Improving Open Set Recognition via Visual Prompts Distilled from Common-Sense Knowledge

Seongyeop Kim[1], **Hyung-Il Kim**[2], **Yong Man Ro**[1]*

[1]Integrated Vision Language Lab., KAIST, South Korea
[2]ETRI, South Korea
seongyeop@kaist.ac.kr, hikim@etri.re.kr, ymro@kaist.ac.kr

Abstract

Open Set Recognition (OSR) poses significant challenges in distinguishing known from unknown classes. In OSR, the overconfidence problem has become a persistent obstacle, where visual recognition models often misclassify unknown objects as known objects with high confidence. This issue stems from the fact that visual recognition models often lack the integration of common-sense knowledge, a feature that is naturally present in language-based models but lacking in visual recognition systems. In this paper, we propose a novel approach to enhance OSR performance by distilling common-sense knowledge into visual prompts. Utilizing text prompts that embody common-sense knowledge about known classes, the proposed visual prompt is learned by extracting semantic common-sense features and aligning them with image features from visual recognition models. The unique aspect of this work is the training of individual visual prompts for each class to encapsulate this common-sense knowledge. Our methodology is model-agnostic, capable of enhancing OSR across various visual recognition models, and computationally light as it focuses solely on training the visual prompts. This research introduces a method for addressing OSR, aiming at a more systematic integration of visual recognition systems with common-sense knowledge. The obtained results indicate an enhancement in recognition accuracy, suggesting the applicability of this approach in practical settings.

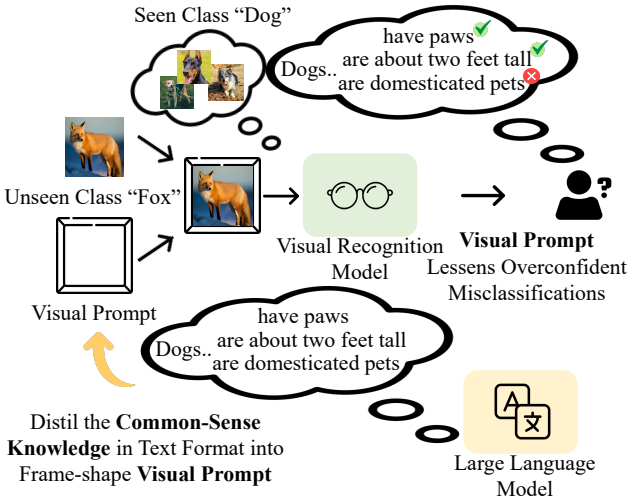

Figure 1: Depiction of the proposed framework, where a large language model distills common-sense knowledge into a visual prompt. This process enables the visual recognition model to adapt and respond effectively when confronted with unfamiliar instances, thereby enhancing its performance in OSR tasks.

Introduction

Open Set Recognition (OSR) is a critical domain in computer vision that focuses on the ability of a system to classify known objects while also recognizing and rejecting unknown ones. This OSR problem reflects real-world scenarios, where the models are expected to encounter objects not belonging to the known classes in the training dataset. Despite the advances in deep learning techniques for closed-set recognition, OSR remains a challenging problem, particularly due to an overconfidence issue in classifying unknown objects (Zhou, Ye, and Zhan 2021).

In traditional OSR methods, visual recognition models often misclassify unknown objects with excessive certainty, so-called overconfidence. (Nguyen, Yosinski, and Clune 2015), (Hein, Andriushchenko, and Bitterwolf 2019), (Padhy et al. 2020), (Wang et al. 2021) For instance, a visual recognition model trained to recognize various breeds of dogs might confidently misclassify a fox, which it has never seen before, as a type of dog. This reflects what is referred to as the overconfidence problem, stemming from the fact that visual classifiers often lack common-sense knowledge (Zellers et al. 2019), (Park et al. 2020), (Shen et al. 2022). Such knowledge, like understanding that foxes and dogs are distinct categories despite sharing similar features, is typically found in language-based models but is missing in visual recognition systems. In other words, while language models can reason about these different contexts and relationships, visual recognition models might encounter limitations, particularly when encountering classes not seen during training.

Instead of merely learning a discriminative model for recognizing known objects, the proposed method enables vi-

*Corresponding author

sual recognition models to differentiate objects using the underlying logic and categorization that humans naturally apply (*i.e.*, common-sense knowledge). To achieve this, we introduce a novel approach distilling common-sense knowledge into structured noise patterns, which we refer to as "visual prompts". These are specially designed to encapsulate common-sense knowledge about the known classes of visual recognition models. As an illustration, Figure 1 shows a conceptual figure for the proposed method. By integrating the knowledge—that foxes are wild animals with distinct behaviors, while dogs are domesticated pets—into these visual prompts, the model is conditioned to perceive a never-before-seen fox as separate from the dog category.

This paper presents a pioneering approach to OSR by integrating text-based common-sense knowledge into visual prompts, aiming to address the inherent limitations in visual recognition models. This distillation of common-sense knowledge serves as a bridge between text-based understanding and visual recognition, enabling models to approach classification challenges with the depth of reasoning typically found in human cognition. This distillation of common-sense knowledge into visual prompts mimics the human ability to integrate information from various senses forming a comprehensive understanding that influences our knowledge and awareness of the world (Iordan et al. 2022). Individual visual prompts are trained for each class, encapsulating a rich understanding of common-sense knowledge. As a result, visual recognition models can harness the complex reasoning traditionally reserved for language-based systems, enhancing accuracy for known classifications and providing a more human-like discernment when encountering unknown objects.

