

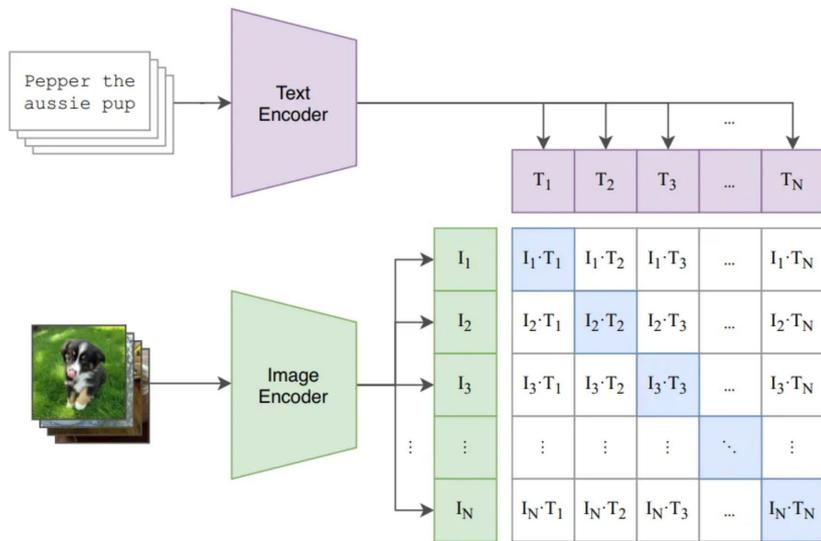
ChatGPT-Powered Hierarchical Comparisons for Image Classification

(Neurips 2023)

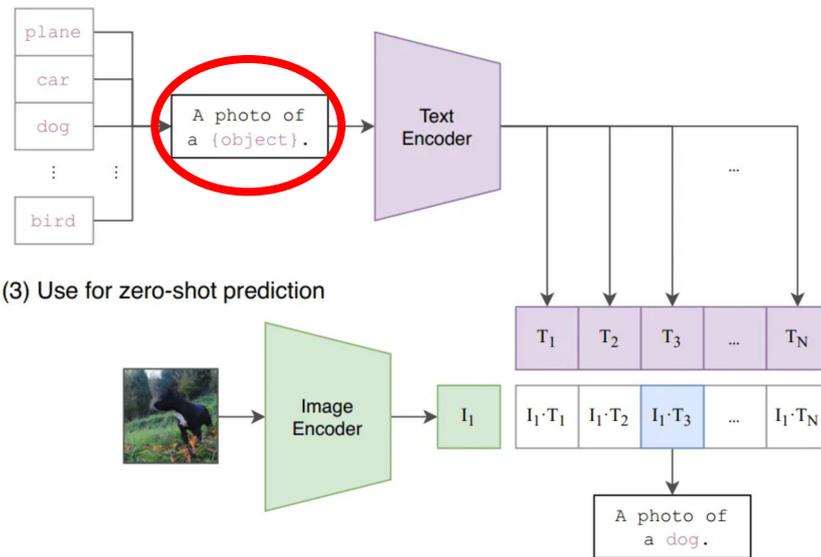
Zhiyuan Ren, Yiyang Su and Xiaoming Liu

CLIP[1] Model

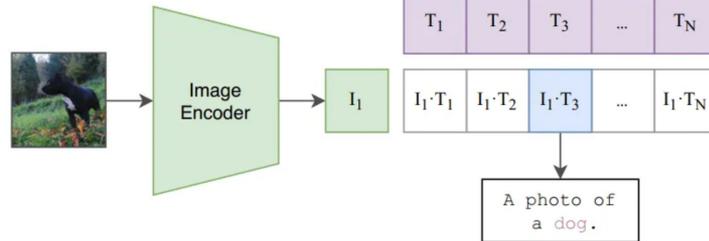
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



Importance of Prompt

Another issue we encountered is that it's relatively rare in our pre-training dataset for the text paired with the image to be just a single word. Usually the text is a full sentence describing the image in some way. To help bridge this distribution gap, we found that using the prompt template "A photo of a {label}." to be a good default that helps specify the text is about the content of the image. This often improves performance over the baseline of using only the label text. For instance, just using this prompt improves accuracy on ImageNet by 1.3%.

