Delving into OOD Detection with Vision-Language Representations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recognizing out-of-distribution (OOD) samples is critical for machine learning systems deployed in the open world. The vast majority of OOD detection methods are driven by a single modality (*e.g.*, either vision or language), leaving the rich information in multi-modal representations untapped. Inspired by the recent success of vision-language pre-training, this paper enriches the landscape of OOD detection from a single-modal to a multi-modal regime. Particularly, we propose Maximum Concept Matching (MCM), a simple yet effective zero-shot OOD detection method based on aligning visual features with textual concepts. We contribute in-depth analysis and theoretical insights to understand the effectiveness of MCM. Extensive experiments demonstrate that our proposed MCM achieves superior performance on a wide variety of real-world tasks. MCM with vision-language features outperforms a common baseline with pure visual features on a hard OOD task with semantically similar classes by 56.60% (FPR95).

1 Introduction

Out-of-distribution (OOD) detection is critical for deploying machine learning models in the wild, where samples from novel classes can naturally emerge and should be flagged for caution. Despite increasing attention, the vast majority of OOD detection methods are driven by single-modal learning [22, 80, 59, 78, 20, 83, 76, 27]. For example, labels are typically encoded as one-hot vectors in image classification, leaving the semantic information encapsulated in texts largely unexploited. OOD detection relying on pure visual information can inherit the limitations, *e.g.*, when an OOD input may be visually similar to in-distribution (ID) data yet semantically different from any ID class.

In this paper, we delve into a new landscape for OOD detection, departing from the classic single-modal toward a *multi-modal* regime. While the motivation is appealing, a core challenge remains: *how to effectively utilize joint vision-language features for OOD detection?* In the visual domain, existing methods typically require good feature representations [61, 57], and a distance metric under which OOD data points are relatively far away from the in-distribution (ID) data [35]. These approaches, however, do not directly translate into the multi-modal regime. On the representation learning side, recent vision-language pre-training schemes such as CLIP [51] and ALIGN [26] have emerged as promising alternatives for visual representation learning. The main idea is to align an image with its corresponding textual description in the feature space. While the resulting representations are powerful, OOD detection based on such aligned multi-modal features is still in its infancy.

We bridge the gap by exploring a distance-based OOD detection approach, leveraging the joint vision-language representations. Our method capitalizes on the compatibility between visual features and textual features. By defining the textual features as the "concept prototypes" for each ID class, we characterize OOD uncertainty by the distance from the visual feature to the closest ID prototype. That is, images closer to one of the textual embeddings of ID classes are more likely to be ID and

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

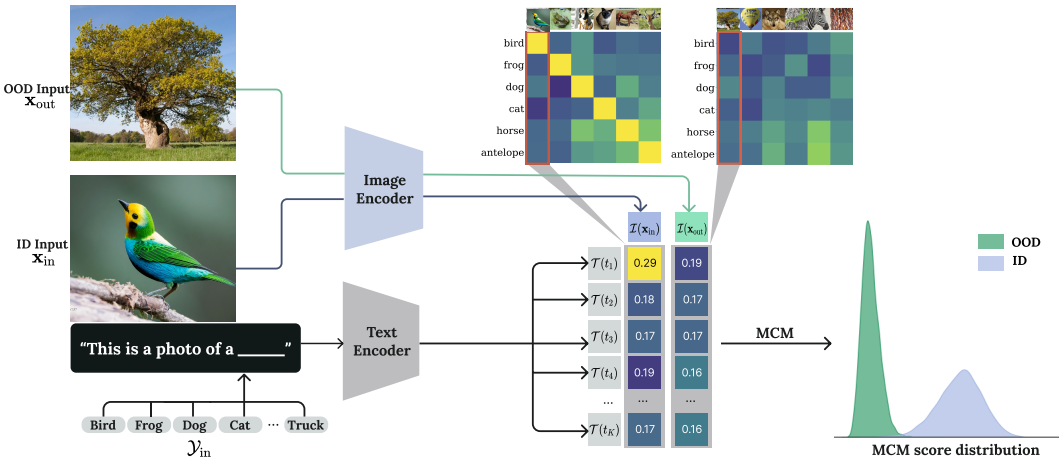

Figure 1: Overview of the proposed zero-shot OOD detection framework. The ID classification task is defined by a set of class labels $\mathcal{Y}_{\text{in}}$. The goal of OOD detection is to detect samples that do not belong to $\mathcal{Y}_{\text{in}}$. We view the textual embeddings of ID classes (wrapped by text template) as concept prototypes. The OOD uncertainty of an input image can be characterized by the distance from visual features to the closest ID prototype. By properly scaling the distance, the MCM score achieves strong ID-OOD separability. See Section 3 for details.

vice versa. By a proper scaling of the distance, our proposed Maximum Concept Matching (**MCM**) score achieves strong ID-OOD separability (see Figure 1). MCM stands in contrast with the previous distance-based approaches, such as Mahalanobis [35], which defines class prototypes based on pure visual embeddings. Indeed, we show later in Section 5 that MCM (with multi-modal vision-language features) is far more competitive than Mahalanobis (with single-modal visual features). Moreover, while prior works of CLIP-based OOD detection [14, 12] rely on a set of candidate OOD labels, MCM is OOD-agnostic and alleviates the need for any prior information about test inputs.

Our work also advances the field by showcasing the promise of zero-shot OOD detection, which offers strong performance and generality without training on the ID samples. In particular, classic OOD detection methods often require training from scratch [21, 7] or fine-tuning [25, 14] on a given ID dataset. In this setting, a classifier and its companion OOD detector are good at only one task. Every new task (ID dataset) requires additional training and brings additional computation and storage costs. In contrast, we show for the first time that: **(1)** MCM achieves superior performance across a wide variety of real-world tasks—with just *one single pre-trained model*. This is encouraging given that there is no training or any OOD information involved. **(2)** On the challenging ImageNet-1k benchmark, MCM's zero-shot OOD detection performance favorably matches and even outperforms strong task-specific baselines fine-tuned on BiT [25] and ViT models [14]. **(3)** MCM remains robust against hard OOD inputs, including both semantically hard OODs [72] and spurious OODs [42].

We summarize our main contributions as follows:

1. We propose MCM, a simple yet effective OOD detection method based on aligned vision-language features. MCM offers several compelling advantages over other OOD detection methods: generalizable (one model supports many tasks), OOD-agnostic (no information required from OOD data), training-free (no downstream fine-tuning required), and scalable to large real-world tasks.

2. We conduct extensive experiments and show that MCM achieves superior performance on a wide range of real-world tasks. On ImageNet-1k, MCM achieves an average AUROC of 91.78%, outperforming most recent methods that require training. Moreover, we show that MCM remains competitive under challenging hard OOD evaluation tasks.

3. We provide in-depth empirical and theoretical analysis, providing insights to understand the effectiveness of MCM. We hope that this work will serve as a springboard for future works on OOD detection with multi-modal features.

2 Preliminaries

Contrastive vision-language pre-training. Compared to visual representation learning models such as ViT [11], vision-language representation learning demonstrates superior performance on image classification tasks. For instance, CLIP [51] adopts a self-supervised contrastive objective (*i.e.*, InfoNCE loss [64]) to align an image with its corresponding textual description in the feature space. Specifically, CLIP adopts a simple dual-stream architecture with one text encoder $\mathcal{T} : t \rightarrow \mathbb{R}^d$ (e.g. Transformer [66]) and one image encoder $\mathcal{I} : \mathbf{x} \rightarrow \mathbb{R}^d$ (*e.g.*, ViT [11]). After pre-training on a dataset of 400 million text-image pairs, the joint vision-language embeddings of CLIP well associate objects in different modalities. Despite the promise, existing CLIP-like models perform zero-shot classification in a *closed-world* setting. That is, it will match an input into a fixed set of categories, even if it is irrelevant (*e.g.*, a tree being predicted as a bird in Figure 1). This motivates our work to leverage the multi-modal representation for OOD detection, which is largely unexplored.

Zero-shot OOD detection. Given a pre-trained model, a classification task of interest is defined by a set of class labels/names $\mathcal{Y}_{\text{in}}$, which we refer to as the known (ID) classes. Here ID classes are defined *w.r.t.* the classification task of interest, instead of the classes used in pre-training. Accordingly, OOD is defined *w.r.t.* the ID classes, not the data distribution during pre-training. The goal of OOD detection is to (1) detect samples that do not belong to any of the known classes; (2) otherwise, assign test samples to one of the known classes. Therefore, the OOD detector can be viewed as a "safeguard" for the classification model. Formally, we denote the OOD detector as a binary function mapping: $G(\mathbf{x}; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) : \mathcal{X} \rightarrow \{\text{in}, \text{out}\}$, where $\mathbf{x} \in \mathcal{X}$ denote a test image. Our method is based on only the names of the given classes in $\mathcal{Y}_{\text{in}}$, and a pre-trained model. *Different from standard supervised learning, there is no training on the ID samples involved, hence zero-shot*.

3 OOD Detection via Concept Matching

We illustrate our approach in Figure 1, which derives the OOD detector $G(\cdot)$ based on *concept matching*. For a given task with label set $\mathcal{Y}_{\text{in}} = \{y_1, y_2, ..., y_K\}$, we can construct a collection of concept vectors $\mathcal{T}(t_i), i \in \{1, 2, ..., K\}$, where t_i is the text prompt "`this is a photo of a` $\langle y_i \rangle$" for a label y_i. The concept vectors are represented by the embeddings of the text prompts.

