Paper Presentation 2

Sheikh Shafayat

Today’s paper

Published as a conference paper at ICLR 2024

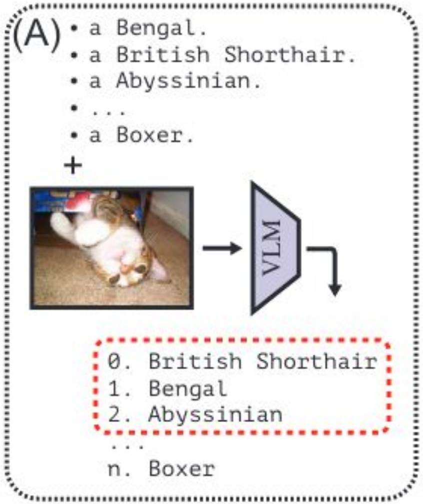
FOLLOW-UP DIFFERENTIAL DESCRIPTIONS: LANGUAGE MODELS RESOLVE AMBIGUITIES FOR IMAGE CLASSIFICATION

Reza Esfandiarpoor & Stephen H. Bach
Department of Computer Science
Brown University
Providence, RI 02906, USA
{reza_esfandiarpoor,stephen_bach}@brown.edu

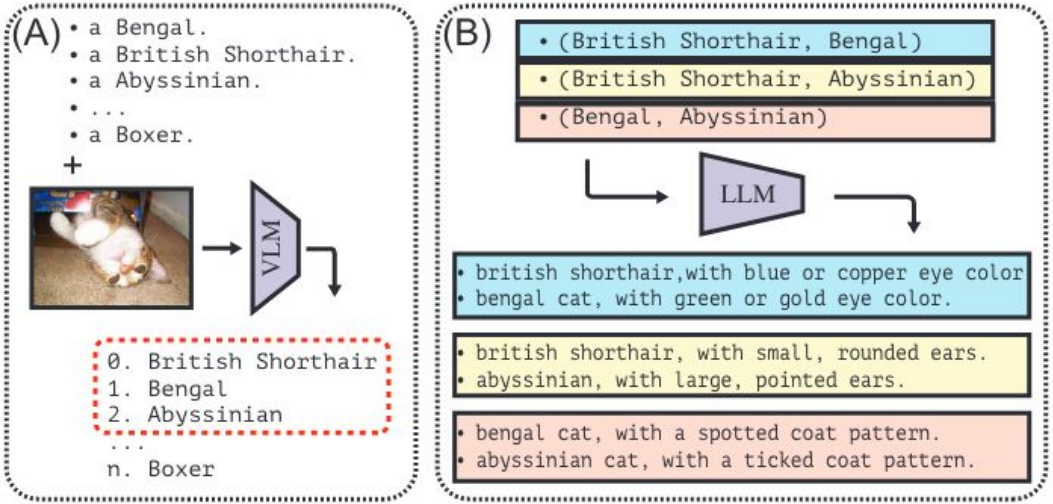
This paper is *very* simple

- It is about classification
- I plan to apply similar idea for my clustering project
- I will first give you a 2 minutes summary