Critically, this approach stands out for several reasons. It introduces a unique method of distilling textual common-sense knowledge into visual prompts, empowering the visual recognition models to detect contradictions between the presented images and their inherent knowledge base. Further, its model-agnostic nature ensures adaptability across a spectrum of visual recognition architectures. Additionally, by focusing on training just the visual prompts, the method saves on computation, making it faster to adapt and use in different situations. These benefits not only improve OSR results but also mark a new path forward in the OSR field.

The contributions of this work can be summarized:

- By introducing the novel framework of distilling text-based common-sense knowledge into visual prompts, this method equips visual recognition models with the deeper insights often reserved for human cognition. This unique bridge between textual understanding and visual recognition is especially vital in OSR: possessing common-sense knowledge about known classes makes it markedly easier for the system to spot inconsistencies when confronted with unknown class samples, thereby mirroring human-like judgment.
- Our approach focuses on training visual prompts, ensuring compatibility across diverse visual recognition models. This flexibility means we can improve OSR performance without altering the base architecture, offering both adaptability and wide-ranging utility.
- By focusing on training visual prompts, our method remains computationally efficient. This streamlined approach speeds up adaptation across various scenarios, making it easier to integrate OSR into current systems without demanding extensive resources.

Related Works

Open Set Recognition

Recent advancements in OSR have primarily revolved around the use of discriminative learning models (Bendale and Boult 2016), (Ge et al. 2017), (Oza and Patel 2019), to model decision boundaries between known classes. In particular, the emergence of Convolutional Neural Networks (CNNs) has spurred notable progress in the OSR field. Pioneering efforts like OpenMax (Bendale and Boult 2016) have calibrated the output of CNNs using novel theories, while other techniques have harnessed generative models (Chen et al. 2021), (Moon et al. 2022) to synthesize examples of unseen classes, enhancing the learning of boundaries between known and unknown classes.

Despite these advancements, a prevalent issue in existing methods is the overconfidence problem, where models can misclassify unknown objects with undue certainty. Such overconfidence often stems from a lack of common-sense knowledge in visual recognition systems. While many solutions in the literature attempt to address OSR challenges through model-specific architectures and specialized learning techniques, these often lead to complicated solutions demanding considerable customization and computational power. Conversely, our proposed method seeks to address the overconfidence issue directly by distilling common-sense knowledge into visual prompts. Rather than resorting to extensive modifications or relying on specialized architectures, we propose a universally applicable visual prompting strategy, potentially providing a more straightforward and efficient means to improve OSR performance.

Prompt Learning

Prompt-based learning, widely recognized in natural language processing (NLP), employs specific text inputs to amplify the proficiency of NLP models (Lester, Al-Rfou, and Constant 2021), (Li and Liang 2021). This concept has been extended to computer vision as visual prompts, showcasing promising applications (Kim, Kim, and Ro 2022), (Lin et al. 2023), (Chen et al. 2023). While methods like fine-tuning pre-trained vision models with visual prompts have demonstrated competitive results across various tasks with minimal resources (Jia et al. 2022). Rather than simply iterating on existing methodologies, we introduce a unique visual prompt learning framework tailored for the OSR task. Our novel approach endeavors to distill common-sense knowledge into visual prompts, thereby enriching pre-trained visual recognition models' performance in OSR.

Large Language Models

Recent advancements in NLP have been driven by the development of Large Language Models (LLMs) such