For any test input image $\mathbf{x}'$, we can calculate the label-wise matching score based on the cosine similarity between the image feature $\mathcal{I}(\mathbf{x}')$ and the concept vector $\mathcal{T}(t_i)$: $s_i(\mathbf{x}') = \frac{\mathcal{I}(\mathbf{x}') \cdot \mathcal{T}(t_i)}{\|\mathcal{I}(\mathbf{x}')\| \cdot \|\mathcal{T}(t_i)\|}$. Formally, we define the maximum concept matching (**MCM**) score as:

$$S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = \max_i \frac{e^{s_i(\mathbf{x}')/\tau}}{\sum_{j=1}^{K} e^{s_j(\mathbf{x}')/\tau}}, \tag{1}$$

where τ is the temperature. Intuitively, for ID data, it will be matched to one of the concept vectors (textual prototypes) with a high score; and vice versa. Formally, our OOD detection function can be formulated as:

$$G(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = \begin{cases} 1 & S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) \geq \lambda \\ 0 & S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) < \lambda \end{cases},$$

where by convention 1 represents the positive class (ID) and 0 indicates OOD. λ is chosen so that a high fraction of ID data (*e.g.*, 95%) is above the threshold. For samples that are classified as ID, one can obtain the class prediction based on the closest concept: $\hat{y} = \arg\max_{i \in [K]} s_i$.

Remark: (1) Our work differs from (and perhaps is complementary to) CLIP by focusing on OOD detection rather than (closed-world) zero-shot classification. We show new theoretical insights next that softmax scaling plays a unique role in zero-shot OOD detection—improving the separability between ID and OOD data. This role has not been studied rigorously for zero-shot OOD detection. Readers familiar with CLIP may notice that MCM can be used for zero-shot classification in the closed world. This also makes MCM practically convenient for dual goals: detect OOD samples and assign ID data to one of the known classes. (2) Our method in principle is not limited to CLIP; it can be generally applicable for contrastive vision-language pre-training models that promote multi-modal feature alignment.

New insights on softmax scaling for zero-shot OOD detection. We provide theoretical justifications that softmax scaling improves the separability between ID and OOD data for CLIP-based OOD detection, which is *contrary* to models trained with cross-entropy (CE) loss. In particular, CLIP-like models are trained with a multi-modal contrastive loss, which maximizes the cosine similarity between an image and its textual description in the feature space. The resulting cosine similarity scores display strong *uniformity*[1] across labels, as evidenced in Figure 2 (right). Compared to OOD inputs, the gap between the maximum cosine similarity and the average is larger for ID inputs. However, the gap can be small when the number of ID classes increases where ID samples occur with lower highest cosine similarity. As a result, the highest cosine similarity for ID samples and OOD samples can be highly close (*c.f.* Figure 2 (left)).

Motivated by these observations, MCM employs softmax as a post hoc mechanism to **magnify** the difference. This is *fundamentally different from the softmax score derived from a model trained with cross-entropy loss*, which inherently maximizes the posterior $p(y|\mathbf{x})$ for the ground-truth label, and minimizes the probability for other labels. Unlike CLIP-like models, logit scores displaying uniformity would be heavily penalized by the CE loss. As a result, the maximum logit score corresponding to the ground truth label can already be significantly higher than other labels. Applying softmax on the logit scores can exacerbate overconfident predictions, and reduce the separability between ID and OOD data [39].

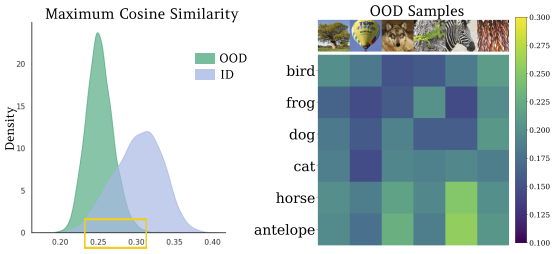

Figure 2: Left: Maximum cosine similarity for ID and OOD inputs. There exists overlapping regions (shown in yellow); Right: Cosine similarities between OOD inputs and ID concept vectors. For OOD inputs, the cosine similarities display uniformity.

Interestingly, for CLIP-like models, the trend is the opposite—applying softmax helps sharpen the uniform-like inner product scores, and increases the separability between ID and OOD data. To help readers better understand the insights, we first formalize our observations that OOD inputs trigger *similar cosine similarities* across ID concepts (Figure 2, right) as the following assumption:

Assumption 3.1. Let $z := \mathbb{1}\{y \in \mathcal{Y}_{\text{in}}\}$ and $Q_{\mathbf{x}}$ denotes the out-of-distribution $\mathbb{P}_{\mathbf{x}|z=0}$ (marginal distribution of $\mathbf{x}$ conditioned on $z = 0$). Assume $\exists\, \delta > 0$ such that

$$Q_{\mathbf{x}}\left(\frac{1}{K-1}\sum_{i\neq\hat{y}}\left[s_{\hat{y}_2}(\mathbf{x}) - s_i(\mathbf{x})\right] < \delta\right) = 1,$$

where $\hat{y} := \text{argmax}_{i\in[K]} s_i(\mathbf{x})$ and $\hat{y}_2 := \text{argmax}_{i\neq\hat{y}, i\in[K]} s_i(\mathbf{x})$ denote the indices of the largest and second largest cosine similarities for an OOD input $\mathbf{x}$.

Now we provide formal guarantees that using softmax can provably reduce the false positive rate (FPR) compared to that without softmax.

Theorem 3.1. Given a task with ID label set $\mathcal{Y}_{\text{in}} = \{y_1, y_2, ..., y_K\}$ and a pre-trained CLIP-like model $(\mathcal{T}, \mathcal{I})$. If $Q_{\mathbf{x}}$ satisfies Assumption 3.1, then there exists a constant $T = \frac{\lambda(K-1)\left(\lambda^{\text{wo}}+\delta-s_{\hat{y}_2}\right)}{K\lambda-1}$ such that for any temperature $\tau > T$, we have

$$\text{FPR}(\tau, \lambda) \leq \text{FPR}^{\text{wo}}(\lambda^{\text{wo}}),$$

where $\text{FPR}(\tau, \lambda)$ is the false positive rate based on softmax scaling *with* temperature τ and detection threshold λ; $\text{FPR}^{\text{wo}}(\lambda^{\text{wo}})$ is the false positive rate *without* softmax scaling based on threshold λ^{wo}. This suggests that applying softmax scaling with a moderate temperature results in superior OOD detection performance compared to that without softmax scaling. The proof is in **Appendix A**. Later in Section 5, we empirically verify on a real-world ImageNet dataset that our bound can indeed be satisfied in CLIP where the thresholds are chosen at 95% true positive rate.

[1]This can be explained both theoretically [71] and empirically [70]. It has been shown that self-supervised contrastive learning with a smaller temperature (*e.g.*, initialized as 0.07 for CLIP) promotes uniform distribution for L2-normalized features. Moreover, as CLIP features lie on a high-dimensional space (512 for CLIP-B/16 and 768 for CLIP-L/14), uniformly distributed points in a high-dimensional sphere tend to be equidistant to each other [68]. Therefore, for OOD inputs, we observe approximately uniform cosine similarity with concept vectors.

What MCM offers: Beyond theoretical insights, we would like to highlight several compelling advantages of our zero-shot OOD detection approach in practice:

- **Generalizable to many tasks**: Traditional OOD detection methods are based on a task-specific model. As a result, the OOD detector is not suitable for a realistic online scenario where the task changes from one to another. In contrast, we will show in Section 4 that MCM can perform a wide variety of OOD detection tasks, with just one single model. For a new task, only the names of the task's visual concepts $\mathcal{Y}_{\text{in}}$ are required.
- **OOD-agnostic**: Our method does not rely on any OOD information, and thus suits many real-world scenarios where one cannot anticipate what the unknowns would be ahead of time. This also mitigates the shortcoming of a recent approach [14], which assumes that a set of unseen labels are given as some weak information about OOD data.
- **Training-free**: MCM enables OOD detection in a completely zero-shot setting. This stands in contrast to the vast majority of OOD detection literature, which often requires training from scratch or fine-tuning to achieve competitive performance.
- **Scalable**: The contrastive vision-language pre-training paradigm makes MCM scalable to a large number of class labels and realistic high-resolution images.

We now proceed to the experimental results, demonstrating these advantages on real-world tasks.

4 A Comprehensive Analysis of MCM

4.1 Datasets and Implementation Details

Datasets. Most previous works on OOD detection only focus on small-scale datasets with blurry images such as CIFAR [33] and TinyImageNet [34]. With pre-trained models such as CLIP, OOD detection can be extended to more realistic and complex datasets. In this work, we scale up evaluations in terms of (1) image resolution, (2) dataset variety, and (3) number of classes. We consider the following ID datasets: CUB-200 [69], STANFORD-CARS [32], FOOD-101 [5], OXFORD-PET [49] and variants of IMAGENET [9]. For OOD test datasets, we use the same ones in [25], including subsets of iNaturalist [65], SUN [73], PLACES [81], and TEXTURE [8]. For each OOD dataset, the categories are not overlapping with the ID dataset. We also use subsets of ImageNet-1k for fine-grained analysis. For example, we construct ImageNet-10 that mimics the class distribution of CIFAR-10 but with high-resolution images. For hard OOD evaluation, we curate ImageNet-20, which consists of 20 classes semantically similar to ImageNet-10 (*e.g.*, dog (ID) vs. wolf (OOD)).

Model. In our experiments, we adopt CLIP [51] as the target pre-trained model, which is one of the most popular and publicly available vision-language pre-trained models. Note that our method is not limited to CLIP; it can generally be applicable for contrastive vision-language pre-training models that promote multi-modal feature alignment. Specifically, we mainly use CLIP-B/16, which consists of a ViT-B/16 Transformer as the image encoder and a masked self-attention Transformer [66] as the text encoder. To indicate input patch size in ViT models, we append "/x" to model names. We prepend -B, -L to indicate `Base` and `Large` versions of the corresponding architecture. For instance, ViT-B/16 implies the `Base` variant with an input patch resolution of 16×16. We also use CLIP-L/14 which is based on ViT-L/14 as a representative of large models. Unless specified otherwise, the temperature τ is 1 for all experiments. *Details of the datasets, the experimental setup and hyper-parameters are provided in Appendix B.*

Metrics. For evaluation, we use the following metrics: (1) the false positive rate (FPR95) of OOD samples when the true positive rate of in-distribution samples is at 95%, (2) the area under the receiver operating characteristic curve (AUROC), and (3) ID classification accuracy (ID ACC).

