Descriptor and Word Soups 🍜: Overcoming the Parameter Efficiency Accuracy Tradeoff for Out-of-Distribution Few-shot Learning

CVPR 2024

汇报人：胡天乐

2024.3.5

Author

Christopher Liao
Boston University
cliao25@bu.edu

Theodoros Tsiligkaridis
MIT Lincoln Laboratory
ttsili@ll.mit.edu

Brian Kulis
Boston University
bkulis@bu.edu

Christopher Liao

Boston University
在 bu.edu 的电子邮件经过验证
Machine Learning   Deep Learning   Domain Adaptation

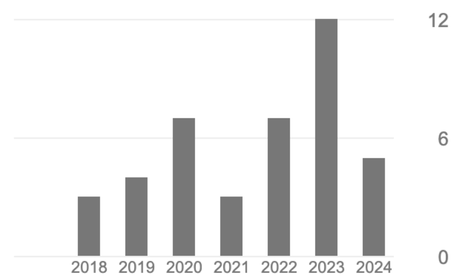
关注

创建我的个人资料

标题	引用次数	年份
Smart parking pricing: A machine learning approach E Simhon, C Liao, D Starobinski 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS …	23	2017
Faster algorithms for learning convex functions A Siahkamari, DAE Acar, C Liao, KL Geyer, V Saligrama, B Kulis International Conference on Machine Learning, 20176-20194	5 *	2022
Pick up the pace: Fast and simple domain adaptation via ensemble pseudo-labeling C Liao, T Tsiligkaridis, B Kulis arXiv preprint arXiv:2205.13508	4	2022
A case study of a shared/buy-in computing ecosystem C Liao, Y Klausner, D Starobinski, E Simhon, A Bestavros Cluster Computing 21, 1595-1606	4	2018
Supervised Metric Learning to Rank for Retrieval via Contextual Similarity Optimization C Liao, T Tsiligkaridis, B Kulis	3 *	2023
Descriptor and Word Soups: Overcoming the Parameter Efficiency Accuracy Tradeoff for Out-of-Distribution Few-shot Learning C Liao, T Tsiligkaridis, B Kulis arXiv preprint arXiv:2311.13612	1	2023

引用次数

	总计	2019 年至今
引用	41	38
h 指数	4	4
i10 指数	1	1



开放获取的出版物数量 查看全部

0 篇文章 4 篇文章

无法查看的文章 可查看的文章

根据资助方的强制性开放获取政策

Author

Christopher Liao
Boston University
cliao25@bu.edu

Theodoros Tsiligkaridis
MIT Lincoln Laboratory
ttsili@ll.mit.edu

Brian Kulis
Boston University
bkulis@bu.edu

Brian Kulis

关注

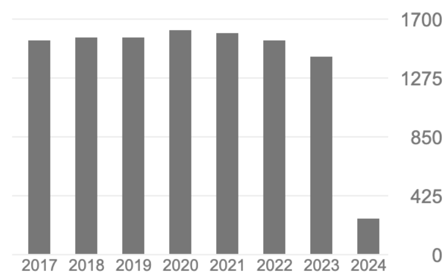
Associate Professor at Boston University & Amazon Scholar
在 bu.edu 的电子邮件经过验证 - 首页
Machine Learning Artificial Intelligence Computer Vision

标题	引用次数	年份
Adapting visual category models to new domains K Saenko, B Kulis, M Fritz, T Darrell Computer Vision–ECCV 2010: 11th European Conference on Computer Vision …	3203	2010
Information-theoretic metric learning JV Davis, B Kulis, P Jain, S Sra, IS Dhillon Proceedings of the 24th international conference on Machine learning, 209-216	2655	2007
Kernel k-means: spectral clustering and normalized cuts IS Dhillon, Y Guan, B Kulis Proceedings of the tenth ACM SIGKDD international conference on Knowledge …	1637	2004
Weighted graph cuts without eigenvectors a multilevel approach IS Dhillon, Y Guan, B Kulis IEEE transactions on pattern analysis and machine intelligence 29 (11), 1944 …	1268	2007
Metric learning: A survey B Kulis Foundations and Trends® in Machine Learning 5 (4), 287-364	1133	2013
Kernelized locality-sensitive hashing for scalable image search B Kulis, K Grauman 2009 IEEE 12th international conference on computer vision, 2130-2137	1116	2009

