinguishing Hooded Oriole from Scott Oriole, Baltimore Oriole in a photo
A: There are several useful visual features to tell there is a Hooded Oriole in a photo:

- distinctive bright orange or yellow and black coloration
- orange or yellow body and underparts
- noticeably curved downwards bill
- a black bib or "hood" that extends up over the head and down the back

[more examples ...]

Q: What are useful features for distinguishing a gnocchi from macaroni and cheese and risotto in a photo?
A: There are several useful visual features to tell there is a gnocchi in a photo:

- White or yellowish in color
- Served with sauces, but not typically covered with cheese like macaroni and cheese
- Individual pieces rather than a creamy, grain-based dish

(a) Direct Comparison (ours).

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

[more examples ...]

Q: What are useful features for distinguishing a gnocchi in a photo?
A: There are several useful visual features to tell there is a gnocchi in a photo:

- Small, rounded or oval shape
- Small ridges or indentations on their surface

(b) Baseline.

Figure 4: Example 2. A detailed illustration for direct comparison.

1.2 Hierarchies

We also visualize the complete tree structure of the Describable Textures dataset in Fig. 5. The visualization shows that our tree can have multiple depths for different nodes and similar nodes will be classified into the same group.

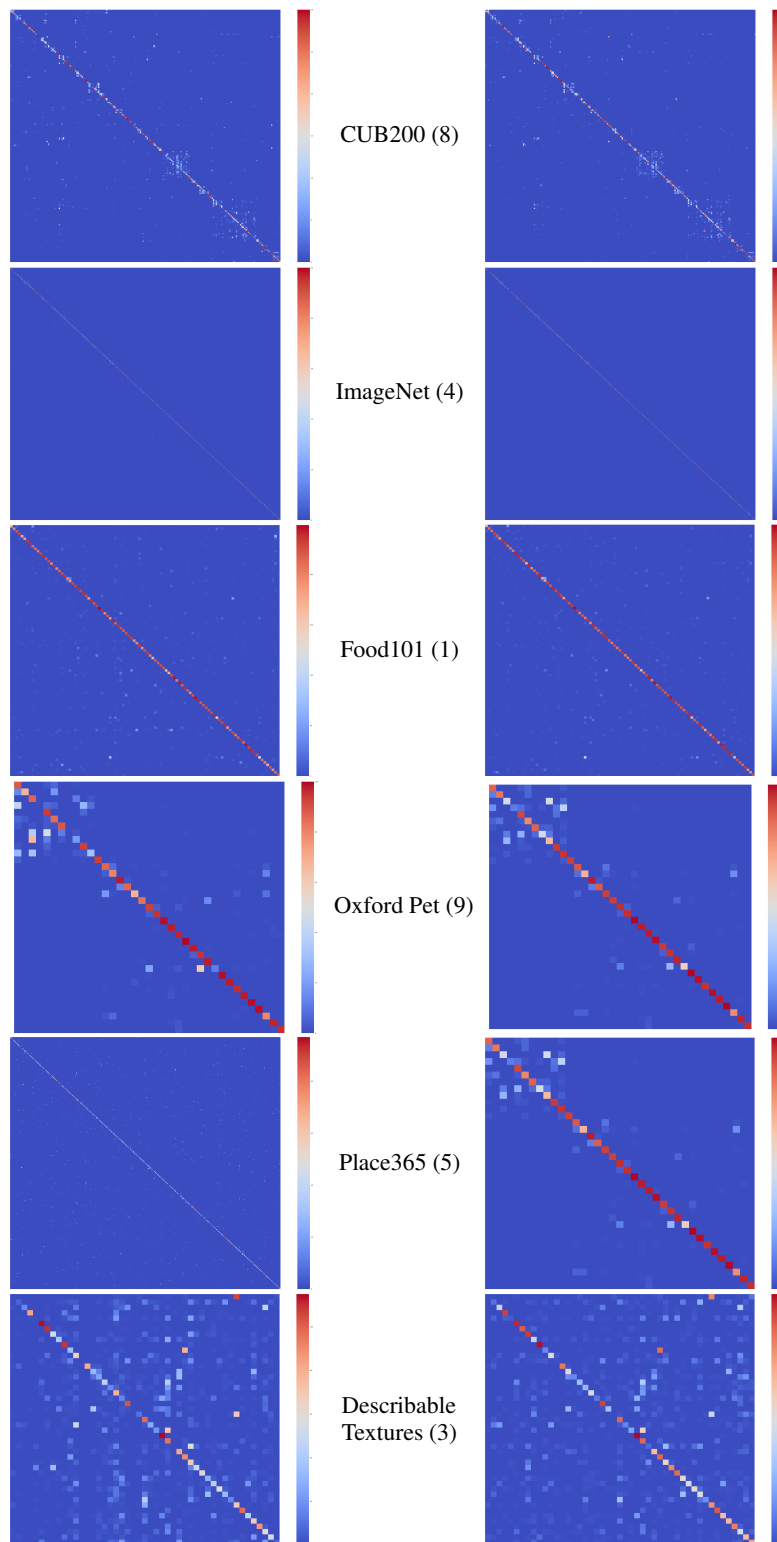


Figure 6: Confusion matrices comparison: the left ones are from the CLIP model and the right ones are from our method.



Figure 7: Visual comparison between a herring gull and a ring-billed gull.

2 Failure Cases

While our approach excels at generating comparative descriptions, it is very hard for CLIP (7) to fully comprehend certain descriptions. As an example, we employ the ViT-B/16 backbone to tackle a binary classification task involving Herring Gull and Ring-billed Gull and compare the basic prompt with some incomprehensible prompts, whose experimental results are shown in Tab. 1. Overall, we find three typical types of descriptions that could hurt the resulting performance:

Table 1: Binary classification accuracy across different prompts.

Base: “Herring Gull” and “Ring-billed Gull”;

