Learning Transferable Negative Prompts for Out-of-Distribution Detection

Tianqi Li[1], Guansong Pang*[2], Xiao Bai*[1], Wenjun Miao[1], and Jin Zheng[1]

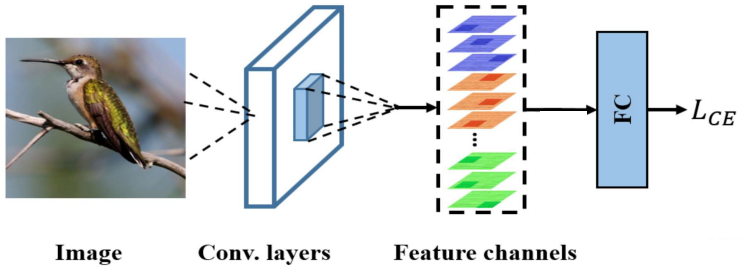
[1]School of Computer Science and Engineering, State Key Laboratory of Complex & Critical Software Environment, Jiangxi Research Institute, Beihang University, China
[2]School of Computing and Information Systems, Singapore Management University

CVPR 2024

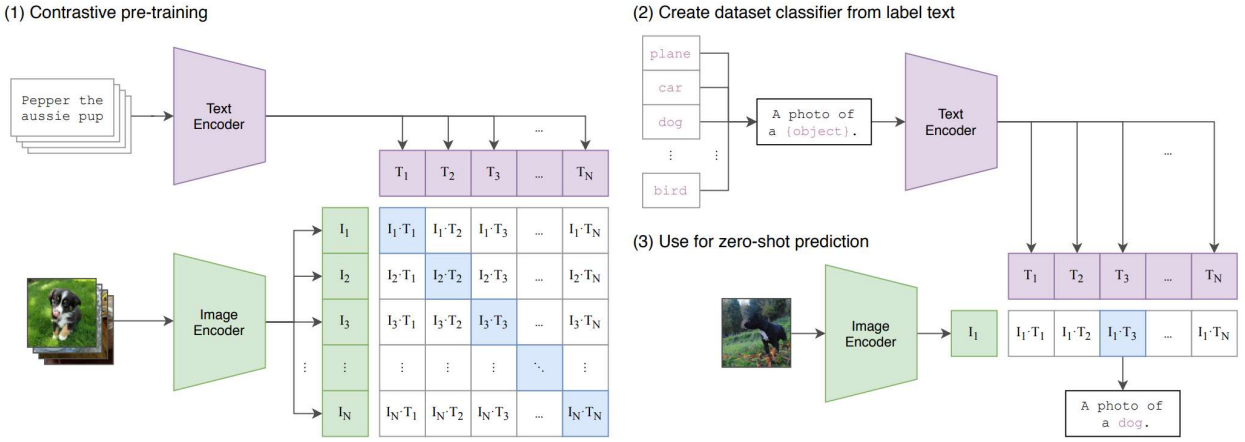
Background

- OOD：正确分类已知类 + **拒绝未知类 (通常来自不同分布)**

- 传统的深度学习方法仅依靠图像特征训练分类器



- 类似CLIP的Vision-Language Models (VLMs)可以将文本与视觉结合，提升模型的语义理解能力



Background -> Prompt Learning (or Prompt Tuning)

- 稍微变动prompt就会得到截然不同的结果

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)

Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

(c)

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

Learning to Prompt for Vision-Language Models IJCV 2022

Background -> Prompt Learning (or Prompt Tuning)

NLP领域中的Prompt Learning

- 核心思想是通过设计和优化输入提示 (prompts) 来引导预训练模型生成所需的输出

请举一个prompt tuning的例子，要求简单明了

好的，这里是一个关于Prompt Tuning的简单明了的例子，展示如何通过调整Prompt来优化模型的输出。

初始Prompt:

```vbnet
请判断以下句子的情感（积极、消极、中性）：
I love the new design of your website! It's user-friendly and looks amazing.

情感：
```

输出:

```
情感：积极
```

Prompt Tuning

调整后的Prompt:

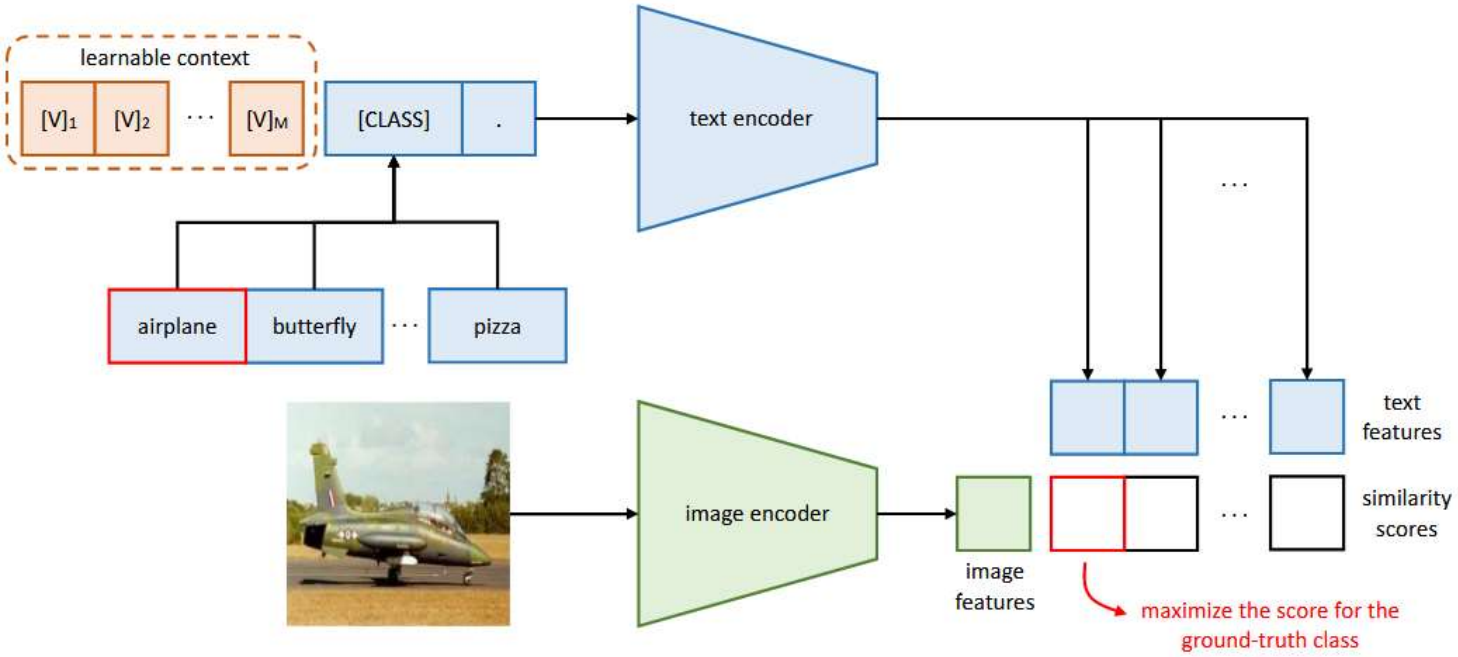
```vbnet
请判断以下句子的情感（积极、消极、中性），并解释你的判断：
I love the new design of your website! It's user-friendly and looks amazing.

情感及理由：
```

输出:

```
情感：积极
理由：句子中使用了“love”和“amazing”这样的积极词汇，表达了对新设计的喜爱和赞赏。
```

Background -> Prompt Learning (or Prompt Tuning)