创建我的个人资料

引用次数 查看全部

	总计	2019 年至今
引用	19392	8022
h 指数	37	28
i10 指数	48	45



开放获取的出版物数量 查看全部

0 篇文章 8 篇文章

无法查看的文章 可查看的文章

根据资助方的强制性开放获取政策

Motivation

- 过去几年，大量多模态研究使用 GPT 生成的描述符进行零样本评估，这些研究通过 GPT 生成的标签特定文本的集合提高了预训练 VL 模型的零样本准确性。

- 最近的一项研究 WaffleCLIP 表明，可以通过一组随机描述符来实现类似的零样本精度。但是，以上两种方法针对OOD数据都存在泛化性不足的问题。

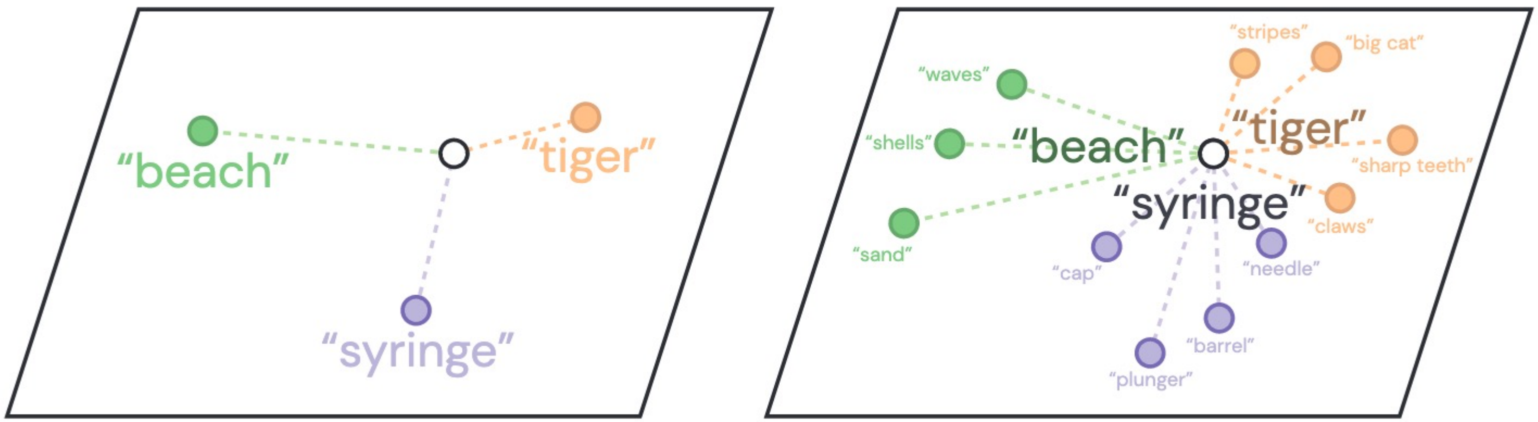
- 本文提出了descriptor and word soups，即选择质量更高的描述符/单词组成集合(soups)来作为prompt，能以更少的参数量实现更高的OOD精度。

Method

VISUAL CLASSIFICATION VIA DESCRIPTION FROM LARGE LANGUAGE MODELS

Sachit Menon, Carl Vondrick
Department of Computer Science
Columbia University

- ICLR2023,DCLIP



(a) (b)

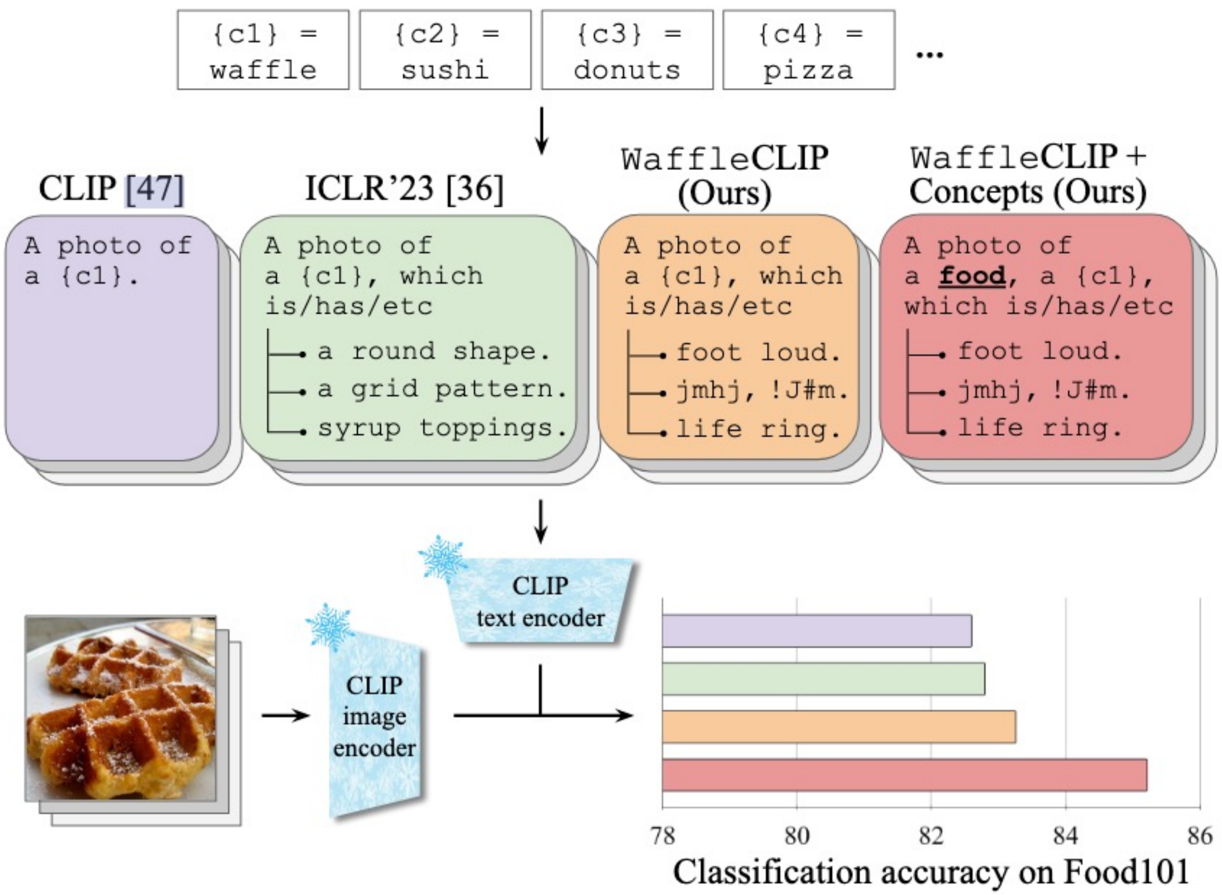
Jackfruit, which (has/is/etc)

- large, round fruit
- green or yellow skin
- white flesh with black seeds
- sweet and sticky taste
- strong smell

