Waffling around for Performance: Visual Classification with Random Words and Broad Concepts

ICCV 2023

汇报人：胡天乐

2023.11.7

Author

Karsten Roth[1,*], Jae Myung Kim[1,*], A. Sophia Koepke[1], Oriol Vinyals[2], Cordelia Schmid[3], Zeynep Akata[1,4]
[1]University of Tübingen, Tübingen AI Center, [2]Google DeepMind,
[3]Inria, Ecole normale supérieure, CNRS, PSL Research University, [4]MPI for Intelligent Systems
[*]equal contribution

Karsten Roth

FOLLOW

ELLIS, IMPRS-IS, University of Tübingen
Verified email at uni-tuebingen.de - Homepage

Continual Learning Generalization Foundation Models Contrastive Learning
Deep Metric Learning

TITLE	CITED BY	YEAR
Covid-19 image data collection: Prospective predictions are the future JP Cohen, P Morrison, L Dao, K Roth, TQ Doung, M Ghassemi Journal of Machine Learning for Biomedical Imaging (MELBA)	812 *	2020
The liver tumor segmentation benchmark (lits) P Bilic, P Christ, HB Li, E Vorontsov, A Ben-Cohen, G Kaissis, A Szeskin, ... Medical Image Analysis 84, 102680	772	2023
Towards total recall in industrial anomaly detection K Roth, L Pemula, J Zepeda, B Schölkopf, T Brox, P Gehler Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern …	320	2022
Predicting covid-19 pneumonia severity on chest x-ray with deep learning JP Cohen, L Dao, K Roth, P Morrison, Y Bengio, AF Abbasi, B Shen, ... Cureus 12 (7)	235	2020
Revisiting Training Strategies and Generalization Performance in Deep Metric Learning K Roth, T Milbich, S Sinha, P Gupta, B Ommer, JP Cohen ICML 2020	154	2020
MIC: Mining Interclass Characteristics for Improved Metric Learning K Roth, B Brattoli, B Ommer Proceedings of the IEEE International Conference on Computer Vision, 8000-8009	98	2019

Motivation

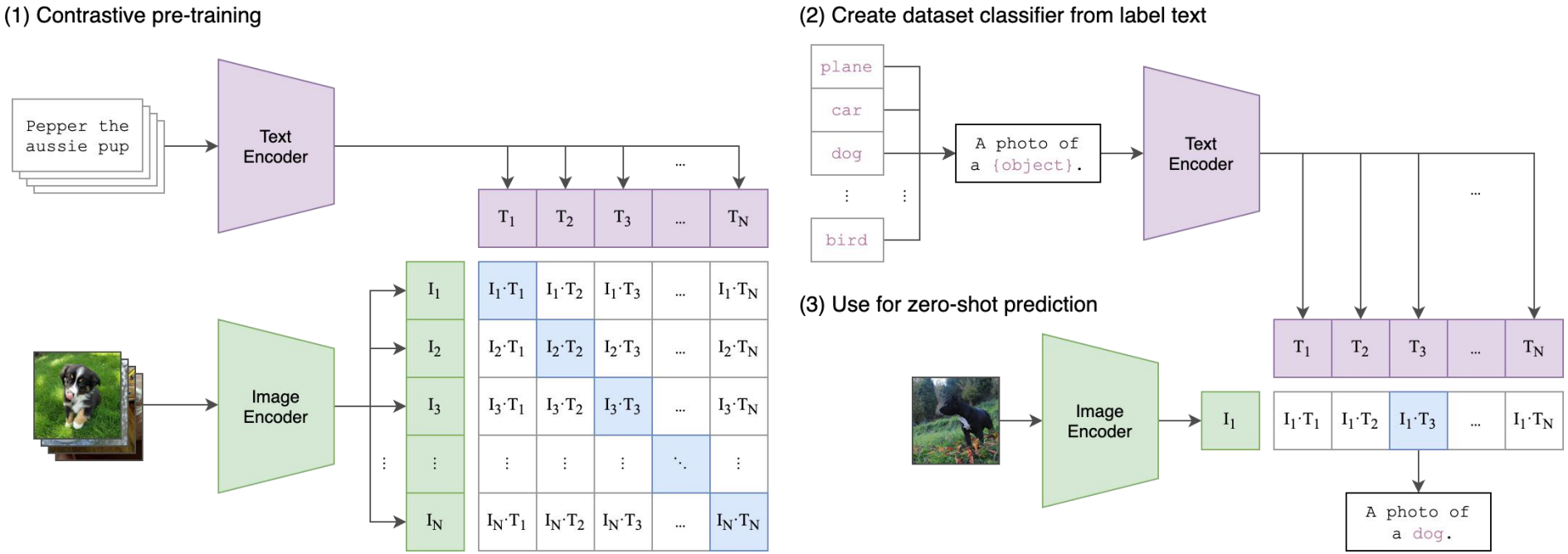
- 自然语言提示的特定任务fine-tuning可以提高VLM的性能，但需要访问额外的训练数据；一种有效的替代方法是通过访问LLM(如GPT-3)来生成类别描述符(descriptors)以提高性能。