- 设计可学习的prompts

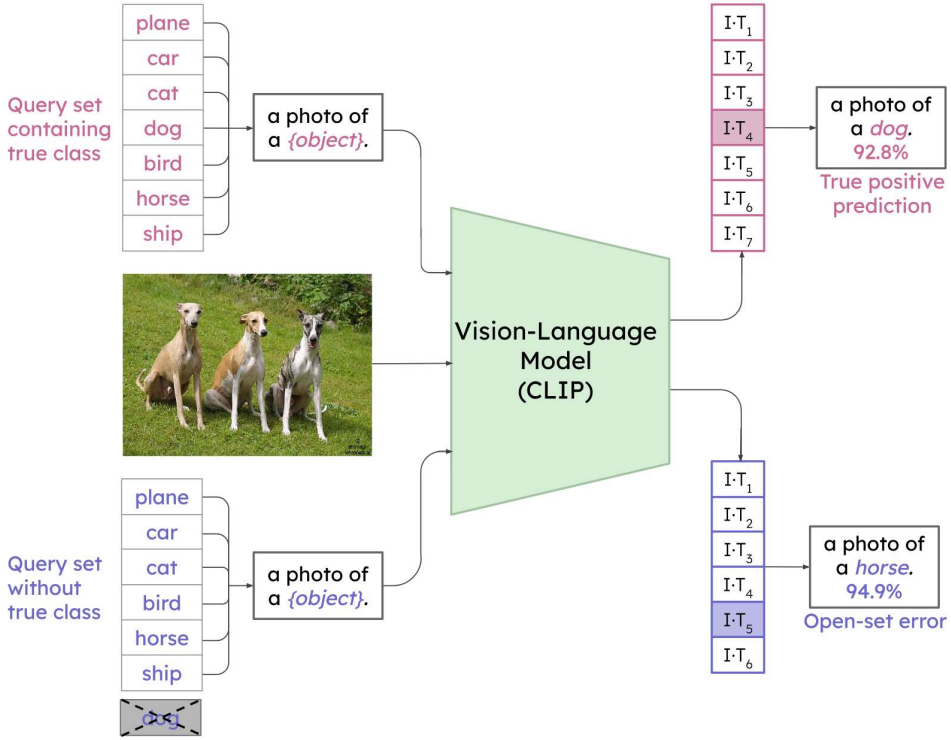


Learning to Prompt for Vision-Language Models IJCV 2022

Motivation

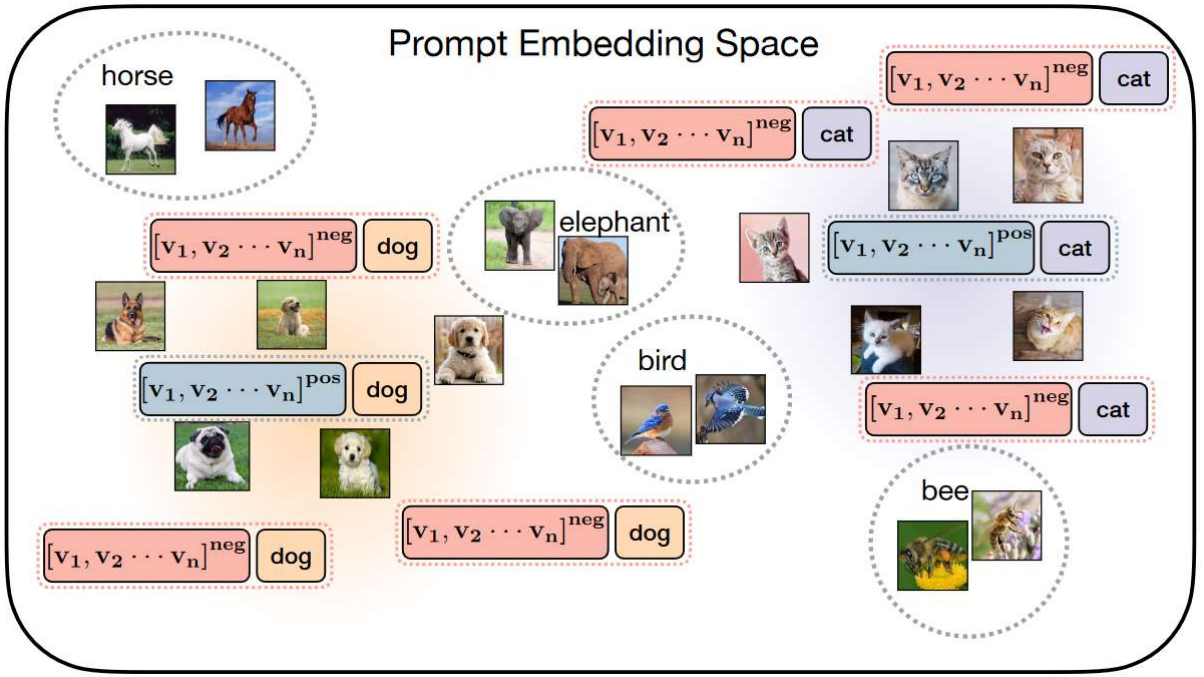
Prompt Learning 提升了CLIP的能力，但仍然无法解决Out-of-Distribution (OOD) 问题