Learnable Prompt

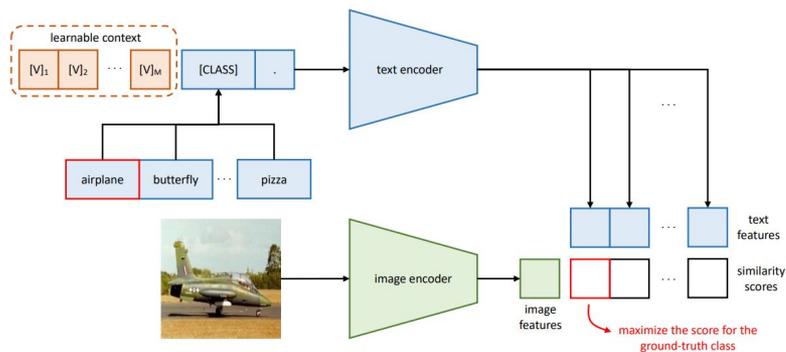


Fig 1: CoOp[2]

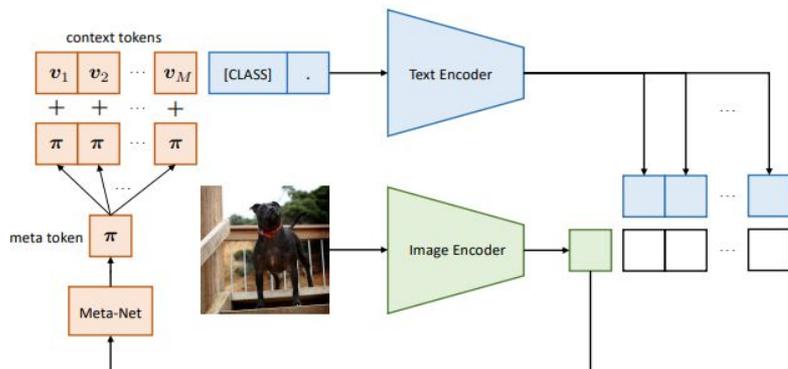


Fig 2: CoCoOp[3]

(a) Average over 11 datasets.

	Base	New	H
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83

Take-away:

1. Prompt is very important in terms of the performance of vision-language model;
2. Learnable prompt is good but it requires to train again for each dataset.

[2] Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *IJCV*. 2021

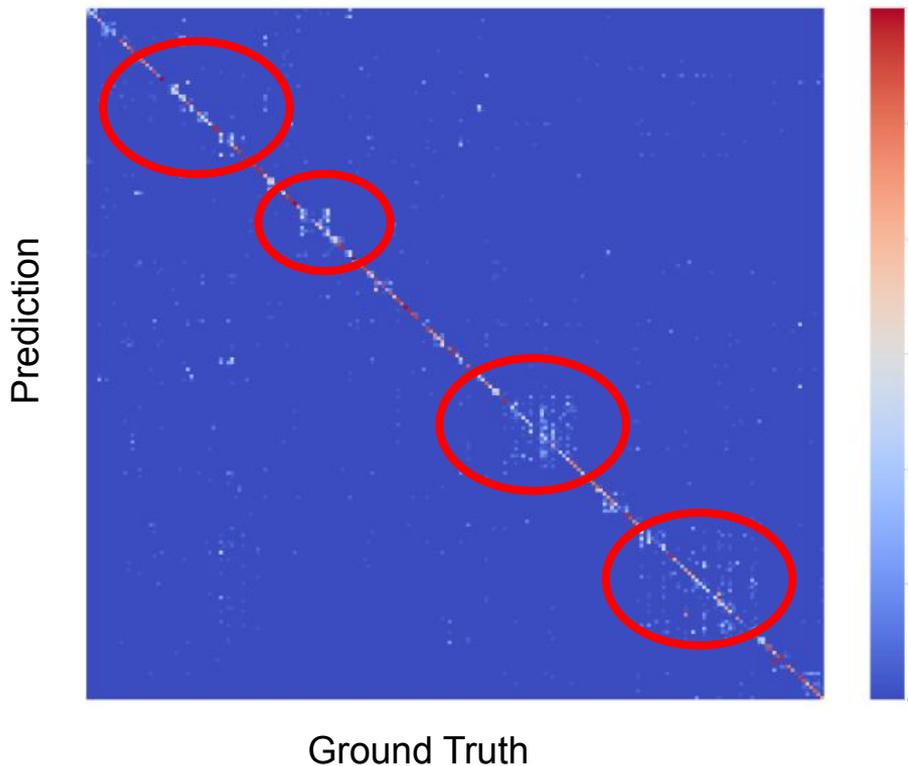
[3] Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." *CVPR*. 2022.