- 然而，仔细检查 GPT-3 生成的语义描述符，发现存在高度的多样性、有限的视觉相关性和歧义，且实验表明，用随机不相关类描述符替换这些LLM生成的文本，性能所受影响不大。

- 本文提出了WaffleCLIP，即用随机词或字符替换LLM生成的描述符，以更低成本产生可比较的zero-shot性能，同时提出加入 LLM 生成的宽泛concept以更好地利用语义。

Method

- CLIP

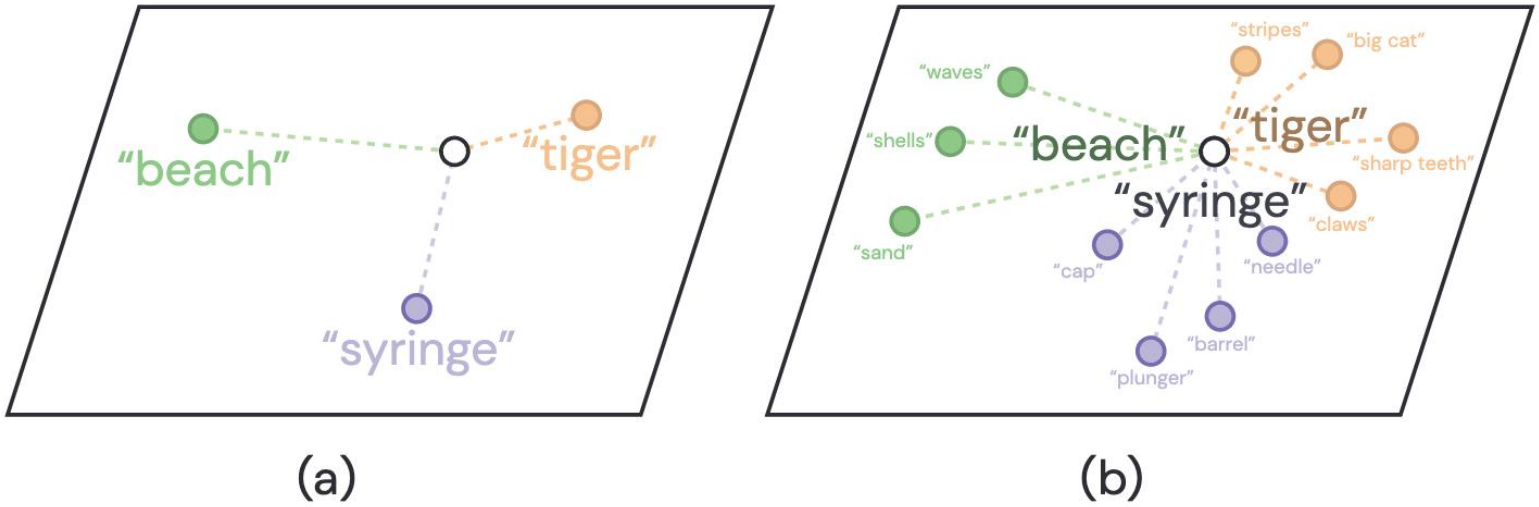

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
A photo of a dog.

Method

VISUAL CLASSIFICATION VIA DESCRIPTION FROM LARGE LANGUAGE MODELS

Sachit Menon, Carl Vondrick
Department of Computer Science
Columbia University

- ICLR2023,DCLIP



Jackfruit, which (has/is/etc)
- large, round fruit
- green or yellow skin
- white flesh with black seeds
- sweet and sticky taste
- strong smell

Method

- DCLIP

Our top prediction: Airliner
and we say that because...

Average	23.49
a livery or paint scheme	25.36
engines mounted on the wings ...	23.58
landing gear with wheels and tires	23.28
large, metal aircraft	23.19
a fuselage with a pointed nose ...	23.00
wings and tail fin	22.48

CLIP's top prediction: Albatross
but we don't say that because...