4.2 Main Results

MCM supports a diverse collection of tasks while being zero-shot. We first show that zero-shot OOD detection with MCM is effective across a wide variety of tasks—with just *one single pre-trained model*. To showcase the versatility of MCM, we consider the seven ID datasets here. To the best of our knowledge, this is among the first attempts to showcase the efficacy under an expansive and

Table 1: Zero-shot OOD detection with MCM score based on CLIP-B/16 with various datasets as ID.

ID Dataset	**OOD Dataset** iNaturalist		SUN		Places		Texture		**Average**	
	FPR95↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑
CUB-200 [69]	8.99	98.37	6.22	98.67	4.58	99.13	6.80	98.78	6.65	98.74
Stanford-Cars [32]	0.04	99.76	0.28	99.89	0.28	99.89	0.02	99.96	0.09	99.89
Food-101 [5]	0.74	99.75	1.93	99.54	1.23	99.72	4.33	98.50	2.06	99.38
Oxford-Pet [49]	2.58	99.41	2.14	99.55	1.06	99.72	0.93	99.80	1.68	99.62
ImageNet-10	0.16	99.79	0.37	99.77	0.97	99.62	0.04	99.90	0.38	99.77
ImageNet-20	1.20	99.64	2.92	99.48	4.72	99.11	2.36	99.02	2.80	99.31
ImageNet-100	15.23	97.30	25.05	95.95	24.91	95.66	33.68	94.11	24.72	95.76

Table 2: OOD detection performance for ImageNet-1k [9] as ID.

Method	**OOD Dataset** iNaturalist		SUN		Places		Texture		**Average**	
	FPR95↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑
Requires training (or w. fine-tuning)										
MOS [25] (BiT)	**9.28**	**98.15**	40.63	92.01	49.54	89.06	60.43	81.23	39.97	90.11
Fort et al. [14] (ViT-B)	15.07	96.64	54.12	86.37	57.99	85.24	53.32	84.77	45.12	88.25
Fort et al. [14] (ViT-L)	15.74	96.51	52.34	87.32	55.14	86.48	**51.38**	85.54	43.65	88.96
MSP [19] (CLIP-B)	40.89	88.63	65.81	81.24	67.90	80.14	64.96	78.16	59.89	82.04
MSP [19] (CLIP-L)	34.54	92.62	61.18	83.68	59.86	84.10	59.27	82.31	53.71	85.68
Zero-shot (no training required)										
MCM (CLIP-B)	32.08	94.41	39.21	92.28	44.88	89.83	58.05	**85.96**	43.55	90.62
MCM (CLIP-L)	26.65	95.25	**29.80**	**94.08**	**33.96**	**92.38**	58.35	85.42	**37.19**	**91.78**

diverse collection of ID datasets. The zero-shot OOD detection performance is summarized in Table 1. A salient observation is that MCM can achieve superior detection performance on many tasks. For example, using STANFORD-CARS as ID, MCM yields an average FPR95 of **0.09**%. Considering that there are no training samples or OOD information involved, these results are very encouraging.

It can be also seen from Table 1 that MCM is promising, especially when the number of samples per ID class is limited in the training set. For example, there are only around 40 samples per class for Stanford-Cars, 100 for Oxford-Pet, and 30 for CUB-200. The sample scarcity makes OOD detection methods that rely on fine-tuning difficult. For example, after fine-tuning on Food-101, while the ID accuracy is increased from 86.3% to $92.5\% \uparrow$, OOD detection based on MSP is on par with MCM (99.5% vs. 99.4% in AUROC).

MCM scales effectively to large datasets. To examine the scalability of MCM, we compare it with recent competitive OOD detection methods [25, 14] on the ImageNet-1k dataset (ID) in Table 2. We observe the following trends:

- Larger model leads to superior performance. Compared with CLIP-B, MCM based on CLIP-L reduces FPR95 by 6.36%. Zero-shot ID classification accuracy is also improved by 7% with the larger model, reaching 75.53% (see Appendix D). This suggests that larger models are endowed with a better representation quality which benefits both ID classification and OOD detection with MCM. Our finding echos with the recent observations [67] that higher ID classification accuracy is correlated with stronger OOD detection performance.
- MOS [25] recently demonstrated competitive performance on ImageNet-1k, which requires model fine-tuning based on BiT [31]. In contrast, we show that MCM (CLIP-L) outperforms MOS by 1.67% in AUROC while being zero-shot (training-free).
- MCM shares a softmax scaling function with the classic (visual) confidence-based score MSP [19]. To implement MSP, we adopt the commonly used linear probe approach by fine-tuning a linear layer on frozen visual features of CLIP. After fine-tuning, ID accuracy significantly improves, reaching 84.12% (CLIP-L). Interestingly, the OOD detection performance of MSP is worse than MCM by 16.52% in FPR95. We investigate more in Section 5.
- Recently, Fort *et al.* [14] explore small-scale OOD detection by fine-tuning the full ViT model. When extended to large-scale tasks, we find that MCM still yields superior performance under the same image encoder configuration (ViT-B or ViT-L). This further highlights the advantage of utilizing vision-language joint embeddings for large-scale visual OOD detection.

Table 3: Performance comparison on **hard OOD detection** tasks. MCM is competitive on all three hard OOD tasks without training involved. MSP (based on fine-tuned CLIP) does not further improve performance.

Method	ID OOD	ImageNet-10 ImageNet-20	ImageNet-20 ImageNet-10	Waterbirds Spurious OOD
		FPR95 / AUROC	FPR95 / AUROC	FPR95 / AUROC
MSP [19] (fine-tuning)		9.38 / 98.31	12.51 / 97.70	39.57 / 90.99
Mahalanobis [35] (visual only)		62.50 / 86.95	51.99 / 88.57	2.97 / 98.56
MCM (zero-shot)		6.00 / 98.60	13.04 / 98.09	5.85 / 98.41

MCM benefits hard OOD detection. Going beyond, we investigate whether MCM is still effective for hard OOD inputs. We consider the following two categories of hard OOD:

- **Semantically hard OOD**: OOD samples that are semantically similar to ID samples are particularly challenging for OOD detection algorithms [72]. To evaluate hard OOD detection tasks in realistic settings, here we consider ImageNet-10 (ID) vs. ImageNet-20 (OOD) and vice versa. The pair consists of high-resolution images with semantically similar categories such as dog versus wolf. As shown in Table 3, MCM outperforms Mahalanobis [35] by **56.50%** in FPR95 for ImageNet-10 (ID) vs. ImageNet-20 (OOD) and **38.95%** vice versa.
- **Spurious OOD**: Modern neural networks can exploit spurious correlations for predictions [3]. For example, in the Waterbirds dataset [55], there are spurious correlations between the habitat (*e.g.*, water) and bird types. A recent work [42] proposes a new type of hard OOD named spurious OOD and shows that most OOD detection methods perform much worse for spurious OOD inputs compared to non-spurious inputs. The spurious OOD inputs are created to share the same background (*i.e.*, water) as ID data but have different object labels (*e.g.*, a boat rather than a bird). See Appendix C for illustrations. The results are shown in Table 3. It has been shown that CLIP representations are robust to distributional shifts [51]. Therefore, while prior works [42] show that spurious OOD inputs are challenging for methods based on ResNet [18], MCM and Mahalanobis scores based on pre-trained CLIP perform much better. On the other hand, fine-tuning exposes the model to the training set containing spurious correlation. As a result, MSP performs much worse than MCM (39.57% vs. 5.85% in FPR95).

MCM outperforms CLIP-based baselines. Two recent works also use CLIP embeddings for OOD detection [14, 12]. However, fundamental limitations exist for both works. Fort *et al.* [14] assume that a candidate OOD label set $\mathcal{Y}_C$ is known, and used $\sum_{y \in \mathcal{Y}_C} \hat{p}(y|\mathbf{x})$ for OOD detection. Here the predictive probability $\hat{p}(y|\mathbf{x})$ is obtained by normalizing the inner products over $|\mathcal{Y}_{\text{in}}| + |\mathcal{Y}_C|$ classes. While applying softmax converts any vector to probabilities, as we show in Section 3, the converted probabilities do not necessarily correspond to $\mathbb{P}(\text{OOD}|\mathbf{x})$. Moreover, obtaining such an OOD label set is typically not feasible, which fundamentally limits the applicability. A latest work [12] realizes this idea by training an extra text decoder on top of CLIP's image encoder to generate candidate labels. However, [12] cannot guarantee the generated labels are non-overlapping with ID labels.

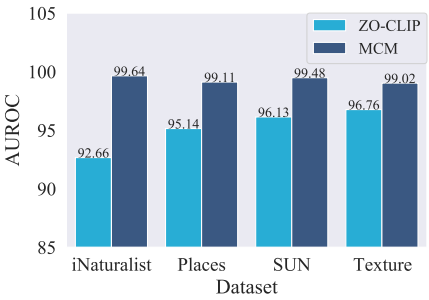

Figure 3: Comparison with candidate label-based score on ImageNet-20. ZO-CLIP scores are our we re-implement of [12]. Implementation details are deferred to Appendix E.1.

We enhance the baseline with a stronger decoder and a filter module (see Appendix E.1). As shown in Figure 3, MCM outperforms the enhanced baseline on all OOD datasets. Moreover, MCM is much simpler to use—alleviating the need for an OOD label set or training an additional caption generator. In contrast, the caption generator's performance largely affects the OOD detection performance. Poor caption quality degenerates the OOD detection performance of candidate label-based methods. Moreover, obtaining a reliable caption generator for *any input image* can significantly increase the computational overhead.