Motivation

Could we design a **generalized**
prompt generating method **without**
training?



Analysis from Confusion Matrix



Song Sparrow

Tree Sparrow

Could we highlight the feature which distinguishes two classes?

And we hope these feature is more easy to understand for CLIP model than the class name.

ChatGPT-Based Prompt Generation

what is the visual difference between tree sparrow and song sparrow

Both the Tree Sparrow (often referred to as the American Tree Sparrow to differentiate it from the Eurasian Tree Sparrow) and the Song Sparrow are small, North American birds, but they have distinct visual differences:

1. Head:

- **Tree Sparrow:** It has a rusty-red cap and a gray eyebrow (supercilium). The lores (the region between the eye and bill) are typically gray without any dark marking.
- **Song Sparrow:** The head markings can vary across different populations, but typically they don't have a distinct cap like the Tree Sparrow. They usually display streaking on the crown and a long, broad, gray eyebrow. They also have dark lines (malar stripes) that extend from the bill and through the eye.

2. Breast:

- **Tree Sparrow:** The breast is mostly clear and unmarked, but it has a small, dark spot in the center (central breast spot).
- **Song Sparrow:** The breast and flanks have thick streaks and spots. Some populations might have a central spot, but it is usually surrounded by other streaks and markings, unlike the clean breast of the Tree Sparrow.

3. Back and Wings:

- **Tree Sparrow:** The wings display two clear wing bars. The back is streaked with brown and gray.
- **Song Sparrow:** The wings often lack clear wing bars and the back is heavily streaked.

4. Size and Shape:

- **Tree Sparrow:** Generally has a rounded head and a moderately forked tail.
- **Song Sparrow:** Possesses a more rounded and fuller tail and often appears slightly larger than the Tree Sparrow.

5. Bill:

- **Tree Sparrow:** The bill is bicolored with a dark upper mandible and a yellow or lighter lower mandible.
- **Song Sparrow:** Typically has a darker, conical bill, uniformly colored.

Input the two class names that we want to compare, and append all the prompt after the class names.

E.g. Tree sparrow with **gray eyebrow, clear breast, clear wing bars and a dark upper mandible.**

Complexity Brought by More Comparison

For ImageNet, we have
1000 class names.

Comparing them to each
other at the same time will
require generating a
1000x999 prompt matrix!!!

what is the visual difference between tiger, dog, cat, lion and rabbit

Sure, let's break down the visual differences between these animals:

Tiger:

- **Size:** Large, typically around 9-12 feet in length including their tail for adults.
- **Appearance:** Orange coat with black stripes. White belly and distinctive white "mask" patterns around the eyes. Eyes are usually amber.
- **Limbs:** Four muscular legs with retractable claws.
- **Tail:** Long with occasional black bands.
- **Teeth:** Carnivorous with large, prominent canine teeth.