Average	19.90
slow, powerful flight	22.73
long, hooked bill	21.48
long, narrow wings	21.23
black wingtips	20.19
large, long-winged bird	19.41
white or grey plumage	17.11
webbed feet	17.11

Our top prediction: Rapeseed
and we say that because...

Average	27.76
petals arranged in a cross-shape	28.86
yellow or greenish-yellow flower	28.25
stem with small, sharp thorns	27.69
hairy leaves	27.42
small, round seedpod	26.59

CLIP's top prediction: Bee
but we don't say that because...

Average	25.85
black and yellow striped body	27.08
two pairs of wings	26.44
mouthparts for chewing	26.08
hairy body	26.03
small, flying insect	25.95
compound eyes	24.84
antennae	24.61

Our top prediction: Valley
and we say that because...

Average	26.54
flanked by mountains or hills	27.61
a river or stream running through it	27.39
a depression in the earth's surface	27.09
lush vegetation	25.83
often with a V-shaped profile	24.70

CLIP's top prediction: Alpine ibex
but we don't say that because...

Average	24.40
four-limbed mammal	24.53
long, curved horns	24.53
hooves	24.44
black, grey, or brown fur	24.41
short tail	24.08

Our top prediction: Goldfish
and we say that because...

Average	25.15
a long, flowing tail	26.53
scales that shimmer in the light	25.31
a fish with a bright orange color	24.84
small, black eyes	24.66
a small mouth	24.36

CLIP's top prediction: Ibizan hound
but we don't say that because...

Average	23.79
long, thin legs	24.23
a lean, athletic build	24.19
a short, smooth coat ...	23.97
a long, narrow head	23.89
large, pointy ears	23.75
a medium-sized dog	23.75
brown or hazel eyes	22.75

Our top prediction: Cloak
and we say that because...

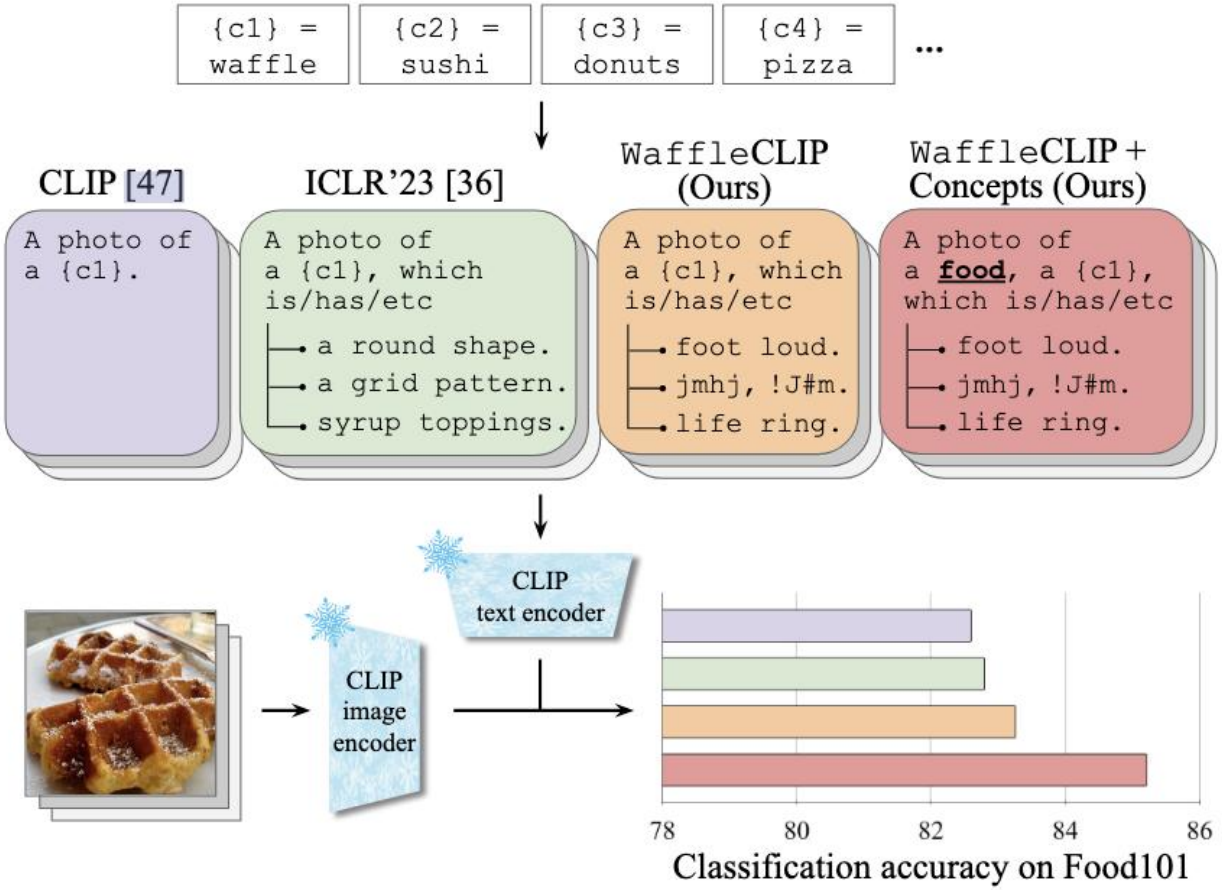
Average	25.20
has a hood	27.95
typically black or dark in color	26.53
a piece of clothing	25.44
often worn by wizards ...	24.97
fastens at the neck	23.94
often made of wool ...	22.34

