How well does CLIP understand texture?

Chenyun Wu Subhransu Maji

University of Massachusetts Amherst

Motivation

We focus on **zero-shot learning** capabilities of CLIP on **textures**, motivated by the following:

- Textures are useful describe object categories, especially in fine-grained domains.
- The vocabulary for describing textures are rich and unique (including color, pattern, periodicity, stochasticity, etc.)

Texture classification

We apply **zero-shot classification** with CLIP on 4 texture datasets: DTD, FMD, KTH, and KTH2a, and report per-image accuracy, using prompt “a photo of a [c] pattern”.

Model	DTD	FMD	KTH	KTH2a	Average
RN50	40.7	83.4	49.1	62.8	59.0
RN101	42.0	79.0	48.5	51.3	55.2
ViT-B/32	41.1	83.8	58.4	59.5	60.7
ViT-B/16	44.7	87.9	57.4	61.1	62.8
ViT-L/14	50.4	89.5	63.5	64.5	67.0
ViT-L/14@336	50.7	90.5	63.9	66.0	67.8

We also experimented with different prompts:

	ViT-B/32				ViT-L/14			
Prompt	DTD	FMD	KTH	KTH2a	DTD	FMD	KTH	KTH2a
[c]	41.1	80.0	48.6	46.7	50.4	88.7	58.3	68.0
a photo of a [c]	43.1	79.9	50.4	49.9	52.3	89.0	61.0	69.4
a photo of a [c] background	43.1	79.9	50.4	49.9	50.4	89.3	59.3	69.8
a photo of a [c] object	42.3	83.2	56.3	59.7	53.0	92.3	59.6	70.0
a photo of a [c] pattern	41.1	83.8	58.4	59.5	50.4	89.5	63.5	64.5
std. dev.	±1.0	±2.0	±4.3	±6.0	±1.3	±1.5	±2.0	±2.3

- CLIP variants with **transformer backbones** are better than those using ResNet.
- Best model: ViT-L/14@336 trained on **larger images** (336×336 vs. 224×224).
- **Prompts** have a larger impact on smaller models (ViT-B/32 compared against ViT-L/14), indicated by larger variance in performance across prompts.

Texture Retrieval

DTD2 is a dataset with texture images and open vocabulary descriptions (an example shown as below).

[1] animal print, zebra, white and black stripes with blue body
[2] black stripes on blue, yellow, and green background
[3] spiral, blue and yellow with black stripes, zebralike, spherical, smooth
[4] striped, blue, yellow, lined, black

On DTD2, we compare 2 models:

- **CLIP**: ViT-B/32, zero-shot.
- **DTML**: smaller metric learning model trained on DTD2.

Task	Model	MAP	MRR	P@5	P@20	R@5	R@20
Phrase retrieval	DTML	**31.6**	**72.5**	**40.6**	**22.9**	**20.2**	**44.5**
	CLIP	12.2	40.0	17.6	11.4	8.4	21.5
Image retrieval	DTML	**13.5**	31.1	16.5	**14.5**	5.2	**17.3**
	CLIP	12.7	**32.1**	**16.9**	13.2	**6.1**	17.3

We also show the best and worst performing attributes in image retrieval:

- CLIP performs similar with DTML on image retrieval but worse on phrase retrieval.
- Both models are good at common colors and patterns.
- CLIP is better at relatively rare colors (“orange”, “pink”, “purple”) and materials (“wood”, “marble”, “glass”).
- CLIP is worse at words only frequent in DTD2 (“lined”, “bumpy”, “rough”).

Texture on Birds

CUB is a fine-grained bird dataset with localized attributes on body parts (an example shown as below).

Red-bellied Woodpecker
Species: Melanerpes carolinus
Head: red; gray; white; multi-color
Belly: white; buff; solid
Wing: white; black; spotted

On CUB, we conduct **zero-shot** phrase/image retrieval on texture attributes with both models. **CLIP is better on image retrieval** than DTML but similar on phrase retrieval.

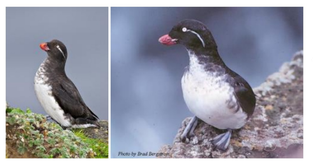
Task	Model	MAP	MRR	P@5	P@20	R@5	R@20
Phrase retrieval	DTML	52.6	68.6	**46.4**	-	**45.8**	-
	CLIP	**54.1**	**75.9**	43.2	-	43.4	-
Image retrieval	DTML	35.3	53.7	44.7	43.8	0.2	0.7
	CLIP	**50.1**	**91.7**	**72.9**	**71.8**	**0.5**	**1.6**

We also conduct **zero-shot classification with CLIP** on fine-grained bird species.

We improve the classification accuracy by **adding localized texture attributes to the prompts**, especially when using scientific species names rather than common names of birds.

category	**name**(200)		**species**(200)		**genus**(115)		**family**(39)		**order**(12)		**“bird”**(1)
#attribute	0	15	0	15	0	15	0	15	0	15	15
top-1	51.8	50.2	6.6	14.3	15.4	14.4	10.8	12.9	6.6	12.5	12.1
top-5	82.7	81.6	19.2	41.0	18.8	40.5	11.5	36.0	6.7	35.4	35.8
top-10	91.0	91.3	24.9	56.6	20.8	53.3	17.3	51.9	11.3	51.5	52.3

Below are examples of two similar species and their generated prompts with detailed texture attributes:



*“An image of a medium size black **Parakeet Auklet** with white eye, specialized red short bill, white belly, black nape, white underparts, black back, black crown, black forehead, grey leg, white breast, black upper tail.”*

*“An image of a medium size **Crested Auklet** with white eye, crested head, specialized orange bill, solid wing, black nape, black forehead, black upper tail, black throat, black solid back, black crown, black under tail, black upperparts.”*