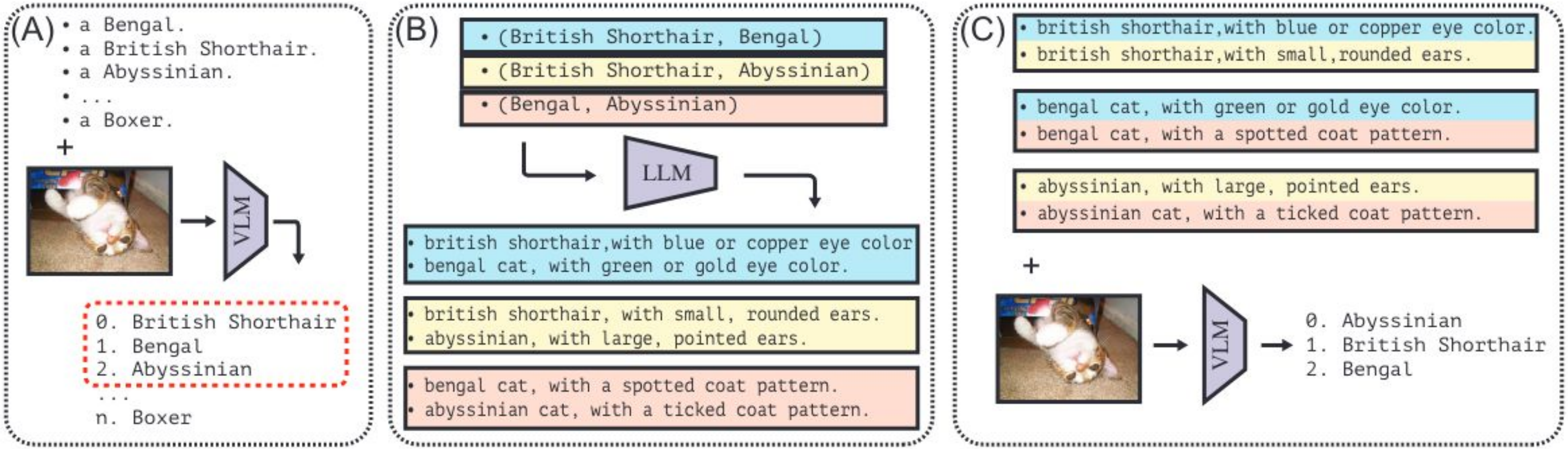
Two minutes summary



Two minutes summary



Two minutes summary



Example of generated attributes

Black-footed Albatross

Attribute: size
0: A photo of a tennessee warbler, a small songbird that is only about 4 inches long.
1: A photo of a black-footed albatross, a large seabird with a wingspan of up to 7 feet.

Attribute: coloration
0: A photo of a tennessee warbler, a bright yellow bird with olive-green wings and back.
1: A photo of a black-footed albatross, a dark-colored bird with a white head and underparts.

Attribute: bill shape
0: A photo of a tennessee warbler, a bird with a small, pointed bill.
1: A photo of a black-footed albatross, a bird with a large, hooked bill.

In summary

- We make initial predictions using CLIP
    - We take the ambiguous classes
- We ask an LLM to write descriptions about those confusing classes

```
For the following objects, generate captions that represent the
    distinguishing visual differences between the photos of the two
    objects. Generate as many captions as you can.
Object 1: {class name 1}
Object 2: {class name 2}
```

- Then we prompt again with those description

More details

- We actually do the comparison for k classes
  - The papers also experiment with all classes

Results

Table 1: Accuracy of FuDD in comparison with baselines. B/32 and L/14* represent the ViT-B/32 and ViT-L/14@336px vision backbones. ΔNaive(k) is the improvement of FuDD with k ambiguous classes over the Naive LLM-generated descriptions proposed by Menon & Vondrick (2023).

Description	Cub		DTD		EuroSAT		FGVCAircraft		Flowers102		Food101	
	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*
Single Template	51.21	63.48	43.14	54.04	40.87	56.82	20.88	37.08	63.80	75.12	82.63	93.49
Template Set	51.52	64.07	42.71	55.32	46.76	54.27	21.15	38.31	63.44	74.14	83.16	93.77
Naive LLM	52.92	65.15	45.90	55.37	44.18	46.69	21.09	38.79	66.12	75.98	84.02	94.26
FuDD (k=10)	53.97	65.90	45.43	57.66	45.18	60.64	21.87	38.82	67.80	78.76	84.05	94.05
FuDD (k=$\|C\|$)	54.30	66.03	44.84	57.23	45.18	60.64	22.32	39.63	67.62	79.67	84.36	94.27
Δ Naive (k=10)	↑1.05	↑0.75	↓-0.47	↑2.29	↑1.00	↑13.95	↑0.78	↑0.03	↑1.68	↑2.78	↑0.03	↓-0.21
Δ Naive (k=$\|C\|$)	↑1.38	↑0.88	↓-1.06	↑1.86	↑1.00	↑13.95	↑1.23	↑0.84	↑1.50	↑3.69	↑0.34	↑0.01

	ImageNet		ImageNet V2		Oxford Pets		Places365		Stanford Cars		Stanford Dogs	
	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*
Single Template	62.04	74.85	54.77	68.79	84.98	92.86	39.10	40.70	60.37	78.06	58.01	73.61
Template Set	63.37	76.54	55.91	70.85	84.55	92.70	40.91	42.54	60.38	79.12	57.79	74.01
Naive LLM	63.52	76.37	55.96	70.47	83.76	93.08	40.58	41.43	59.63	77.90	57.86	74.02
FuDD (k=10)	64.05	76.70	56.62	70.60	86.92	93.40	42.12	43.95	60.86	78.25	60.03	75.99
FuDD (k=$\|C\|$)	64.19	77.00	56.75	71.05	89.34	93.51	42.17	44.09	61.46	78.96	60.28	76.34
Δ Naive (k=10)	↑0.53	↑0.33	↑0.66	↑0.13	↑3.16	↑0.32	↑1.54	↑2.52	↑1.23	↑0.35	↑2.17	↑1.97
Δ Naive (k=$\|C\|$)	↑0.67	↑0.63	↑0.79	↑0.58	↑5.58	↑0.43	↑1.59	↑2.66	↑1.83	↑1.06	↑2.42	↑2.32

Ablation

Table 2: Accuracy of differential and non-differential descriptions for ambiguous classes. B/32 and L/14* represent the ViT-B/32 and ViT-L/14@336px vision backbones. Δ is the improvement of differential over non-differential descriptions.