Method

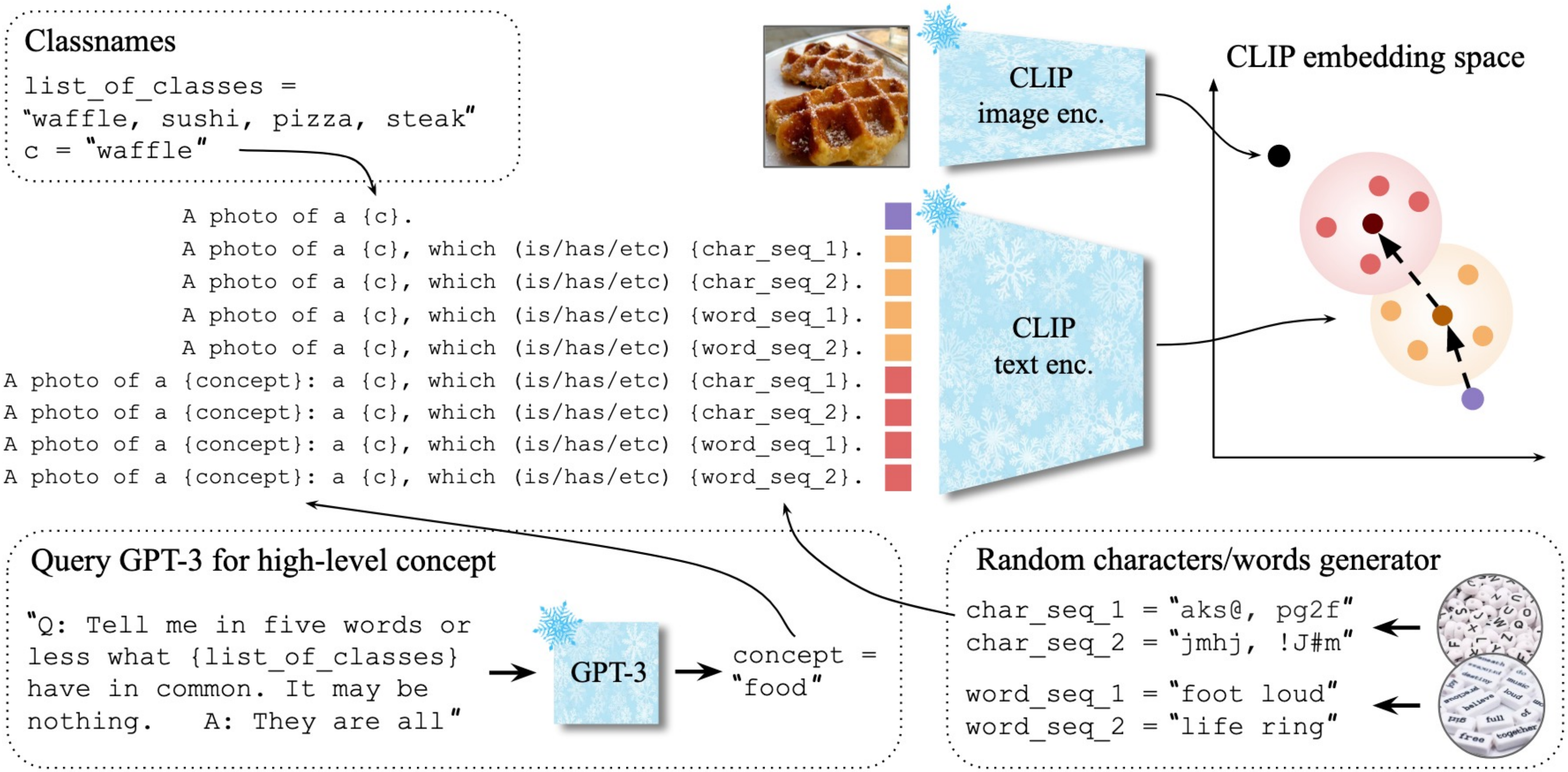
Waffling around for Performance: Visual Classification with Random Words and Broad Concepts