CLIP's top prediction: Southern Black Widow
but we don't say that because...

Average	23.95
a small head	24.55
black with a red hourglass	24.25
long, black legs	23.95
a round, bulbous abdomen	23.05

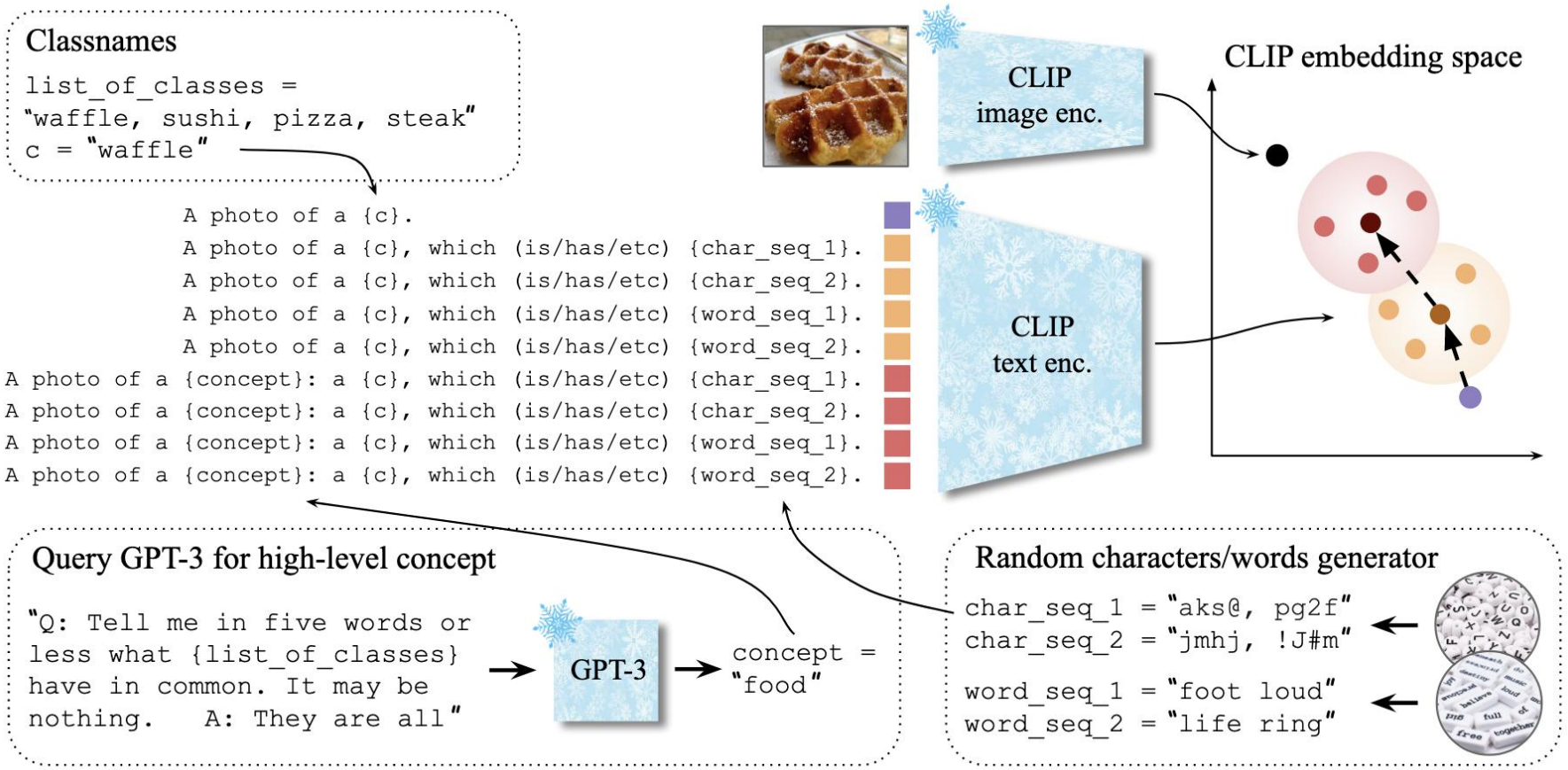
Method

- 本文思路


{c1} = waffle
{c2} = sushi
{c3} = donuts
{c4} = pizza
...
CLIP [47]
A photo of a {c1}.
ICLR'23 [36]
A photo of a {c1}, which is/has/etc
a round shape.
a grid pattern.
syrup toppings.
WaffleCLIP (Ours)
A photo of a {c1}, which is/has/etc
foot loud.
jmhj, !J#m.
life ring.
WaffleCLIP + Concepts (Ours)
A photo of a food, a {c1}, which is/has/etc
foot loud.
jmhj, !J#m.
life ring.
CLIP text encoder
CLIP image encoder
78
80
82
84
86
Classification accuracy on Food101

Method

- WaffleCLIP



Experiment

- 在8个数据集上进行图像分类性能测试

ViT-B/32	ImageNetV2	ImageNet	CUB200	EuroSAT	Places365	Food101	Oxford Pets	DTD	Avg
CLIP [47]	54.71	62.01	51.28	40.78	39.12	82.59	85.06	43.18	57.34
+ Concepts	↓	↓	52.23	48.86	39.31	84.66	86.73	↓	58.96