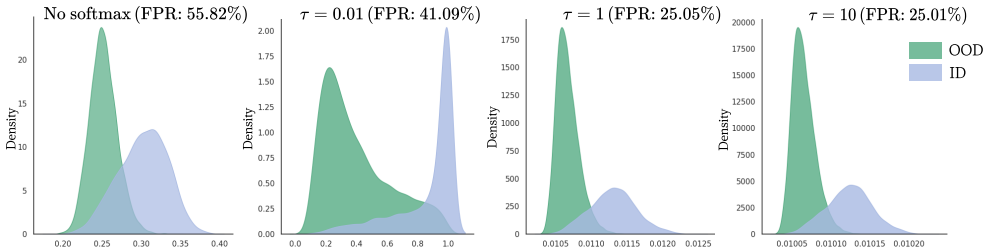

Figure 4: The influence of softmax scaling and temperature. We use ImgeNet-100 (ID) vs. SUN (OOD). Softmax scaling with a moderate temperature significantly improves FPR95.

5 Discussion: A Closer Look at MCM

Empirical verification on the role of softmax. In Section 3, we prove that softmax scaling on cosine similarity scores with a moderate τ improves the ID-OOD separability. Here we empirically verify our theoretical results. As shown in Figure 4, compared to directly using the maximum cosine similarity without softmax (leftmost figure), softmax scaling with a temperature $\tau = 1$ significantly improves the performance by 30.8% in FPR95, and further increasing τ (*e.g.*, $\tau = 10$) leads to similar performance. The results are based on ImageNet-100 (ID) versus SUN (OOD).

Now, we verify if our theoretical bound (*c.f.* Theorem 3.1) is satisfied empirically as well. From the leftmost of Figure 4, we can estimate $\lambda^{\text{wo}} \approx 0.26$, $\delta \approx 0.03$, and $s_{\hat{y}_2} \approx 0.23$. By checking the third figure ($\tau = 1$ is the temperature value we use for most experiments), we approximate $\lambda \approx 0.011$. As $K = 100$, we plug in the values and obtain the lower bound $T = \frac{\lambda(K-1)\left(\lambda^{\text{wo}}+\delta-s_{\hat{y}_2}\right)}{K\lambda-1} \approx 0.65$. Since $\tau = 1 > 0.65$, by Theorem 3.1, applying softmax scaling with $\tau = 1$ is provably superior than without softmax scaling for OOD detection.

Are vision-language features better than visual feature alone? MCM can be interpreted as a distance-based approach—images that are closer to one of the K class prototypes are more likely to be ID and vice versa. Here the class prototypes are defined based on a textual encoder. Alternatively, one could define the class prototype based on visual features. For example, Mahalanobis [35] defines the class prototype as the average of visual embeddings for images belonging to the same class. This raises the question of whether MCM (with *multi-modal* vision-language features) is better than Mahalanobis (with *single-modal* visual feature). For a fair comparison, we use the same ViT image encoder from CLIP-B. Both MCM and Mahalanobis extract visual features from the penultimate layer. On ImageNet-1k, Mahalanobis displays a limited performance, with 73.87% AUROC averaged across four OOD test datasets (**90.62%** for MCM), as shown in Figure 5. From a practical perspective, Mahalanobis requires computing the inverse covariance matrix, which can be both computationally expensive and inaccurate when the number of samples is scarce and the number of ID classes grows. In contrast, MCM is much easier to use and more robust.

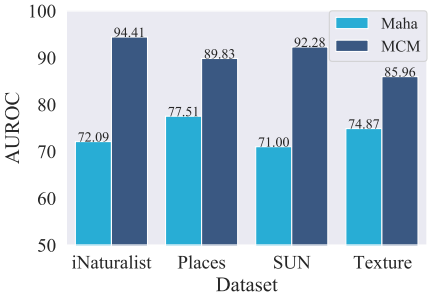

Figure 5: Comparison with Mahalanobis (Maha) score on ImageNet-1k.

Effect of prompt ensembling. We examine MCM's performance with prompt ensembling. For example, Radford *et al.* [51] create 80 possible prompts according to the image modalities and nuances in ImageNet. We experiment with the two prompt sets, one of size 80 as in [51], and our own set of 5 prompts. Ensembles are obtained by averaging the textual features. As expected, using ensembles increases the ID classification accuracy on ImageNet-1k (2% with CLIP-B and 3% with CLIP-L). For OOD detection, the average FPR95 is reduced from 37.19% with the default prompt to 35.17%↓ with an ensemble of five prompts shown in Table 4. In addition, the detection performance with 5 prompts is slightly better than with 80 prompts. Note that

A photo of a <label>.
A blurry photo of a <label>.
A photo of many <label>.
A photo of of the large <label>.
A photo of of the small <label>.

Table 4: The five prompt templates.

prompt ensembling does not increase the inference-time cost, as the textual embeddings (across many prompts) can be pre-calculated and averaged into a single embedding.

6 Related Works

OOD detection in computer vision. For open-world multi-class classification, the goal of OOD detection is to derive a binary ID-OOD classifier along with a multi-class classification model for visual inputs. A plethora of methods has been proposed for deep neural networks, including generative model-based methods [15, 46, 48, 29, 45, 53, 58, 75], and discriminative-model based methods. For the latter category, an OOD score can be derived based on the softmax output or logit layer [4, 10, 19, 22, 25, 39, 41, 60], gradient information [24], or the feature embeddings [35, 56, 57, 61, 72] of a classifier. Morteza *et al.* [44] recently provided theoretical guarantees for existing OOD detection scores. Works insofar have mostly focused on OOD detection for a task-specific model using only the visual information. In contrast, we explore a novel paradigm of zero-shot OOD detection that incorporates rich textual information and can perform a wide variety of tasks.

OOD detection in natural language processing. Distribution shifts can occur due to the change of topics and domains, unexpected user utterances, *etc*. Challenging benchmarks [30] and characterization of distributional shifts [1] have been proposed in recent years. Compared to early language models such as ConvNets and LSTMs, pre-trained language models are more robust to distribution shifts and more effective at identifying OOD instances [20, 50, 76]. Various algorithmic solutions are proposed to handle OOD detection, including outlier exposure [23], model ensembling [37], data augmentation [6, 80, 78], contrastive learning [83, 27], and an auxiliary module that incorporates domain labels [59]. Tan *et al.* [62] also explore zero-shot OOD detection for text classification tasks. However, prior works focus on pure NLP settings while we explore utilizing textual embeddings for zero-shot *visual* OOD detection.

Vision-language models. Utilizing large-scale pre-trained vision-language models for multimodal downstream tasks has become an emerging paradigm with remarkable performance [63, 17]. In general, two types of architectures exist: single-stream models like VisualBERT [36] and ViLT [28] feed the concatenated text and visual features into a single transformer-based encoder; dual-stream models such as CLIP [51], ALIGN [26], and FILIP [77] use separate encoders for text and image and optimize with contrastive objectives to align semantically similar features in different modality. In particular, CLIP enjoys popularity due to its simplicity and strong performance. CLIP-like models inspire numerous follow-up works [79, 38, 82], which aim to improve data efficiency and better adaptation to downstream tasks. This paper adopts CLIP as the target pre-trained model, but our approach can be generally applicable for contrastive models that promote vision-language alignment.

Multi-modal OOD detection. Exploring textual information for visual OOD detection is a new area with limited existing works. Fort *et al.* [14] propose to feed the potential OOD labels to the textual encoder of CLIP [51]. Recently, Esmaeilpour *et al.* [12] propose to train a label generator based on the visual encoder of CLIP and use the generated labels for OOD detection. While both works rely on a set of candidate OOD labels, MCM is OOD-agnostic and alleviates the need for any prior information about OOD. Moreover, prior works [51, 12] only focus on small-scale inputs. We largely expand the scope to a wide range of large-scale realistic datasets, and show new theoretical insights.

7 Conclusion

In this work, we delve into a new landscape for OOD detection, departing from the classic single-modal toward a multi-modal regime. By viewing the textual features as the "concept prototypes", we explore a new OOD detection approach MCM, based on the joint vision-language representations. Unlike the majority of OOD detection methods, MCM offers several compelling advantages: training-free, generalizable to many tasks, scalable to hundreds of classes, and does not require any prior information on OOD inputs. Moreover, we provide theoretical guarantees on how softmax scaling provably improves zero-shot OOD detection. We investigate the effectiveness of MCM on a wide range of large-scale realistic tasks, including several types of hard OOD datasets. Lastly, we demonstrate the advantage of vision-language features over pure visual features for OOD detection. We hope our work will inspire future research toward multi-modal OOD detection.

References

[1] Udit Arora, William Huang, and He He. Types of out-of-distribution texts and how to detect them. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *The European Conference on Computer Vision (ECCV)*, 2018.

[4] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2016.

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *The European Conference on Computer Vision (ECCV)*, 2014.

[6] Derek Chen and Zhou Yu. Gold: improving out-of-scope detection in dialogues using data augmentation. *arXiv preprint arXiv:2109.03079*, 2021.

[7] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2021.

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2014.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2009.

[10] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[12] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot open set detection by extending clip. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[13] Christiane Fellbaum. Wordnet. In *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer, 2010.

[14] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[15] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017.

[16] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

[17] Jiuxiang Gu, Jason Kuen, Shafiq Joty, Jianfei Cai, Vlad Morariu, Handong Zhao, and Tong Sun. Self-supervised relationship probing. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2016.

[19] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[20] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Association for Computational Linguistics (ACL)*, 2020.

[21] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2018.

[22] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[23] Yibo Hu and Latifur Khan. Uncertainty-aware reliable text classification. In *SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2021.