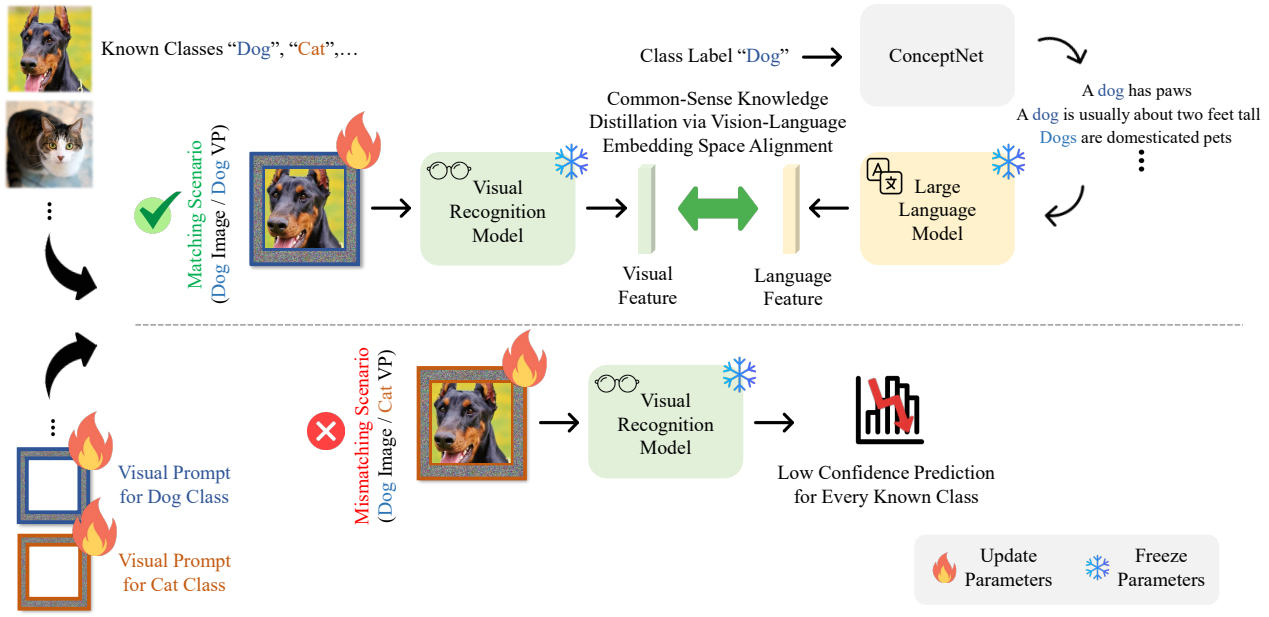

Figure 2: Schematic of the training process for visual prompts. The figure illustrates the extraction of common-sense text prompts for each known class (*e.g.*, dogs, cats) and their encoding by a large language model. The visual encoder processes two scenarios: 1) When an image (*e.g.*, a dog) is paired with its corresponding visual prompt (*e.g.*, dog), the training aligns the image feature with the text feature that includes the common-sense knowledge; 2) When an image (*e.g.*, a dog) is paired with a non-corresponding visual prompt (*e.g.*, cat), the visual prompt is trained to suppress overconfident classification, promoting uniform prediction probability across classes. The process results in a set of multiple distinct visual prompts, one for each known class, fostering improved OSR performance.

as Transformer-based architectures (Vaswani et al. 2017). These models, encapsulating billions of parameters, have demonstrated a profound ability to grasp linguistic patterns and common-sense knowledge (Brown et al. 2020), (Devlin et al. 2018). Alongside, vision-language alignment models like CLIP have endeavored to bridge the gap between visual knowledge space and language semantics, laying a foundational framework for integrated understanding (Radford et al. 2021), (Li et al. 2023a). A notable extension in this domain is BLIP-2, which introduces an efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models (Li et al. 2023a). The true potential of LLMs in our context lies in their ability to store vast amounts of textual common-sense knowledge. Our research leverages these models to distill this knowledge into a visual prompt format, bridging textual understanding with visual recognition. This approach of using LLMs for OSR provides a new perspective to improve system performance for identifying unseen classes and samples in real-world scenarios.

Proposed Method

Consider an input image $x \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and channels of the image, respectively. The associated label for this image is y. To allow the visual recognition model $f(\cdot)$ to incorporate common-sense knowledge, we introduce a unique structure: a visual prompt ϕ that has the same spatial dimensions as the input image, $\mathbb{R}^{H\times W\times C}$. However, ϕ is structured as a frame with a width of P pixels. The central region of ϕ is void, measuring $(H-2P)\times(W-2P)\times C$, ensuring that only the border of width P carries the information. This design is crafted to utilize minimal parameters without obscuring or compromising the critical information within the image, ensuring an efficient yet effective fusion of the prompt with the image.

The aim of this visual prompt is to infuse the model $f(\cdot)$ with common-sense knowledge, particularly aiding in addressing the OSR problem. By training ϕ and preserving the parameters of the visual recognition model, we ensure that the inherent model functionalities remain undisturbed. For the C known classes in our dataset, we devise C distinct visual prompts, with each crafted for a specific class. We define the set of the visual prompts as $\Phi = \{\phi_1, \phi_2, \cdots, \phi_C\}$. This facilitates the model to grasp and apply the unique common-sense knowledge associated with each class, refining its ability to act robustly in open set situations.

Textual Common-Sense Data Retrieval

To provide visual recognition models with a richer understanding grounded in common-sense knowledge, we leverage the expansive knowledge base of ConceptNet 5.5 (Speer, Chin, and Havasi 2017). Given the C known classes in our dataset, each class is associated with a specific name

or label. Using these class names as our starting point, we query ConceptNet for relevant sentences that encapsulate the common-sense knowledge associated with each class.

Upon querying, ConceptNet returns sentences or textual prompts that are intrinsically tied to the essence of the class names. However, not all returned sentences carry equal relevance or importance. To filter and retain only the most pertinent information, we take advantage of a weight threshold (ω). Sentences in ConceptNet are associated with weights that signify their relevance and importance to the queried term. By setting up a minimum weight, we ensure that only sentences surpassing this threshold are extracted. This ensures that our model gets useful and relevant common-sense knowledge, strengthening its ability to make better decisions in OSR situations.

Distillation of Common-Sense Knowledge via Contrastive Learning

The core of our approach lies in effectively bridging the knowledge from textual data to our visual recognition model. This is achieved using contrastive learning, ensuring the visual and textual features are accurately aligned.