DCLIP [36]	55.82	63.12	52.47	43.29	40.47	82.79	86.54	**43.99**	58.56
WaffleCLIP (ours)	**55.92** ±0.08	**63.31** ±0.09	**52.38** ±0.12	44.31 ±1.07	40.56 ±0.07	83.25 ±0.21	85.70 ±0.25	43.16 ±0.25	58.57 ±0.41
+ Concepts	↓	↓	52.83 ±0.19	48.51 ±0.70	40.97 ±0.08	**85.21** ±0.06	87.52 ±0.10	↓	59.47 ±0.42
+ GPT descr. + Concepts	↓	↓	**52.77** ±0.26	**51.64** ±0.25	**41.35** ±0.09	84.87 ±0.05	**87.71** ±0.18	↓	**60.21** ±0.20

Table 2: **Image classification with WaffleCLIP** which extends input prompts with random word and character sequences and matches the performance of DCLIP [36] using GPT-generated class descriptors. Additional semantic context through high-level concepts (+ *Concepts*) can offer further boosts, particularly on benchmarks where classnames can be generic or ambiguous. We further find that WaffleCLIP complements the use of GPT-generated descriptors (+ *GPT descr.*). (↓) denotes same results as previous lines where high-level concept guidance is not applicable. For ViT-L/14 and RN50, see Supp.