Dog:

- **Size:** Varies greatly depending on breed, from small breeds like Chihuahuas to large breeds like Great Danes.
- **Appearance:** Diverse range of coat colors, lengths, and textures. The shape of the face, snout, and ears also varies among breeds.
- **Limbs:** Four legs with non-retractable claws.
- **Tail:** Can be long, short, curly, or straight, depending on the breed.
- **Teeth:** Omnivorous teeth suitable for a varied diet.

Cat (Domestic Cat):

- **Size:** Much smaller than a tiger, generally weighing between 5 to 20 pounds.
- **Appearance:** Various coat colors and patterns including solid, tabby, calico, and more. Generally, cats have pointed ears and whiskers.
- **Limbs:** Four legs with retractable claws.
- **Tail:** Typically long and slender.
- **Teeth:** Carnivorous with sharp, pointed teeth.

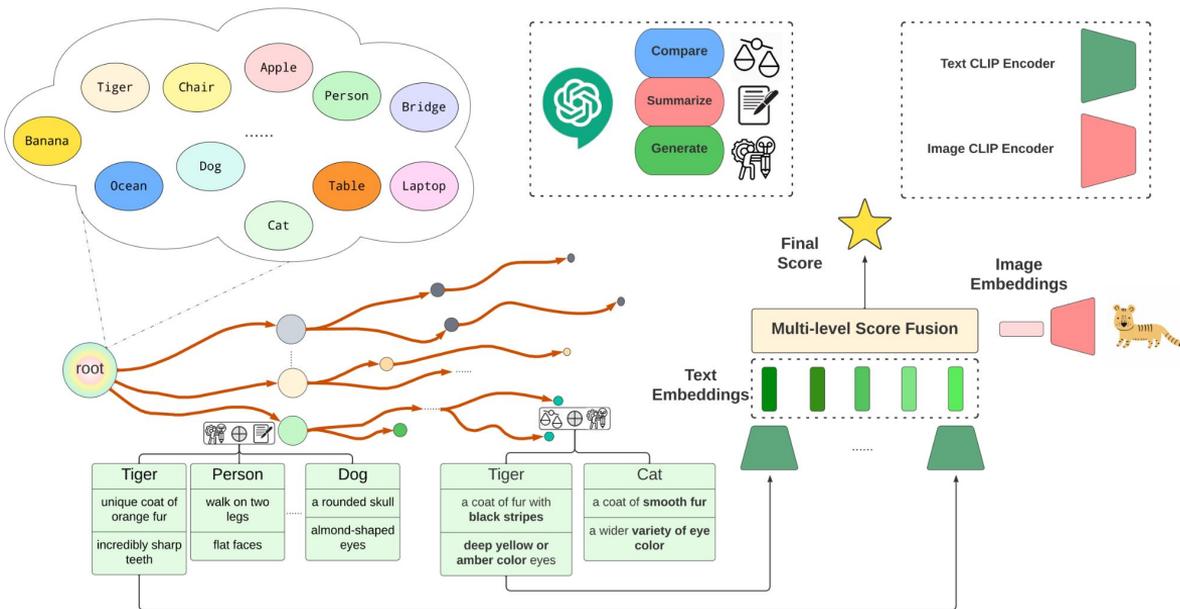
Lion:

- **Size:** Large, but usually a bit smaller than the tiger. Males are typically larger than females.
- **Appearance:** Uniform light brown or golden coat. Males have a distinctive mane around their head and neck, which can range in color from blonde to black.
- **Limbs:** Four muscular legs with retractable claws.
- **Tail:** Has a tuft or a tassel at the end, which is unique among big cats.
- **Teeth:** Carnivorous with large, prominent canine teeth.