P1: “Herring Gull with **a robust and heavier build**” and “Ring-billed Gull with **a stocky appearance overall**”;

P2: “Herring Gull with **pale yellow eyes**” and “Ring-billed Gull with **dark or blackish-brown eyes**”;

P3: “Herring Gull with **larger and rounder head**” and “Ring-billed Gull with **the large and round head**”.

Prompt	Base	P1	P2	P3
Accuracy	63.33%	18.33%	58.33%	53.33%

2.1 Size Ambiguity

Describing the size of objects in an image becomes inherently challenging when using text alone, primarily due to the loss of depth information after projection. As an example, let’s consider two prompts generated by ChatGPT (2): “Herring Gull with **a robust and heavier build**” and “Ring-billed Gull with **a stocky appearance overall**”. But it does not match the display in Fig. 7.

2.2 Similar Color

Although ChatGPT can provide information about color differences, accurately representing these differences in pixel space poses significant challenges. Factors such as category variation, lighting conditions, and other environmental factors make it difficult to precisely convey color distinctions. Distinguishing between a “herring gull with pale yellow eyes” and a “ring-billed gull with dark or blackish-brown eyes” solely based on textual descriptions from ChatGPT is quite challenging when observing the two birds in Fig. 7.

2.3 Comparative Adjective

Encoding absolute content with CLIP is quite straightforward, but encoding relative content is challenging. This is because CLIP’s references are biased towards the training dataset, rather than contrasting images.

References

- [1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *ECCV*, 2014.
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.
- [3] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [5] Alejandro López-Cifuentes, Marcos Escudero-Vinolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognition*, 2020.
- [6] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023.
- [7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [8] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [9] Hui Zhang, Shenglong Zhou, Geoffrey Ye Li, and Naihua Xiu. 0/1 deep neural networks via block coordinate descent. *arXiv preprint arXiv:2206.09379*, 2022.