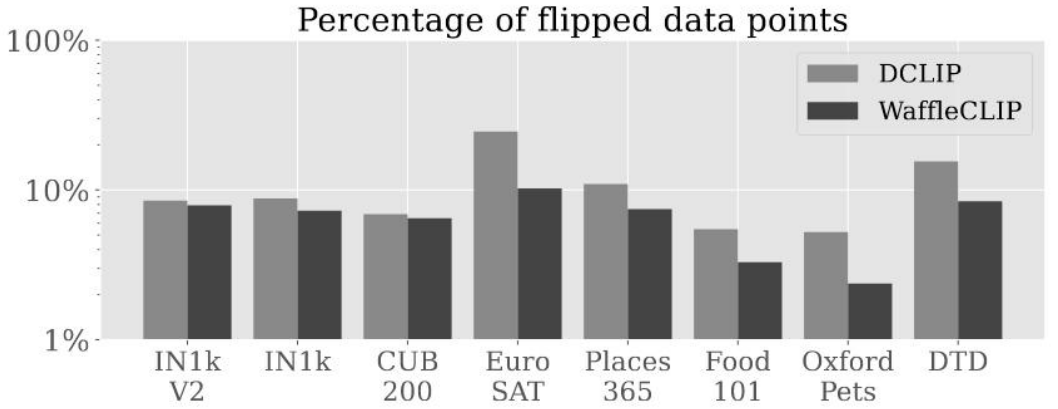
Experiment

- Dclip性能的提升与语义的关系不大

ViT-B/32	ImageNetV2	ImageNet	CUB200	EuroSAT	Places365	Food101	Oxford Pets	DTD	Avg
CLIP [47]	54.71	62.01	51.28	40.78	39.12	82.59	85.06	43.18	57.34
DCLIP [36]	**55.82**	**63.12**	52.47	**43.29**	40.47	82.79	86.54	**43.99**	**58.56**
DCLIP (same, 1x)	55.47 ±0.24	62.89 ±0.19	52.64 ±0.28	39.74 ±2.69	40.29 ±0.47	83.82 ±0.48	87.04 ±0.27	43.35 ±0.41	58.16 ±1.01
DCLIP (same, 2x)	55.75 ±0.21	63.10 ±0.19	**52.72** ±0.23	39.73 ±1.66	**40.61** ±0.22	**84.01** ±0.23	**87.10** ±0.14	43.29 ±0.22	58.29 ±0.62

Table 1: **Motivating random class descriptors.** Comparing CLIP [47] and the GPT-descriptor-extended CLIP [36] (DCLIP) with the same set of randomly sampled descriptors for each class, where the set size is either the average number of descriptors per class in DCLIP (*same, 1x*), or twice that (*same, 2x*). A random set of descriptors per class can match or even outperform DCLIP across backbone architectures (results for ViT-L/14 and ResNet50 are included in the suppl. material) confirming that randomized prompt averaging leads to higher performance.

Experiment



- max效果不如mean好
- 替换的词属于大类中，效果相当于mean，会有提升，否则下降
- dclip性能提升主要来源于prompt结构

ViT-B/32	ImageNetV2	ImageNet	CUB200	EuroSAT	Places365	Food101	Oxford Pets	DTD	Flowers102	FGVCAircraft	Stanford Cars	Avg
CLIP [47]	54.71	62.01	51.28	40.78	39.12	82.59	85.06	43.18	62.89	24.99	58.54	55.01
DCLIP [36] (*mean*)	**55.82**	**63.12**	**52.47**	**43.29**	**40.47**	**82.79**	86.54	**43.99**	**64.01**	**26.94**	**57.08**	**56.05**
DCLIP [36] (*max*)	54.41	61.67	52.40	37.11	37.21	82.37	**88.03**	43.35	63.62	25.77	56.21	54.74

ViT-B/32	ImageNetV2	ImageNet	CUB200	EuroSAT	Places365	Food101	Oxford Pets	DTD	Avg
DCLIP [36]	**55.82**	**63.12**	52.47	**43.29**	40.47	82.79	86.54	**43.99**	**58.56**
DCLIP (interchanged)	52.51 ±0.42	59.62 ±0.13	52.52 ±0.41	33.63 ±4.16	35.52 ±0.32	81.71 ±0.35	86.28 ±0.50	38.42 ±1.14	55.03 ±1.56
DCLIP (scrambled)	55.12 ±0.12	62.57 ±0.12	52.18 ±0.28	40.48 ±2.52	39.91 ±0.08	82.46 ±0.13	86.10 ±0.40	41.58 ±0.31	57.55 ±0.92
DCLIP (random, 1x)	54.11 ±0.28	61.37 ±0.18	52.42 ±0.19	36.83 ±4.27	38.80 ±0.26	82.86 ±0.23	85.99 ±0.62	42.20 ±0.85	56.82 ±1.57
DCLIP (random, 5x)	55.43 ±0.12	62.81 ±0.05	52.66 ±0.17	38.57 ±1.52	40.54 ±0.05	84.03 ±0.11	86.75 ±0.21	43.41 ±0.74	58.02 ±0.61