[24] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[25] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2021.

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[27] Di Jin, Shuyang Gao, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. Towards textual out-of-domain detection without in-domain labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[29] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[30] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

[31] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *The European Conference on Computer Vision (ECCV)*, 2020.

[32] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[34] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[35] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[36] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[37] Xiaoya Li, Jiwei Li, Xiaofei Sun, Chun Fan, Tianwei Zhang, Fei Wu, Yuxian Meng, and Jun Zhang. kfolden: k-fold ensemble for out-of-distribution detection. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[38] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations (ICLR)*, 2022.

[39] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *The European Conference on Computer Vision (ECCV)*, 2014.

[41] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[42] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. *The AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[43] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[44] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. *The AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[45] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations (ICLR)*, 2019.

[46] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *The European Conference on Computer Vision (ECCV)*, 2018.

[47] Edwin G. Ng, Bo Pang, Piyush Sharma, and Radu Soricut. Understanding guided image captioning performance across domains. *arXiv preprint arXiv:2012.02339*, 2020.

[48] Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[49] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2012.

[50] Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[52] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[53] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[55] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2019.

[56] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In *International Conference on Machine Learning (ICML)*, 2020.

[57] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations (ICLR)*, 2021.

[58] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations (ICLR)*, 2020.

[59] Yilin Shen, Yen-Chang Hsu, Avik Ray, and Hongxia Jin. Enhancing the generalization for intent classification and out-of-domain detection in slu. *arXiv preprint arXiv:2106.14464*, 2021.

[60] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[61] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[62] Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. Out-of-domain detection for low-resource text classification tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[63] Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171, 2022.

[64] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.

[65] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Conference on Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[67] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations (ICLR)*, 2021.

[68] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[69] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[70] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2021.

[71] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

[72] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[73] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2010.

[74] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations (ICLR)*, 2021.

[75] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[76] Keyang Xu, Tongzheng Ren, Shikun Zhang, Yihao Feng, and Caiming Xiong. Unsupervised out-of-domain detection via pre-trained transformers. In *Association for Computational Linguistics (ACL)*, 2021.

[77] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *International Conference on Learning Representations (ICLR)*, 2021.

[78] Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Lam. Out-of-scope intent detection with self-supervision and discriminative training. *Association for Computational Linguistics (ACL)*, 2021.

[79] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[80] Yinhe Zheng, Guanyi Chen, and Minlie Huang. Out-of-domain detection for natural language understanding in dialog systems. *TASLP*, 2020.

[81] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

[82] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2022.

[83] Wenxuan Zhou, Fangyu Liu, and Muhao Chen. Contrastive out-of-distribution detection for pretrained transformers. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[84] Zhuotun Zhu, Lingxi Xie, and Alan Yuille. Object recognition with and without objects. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [N/A]

(c) Did you discuss any potential negative societal impacts of your work? [N/A]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3 and Section A

(b) Did you include complete proofs of all theoretical results? [Yes] See Section A

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section B

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4 and Section B

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] There is no training involved in MCM. For fair comparison and reproducibility, we use the publicly available checkpoints of CLIP from OpenAI. The test OODs do not have randomness either.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section B

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See Section B

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section B

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Theoretical Justification: Softmax Scaling for Zero-Shot OOD Detection

In this section, we provide the proof for Theorem 3.1 in Section 3, which states the benefits of applying softmax scaling to inner products for OOD detection. We begin with a review of notations.

Notations. We denote the text encoder of a pre-trained CLIP-like model as $\mathcal{T}: t \to \mathbb{R}^d$ and the image encoder $\mathcal{I}: \mathbf{x} \to \mathbb{R}^d$. For a given task with label set $\mathcal{Y}_{\text{in}} = \{y_1, y_2, ..., y_K\}$, we construct a collection of concept vectors $\mathcal{T}(t_i)$. For a given input $\mathbf{x}'$, we denote the cosine similarity *w.r.t* concept vectors as $s_i(\mathbf{x}') = \frac{\mathcal{I}(\mathbf{x}')\cdot\mathcal{T}(t_i)}{\|\mathcal{I}(\mathbf{x}')\|\cdot\|\mathcal{T}(t_i)\|}\,\forall i \in [K]$, where $|s_i(\mathbf{x}')| \leq B$ for all $\mathbf{x}' \in \mathcal{X}$.[2]. We define the maximum concept matching (MCM) score as: $S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = \max_{i\in[K]} \frac{e^{s_i(\mathbf{x}')/\tau}}{\sum_{j=1}^{K} e^{s_j(\mathbf{x}')/\tau}}$. We denote the maximum inner product without applying softmax scaling as $S^{\text{wo}}_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = \max_{i\in[K]} s_i(\mathbf{x}')$. By convention, the OOD detection functions are given by:

$$G^{\text{wo}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = \begin{cases} 1 & S^{\text{wo}}_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) \geq \lambda^{\text{wo}} \\ 0 & S^{\text{wo}}_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) < \lambda^{\text{wo}} \end{cases},$$

$$G(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = \begin{cases} 1 & S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) \geq \lambda \\ 0 & S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) < \lambda \end{cases},$$

Remarks: By convention, 1 represents the positive class (ID) and 0 indicates OOD; λ and λ^{wo} are typically chosen such that the true positive rate is at 95%.

For convenience, we paste the assumptions and the theorem in Section 3 below,

Assumption A.1. Let $z := \mathbb{1}\{y \in \mathcal{Y}_{\text{in}}\}$ and $Q_{\mathbf{x}}$ denotes the out-of-distribution $\mathbb{P}_{\mathbf{x}|z=0}$ (marginal distribution of $\mathbf{x}$ conditioned on $z = 0$). Assume $\exists\, \delta > 0$ such that

$$Q_{\mathbf{x}}\left(\frac{1}{K-1}\sum_{i\neq\hat{y}}\left[s_{\hat{y}_2}(\mathbf{x}) - s_i(\mathbf{x})\right] < \delta\right) = 1,$$

where $\hat{y} := \text{argmax}_{i\in[K]} s_i(\mathbf{x})$ and $\hat{y}_2 := \text{argmax}_{i\neq\hat{y}, i\in[K]} s_i(\mathbf{x})$ denote the indices of the largest and second largest cosine similarities for an OOD input $\mathbf{x}$.

Theorem A.1. Given a pre-trained CLIP-like model $(\mathcal{T}, \mathcal{I})$ and a task with label set $\mathcal{Y}_{\text{in}} = \{y_1, y_2, ..., y_K\}$. If $Q_{\mathbf{x}}$ satisfy Assumption A.1, Then there exists a constant $T = \frac{\lambda(K-1)\left(\lambda^{\text{wo}} + \delta - s_{\hat{y}_2}\right)}{K\lambda - 1}$ such that for any temperature $\tau > T$, we have:

$$\text{FPR}(\tau, \lambda) \leq \text{FPR}^{\text{wo}}(\lambda^{\text{wo}}),$$

where $\text{FPR}(\tau, \lambda)$ is the false positive rate based on softmax scaling with temperature τ and threshold λ; $\text{FPR}^{\text{wo}}(\lambda^{\text{wo}})$ is the false positive rate without softmax scaling based on threshold λ^{wo}. This suggests that applying softmax scaling with temperature results in superior OOD detection performance compared to without softmax scaling.

Proof. By definition, we express the false positive rate $\text{FPR}(\tau, \lambda)$ as follows,

$$\begin{aligned}
\text{FPR}(\tau, \lambda) &= \mathbb{P}\left(G(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = 1 \mid z = 0\right) \\
&= Q_{\mathbf{x}'}\left(G(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = 1\right) \\
&= Q_{\mathbf{x}'}\left(p_{\hat{y}}\left(\mathbf{x}'; \tau\right) > \lambda\right) \\
&= Q_{\mathbf{x}'}\left(\frac{e^{s_{\hat{y}}(\mathbf{x}')/\tau}}{\sum_{j=1}^{K} e^{s_j(\mathbf{x}')/\tau}} > \lambda\right) \\
&= Q_{\mathbf{x}'}\left(\frac{1}{\lambda} > \sum_{i=1}^{K} \exp\left(\frac{s_i(\mathbf{x}') - s_{\hat{y}}(\mathbf{x}')}{\tau}\right)\right)
\end{aligned}$$

[2] In practice, we observe that $s_i \in [0.1, 0.3]$ for CLIP with high probability.

By inequality $e^x \geq 1 + x$, we have,

$$Q_{\mathbf{x}'}\left(\frac{1}{\lambda} > \sum_{i=1}^{K} \exp\left(\frac{s_i(\mathbf{x}') - s_{\hat{y}}(\mathbf{x}')}{\tau}\right)\right) \leq Q_{\mathbf{x}'}\left(\frac{1}{\lambda} > \sum_{i=1}^{K}\left[1 + \frac{s_i(\mathbf{x}') - s_{\hat{y}}(\mathbf{x}')}{\tau}\right]\right)$$

This indicates

$$\begin{aligned}
&Q_{\mathbf{x}'}\left(\frac{1}{\lambda} > \sum_{i=1}^{K} \exp\left(\frac{s_i(\mathbf{x}') - s_{\hat{y}}(\mathbf{x}')}{\tau}\right)\right) \leq Q_{\mathbf{x}'}\left(\frac{1}{\lambda} > \sum_{i=1}^{K}\left[1 + \frac{s_i(\mathbf{x}') - s_{\hat{y}}(\mathbf{x}')}{\tau}\right]\right) \\
&= Q_{\mathbf{x}'}\left(\sum_{i=1}^{K}\left(s_{\hat{y}}(\mathbf{x}') - s_i(\mathbf{x}')\right) > \left(K - \frac{1}{\lambda}\right)\tau\right)
\end{aligned}$$