Rabbit:

- **Size:** Much smaller compared to the aforementioned animals. Size can vary among species but is generally between 1 to 20 pounds.
- **Appearance:** Soft fur that comes in a variety of colors. Round bodies, large eyes (usually dark), and long ears (though some breeds have short ears). Whiskers are also present.
- **Limbs:** Four legs. Hind legs are much stronger and larger, adapted for hopping.
- **Tail:** Short and fluffy, often referred to as a "cotton tail".
- **Teeth:** Herbivorous. They have large incisors that continually grow throughout their life.

Complete Pipeline



```

# T_init[n, d]           - minibatch of initial text embeddings
# Text_encoder          - CLIP Text Encoder
# LLM(prompt)->description - ChatGPT model

# Target: collect tree-level description

def build_tree_in_loop(T_emb):
    new_groups = K-means(T_emb)

    for group in new_groups:
        if len(group) == 1:
            # Skip processing single-item group
            pass

        elif len(group) > thres:
            # Generate more fine-grained descriptions
            summary = LLM("summarize the following categories(...)")
            descriptions = LLM(group, summary)
            collect(descriptions)

            # Keep Grouping into smaller ones
            T_new = Text_encoder(description)
            build_tree_in_loop(T_new)

        elif 1 < len(group) <= thres:
            # Generate comparative descriptions
            descriptions = LLM("compare the following categories(...)")
            collect(descriptions)

# Starting point
build_tree_in_loop(T_emb=T_init)

```

Quantitative Results

Table 1: Zero-shot classification accuracy gains over CLIP and OpenCLIP category name embedding baseline. We see significant increases across all the settings except Food101 with the OpenCLIP.

Dataset	ImageNet			CUB			Food101		
	Architecture	CLIP	Ours	Δ	CLIP	Ours	Δ	CLIP	Ours
Res-50	54.85	60.63	+5.78	49.02	49.86	+0.74	71.61	75.44	+3.83
ViT-B/32	58.47	63.88	+5.41	52.23	54.18	+1.95	78.83	83.02	+4.19
ViT-B/16	63.53	68.95	+5.42	56.66	59.25	+2.59	86.17	88.13	+1.96
ViT-L/14	76.60	79.64	+3.04	63.46	65.45	+1.99	91.86	93.17	+1.31

Dataset	Places365			Oxford Pets			Describable Textures		
	Architecture	CLIP	Ours	Δ	CLIP	Ours	Δ	CLIP	Ours
Res-50	32.48	36.83	+4.35	74.60	79.99	+5.39	37.63	45.67	+8.04
ViT-B/32	37.23	40.73	+3.50	79.80	82.69	+2.89	40.82	49.19	+7.37
ViT-B/16	38.21	41.52	+3.31	81.68	87.19	+5.51	43.56	51.26	+7.70
ViT-L/14	38.65	41.13	+2.48	87.92	92.84	+4.92	51.01	58.04	+7.03

Dataset	ImageNet			CUB			Food101		
	Architecture	OpenCLIP	Ours	Δ	OpenCLIP	Ours	Δ	OpenCLIP	Ours
ViT-B/16	62.96	67.08	+4.12	66.81	68.20	+1.39	84.85	84.59	-0.26
ViT-L/14	68.34	72.69	+4.35	75.32	75.80	+0.48	89.40	89.35	-0.05
ViT-G/14	70.94	75.90	+4.96	84.19	85.19	+1.00	93.41	92.68	-0.73

Dataset	Places365			Oxford Pets			Describable Textures		
	Architecture	OpenCLIP	Ours	Δ	OpenCLIP	Ours	Δ	OpenCLIP	Ours
ViT-B/16	38.80	42.52	+3.72	84.03	87.47	+3.44	46.72	54.01	+7.29
ViT-L/14	38.46	44.12	+5.66	87.13	89.62	+2.49	51.69	61.79	+10.10
ViT-G/14	41.20	45.61	+4.41	92.37	95.07	+2.70	63.70	69.69	+5.99