- 基于CLIP的分类通常是将prompt中的 “[class name]” 替换为每个类别的名称
  Closed Set
- 然而在OOD问题中，模型无法得知未知类的名称
  Open Set



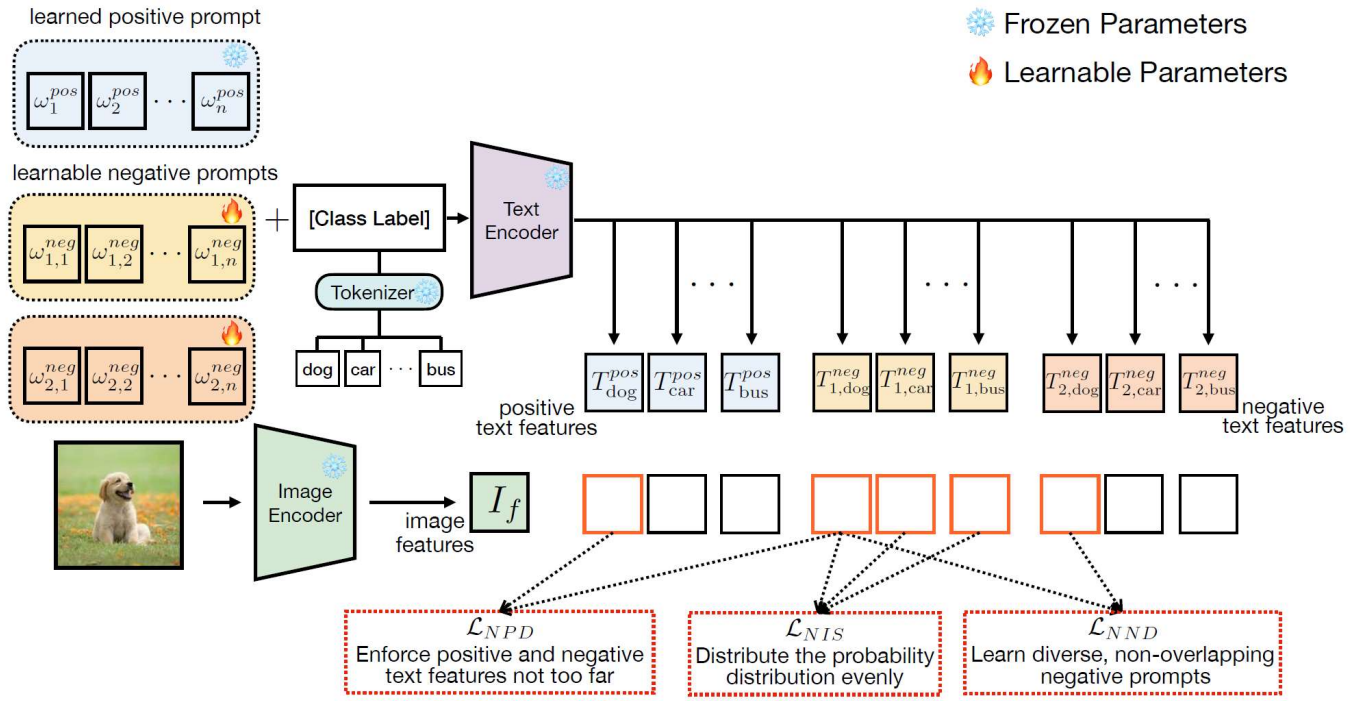
Method

NegPrompt



- 对于每个ID class，NegPrompt训练若干negative prompts，negative prompts与对应类别positive prompts的语义是相反的
- OOD样本与negative prompts的相似性更高