For a given known class label y_i, where $i \in \{1, 2, ..., C\}$, the textual common-sense prompt derived from ConceptNet for that class is represented as t_{y_i}. We first incorporate the corresponding visual prompt ($\phi_{y_i=y^*}$) into each class image. For instance, for the "Dog" class, the visual prompt tailored for "Dog" is added on the image of a "Dog" in a pixel-wise manner. Denoting this visual prompt augmented image as $x' = x + \phi_{y_i=y^*}$, it is then fed into the visual recognition model $f(\cdot)$ to extract its feature, represented as $v_{y_i} = f(x')$. Simultaneously, each of these textual common-sense prompts t_{y_i} is processed by an LLM $g(\cdot)$ to extract its feature, denoted as $u_{y_i} = g(t_{y_i})$. The objective is to align features v_{y_i} and u_{y_i} for matching pairs and create a divergence for mismatching pairs. The contrastive learning loss (Radford et al. 2021) can be defined as:

$$\mathcal{L}_{Align} = -\sum_{i=1}^{C} \log \left(\frac{\exp(sim(v_{y_i}, u_{y_i})/\tau)}{\sum_{j=1}^{C} \exp(sim(v_{y_i}, u_{y_j})/\tau)} \right), \tag{1}$$

where $sim(a, b)$ computes the similarity between vectors a and b, often using the cosine similarity measure, and τ is a temperature parameter that scales the logits before applying the softmax function.

Through this contrastive learning mechanism, matching pairs' features are drawn closer in the feature space, while mismatching pairs are pushed apart, thereby infusing the visual recognition model with common-sense knowledge from the textual domain driven by the visual prompt.

Suppressing Overconfident Predictions

A vital component of our methodology is to prevent overconfidence when visual prompts are mismatched with image classes. For instance, if the "Dog" visual prompt is applied to a "Cat" image, the prediction should ideally be a uniform distribution across all classes, reflecting uncertainty and avoiding biased predictions.

Given an image augmented with a mismatched visual prompt, $x'' = x + \phi_{\tilde{c}}$, we first extract features using our visual recognition model, denoted by $f(x'')$, where $\phi_{\tilde{c}}$ is sampled from $\Phi \setminus \{\phi_{y_c=y^*}\}$). These features are then passed through a classifier h, which produces the final prediction output $h(f(x''))$. Our objective is to make this final prediction resemble a uniform distribution U, where $U = \frac{1}{C}$ for a scenario with C classes. To accomplish this, we utilize the Kullback-Leibler (KL) divergence to measure the disparity between the predicted distribution and the uniform distribution. The loss is given by:

$$\mathcal{L}_{Supp} = KL(h(f(x''))||U), \tag{2}$$

where $KL(a||b)$ computes the KL divergence between distributions a and b. This loss mechanism ensures that while the model remains accurate for standard classification scenarios, it exhibits desirable uncertainty when faced with mismatched visual prompts.

Strategic Optimization for OSR

Label smoothing is a well-known technique used to prevent over-fitting in deep learning models. Its significance in OSR arises from the need to avoid overconfidence in predictions, which could mislead the model when faced with unfamiliar data. To this end, for scenarios where the visual prompt aligns with the image's true class, we utilize label smoothing to produce more generalized predictions, thus providing more adaptable decision boundaries for open set inputs. This is complemented by the RandAugment technique (Cubuk et al. 2020), applying a varied augmentation strategy in each iteration to further enrich our model's robustness. The combined loss for these matching scenarios can be expressed as:

$$\mathcal{L}_{LS} = (1-\epsilon) \cdot CE(h(f(x')), y) + \epsilon \cdot CE(h(f(x')), U), \tag{3}$$

where ϵ represents the smoothing parameter and $CE(a, b)$ denotes the cross-entropy loss between prediction a and true label b. These techniques collaboratively fine-tune the visual prompts, enhancing their performance in OSR. In order to integrate the various components of our strategy, we compute the final loss as a weighted sum of the individual losses:

$$\mathcal{L} = \alpha \mathcal{L}_{Align} + \beta \mathcal{L}_{LS} + \gamma \mathcal{L}_{Supp}, \tag{4}$$

where α, β, and γ are hyperparameters that control the contribution of each loss to the final loss. The procedure of training visual prompts is described in Algorithm 1.

Visual Prompt Selection for Inference

In the inference phase, the optimal selection of the visual prompt plays a critical role in harnessing the trained capabilities of the visual recognition model for OSR tasks. While the model has been trained with diverse prompts tailored to each class, determining the most suitable prompt for a given test image is crucial.

The proposed inference procedure unfolds in two sequential stages: Preliminary Classification stage and Refined Prediction with Visual Prompt stage.

Initially, in the Preliminary Classification stage, the test image x, without any visual prompt, is first fed into the visual recognition model $f(\cdot)$. The model generates a probability distribution over the known classes. We select the

Algorithm 1: Training for OSR with Visual Prompts Distilled from Common-Sense Knowledge

Input: Dataset with training images $X = \{x_j\}_{j=1}^{N}$ and corresponding labels $Y = \{y_c\}_{c=1}^{C}$
Parameters: Set of visual prompts $\Phi = \{\phi_1, \cdots, \phi_C\}$, Pre-trained visual recognition model $f(\cdot)$, Classifier $h(\cdot)$, Large language model $g(\cdot)$, Smoothing parameter ϵ, Weight threshold ω, Hyperparameters α, β, γ, Temperature parameter τ
Output: Trained visual prompts Φ