Since

$$\begin{aligned}
\sum_{i=1}^{K}\left(s_{\hat{y}}(\mathbf{x}') - s_i(\mathbf{x}')\right) &= \sum_{i \neq \hat{y}}\left(s_{\hat{y}}(\mathbf{x}') - s_{\hat{y}_2}(\mathbf{x}') + s_{\hat{y}_2}(\mathbf{x}') - s_i(\mathbf{x}')\right) \\
&= \sum_{i \neq \hat{y}}\left(s_{\hat{y}}(\mathbf{x}') - s_{\hat{y}_2}(\mathbf{x}')\right) + \sum_{i \neq \hat{y}}\left(s_{\hat{y}_2}(\mathbf{x}') - s_i(\mathbf{x}')\right) \\
&= (K-1)\left(s_{\hat{y}}(\mathbf{x}') - s_{\hat{y}_2}(\mathbf{x}')\right) + \sum_{i \neq \hat{y}}\left(s_{\hat{y}_2}(\mathbf{x}') - s_i(\mathbf{x}')\right)
\end{aligned}$$

By Assumption 3.1, we have

$$Q_{\mathbf{x}'}\left(\sum_{i=1}^{K}\left(s_{\hat{y}}(\mathbf{x}') - s_i(\mathbf{x}')\right) < (K-1)\left(s_{\hat{y}}(\mathbf{x}') - s_{\hat{y}_2}(\mathbf{x}')\right) + (K-1)\delta\right) = 1.$$

Therefore,

$$\begin{aligned}
Q_{\mathbf{x}'}\left(\sum_{i=1}^{K}\left(s_{\hat{y}}(\mathbf{x}') - s_i(\mathbf{x}')\right) > \left(K - \frac{1}{\lambda}\right)\tau\right) &\leq Q_{\mathbf{x}'}\left(s_{\hat{y}}(\mathbf{x}') - s_{\hat{y}_2}(\mathbf{x}') > -\delta_2 + \left(K - \frac{1}{\lambda}\right)\frac{\tau}{K-1}\right) \\
&= Q_{\mathbf{x}'}\left(s_{\hat{y}}(\mathbf{x}') - s_{\hat{y}_2}(\mathbf{x}') > -\delta_2 + \lambda'\right) \\
&= Q_{\mathbf{x}'}\left(s_{\hat{y}}(\mathbf{x}') > s_{\hat{y}_2}(\mathbf{x}') - \delta_2 + \lambda'\right)
\end{aligned}$$

where $\lambda' = \left(K - \frac{1}{\lambda}\right)\frac{\tau}{K-1}$ is a monotonic function of λ (*i.e.*, minimizing false positive rate w.r.t. λ is equivalent to minimizing *w.r.t* λ'.)

For $\tau > 0$, we can rewrite the MCM score as

$$\begin{aligned}
S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) &= \max_{i \in [K]} \frac{e^{s_i(\mathbf{x}')/\tau}}{\sum_{j=1}^{K} e^{s_j(\mathbf{x}')/\tau}} = \frac{e^{s_{\hat{y}}(\mathbf{x}')/\tau}}{\sum_{j=1}^{K} e^{s_j(\mathbf{x}')/\tau}} \\
&= \frac{1}{1 + \sum_{j=1, j \neq \hat{y}}^{K} e^{(s_j(\mathbf{x}') - s_{\hat{y}}(\mathbf{x}'))/\tau}}
\end{aligned}$$

As $\hat{y} := \text{argmax}_{i \in [K]} s_i(\mathbf{x})$, $s_j(\mathbf{x}') - s_{\hat{y}}(\mathbf{x}') \leq 0$, $S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I})$ is a monotonically decreasing function of τ, we have:

$$S_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) > \lim_{\tau \to \infty} \frac{1}{1 + \sum_{j=1, j \neq \hat{y}}^{K} e^{(s_j(\mathbf{x}') - s_{\hat{y}}(\mathbf{x}'))/\tau}} = \frac{1}{K}$$

Therefore by the definition of λ, we have $\lambda > \frac{1}{K}$, $\lambda' = \left(K - \frac{1}{\lambda}\right)\frac{\tau}{K-1} > 0$

For moderately large $\tau > T$ where $T = \frac{\lambda(K-1)\left(\lambda^{\text{wo}} + \delta - s_{\hat{y}_2}\right)}{K\lambda - 1}$, we always have $s_{\hat{y}_2}(\mathbf{x}') - \delta + \lambda' > \lambda^{\text{wo}}$. Therefore, we obtain the following inequality,

$$\text{FPR}(\tau, \lambda) \leq Q_{\mathbf{x}'}\left(s_{\hat{y}}(\mathbf{x}') > s_{\hat{y}_2}(\mathbf{x}') - \delta_2 + \lambda'\right) \leq Q_{\mathbf{x}'}\left(s_{\hat{y}}(\mathbf{x}') > \lambda^{\text{wo}}\right) := \text{FPR}^{\text{wo}}(\lambda^{\text{wo}})$$

which means that the FPR without softmax scaling is larger than that with softmax scaling and a moderately large temperature. We show in Section 5 that the bound is indeed satisfied in practice with a large-scale ID dataset. □

B Experimental Details

B.1 Software and Hardware

All methods are implemented in Pytorch 1.10. We run all OOD detection experiments on NVIDIA GeForce RTX-2080Ti GPU and use NVIDIA A100 GPU for fine-tuning CLIP and ViT.

B.2 Hyperparameters

The only hyperparameter in MCM is the (temperature) scaling factor τ. We use $\tau = 1$ by default unless otherwise specified. Our experiments suggest that MCM is insensitive to the scaling factor, where τ in a wide range of $[0.5, 100]$ shares similar performance.

B.3 Datasets

ImageNet-10 We create ImageNet-10 that mimics the class distribution of CIFAR-10 but with high-resolution images. It contains the following categories (with class ID): warplane (n04552348), sports car (n04285008), brambling bird, (n01530575), Siamese cat (n02123597), antelope (n02422699), Swiss mountain dog (n02107574), bull frog (n01641577), garbage truck (n03417042), horse (n02389026), container ship (n03095699).

ImageNet-20 For hard OOD evaluation with realistic datasets, we curate ImageNet-20, which consists of 20 classes semantically similar to ImageNet-10 (*e.g.*, dog (ID) vs. wolf (OOD)). The categories are selected based on the distance in the WordNet synsets [13]. Specifically, it contains the following categories: sailboat (n04147183), canoe (n02951358), balloon (n02782093), tank (n04389033), missile (n03773504), bullet train (n02917067), starfish (n02317335), spotted salamander (n01632458), common newt (n01630670), zebra (n01631663), frilled lizard (n02391049), green lizard (n01693334), African crocodile (n01697457), Arctic fox (n02120079), timber wolf (n02114367), brown bear (n02132136), moped (n03785016), steam locomotive (n04310018), space shuttle (n04266014), snowmobile (n04252077).

ImageNet-100 We randomly sample 100 classes from ImageNet-1k to curate ImageNet-100. It contain the following categories: n01986214, n04200800, n03680355, n03208938, n02963159, n03874293, n02058221, n04612504, n02841315, n02099712, n02093754, n03649909, n02114712, n03733281, n02319095, n01978455, n04127249, n07614500, n03595614, n04542943, n02391049, n04540053, n03483316, n03146219, n02091134, n02870880, n04479046, n03347037, n02090379, n10148035, n07717556, n04487081, n04192698, n02268853, n02883205, n02002556, n04273569, n02443114, n03544143, n03697007, n04557648, n02510455, n03633091, n02174001, n02077923, n03085013, n03888605, n02279972, n04311174, n01748264, n02837789, n07613480, n02113712, n02137549, n02111129, n01689811, n02099601, n02085620, n03786901, n04476259, n12998815, n04371774, n02814533, n02009229, n02500267, n04592741, n02119789, n02090622, n02132136, n02797295, n01740131, n02951358, n04141975, n02169497, n01774750, n02128757, n02097298, n02085782, n03476684, n03095699, n04326547, n02107142, n02641379, n04081281, n06596364, n03444034, n07745940, n03876231, n09421951, n02672831, n03467068, n01530575, n03388043, n03991062, n02777292, n03710193, n09256479, n02443484, n01728572, n03903868.

Conventional (non-spurious) OOD datasets Huang *et al.* [25] curate a diverse collection of subsets from iNaturalist [65], SUN [73], Places [81], and Texture [8] as large-scale OOD datasets for ImageNet-1k, where the classes of the test sets do not overlap with ImageNet-1k. We provide a brief introduction for each dataset as follows.

iNaturalist contains images of natural world [65]. It has 13 super-categories and 5,089 sub-categories covering plants, insects, birds, mammals, and so on. We use the subset that contains 110 plant classes not overlapping with ImageNet-1k.

SUN stands for the Scene UNderstanding Dataset [73]. SUN contains 899 categories that cover more than indoor, urban, and natural places with or without human beings appearing in. We use the subset which contains 50 natural objects not showing in ImageNet-1k.

Places is a large scene photographs dataset [81]. It contains photos that are labeled with scene semantic categories from three macro-classes: Indoor, Nature, and Urban. The subset we use is sampled from 50 categories that are not present in ImageNet-1k.

Texture stands for the Describable Textures Dataset [8]. It contains images of textures and abstracted patterns. As no categories overlap with ImageNet-1k, we use the entire dataset as in [25].

C Spurious OOD Datasets

In general, spurious attributes refer to statistically informative features that co-exist with the majority of ID samples but do not necessarily capture cues related to the labels such as color, texture, background, etc [3, 2, 84, 16, 74]. A recent work [42] investigated a new type of hard OOD samples (called spurious OOD) that contain spurious or environmental features, but no object features related to the ID classes. A concrete example is shown in Figure 6, where images of birds co-occur frequently with either the land background or water background. Modern neural networks can spuriously rely on the image background (*e.g.*, water or land) for classification instead of learning to recognize the actual object [54]. Ming *et al.* [42] show that spurious OOD samples remain challenging for most common OOD detection methods based on pure vision models such as ResNet [18].