Overview of NegPrompt



- NegPrompt希望利用 ID images 和 positive prompts来学习对应的negative prompts

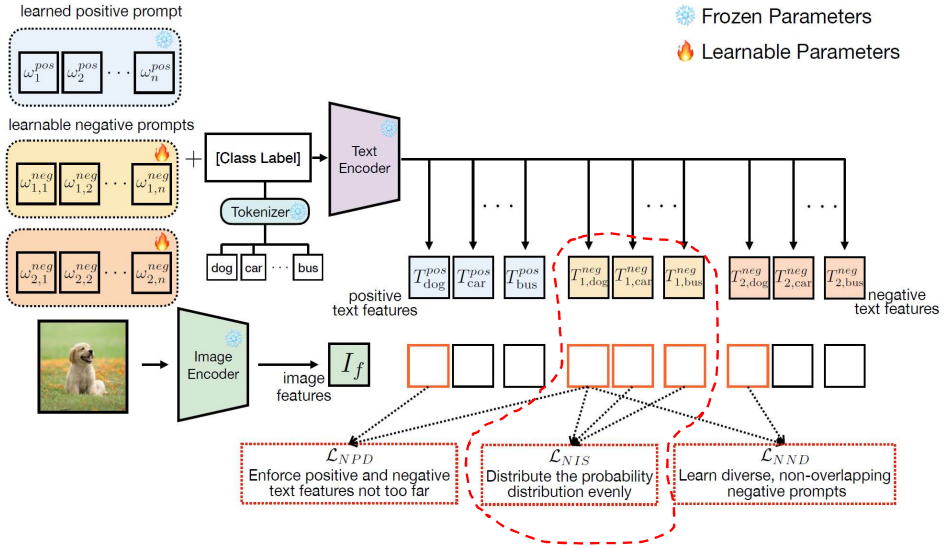
- Negative prompts应该具有这些性质：
  - 与 OOD images 相似性较高
  - 与 positive prompts 分隔开，即与每个positive prompts的相似性都比较低
  - Negative prompts 应具有多样性（互不重叠）

Negative-Image Separation Loss (NIS)

目标： negative prompts向OOD数据靠近

问题： 没有OOD训练数据，只有ID数据

方法： negative prompts远离ID数据



- ID数据与negative prompts的概率分布均匀分配到所有negative prompts，即ID数据与所有negative prompts均一的远

$$\mathcal{L}_{NIS} = \mathbb{E}_{x_{\text{in}} \sim D_{\text{train}}^{\text{in}}} [H(\mathbf{u}; F(x))],$$

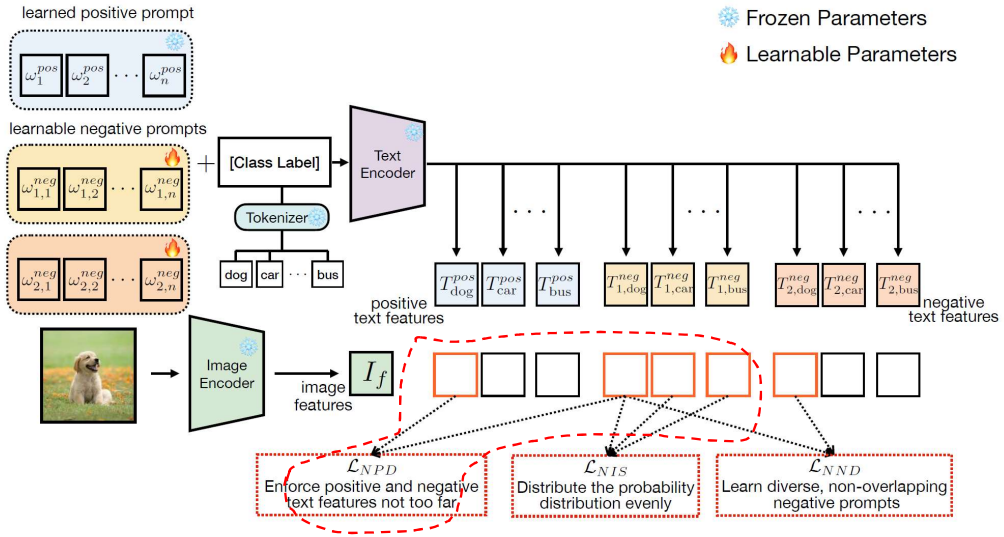
其中，$F(x) = \text{Softmax}(\text{sim}(T_{i,j}^{f,neg}, I^f))$，u是均匀分布，H是交叉熵损失

Negative-Positive Distance Loss (NPD)