Experiment

- concept具有一定的作用

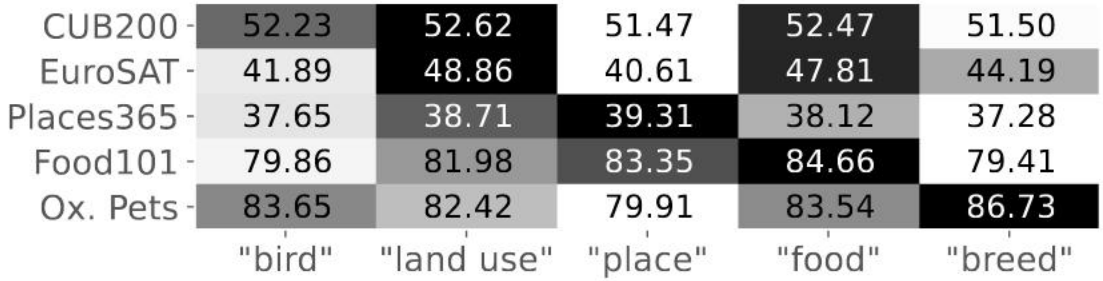


Figure 4: Study of the semantic impact of GPT-3 generated high-level concepts through a semantic confusion matrix, where we cycle high-level concepts between each benchmark. We find that interchanging the concepts generally reduces performance, indicating that high-level concepts provide complementary semantic context.

Experiment

- 额外数据集上评估，结论不变，结构仍然发挥主导作用，concept也有作用但不大

ViT-B/32	Flowers102	FGVCAircraft	Stanford Cars	Avg
CLIP [47]	62.89	24.99	58.54	48.81
DCLIP [36]	64.01	26.94	57.08	49.34
WaffleCLIP	66.27 ±0.26	25.66 ±0.19	58.91 ±0.17	50.28 ±0.21
+ Concepts	**67.19** ±0.19	28.44 ±0.22	**59.70** ±0.12	**51.78** ±0.18
+ GPT dsc. + Conc.	66.71 ±0.39	**28.96** ±0.37	59.33 ±0.14	51.67 ±0.32

Benchmarks	ImageNet-R [24]	ImageNet-S [58]	ImageNet-A [25]
CLIP [47]	65.97	40.73	29.63
DCLIP [36]	65.12	41.09	29.19
WaffleCLIP	**67.31**	**42.00**	**31.52**

Experiment

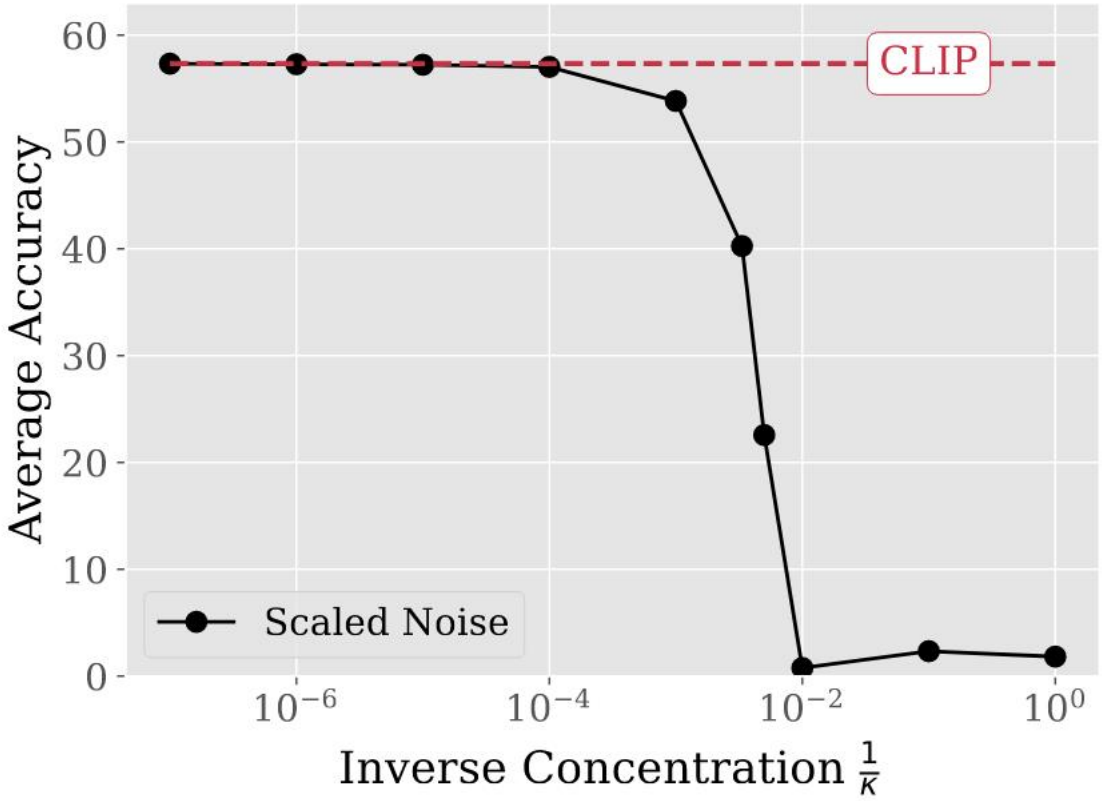
- P.Ensemple保留了基本的结构，prompt种类变多了，性能提高，加入concept会进一步提升性能，据此构思出了从dclip到waffleclip，加入concept性能也会有部分提高