- ICCV2023,WaffleCLIP



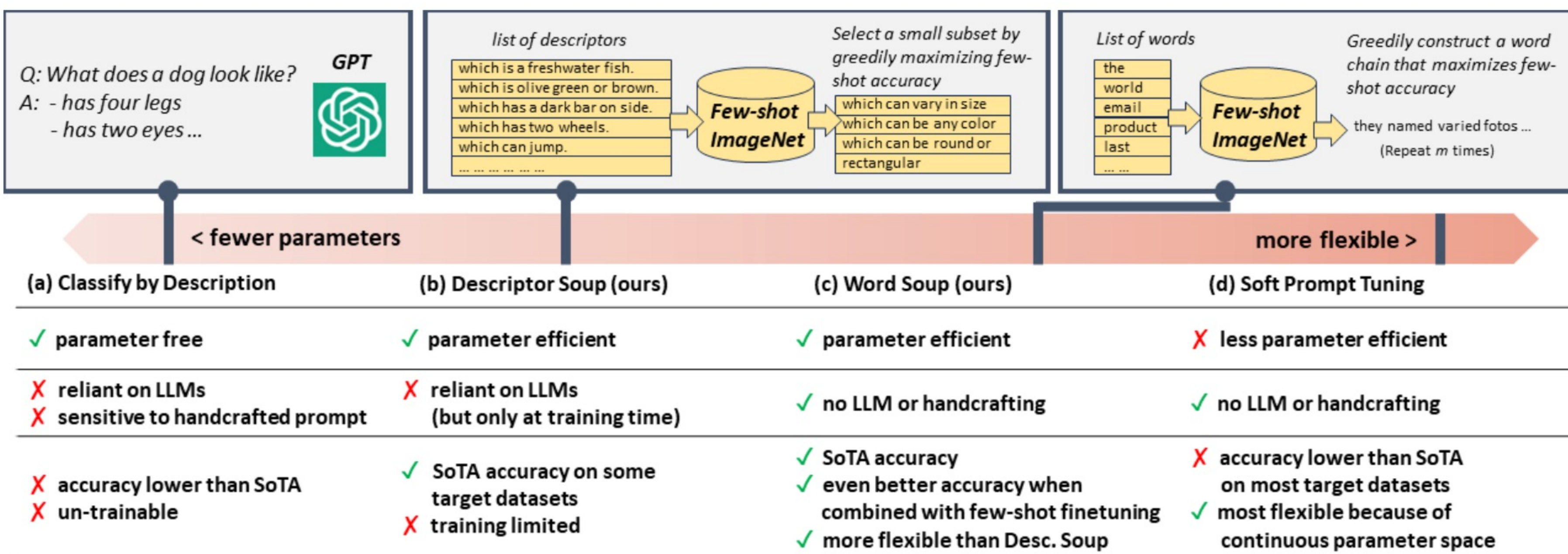
Method

- WaffleCLIP



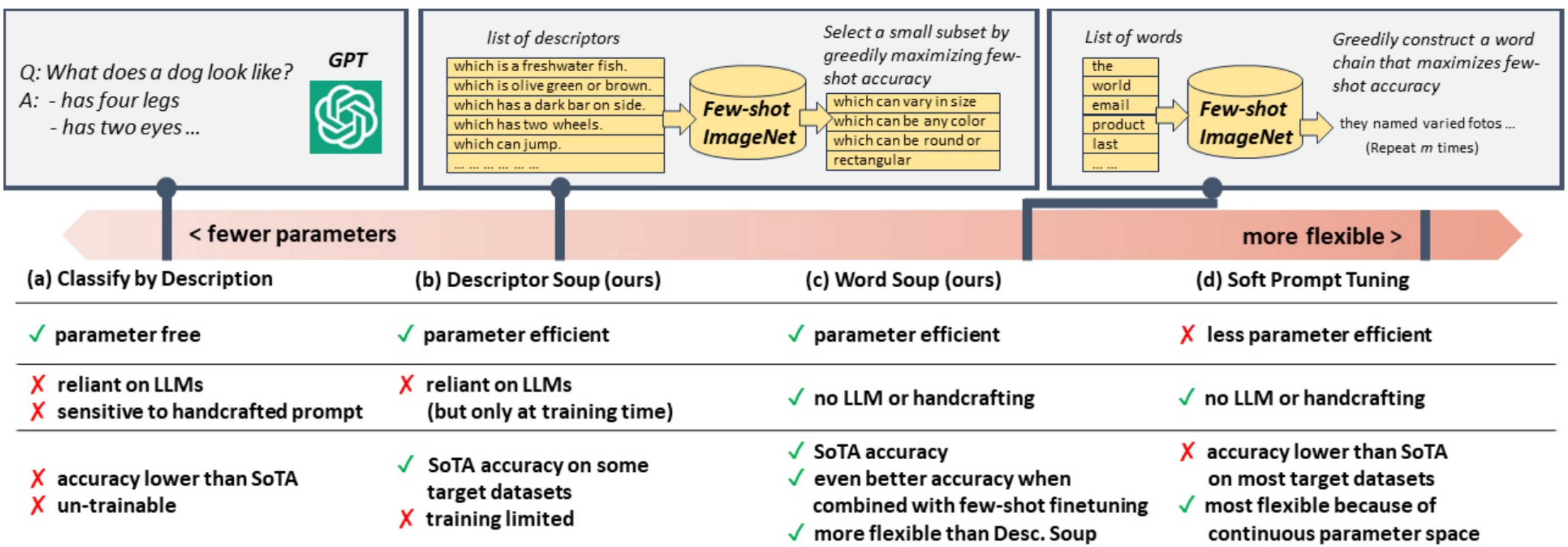
Method

- 本文思路
  - Waffle:集成随机描述符，并使用LLM来发现数据集级别的概念
  - 本文：设计一个优化过程从数据中学习好的描述符



Method

- 本文思路
  - Descriptor soup: 在GPT描述符(DCLIP)中筛选
  - Word soup: 在10000个最常用词中筛选
    - github.com/first20hours/google-10000-english



Method

- Descriptor Soup

1. Calculate $\ell(\mathcal{S}_{\text{train}}, \mathcal{T}_{\text{train}}(d))$ for all $d \in \mathcal{D}$. Sort the descriptors by increasing loss / decreasing accuracy. With slight abuse of notation, denote the sorted list as $\mathcal{D} = [d_0, ..., d_n]$.
2. Initialize the "descriptor soup" $\mathcal{D}^* = \{d_0\}$ with the best descriptor.
3. For i in $1 : n$: Add d_i to $\mathcal{D}^*$ if it decreases the loss of $\mathcal{D}^*$.
4. Return the first m descriptors in $\mathcal{D}^*$.

```
]acc now:  70.68125605583191  which has usually green or yellow.
]acc now:  70.69375514984131  which typically orange or brown.
]acc now:  70.71250081062317  which has long body.
acc now:  70.71875333786011  which is a long, horizontal seat.
]acc now:  70.73750495910645  which can be various colors, patterns, and styles.
]acc now:  70.7437515258789  which has small to medium-sized dog.
]acc now:  70.7562506198883  which has black, blue, brindle, fawn, or harlequin coloration.
```