目标：避免平凡解，即避免negative prompts同时远离 ID 和 OOD

问题：需要控制negative prompts与 ID 和 OOD 的距离

方法：设计loss，使negative prompts 与 positive prompts 距离不要太远



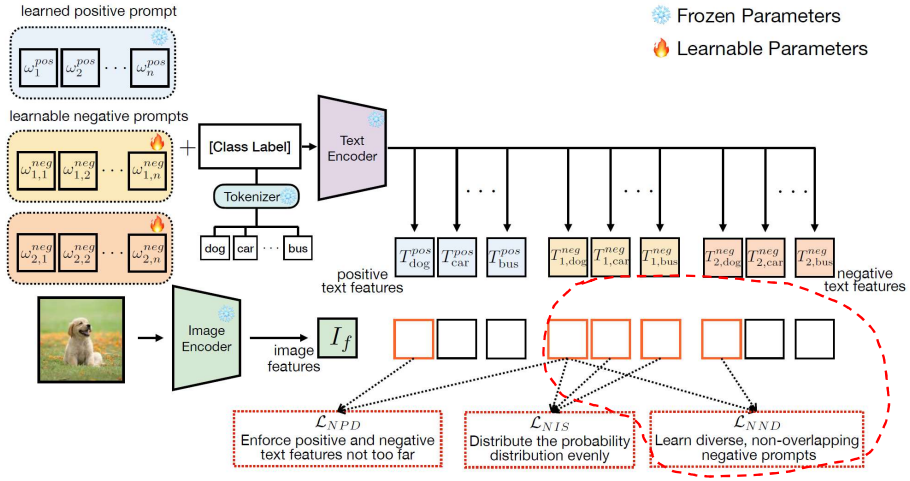
$$\mathcal{L}_{NPD} = -\frac{1}{k*p}\sum_{j=1}^{k}\sum_{i=1}^{p} sim(T_{i,j}^{f,neg}, T_{j}^{f,pos}).$$

Negative-Negative Distance Loss (NND)

目标：使negative prompts多样化，避免重叠

方法：设计loss，增大同一类别negative prompts 之间的距离

$$\mathcal{L}_{NND} = \frac{1}{k * p * (p-1)} \sum_{j=1}^{k} \sum_{i=1}^{p} \sum_{l \neq i} sim(T_{i,j}^{f,neg}, T_{l,j}^{f,neg}).$$



最终损失函数： $\mathcal{L}_{NegativePrompts} = \mathcal{L}_{NIS} + \beta * \mathcal{L}_{NPD} + \gamma * \mathcal{L}_{NND}$

推理：

$$p(y=i|x) = \frac{\exp(S_i^{f,pos})}{\sum_{j=1}^{k} \exp(S_j^{f,pos}) + \sum_{l=1}^{p} \sum_{j=1}^{k} \exp(S_{l,j}^{f,neg})}$$

Experiments

ID dataset: ImageNet-1k **OOD dataset: iNaturalist, SUN, Places and Textures**

Method	Texture		iNaturalist		Places		SUN		Avg	
	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓
Zero-shot methods										
MCM [32]†	86.11	57.77	94.61	30.91	89.77	44.69	92.57	34.59	90.76	42.74
CLIPN [47]†	90.93	40.83	95.27	23.94	92.28	33.45	93.92	26.17	93.10	31.10
CLIP-based posthoc methods										
MSP [10]†	74.84	73.66	77.74	74.57	72.18	79.12	73.97	76.95	74.98	76.22
MaxLogit [12]†	88.63	48.72	88.03	60.88	87.45	55.54	91.16	44.83	88.82	52.49
Energy [27]†	88.22	50.39	87.18	64.98	87.33	57.40	91.17	46.42	88.48	54.80
ReAct [42]†	88.13	49.88	86.87	65.57	87.42	56.85	91.04	46.17	88.37	54.62
ODIN [26] †	87.85	51.67	94.65	30.22	85.54	55.06	87.17	54.04	88.80	47.75
Prompt learning methods										
CoOp [55]	89.47	45.00	93.77	29.81	90.58	40.11	93.29	40.83	91.78	51.68
LoCoOp [33]†	90.19	42.28	96.86	16.05	91.98	32.87	95.07	23.44	93.52	28.66
NegPrompt (Ours)	**91.60**	**35.21**	**98.73**	**6.32**	**93.34**	**27.60**	**95.55**	**22.89**	**94.81**	**23.01**
Open-vocabulary OOD detection										
CoOp (10%)	87.58	50.55	91.08	42.53	89.56	46.12	91.52	41.92	89.94	45.28
LoCoOp (10%)	88.21	47.32	94.47	34.90	91.64	39.85	92.54	26.30	90.15	37.09
NegPrompt (Ours) (10%)	90.30	39.31	98.39	7.48	92.68	29.75	93.70	26.92	93.76	25.86