For ID dataset, we use Waterbirds [55], which combines bird photographs from CUB-200 [69] with water or land background images from PLACES [81]. For the spurious OOD dataset, we use the one created in [42] consisting of land and water background from Places [81].

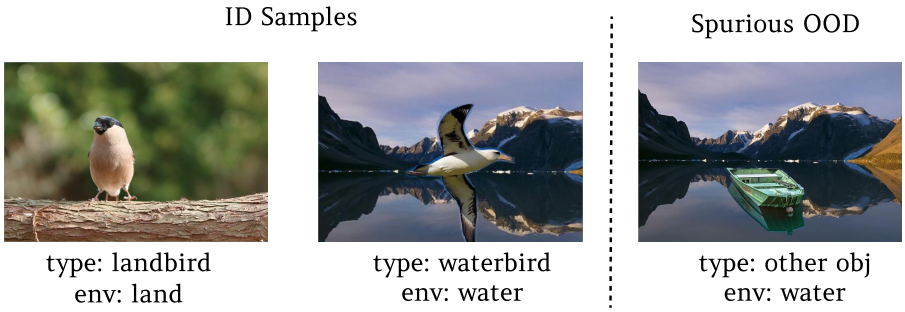

Figure 6: Illustration of spurious OOD samples for Waterbirds [55]. Image sourced from [42].

D ID Classification Accuracy

Table 5 shows the multi-class classification accuracy on ImageNet-1k for methods in Table 2.

Table 5: ID classification accuracy on ImageNet-1k (%)

Method	**ID ACC**
Zero-shot	
MCM (CLIP-B/16)	68.53
MCM (CLIP-L/14)	75.53
w. fine-tuning	
MSP (CLIP-B/16)	79.39
MSP (CLIP-L/14)	84.12
Fort et al. [14] (ViT-B/16)	81.25
Fort et al. [14] (ViT-L/14)	84.05
MOS [25] (BiT)	75.16

E Implementation of CLIP-Based Baselines

E.1 Overview of Baselines

We review two previous works on CLIP-based OOD detection [14, 12] in Figure 7, which derive the scoring function based on candidate OOD labels. For a given task with ID label set $\mathcal{Y}_{\text{in}} =$

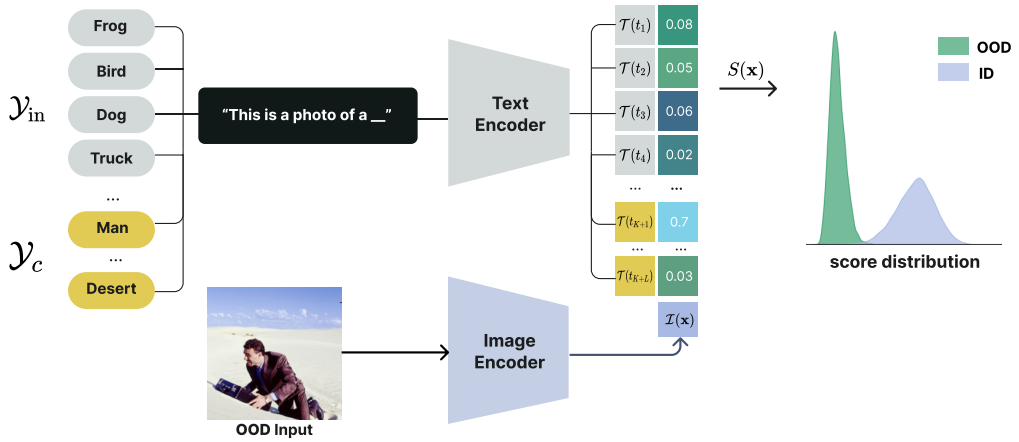

Figure 7: Zero-shot OOD detection with candidate OOD labels. The ID classification task is defined by a set of class labels $\mathcal{Y}_{\text{in}}$. With an additional set of candidate labels $\mathcal{Y}_{\text{C}}$ that describes the contents of the input image, the OOD detection scoring function is calculated by approximating the probability of the input belonging to $\mathcal{Y}_{\text{in}}$.

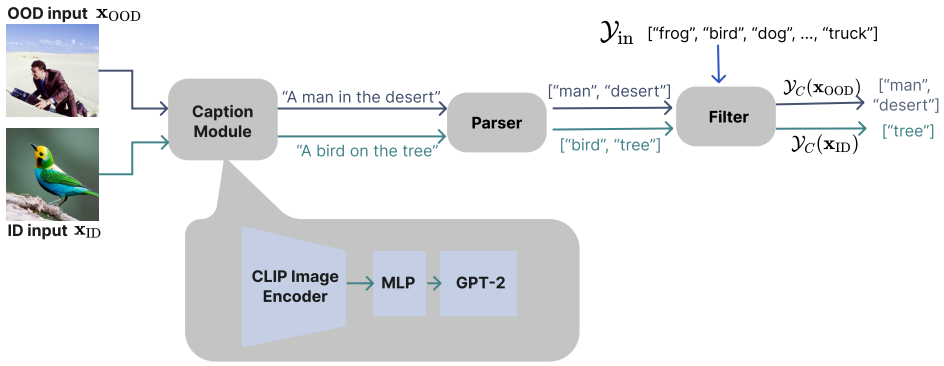

Figure 8: Improved pipeline to generate candidate OOD labels. It consists of three main components: a caption generator, a syntactic parser, and a filtering module to remove candidate labels that overlap with the ID label set.

$\{y_1, y_2, ..., y_K\}$ and candidate labels $\mathcal{Y}_{\text{C}} = \{y_{K+1}, y_{K+2}, ..., y_{K+L}\}$, where ideally $\mathcal{Y}_{\text{in}} \cap \mathcal{Y}_{\text{C}} = \emptyset$, they construct a collection of text embeddings $\mathcal{T}(t_i), i \in \{1, 2, ..., K+L\}$. Here, t_i is the text prompt "`this is a photo of a` $\langle y_i \rangle$" for a label y_i. For any test input image $\mathbf{x}$, we can calculate the label-wise matching score based on the cosine similarity between the image and text features: $s_i(\mathbf{x}) = \frac{\mathcal{I}(\mathbf{x}) \cdot \mathcal{T}(t_i)}{\|\mathcal{I}(\mathbf{x}')\| \cdot \|\mathcal{T}(t_i)\|}$. Therefore, a detection score can be derived as:

$$S(\mathbf{x}; \mathcal{Y}_{\text{in}}, \mathcal{Y}_{\text{C}}, \mathcal{T}, \mathcal{I}) = \sum_{i=1}^{K} \frac{e^{s_i(\mathbf{x})/\tau}}{\sum_{j=1}^{K+L} e^{s_j(\mathbf{x})/\tau}}.$$

where $\tau > 0$ is the temperature scaling hyperparameter.

E.2 Obtaining OOD Candidate Labels

For the baseline methods, obtaining OOD candidate labels is a major challenge and limitation. Recently, [12] propose ZO-CLIP, where a transformer (decoder) based on the image encoder of CLIP is used to generate candidate labels. The transformer is trained from scratch on the COCO dataset [40] with simple teacher forcing algorithms. Although the decoder trained on COCO may work well on CIFAR (ID), it is does not scale up to large-scale datasets such as ImageNet [9] where categories are not covered in COCO. As a result, [12] only test on small-scale datasets with common classes such as CIFAR (ID).

We improve the baseline by using a high-quality caption generator pre-trained on much larger datasets, which not only saves computational overhead but can potentially improve the quality of generated labels. The pipeline involves three components (see Figure 8):

- A caption generator. Given an input image, it generates a caption serving as the textual description of the input. In this work, we consider ClipCap [43], which uses GPT-2 [52] to generate captions based on CLIP's image encoder. ClipCap is pre-trained on a much larger dataset Conceptual Captions [47] compared to COCO, which can be viewed as an enhanced version of the ZO-CLIP baseline [12]. The checkpoints are publicly available[3].
- A syntactic parser. Given a caption, we extract noun objects using a parsing toolkit released by spaCy [4]. Those nouns can be used as candidate labels $\mathcal{Y}_C$ of the input image.
- A filter module. Unlike [12], we further enhance the baseline by adopting a filtering technique to remove overlapping categories in $\mathcal{Y}_C$ with ID labels $\mathcal{Y}_{\text{in}}$, which we detail below.

E.3 Label Filtering

Example. To illustrate the effects of filtering, we begin with a concrete example where ID labels are ["frog","bird"... "truck"], as shown in Figure 8. The generated labels (without filtering) of an ID input of a bird sitting on a tree are ["bird", "tree"]. Therefore, $\mathcal{Y}_{\text{in}} \cup \mathcal{Y}_{\text{C}}$ =["frog","bird"... "truck","bird", "tree"]. Ideally, the softmax probability distribution over the concatenated labels would be $[0, 0.5, 0, \ldots, 0.5, 0]$ and by definition $S(\mathbf{x}) \approx 0.5$. However, if we filter the generated labels to eliminate nouns with similar meaning as ID, our concatenated labels would be ["frog","bird"... "truck","tree"] and the probability vector would be $[0, 1, 0, \ldots, 0]$, which leads to a much higher score $S(\mathbf{x}) = 1$. In contrast, the generated labels for an OOD input with a man in the desert would be ["man", "desert"]. The resulting probability vector would be $[0, 0, 0, \ldots, 1, 0]$ and the score $S(\mathbf{x}) = 0$. Therefore, filtering makes it easier to separate ID inputs from OOD inputs (*c.f.* Figure 9).

String-based filtering. To implement the idea of filtering, we need a measurement of similarity between the generated labels and ID labels. The simplest way is string-based filtering where a generated label is filtered if it matches any ID labels (in the string format), as in the case above.