Experiment

- 描述符的定性比较，Imagenet数据集训练后的描述符泛化能力更强

Color-coded by **source**: ImageNet, Pets, DTD, Random

Target: ImageNet	Alignment	Accuracy
no descriptor	0.301	67.1
which typically brightly colored.	0.305 (+0.004)	68.2 (+1.1)
which has usually white or off-white.	0.310 (+0.009)	68.4 (+1.3)
which is a long, low-slung body.	0.312 (+0.011)	68.3 (+1.2)
which is a curved or rectangular shape.	0.309 (+0.008)	68.6 (+1.5)
which can vary in size from small to large.	0.315 (+0.014)	68.5 (+1.4)
which has reddish brown fur.	0.300 (-0.001)	66.2 (-0.9)
which is a hard skeleton.	0.295 (-0.006)	66.6 (-0.5)
which is a medium-sized, short-haired cat.	0.291 (-0.010)	66.0 (-1.1)
which has sharp claws.	0.299 (-0.002)	66.6 (-0.5)
which is a repeating pattern.	0.295 (-0.006)	66.1 (-1.0)
which is a sign with the shop's name.	0.295 (-0.006)	66.7 (-0.4)

Target: Pets	Alignment	Accuracy
no descriptor	0.322	88.4
a type of pet. (handcrafted; for reference)	0.331 (+0.009)	89.0 (+0.6)
which is a large, powerful cat.	0.321 (-0.001)	89.8 (+1.4)
which has sharp claws.	0.324 (+0.002)	89.9 (+1.5)
which has soulful eyes.	0.317 (-0.005)	89.9 (+1.5)
which is a long arm with a claw ...	0.324 (+0.002)	87.8 (-0.6)
which is a medium-sized, short-haired cat.	0.327 (+0.005)	91.4 (+3.0)
which is a boat with sails.	0.293 (-0.029)	81.5 (-6.9)
which often used by knights and soldiers.	0.315 (-0.007)	80.8 (-7.6)
which can vary in size from small to large.	0.333 (+0.011)	88.6 (+0.2)
which typically has a yellow or brownish color.	0.335 (+0.013)	89.3 (+0.9)

Target: Textures (DTD)	Alignment	Accuracy
no descriptor	0.273	44.3
a type of texture. (handcrafted; for reference)	0.287 (+0.014)	44.1 (-0.2)
which may be decorated with a pattern or logo.	0.286 (+0.013)	47.2 (+2.9)
which is a sign with the shop's name.	0.261 (-0.012)	45.3 (+1.0)
which is a backdrop.	0.280 (+0.007)	46.6 (+2.3)
which is a repeating pattern.	0.283 (+0.010)	46.3 (+2.0)
which typically has a pattern or design.	0.295 (+0.022)	45.5 (+1.2)
which is a guard tower.	0.243 (-0.030)	43.4 (-0.9)
which has loud crow.	0.253 (-0.020)	42.4 (-1.9)
which can be brightly colored or patterned.	0.283 (+0.010)	44.5 (+0.2)
which is a curved or rectangular shape.	0.281 (+0.008)	44.4 (+0.1)

Token offset (令牌偏移)trick

- **original:** a photo of a dog, which may be large or small.
- **augmented:** a photo of a dog, ! ! ! ! ! which may be large or small. ("!" denotes the null token)

Method

- Word Soup

1. **Initialization:** Sort $\mathcal{W}$ by decreasing ZS accuracy to filter out unsuitable words (see Fig. 3 left). For this step, we only consider single word descriptors (e.g. "a photo of a cat, the."). Select the top-k_0 and top-k_1 words, denoted as $\mathcal{W}_{\text{top}k_0}$ and $\mathcal{W}_{\text{top}k_1}$, resp. $k_0 < k_1$.
2. Randomly select a word w from $\mathcal{W}_{\text{top}k_0}$ and initialize the descriptor $d = w$.
3. Shuffle $\mathcal{W}_{\text{top}k_1}$. Then, for $w' \in \mathcal{W}_{\text{top}k_1}$, append w' to d, only if it increases the accuracy of d.
4. return d.

重复2～4

```
]acc now 69.66875195503235, example: a photo of a tench, appearance
]acc now 69.67500448226929, example: a photo of a tench, appearance silly
]acc now 69.93125081062317, example: a photo of a tench, appearance silly similar
]acc now 70.11875510215759, example: a photo of a tench, appearance silly similar webpage
]acc now 70.13750076293945, example: a photo of a tench, appearance silly similar webpage particular
]acc now 70.35625576972961, example: a photo of a tench, appearance silly similar webpage particular weblog
]acc now 70.38750052452087, example: a photo of a tench, appearance silly similar webpage particular weblog familiar
```