Experiments

Top-1 Accuracy

Method	Top-1 Accuracy
CoOp	72.1
LoCoOp†	71.7
CLIPN & MCM	67.0
NegPrompt(Ours)(10%)	71.9
NegPrompt(Ours)(Full)	72.1

Why Does NegPrompt Work?

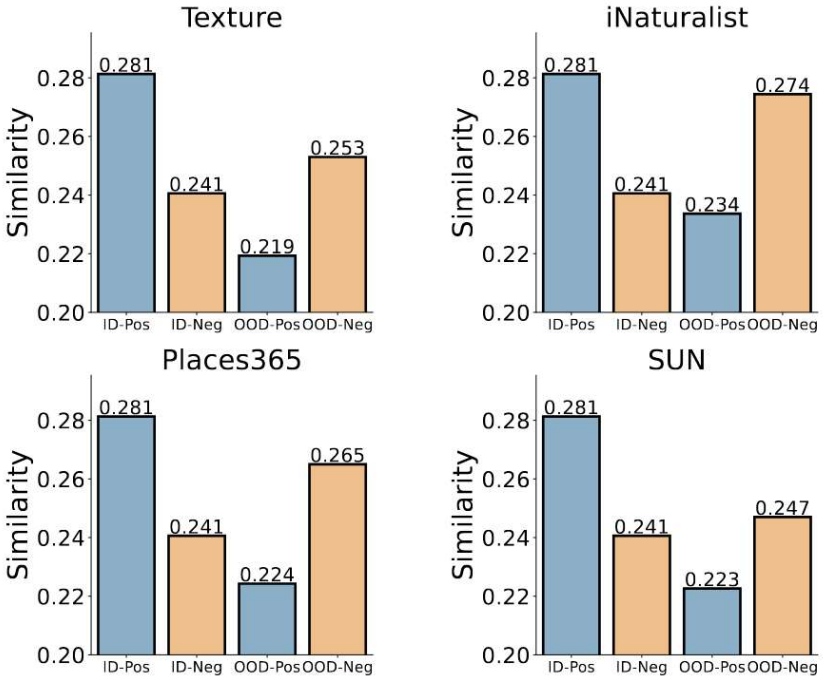


Figure 3. Similarity of ID/OOD and Positive/Negative Prompts.

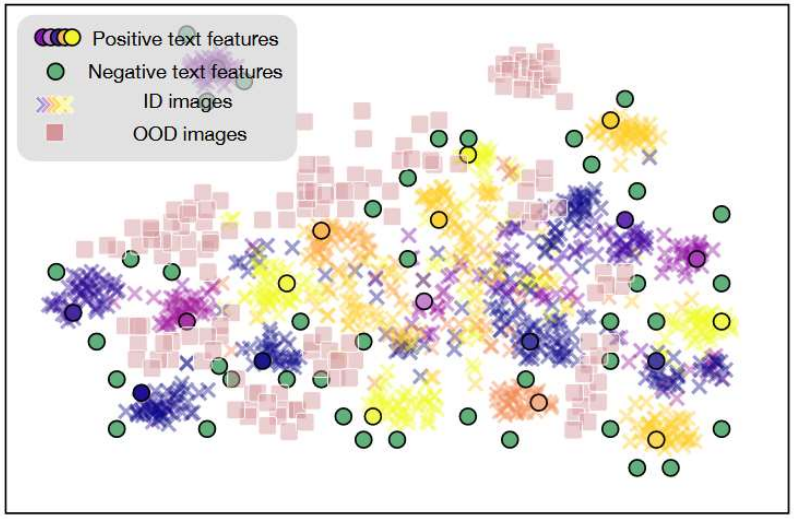


Figure 4. T-SNE visualization of NegPrompt, utilizing a subset of ImageNet - TinyImageNet as the dataset.

Thank you!