Initialize the visual prompts ϕ_c's
$\mathcal{T} \leftarrow$ ConceptNet(class names, ω) {Retrieve texts with weight above ω}
for each epoch **do**
  for each mini-batch B of images from X **do**
    Split B into two halves: B_{match} and B_{mis}
    for each class y_c **do**
      $t_{y_c} \leftarrow \mathcal{T}_{y_c} \subset \mathcal{T}$ {Extract related text}
      $u_{y_c} \leftarrow g(t_{y_c})$ {Compute textual feature}
      Append u_{y_c} into a set to form $\{u_{y_c}\}_{c=1}^{C}$
    end for
    for each image x_j in B_{match} **do**
      $x'_j \leftarrow x_j + \phi_{y_c=y^*}$ {Add matching visual prompt}
      $v_j \leftarrow f(x'_j)$ {Compute visual feature}
      $\mathcal{L}_{Align} \leftarrow$ Eq. (1) {Contrastive loss}
      $\mathcal{L}_{LS} \leftarrow$ Eq. (3) {Label-smoothing loss}
    end for
    for each image x_k in B_{mis} **do**
      Sample random $\phi_{\tilde{c}}$ from $\Phi \setminus \{\phi_{y_c=y^*}\}$
      $x''_k \leftarrow x_k + \phi_{\tilde{c}}$ {Add mismatching visual prompt}
      $\mathcal{L}_{Supp} \leftarrow$ Eq. (2) {Suppression loss}
    end for
    Compute final loss $\mathcal{L}$ by Eq. (4)
    Optimize Φ using $\mathcal{L}$
  end for
end for

visual prompt corresponding to the class with the highest predicted probability. For instance, if the model predicts the test image to be a "Dog" with the highest confidence, then the visual prompt associated with the "Dog" class will be chosen for the subsequent step.

Following this, in the Refined Prediction with Visual Prompt stage, having determined the most likely class for the test image, the corresponding visual prompt is added to the original image, creating an augmented version x'. This augmented image is then fed back into the visual recognition model. The resulting prediction scores, while images being applied with visual prompts, provide a more refined and informed classification, optimized by the additional context the visual prompt offers.

This two-stage inference process, starting with a preliminary classification followed by a refined prediction using the selected visual prompt, ensures that the model not only recognizes the most probable class but also refines its decision using the tailored knowledge encapsulated in the visual prompts. We describe the procedure of applying the proposed visual prompt for inference in Algorithm 2.

Algorithm 2: Inference for OSR with Visual Prompts Distilled from Common-Sense Knowledge

Input: Dataset with test images $X_{test} = \{x_j\}_{j=1}^{N}$, Visual recognition model $f(\cdot)$, Classifier $h(\cdot)$, Trained visual prompts $\Phi = \{\phi_1, \cdots, \phi_C\}$
Output: Prediction for X_{test}

for each test image x_j **do**
  $y_{init} \leftarrow h(f(x_j))$ {Preliminary Classification stage}
  $x' \leftarrow x_j + \phi_{y_c=y_{init}}$ {Apply visual prompt}
  $y_{final} \leftarrow h(f(x'))$ {Refined Prediction with Visual Prompt stage}
end for
return Prediction for X_{test}

Experiments

Dataset

- **CIFAR10 (C10)**: Comprising 10 image classes with 50,000 training and 10,000 testing images, six classes were designated as known for the OSR task, leaving the remaining four as open set (Krizhevsky, Hinton et al. 2009).
- **CIFAR+N (C+N)**: This setup involves four known classes from CIFAR10, with a variable number of unknown classes from CIFAR100, creating a more complex OSR problem as the number of unknown classes increases.
- **TinyImagenet (TI)**: As the downscaled version of the ImageNet dataset, it includes 200 classes. In our experiments, 20 were used as known classes, and the remaining 180 were treated as unknown (Le and Yang 2015), (Deng et al. 2009).

Metrics

- **AUC (Area Under the ROC Curve)**: This metric, ranging from 0 to 100%, assesses the classifier's ability to differentiate between known and unknown classes by quantifying the trade-off between sensitivity and specificity across different thresholds (Phillips, Grother, and Micheals 2011).
- **F1 Score**: As the balance between precision and recall, the F1 score expressed as a percentage is especially valuable in imbalanced datasets, offering insight into the trade-off between false positives and false negatives (Hand and Christen 2018).
- **OSCR (Open Set Classification Rate)**: Adapted from the Detection and Identification Rate (DIR) curve, the OSCR curve plots the Correct Classification Rate (CCR) versus False Positive Rate (FPR) for known and unknown classes, providing a nuanced evaluation of accuracy in open set scenarios (Dhamija, Günther, and Boult 2018).