Descriptor	CUB		DTD		FGVCAircraft		Flowers102		Food101	
	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*
Differential	53.62	65.79	45.37	56.91	22.17	39.06	67.62	79.54	84.17	94.34
Non-Differential	52.28	64.38	42.82	56.44	22.14	36.90	65.73	77.74	83.92	94.02
Δ	↑1.35	↑1.42	↑2.55	↑0.47	↑0.03	↑2.16	↑1.89	↑1.81	↑0.25	↑0.32

Descriptor	Oxford Pets		Places365		Stanford Cars		Stanford Dogs	
	B/32	L/14*	B/32	L/14*	B/32	L/14*	B/32	L/14*
Differential	87.24	93.68	42.45	44.26	60.90	79.39	60.31	75.96
Non-Differential	86.24	93.62	41.73	43.98	60.74	78.55	59.30	75.41
Δ	↑1.01	↑0.06	↑0.73	↑0.28	↑0.16	↑0.85	↑1.01	↑0.55

Ablation: Effect of K

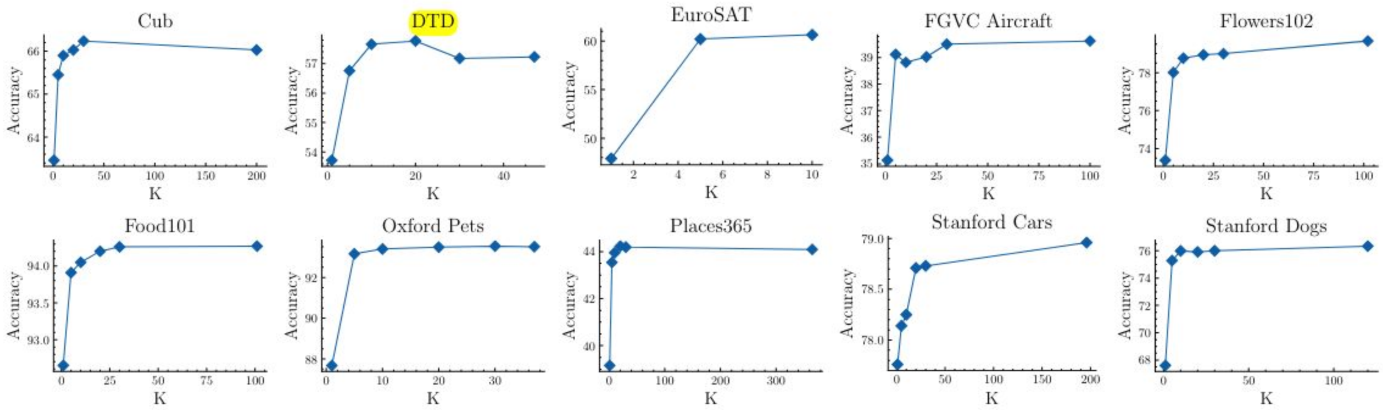


Figure 3: Impact of differential descriptions for k most ambiguous classes with ViT-L/14@336px. k=1 is accuracy with a single template. Providing differentiating details for the most ambiguous classes accounts for most of FuDD's gains, with diminishing gains for less ambiguous classes.

LLM Knowledge Matters

- Open models like LLama-2 doesn’t know much about satellite imageries
  - So their feedback is not very helpful for EuroSAT dataset
- But GPT3.5 knows quite a lot
- Fine-tuning on GPT3.5 output helps

Pros and Cons

Pros of the paper ✅

- Very simple method
- Consistently outperforms other similar methods
- Works across models (CLIP, BLIP2)

Cons of the paper ❌

- Computationally expensive
  - Not very practical for real-time applications
- The accuracy gain is small 2~3%
  - Is it worth it?

Thank You

Quiz

Quiz

What happens if you increase k (the number of ambiguous classes to compare)

too much?

a. Accuracy increase is marginal
b. Accuracy increase is drastic
c. Accuracy decreases significantly
d. Accuracy drops slightly