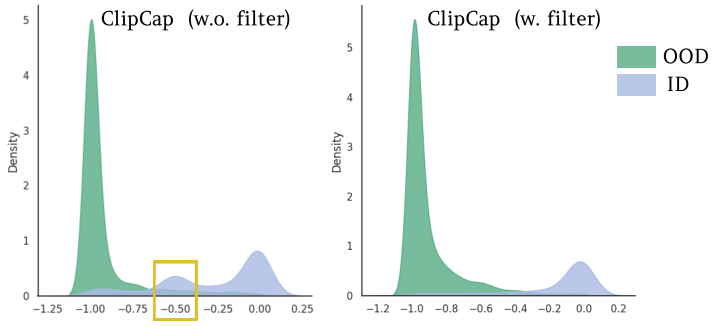

Figure 9: Score distributions for ImageNet-10 (ID) and iNaturalist (OOD) inputs. Simple string-based filtering alleviates the overlap between OOD inputs and ID inputs especially with scores around 0.5 (yellow rectangle), resulting in better ID-OOD separability.

F Alternative Scoring Functions

F.1 Does Energy-Based Score Work Well on Inner Products?

In the visual domain, denote the ID label set $\mathcal{Y}_{\text{in}} = \{1, 2, ..., K\}$ and a discriminative neural classifier $f : \mathbf{x} \mapsto f(\mathbf{x}) \in \mathbb{R}^K$ that maps the input to K logits. After the model is trained with the softmax

[3] https://github.com/rmokady/CLIP_prefix_caption
[4] https://spacy.io/models/en

cross-entropy loss, the logits can be converted to a categorical distribution:

$$\hat{p}(y=j \mid \mathbf{x}) = \frac{e^{f_j(\mathbf{x})/\tau}}{\sum_{i=1}^{K} e^{f_i(\mathbf{x})/\tau}}$$

where $f_j(\mathbf{x})$ is the j-th index of $f(\mathbf{x})$. Given the softmax activations, Liu *et al.* [41] propose to use energy score (*i.e.*, free energy) for OOD detection, defined as:

$$S_{\text{energy}}(\mathbf{x}; \mathcal{Y}_{\text{in}}, f) := -E(\mathbf{x}; \mathcal{Y}_{\text{in}}, f) = \tau \cdot \log \sum_{i}^{K} e^{f_i(\mathbf{x})/\tau}$$

Energy score is shown to better align with the logarithm of the in-distribution density $P_{\mathbf{x}}$ (i.e. $\mathbb{P}_{\mathbf{x}|z=1}$) compared to the classic MSP score and is supported by empirical evidence [41].

As we show in Section 3, although MCM applies the softmax function and converts the cosine similarity between the image and text features to a categorical distribution, it cannot be interpreted as $\hat{p}(y \mid \mathbf{x})$. Next, we empirically verify whether applying the energy function on the inner products $s(\mathbf{x})$ leads to improved OOD detection performance, where the detection score is defined as:

$$S_{\text{energy}}(\mathbf{x}; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) := \tau \cdot \log \sum_{i}^{K} e^{s_i(\mathbf{x})/\tau}$$

The results are shown in Table 6. We use CLIP-B/16 on ImageNet1k (ID). The trend is **contrary** to the observations of pure vision models [41]. With the default temperature $\tau = 1$, the performance of the energy score is significantly worse compared to MCM (AUROC 50.74% vs. 90.62%). With a small temperature ($\tau = 0.005$), the performance of the energy score is largely improved. However, it is still less competitive compared to MCM (AUROC 84.42% vs. 90.62%). This further suggests that while the (softmax) scaling function of MCM is similar to the classic MSP score, they have very different properties.

Table 6: Comparison with the energy score [41] (applied to inner products) on ImageNet1k (ID). We use CLIP-B/16 as the backbone. The results are contrary to the observations for pure vision models where the energy score (applied to logits) demonstrates competitive performance.

Method	**OOD Dataset**								**Average**	
	iNaturalist		SUN		Places		Texture			
	FPR↓	**AUROC**↑	**FPR**↓	**AUROC**↑	**FPR**↓	**AUROC**↑	**FPR**↓	**AUROC**↑	**FPR**↓	**AUROC**↑
Energy ($\tau = 0.005$) [41]	58.66	90.06	66.96	87.39	58.47	87.66	85.87	72.56	67.49	84.42
Energy ($\tau = 1$) [41]	99.33	53.14	96.15	53.51	91.5	61.76	98.15	34.56	96.28	50.74
MCM ($\tau = 1$)	32.08	94.41	39.21	92.28	44.88	89.83	58.05	85.96	43.55	90.62

F.2 Other Scoring Functions

In addition to the energy score, we explore the effectiveness of several alternative scoring functions:

- The (negative) entropy of softmax scaled cosine similarities $S_{\text{entropy}} := -H(s) = \sum_{i=1}^{K} s_i(\mathbf{x}) \log s_i(\mathbf{x})$;
- The variance of the cosine similarities denoted as S_{var};
- The scaled difference between the largest and second largest cosine similarities $S_{\text{scaled}} := e^{s_{\hat{y}}(\mathbf{x}) - s_{\hat{y}_2}(\mathbf{x})}$ where $\hat{y} := \text{argmax}_{i \in [K]} s_i(\mathbf{x})$ and $\hat{y}_2 := \text{argmax}_{i \neq \hat{y}, i \in [K]} s_i(\mathbf{x})$.

Similar to the energy score, for S_{entropy}, we select the hyperparameter $\tau \in \{0.005, 0.01, 0.1, 1, 10\}$ and report the performance with the best τ. As shown in Table 7, MCM still gives the most promising results compared to the other three alternative scores across most OOD test sets.

G MCM for ResNet-based CLIP Models

In the main paper, we demonstrated the effectiveness of the MCM score across a wide range of tasks for CLIP-B and CLIP-L, both of which use ViT as the image encoder. In this section, we investigate

Table 7: Comparison with other scaling functions (applied to inner products) on the large scale benchmark ImageNet-1k (ID). We use CLIP-B/16 as the backbone.

Method	**OOD Dataset**								**Average**	
	iNaturalist		SUN		Places		Texture			
	FPR↓	**AUROC**↑	**FPR**↓	**AUROC**↑	**FPR**↓	**AUROC**↑	**FPR**↓	**AUROC**↑	**FPR**↓	**AUROC**↑
Entropy ($\tau = 0.01$)	44.63	83.76	50.52	81.53	52.75	79.00	64.15	75.94	53.01	80.06
Var	85.45	65.41	70.00	80.32	75.29	76.14	79.79	71.60	77.63	73.37
Scaled	88.90	71.48	90.25	67.86	89.39	69.61	90.25	67.86	89.47	69.87
MCM ($\tau = 1$)	32.08	94.41	39.21	92.28	44.88	89.83	58.05	85.96	43.55	90.62

Table 8: Comparison with ResNet-based CLIP models on ImageNet-1k (ID).

Model	**OOD Dataset**								**Average**	
	iNaturalist		SUN		Places		Texture			
	FPR↓	**AUROC**↑	**FPR**↓	**AUROC**↑	**FPR**↓	**AUROC**↑	**FPR**↓	**AUROC**↑	**FPR**↓	**AUROC**↑
RN50x4	44.51	91.51	35.11	92.84	43.74	89.60	57.73	85.93	45.27	89.97
CLIP-B/16	32.08	94.41	39.21	92.28	44.88	89.83	58.05	85.96	43.55	90.62

the effectiveness of MCM on ResNet-based CLIP. Specifically, we use RN50x4 (178.3M), which shares a similar number of parameters as CLIP-B/16 (149.6M). The results are shown in Table 8.

It can be seen that MCM still shows promising results with ResNet-based CLIP models, and the performance is comparable between RN50x4 and CLIP-B/16 (89.97 vs. 90.62 in AUROC).

H MCM Without Softmax Scaling

In Section 3, we provide theoretical justifications on the necessity of softmax scaling for CLIP-like models and we further verify our theoretical insights based on ImageNet-100 (ID) in Section 5 (Figure 4). To further verify our observations empirically, we show OOD detection performance for a wide range of datasets based on the maximum cosine similarity score $S^{\text{wo}}_{\text{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}, \mathcal{T}, \mathcal{I}) = \max_{i \in [K]} s_i(\mathbf{x}')$. The results are shown in Table 9. For easy tasks such as Food-101 [32], Stanford-Cars [32], and Oxford-Pet [49] as ID, the performance of maximum cosine similarity score is similar to MCM (see Table 1 and Table 2). However, for more challenging tasks such as ImageNet-20 and ImageNet-1k, MCM significantly outperforms that without softmax scaling. For example, the average FPR95 is improved by **11.33%** on ImageNet-20 and **22.02%** on ImageNet-1k, which highlights the necessity of a proper scaling function for CLIP-based OOD detection.

Table 9: Zero-shot OOD detection of $S^{\text{wo}}_{\text{MCM}}$ based on CLIP-B/16.

ID Dataset	**OOD Dataset**								**Average**	
	iNaturalist		SUN		Places		Texture			
	FPR95↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑
Stanford-Cars [32]	0.00	100.00	0.03	99.99	0.29	99.94	0.00	100.0	0.08	99.98
Food-101 [5]	0.65	99.83	0.21	99.92	0.52	99.86	9.31	97.15	2.67	99.19
Oxford-Pet [49]	0.00	99.99	0.04	99.98	0.19	99.95	0.27	99.92	0.12	99.96
ImageNet-10	41.15	95.74	33.31	96.45	30.6	96.6	34.57	96.44	34.91	96.31
ImageNet-20	14.88	97.91	14.74	97.86	13.49	97.84	13.42	97.39	14.13	97.75
ImageNet-1k	56.59	90.27	64.28	87.60	57.23	87.76	84.20	73.50	65.57	84.70