Experiment

- word soup 比descriptor soup能达到更高的精度

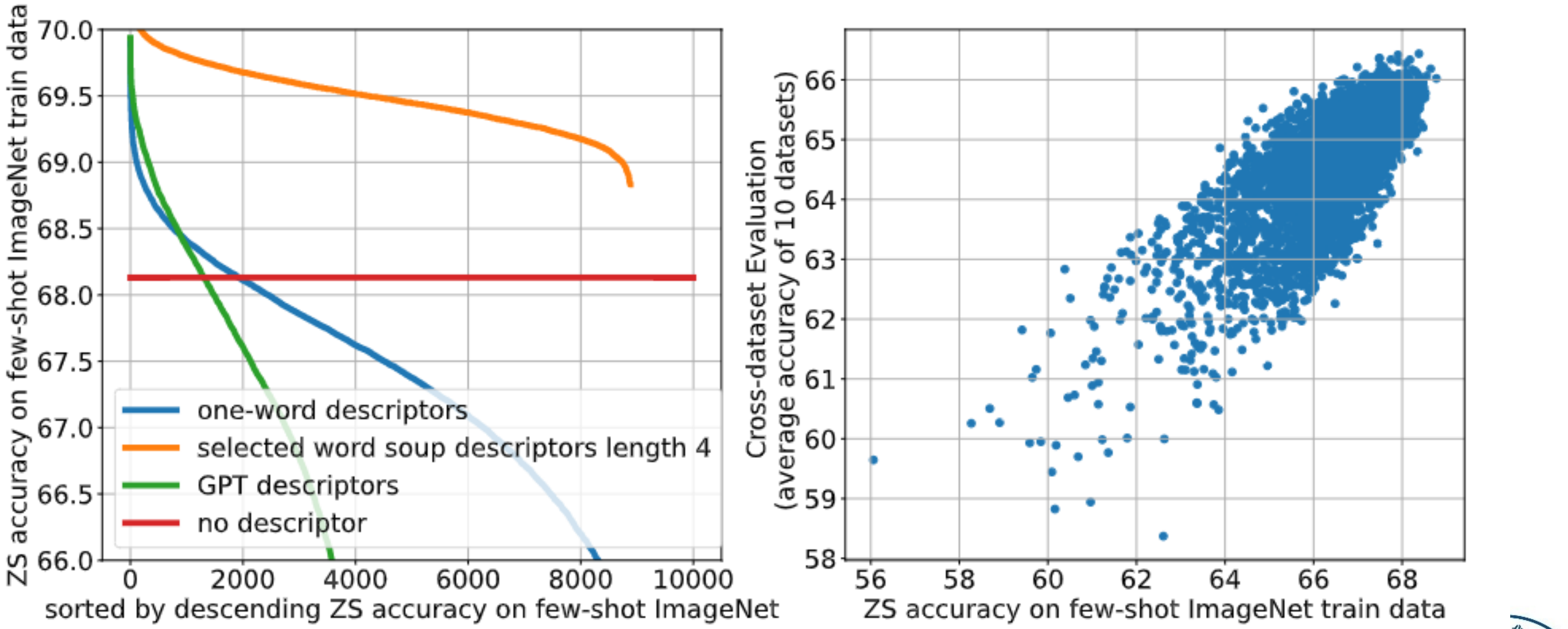
Target: ImageNet	Alignment	Uniformity	Accuracy
no descriptor	0.301	0.173	67.1
dat they ... difficulties.	0.306 (+0.005)	0.174 (+0.001)	68.9 (+1.8)
similar vary ... mention etc.	0.314 (+0.013)	0.183 (+0.010)	69.1 (+2.0)
separately aspects ... adopted.	0.315 (+0.014)	0.181 (+0.008)	69.2 (+2.1)
tue alot ... itself.	0.303 (+0.002)	0.178 (+0.005)	69.0 (+1.9)
bufing beginner ... status.	0.311 (+0.010)	0.181 (+0.008)	68.8 (+1.7)
soviet vbulletin ... inexpensive.	0.320 (+0.019)	0.195 (+0.022)	62.0 (-5.1)
ideal ips ... filename.	0.314 (+0.013)	0.196 (+0.023)	59.7 (-7.4)

Color-coded by **source**: ImageNet, Pets,

Target: ImageNet	Alignment	Accuracy
no descriptor	0.301	67.1
which typically brightly colored.	0.305 (+0.004)	68.2 (+1.1)
which has usually white or off-white.	0.310 (+0.009)	68.4 (+1.3)
which is a long, low-slung body.	0.312 (+0.011)	68.3 (+1.2)
which is a curved or rectangular shape.	0.309 (+0.008)	68.6 (+1.5)
which can vary in size from small to large.	0.315 (+0.014)	68.5 (+1.4)
which has reddish brown fur.	0.300 (-0.001)	66.2 (-0.9)
which is a hard skeleton.	0.295 (-0.006)	66.6 (-0.5)
which is a medium-sized, short-haired cat.	0.291 (-0.010)	66.0 (-1.1)
which has sharp claws.	0.299 (-0.002)	66.6 (-0.5)
which is a repeating pattern.	0.295 (-0.006)	66.1 (-1.0)
which is a sign with the shop's name.	0.295 (-0.006)	66.7 (-0.4)

Experiment

- 单个单词描述符/GPT描述符优于标准 ZS的描述符数量仅有1000+；当使用长度为 4 的word soup时，准确描述符的数量急剧增加



Experiment

- word soup与 zero-shot 方法比较，精度均有提升

		Source	Cross-dataset (XD) Evaluation Targets											Domain Generalization Targets				
	m	INet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	**Mean**	INet-V2	Sketch	INet-A	INet-R	**Mean**
CLIP ZS [68]	1	67.1	93.3	89.0	65.4	71.0	85.7	25.0	63.2	43.6	46.7	67.4	65.02	61.0	46.6	47.2	74.1	57.22
Ensemble [42]	80	68.4	93.5	88.8	66.0	71.1	86.0	24.8	66.0	43.9	45.0	68.0	65.31	61.9	48.5	49.2	**77.9**	59.36
GPT centroids [35]	5.8	68.2	94.1	88.4	65.8	71.5	85.7	24.7	**67.5**	44.7	46.6	67.4	65.63	61.5	48.2	48.9	75.1	58.40
GPT score mean [35]	5.8	68.6	93.7	89.0	65.1	72.1	85.7	23.9	67.4	44.0	46.4	66.8	65.42	61.8	48.1	48.6	75.2	58.42
Random descriptors	16	67.9	94.1	87.6	65.6	71.5	85.6	24.9	66.1	44.7	49.1	67.2	65.65	61.6	48.7	50.0	76.7	59.22
+ offset trick (ours)	96	68.5	93.5	89.2	65.8	72.0	85.7	25.2	66.1	44.4	53.0	68.2	66.29	61.9	48.9	**50.6**	77.5	59.76
Waffle CLIP [44]	16	68.1	93.5	88.4	65.4	72.0	85.9	25.9	66.2	44.1	46.3	68.0	65.58	61.8	48.6	49.8	76.2	59.08
+ offset trick (ours)	96	68.6	93.1	89.5	65.9	72.1	86.1	**26.3**	66.2	44.2	52.5	68.8	66.49	62.1	48.9	50.2	77.1	59.59
Descriptor soup (ours)	16.7	68.9	**94.7**	89.4	**66.2**	72.2	86.2	25.5	67.3	45.1	46.6	68.7	66.18	62.1	48.7	49.7	76.4	59.25
+ offset trick (ours)	100	69.1	93.8	**89.8**	66.0	**72.9**	**86.2**	25.4	66.8	45.0	51.6	**69.1**	66.67	62.6	**49.0**	50.5	77.2	59.82
Word soup (ours)	8	69.2	94.4	89.5	65.4	72.3	85.8	25.8	67.4	44.7	53.5	68.4	66.72	62.9	48.7	50.2	77.0	59.69
Word soup score mean (ours)	8	**69.4**	94.3	89.6	65.4	72.4	85.9	25.9	67.3	**45.2**	**55.8**	68.5	**67.03**	**63.0**	49.0	50.4	77.2	**59.90**
gain over GPT		+0.8	+0.6	+0.6	+0.3	+0.3	+0.2	+2.0	-0.1	+1.2	+9.4	+1.7	+1.6	+1.2	+0.9	+1.8	+2.0	+1.5
gain over Waffle		+1.3	+0.8	+1.2	+0.0	+0.4	+0.0	-0.0	+1.1	+1.1	+9.5	+0.5	+1.5	+1.2	+0.4	+0.6	+1.0	+0.8