Method	CIFAR10 (C10)	CIFAR+10 (C+10)	CIFAR+50 (C+50)	TinyImageNet (TI)
Baseline (Softmax)	66.7 / 69.4 / 82.9	81.3 / 77.5 / 91.1	79.8 / 64.7 / 88.3	58.1 / 51.7 / 61.9
OpenMax (Bendale and Boult 2016)	69.5 / 71.4 / -	81.7 / 78.7 / -	79.6 / 67.7 / -	57.6 / 50.7 / -
G-OpenMax (Ge et al. 2017)	67.5 / - / -	82.7 / - / -	81.9 / - / -	58.0 / - / -
OSRCI (Neal et al. 2018)	69.9 / - / -	83.8 / - / -	82.7 / - / -	58.6 / - / -
C2AE (Oza and Patel 2019)	89.5 / - / -	95.5 / - / -	93.7 / - / -	74.8 / - / -
PROSER (Geng, Tao, and Chen 2020)	89.1 / - / -	96.0 / - / -	95.3 / - / -	69.3 / - / -
OpenHybrid (Zhang et al. 2020)	95.0 / - / -	96.2 / - / -	95.5 / - / -	79.3 / - / -
RPL (Chen et al. 2020)	86.1 / - / 85.2	85.6 / - / 91.8	85.0 / - / 89.6	70.2 / - / 53.2
CSSR (Huang et al. 2022)	91.3 / - / -	96.3 / - / -	96.2 / - / -	82.3 / - / -
RCSSR (Huang et al. 2022)	91.5 / - / -	96.0 / - / -	96.3 / - / -	81.9 / - / -
OpenGAN (Kong and Ramanan 2021)	- / - / -	- / - / -	- / - / -	- / 58.5 / -
CGDL (Sun et al. 2020)	- / 71.0 / -	- / 77.9 / -	- / 71.0 / -	- / - / -
GCM-CF (Yue et al. 2021)	- / 72.6 / -	- / 79.4 / -	- / 74.6 / -	- / - / -
GCPL (Yang et al. 2018)	- / - / 84.3	- / - / 91.0	- / - / 88.3	- / - / 59.3
AMPF++ (Xia et al. 2023)	- / - / 89.0	- / - / 95.1	- / - / 93.3	- / - / 69.0
ARPL+CS (Chen et al. 2021)	91.0 / 75.3 / 87.9	97.1 / 82.7 / 94.7	95.1 / 75.3 / 92.9	78.2 / 67.6 / 65.9
DIAS (Moon et al. 2022)	85.0 / 80.9 / 84.2†	92.0 / 85.9 / 93.1†	91.6 / 82.9 / 91.8†	73.1 / 65.4 / 66.4†
Baseline + Visual Prompt	91.2 / 80.0 / 89.0	94.5 / 85.0 / 95.8	94.9 / 77.7 / 92.3	77.4 / 67.6 / 67.9
(ARPL + CS) + Visual Prompt	**95.2 / 83.3 / 89.6**	**97.9 / 87.5 / 96.6**	**97.1** / 80.1 / **94.5**	**83.1 / 70.2** / 68.5
DIAS + Visual Prompt	93.9 / 83.0 / 86.6	95.9 / 86.0 / 94.7	96.3 / **83.6** / 93.1	80.9 / 68.1 / **69.1**

Table 1: Comparison of AUC / F1 / OSCR metric results (%) with state-of-the-art OSR methods, using the same baseline architectures as VGG32. Higher values indicate superior performance for all three metrics. Note that "Visual Prompt" denotes the proposed visual prompts distilled from common-sense knowledge, and † denotes results reproduced by us.

Experimental Settings

VGG32 architecture, popularly employed as a backbone by state-of-the-art methods, served as our foundation for visual recognition tasks. VGG32, denoted as Baseline in the experimental results, trained initially with a softmax loss for closed-set classification, was subsequently augmented by our visual prompt approach. Our experimental setup began with the reproduction and optimization of two notable OSR models: ARPL+CS (Chen et al. 2021) and DIAS (Moon et al. 2022). Once optimized, their weights were frozen, after which our proposed method was applied. For the extraction of common-sense knowledge features, we incorporated the OPT-2.7B model from BLIP2 (Li et al. 2023b). The threshold for extracting common-sense text prompts from ConceptNet 5.5 varied per dataset and will be elaborated upon later in this section. We have utilized a constant visual prompt size of P set at 30.

Enhancing OSR Performance with the Proposed Approach

Our experimentation shows that applying our strategy significantly enhances the OSR capacities of visual recognition models across various metrics. This improvement is largely attributed to the infusion of common-sense knowledge carried by the visual prompt with the visual recognition models.

From Table 1, we observe significant enhancements in AUC, a metric measuring the model's ability to distinguish between in-distribution and out-of-distribution samples. The Baseline model showcased a robust boost of +24.5% in AUC on the CIFAR10 dataset post the integration of our approach, indicating significant improvement in recognizing out-of-distribution samples. Our experiments also presents results in terms of the F1 score, a metric sensitive to data imbalance, which is inherent in OSR due to the undefined nature of out-of-distribution samples. For instance, the ARPL+CS model's performance on the CIFAR10 dataset saw an increase of +8.0% with our approach, indicating effective handling of data imbalance. The OSCR scores in Table 1 capture both open-set and closed-set recognition capabilities. Our method's consistent enhancement, such as the +1.7% OSCR improvement for the ARPL+CS model on CIFAR10, underscores that our approach enhances OSR without compromising closed-set task proficiency.

In summary, our method enhances performance across various metrics by incorporating common-sense knowledge.