ViT-B/32	IN1k-V2	IN1k	CUB	Euro	Places	Food
CLIP [47]	54.71	62.01	51.28	40.78	39.12	82.59
DCLIP [36]	55.82	63.12	52.47	43.29	40.47	82.79
P. Ensemble	55.49 ±0.21	62.79 ±0.29	51.46 ±0.43	45.76 ±0.49	40.58 ±0.06	82.67 ±0.37
+ Concepts	↓	↓	**52.08** ±0.17	**49.80** ±0.66	**40.61** ±0.14	**84.45** ±0.15
WaffleCLIP	**55.92** ±0.08	**63.31** ±0.09	52.38 ±0.12	44.31 ±1.07	40.56 ±0.07	83.25 ±0.21
+ Concepts	↓	↓	**52.83** ±0.19	**48.51** ±0.70	**40.97** ±0.08	**85.21** ±0.06

Pets	DTD	Flowers	FGVC	Cars	**Avg**
85.06	43.18	62.89	24.99	58.54	55.01
86.54	**43.99**	64.01	26.94	57.08	56.05
83.26 ±0.72	42.53 ±0.54	63.30 ±0.33	25.14 ±0.45	58.38 ±0.29	55.58 ±0.42
87.42 ±0.20	↓	**65.38** ±0.27	**26.64** ±0.50	**59.12** ±0.14	**56.94** ±0.34
85.70 ±0.25	43.16 ±0.25	66.27 ±0.26	25.66 ±0.19	58.91 ±0.17	56.31 ±0.37
87.52 ±0.10	↓	**67.19** ±0.19	**28.44** ±0.22	**59.70** ±0.12	**57.52** ±0.26

Experiment

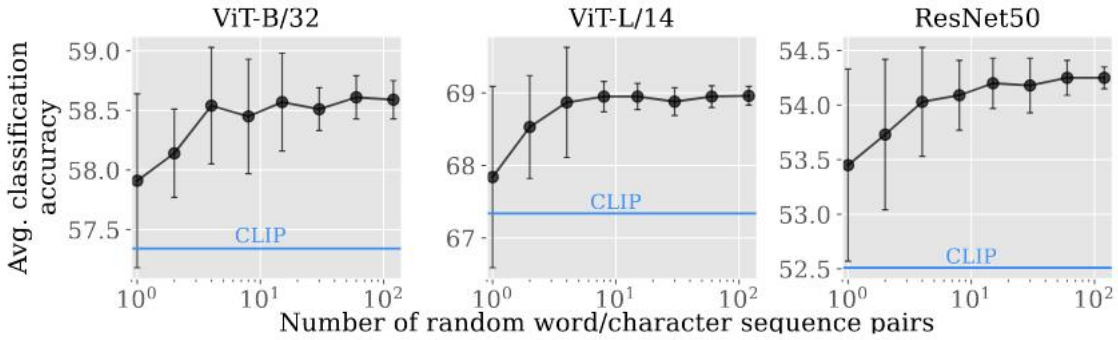
- 噪声浓度要限定在一个范围之内，太小->clip，太大->性能剧烈下降



Experiment

- 消融实验：描述符的数量和种类的影响

Avg.	ViT-B/32	ViT-L/14	RN50
Joint	58.57 ±0.41	68.95 ±0.18	54.20 ±0.23
Random Words	58.18 ±0.44	68.73 ±0.58	55.24 ±0.41
Random Characters	58.59 ±0.27	68.02 ±0.14	53.79 ±0.16



Summary

- 本文发现用随机描述符替换LLM生成的描述符可以获得类似的性能收益，从而提出WaffleCLIP，在不访问外部LLM的情况下，能获得与LLM描述符相当或更好的结果。
- VLM 很难利用细粒度语义描述符引入的实际语义，如果可以访问外部 LLM，通过粗略、高级的concept则能更好地利用语义。
- 启示：尽管本文研究方法和理论框架相对简洁，但拥有充分的实验验证和坚实的理论基础，值得借鉴学习。