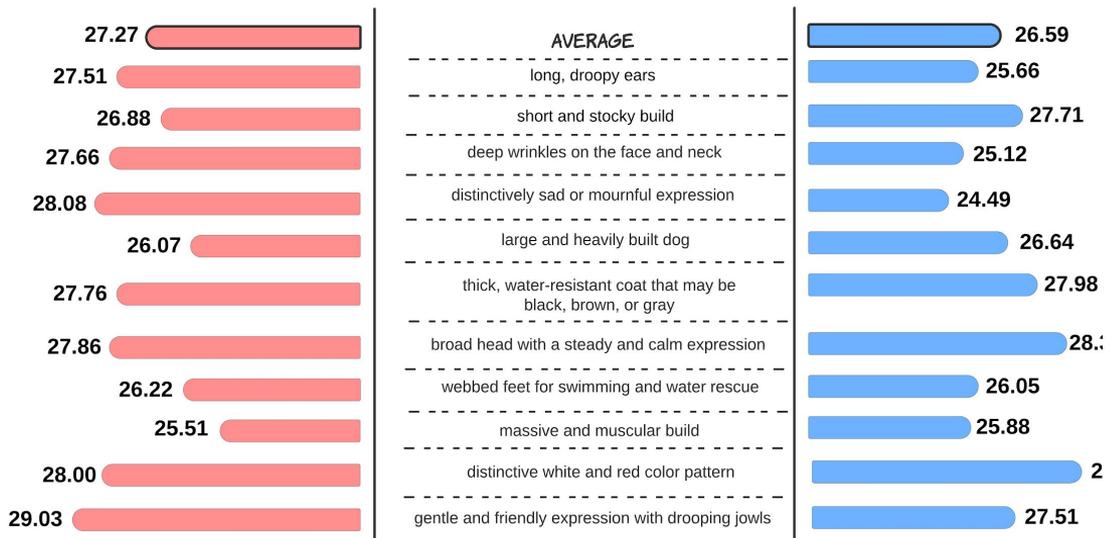
saint bernard



newfoundland



Go to 'saint bernard' description



Qualitative Results

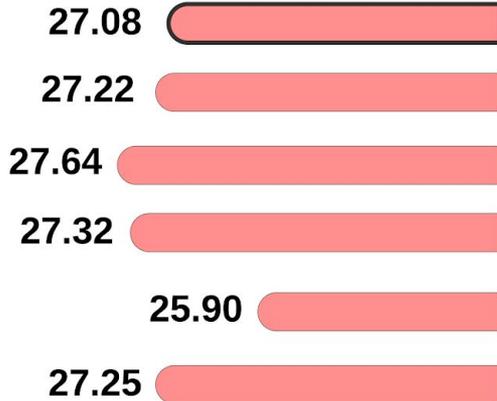
saint bernard



newfoundland



Go to 'saint bernard' description



AVERAGE

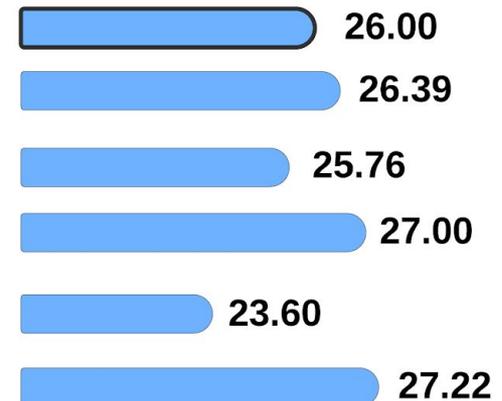
a thick and fluffy coat, typically with a white and brown or red pattern

droopy eyes with a calm expression

a broad head with a deep and powerful jaw

a folded or floppy ear shape

a friendly and gentle disposition, often seen with a tongue sticking out



Conclusion

What we want to inspire the field is that LLM is an expert in different fields therefore it can extend simple semantics to rich and diverse semantics.

(Distill the knowledge from LLM)