$\mathcal{L}_{Align}$	$\mathcal{L}_{Supp}$	$\mathcal{L}_{LS}$	C10	C+10	C+50	TI
✗	✗	✗	66.7	81.3	79.8	58.1
✓	✗	✗	76.0	87.1	85.5	67.2
✓	✗	✓	77.5	87.6	86.0	68.0
✓	✓	✗	89.0	93.5	94.0	75.6
✓	✓	✓	**91.2**	**94.5**	**94.9**	**77.4**

Table 2: Effects of loss functions devised for training the proposed visual prompts on AUC performance (%). The first row represents the model's performance without the proposed method (*i.e.*, Baseline (Softmax) in Table. 1).

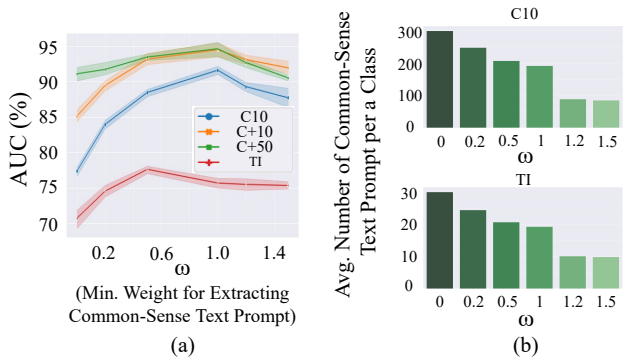

Figure 3: Trade-off between the quality and quantity of common-sense text prompts for training visual prompts. (a) depicts AUC of the Baseline model while varying minimum weight ω to extract common-sense text prompt. (b) illustrates the average number of common-sense text prompts per class used for training the visual prompt for two datasets.

This leads to improvements in OSR, enabling models to effectively identify out-of-distribution samples, handle data imbalance, and maintain closed-set classification abilities.

Ablation Study on Training Strategies

In this ablation study, we evaluate the impact of different Visual Prompt training strategies on the AUC performance of our model, using the baseline architecture VGG32. The results are presented in Table 2. We varied the use of objective functions—$\mathcal{L}_{Align}$, $\mathcal{L}_{Supp}$, and $\mathcal{L}_{LS}$—to train our model with Visual Prompts. The first row of Table 2 illustrates the performance without any visual prompt application. Upon introducing $\mathcal{L}_{Align}$, designed to distill common-sense knowledge, we observed a significant increase in AUC. The $\mathcal{L}_{Supp}$ loss, aimed to suppress predictions when encountering mismatched visual prompts, further enhanced performance. Notably, the AUC on CIFAR10 increased to 89.0% with both $\mathcal{L}_{Align}$ and $\mathcal{L}_{Supp}$ applied. The addition of $\mathcal{L}_{LS}$ led to a modest, but consistent improvement, indicating its positive effect on stabilizing visual prompt training. Note that while $\mathcal{L}_{LS}$ is not used, we utilize softmax cross entropy loss instead, where ϵ equals to 0. In summary, our study shows that each loss function contributes uniquely to enhancing the performance, confirming the effectiveness of visual prompts trained with these objective functions.

Trade-Off between Common-Sense Text Quality and Number of Text Prompts

In this subsection, we analyze a graph depicted in Figure 3, which explores the trade-off between the quality and quantity of common-sense text prompts, as measured by the minimum weight set on the edges of ConceptNet, and the number of such text prompts used for training the visual prompts. A higher weight in ConceptNet signifies more reliable and informative common-sense knowledge, while a lower weight suggests potential for less recognized or lower-quality information.

Size of P	# of Parameters	C10	C+10	C+50	TI
-	-	66.7	81.3	79.8	58.1
10	26K	80.4	88.7	87.4	66.2
20	49K	86.3	91.9	91.2	73.4
30	70K	**91.2**	**94.5**	**94.9**	**77.4**
40	88K	88.3	94.2	94.1	76.2
50	104K	85.4	92.4	92.0	73.8

Table 3: Performance comparison of the model under varying trainable parameters (Prompt size P). The table shows the AUC performance (%) across different datasets. The first row represents the model's performance without the proposed visual prompts (*i.e.*, Baseline (Softmax) in Table. 1).

We have conducted experiments on VGG32, which is described as the Baseline model in the previous experiments. For the CIFAR10 dataset, with general classes like 'dog' and 'cat', a high ConceptNet weight threshold (*e.g.*, 1.0) is optimal, indicating that fewer but high-quality text prompts suffice for effective visual prompt training. In contrast, for TinyImageNet, which includes specific classes like 'gazelle' and 'espresso', a lower threshold (*e.g.*, 0.5) yields the best results. This strategy embraces a larger volume of prompts, potentially of lower quality, proving beneficial for the training of visual prompts for specific and uncommon classes.

Computational Efficiency of the Proposed Method

One of the contributions of our proposed method is its computational efficiency of the visual prompt. As depicted in Table 3, we have conducted experiments on VGG32 while varying the number of trainable parameters of the visual prompt, which corresponds to the prompt size P. We achieve optimal OSR performance by setting the visual prompt size P as 30, which corresponds to approximately 70,000 parameters for each class. As observed, the AUC performance improves as the prompt size increases from 10 to 30. Beyond this point, we notice that the performance exhibits a slight decline as the prompt size continues to increase towards 50. Unlike methods that require extensive optimization of the visual recognition model, our strategy focuses on refining a relatively lightweight, cost-efficient visual prompt.