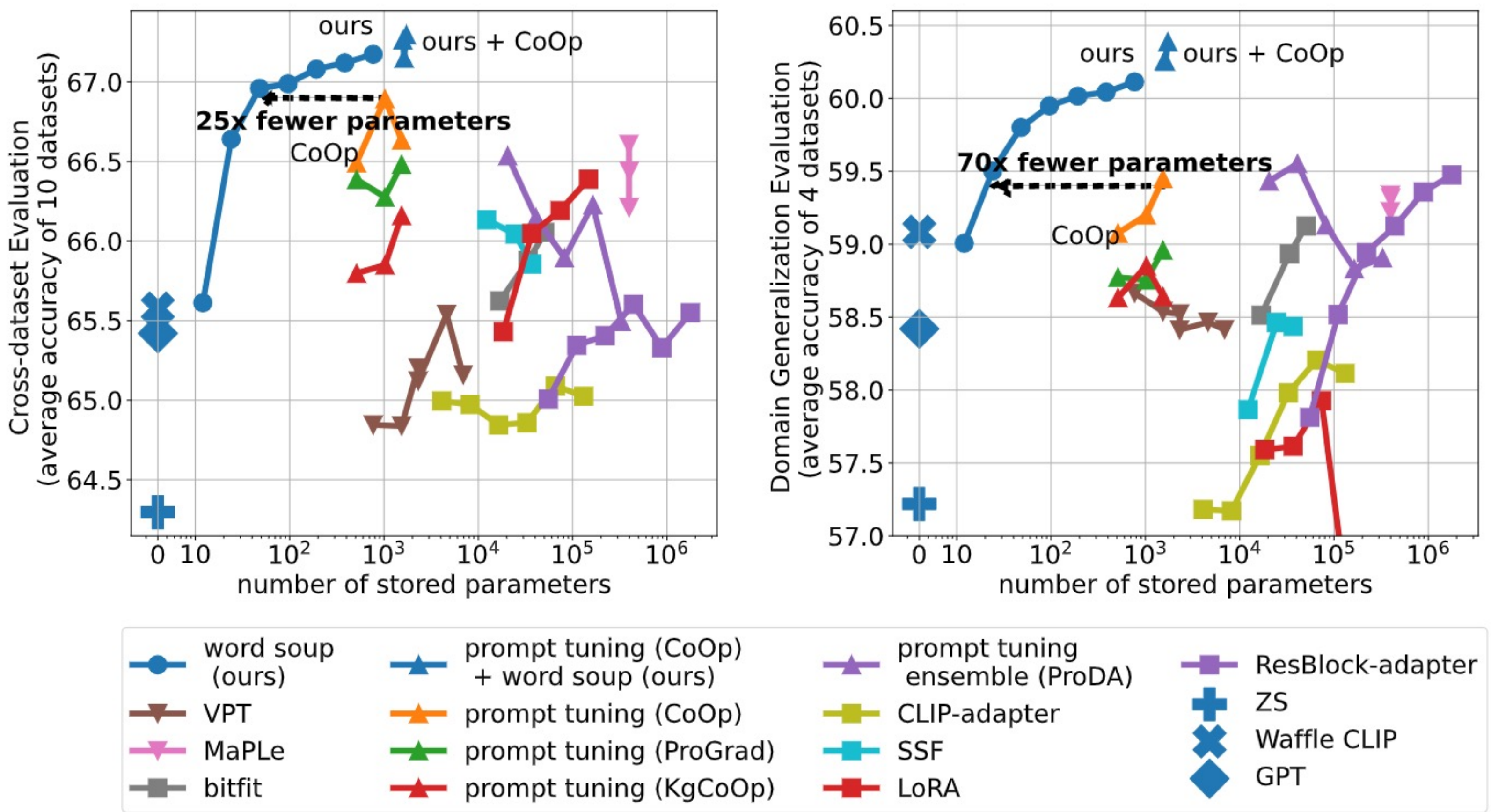
Experiment

- word soup与 few-shot 方法(叠加)比较，精度也均有提升

	m	Source INet	**XD Mean** (10 datasets)	**DG Mean** (4 datasets)
CLIP ZS [42]	1	67.1	65.02	57.22
CoOp [68]†		71.5	63.88	59.3
Co-CoOp [67]†		71.0	65.74	59.9
MaPLe [25]†		70.7	66.30	60.3
CLIPood [50]†		71.6		60.5
Cross Entropy (CE)	1	**72.3**	66.80	60.39
+ GPT score mean [35]	5.8	71.7	66.86	59.92
+ Random descriptors	32	71.6	66.89	60.69
+ Waffle CLIP [44]	32	71.6	66.58	60.65
+ Descriptor soup (ours)	16.7	72.1	67.10	60.70
+ offset trick (ours)	100	72.1	**67.51**	61.01
+ Word soup centroids (ours)	8	71.8	67.16	61.22
+ Word soup score mean (ours)	8	71.7	67.43	**61.32**
+ Descriptor soup upper bound	11	71.7	67.62	61.01
ProGrad [69]	1	69.8	66.48	58.96
KgCoOp [22]	1	69.2	66.16	58.64
ProDA [31]	32	70.0	66.23	58.83
Vanilla CoOp [68]	1	70.0	66.52	59.25
+ Word soup score mean (ours)	8	**70.2**	**67.30**	**60.25**
Vanilla MaPLe [25]	1	70.7	66.44	59.32
+ Word soup score mean (ours)	8	**70.8**	**66.65**	**60.20**
Vanilla CLIPood [50]	1	**72.9**	66.50	60.47
+ Word soup score mean (ours)	8	72.0	**67.42**	**61.23**

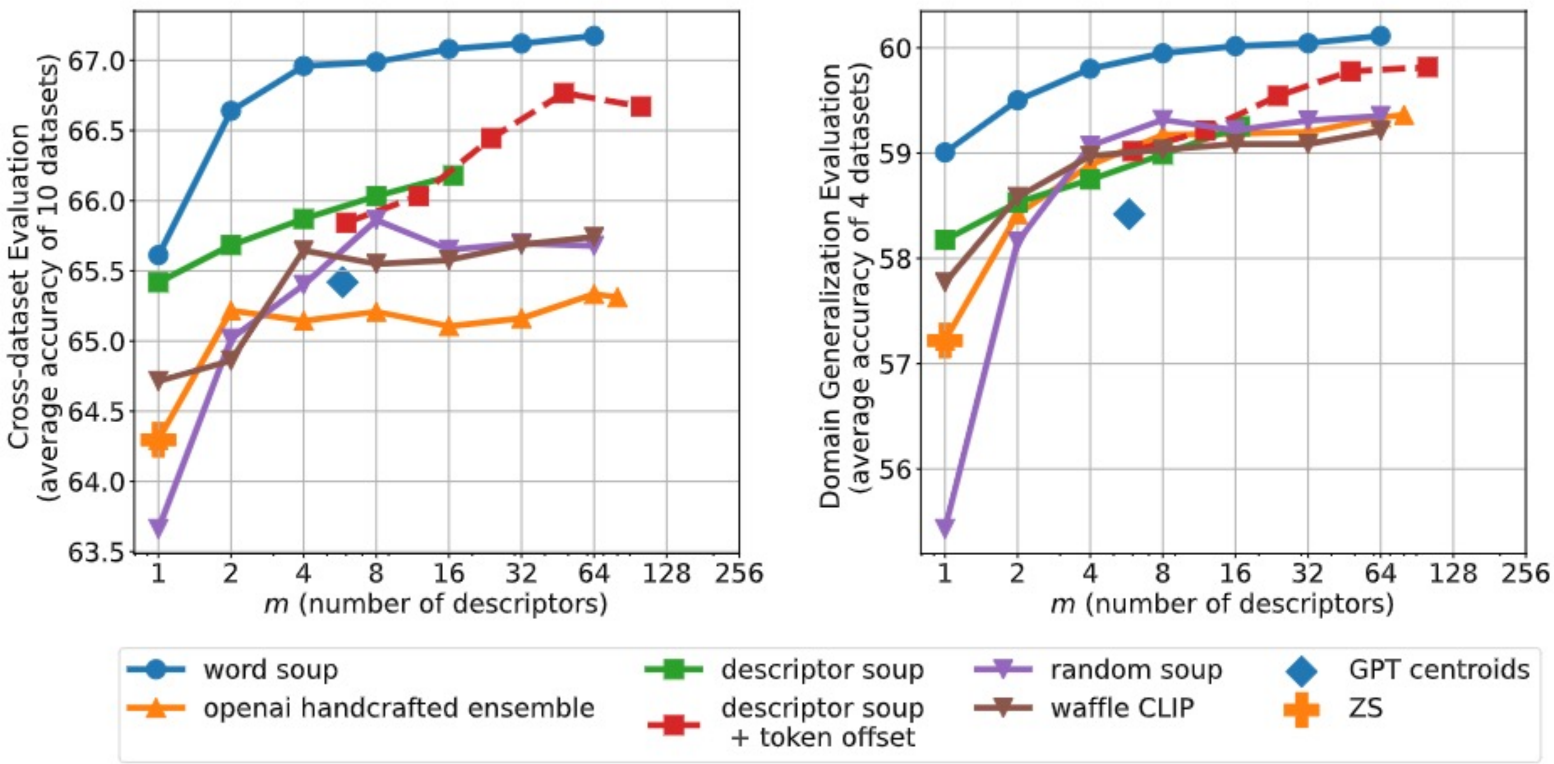
Experiment

- word soup可以在 XD 和 DG 基准上分别使用 25 倍和 70 倍更少的参数来实现最大的 CoOp 精度。



Experiment

- 消融实验
- word soup仅用更少的描述符在XD和DG上分别达到更高的精度



Summary

- 本文提出了descriptor and word soups来解决跨数据集和域泛化问题。

- 通过最大化源数据集的训练精度，使用贪心算法选择一组描述符/构建一系列单词。这些soup方法通过显式最大化训练精度，实现了比以前基于描述符的方法更高的目标分类精度。

- 与所有基线相比，word soup在参数效率和目标精度之间实现了最佳权衡。