Conclusion

In this work, we presented a novel approach to OSR, by integrating text-based common-sense knowledge into visual prompts, our methodology equips visual recognition models with a depth of understanding akin to human cognition. These prompts, rich in contextual insights, enable models to discriminate known from unknown objects, addressing the pervasive overconfidence problem inherent in conventional systems. Notably, the flexibility and model-agnostic nature of our solution signify its adaptability across various architectures. By focusing on computational efficiency, our method makes it easier to use in various applications without using a lot of resources. Basically, this paper shows how combining language and vision models can help improve OSR performance.

Acknowledgments

This work was partially supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.NRF-2022R1A2C2005529).

References

Bendale, A.; and Boult, T. E. 2016. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1563–1572.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Chen, G.; Peng, P.; Wang, X.; and Tian, Y. 2021. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8065–8081.

Chen, G.; Qiao, L.; Shi, Y.; Peng, P.; Li, J.; Huang, T.; Pu, S.; and Tian, Y. 2020. Learning open set network with discriminative reciprocal points. In *Proceedings of the European Conference on Computer Vision*, 507–522. Springer.

Chen, Z.; Zhou, Q.; Shen, Y.; Hong, Y.; Zhang, H.; and Gan, C. 2023. See, Think, Confirm: Interactive Prompting Between Vision and Language Models for Knowledge-based Visual Reasoning. *arXiv preprint arXiv:2301.05226*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 702–703.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhamija, A. R.; Günther, M.; and Boult, T. 2018. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31.

Ge, Z.; Demyanov, S.; Chen, Z.; and Garnavi, R. 2017. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*.

Geng, C.; Tao, L.; and Chen, S. 2020. Guided CNN for generalized zero-shot and open-set recognition using visual and semantic prototypes. *Pattern Recognition*, 102: 107263.

Hand, D.; and Christen, P. 2018. A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28: 539–547.

Hein, M.; Andriushchenko, M.; and Bitterwolf, J. 2019. Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, H.; Wang, Y.; Hu, Q.; and Cheng, M.-M. 2022. Class-specific semantic reconstruction for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4214–4228.

Iordan, M. C.; Giallanza, T.; Ellis, C. T.; Beckage, N. M.; and Cohen, J. D. 2022. Context Matters: Recovering Human Semantic Structure from Machine Learning Analysis of Large-Scale Text Corpora. *Cognitive science*, 46(2): e13085.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision*, 709–727. Springer.

Kim, M.; Kim, H.; and Ro, Y. M. 2022. Speaker-adaptive lip reading with user-dependent padding. In *European Conference on Computer Vision*, 576–593. Springer.

Kong, S.; and Ramanan, D. 2021. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE International Conference on Computer Vision*, 813–822.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Lin, Y.; Zhao, Z.; ZHU, Z.; Wang, L.; Cheng, K.-T.; and Chen, H. 2023. Exploring Visual Prompts for Whole Slide Image Classification with Multiple Instance Learning. *arXiv preprint arXiv:2303.13122*.

Moon, W.; Park, J.; Seong, H. S.; Cho, C.-H.; and Heo, J.-P. 2022. Difficulty-Aware Simulator for Open Set Recognition. In *Proceedings of the European Conference on Computer Vision*, 365–381. Springer.

Neal, L.; Olson, M.; Fern, X.; Wong, W.-K.; and Li, F. 2018. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision*, 613–628.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Oza, P.; and Patel, V. M. 2019. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2307–2316.

Padhy, S.; Nado, Z.; Ren, J.; Liu, J.; Snoek, J.; and Lakshminarayanan, B. 2020. Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks. *arXiv preprint arXiv:2007.05134*.

Park, J. S.; Bhagavatula, C.; Mottaghi, R.; Farhadi, A.; and Choi, Y. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 508–524. Springer.

Phillips, P. J.; Grother, P.; and Micheals, R. 2011. Evaluation methods in face recognition. *Handbook of face recognition*, 551–574.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Shen, S.; Li, C.; Hu, X.; Xie, Y.; Yang, J.; Zhang, P.; Gan, Z.; Wang, L.; Yuan, L.; Liu, C.; et al. 2022. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35: 15558–15573.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Sun, X.; Yang, Z.; Zhang, C.; Ling, K.-V.; and Peng, G. 2020. Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13480–13489.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, Y.; Li, B.; Che, T.; Zhou, K.; Liu, Z.; and Li, D. 2021. Energy-Based Open-World Uncertainty Modeling for Confidence Calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9302–9311.

Xia, Z.; Wang, P.; Dong, G.; and Liu, H. 2023. Adversarial Kinetic Prototype Framework for Open Set Recognition. *IEEE Transactions on Neural Networks and Learning Systems*.

Yang, H.-M.; Zhang, X.-Y.; Yin, F.; and Liu, C.-L. 2018. Robust classification with convolutional prototype learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3474–3482.

Yue, Z.; Wang, T.; Sun, Q.; Hua, X.-S.; and Zhang, H. 2021. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15404–15414.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, H.; Li, A.; Guo, J.; and Guo, Y. 2020. Hybrid models for open set recognition. In *Proceedings of the European Conference on Computer Vision*, 102–117. Springer.

Zhou, D.-W.; Ye, H.-J.; and Zhan, D.-C. 2021. Learning Placeholders for Open-Set Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4401–4410.