Memory-Modular Classification: Learning to Generalize with Memory Replacement

Anonymous authors
Paper under double-blind review

Abstract

We propose a novel memory-modular learner for image classification that separates knowledge memorization from reasoning. Our model enables effective generalization to new classes by simply replacing the memory contents, without the need for model retraining. Unlike traditional models that encode both world knowledge and task-specific skills into their weights during training, our model stores knowledge in the external memory of web-crawled image and text data. At inference time, the model dynamically selects relevant content from the memory based on the input image, allowing it to adapt to arbitrary classes by simply replacing the memory contents. The key differentiator is that our learner meta-learns to perform classification tasks with noisy web data from unseen classes, resulting in robust performance across various classification scenarios. Experimental results demonstrate the promising performance and versatility of our approach in handling diverse classification tasks, including zero-shot/few-shot classification of unseen classes, fine-grained classification, and class-incremental classification.

1 Introduction

Large-scale neural models have achieved remarkable results when fine-tuned and applied to downstream tasks in computer vision (Kolesnikov et al., 2020; Yuan et al., 2021; Alayrac et al., 2022) and natural language processing (Brown et al., 2020; Touvron et al., 2023). These models are trained on massive datasets using immense computational resources, resulting in a vast number of model parameters that encapsulate both world knowledge and task-specific skills. This complexity poses two challenges. First, it is difficult to determine which knowledge in the training data or learned skills contributes to the model output for a specific task. Second, models cannot directly reflect changes in the ever-growing real world, such as updates to data sources relevant to the target task, without undergoing additional training.

To flexibly adapt to the external world knowledge, recent zero-shot image recognition models (Guu et al., 2020; Hu et al., 2023b) enhances image representations with their relevant data retrieved from an external knowledge source. Such learning method is often called retrieval-augmented learning. This approach allows models to leverage external knowledge sources and efficiently allocate model parameters to focus on reasoning tasks. Although these models have shown promising results in knowledge-intensive applications, such as question answering (Gao et al., 2022) and long-tailed classification (Long et al., 2022), their capabilities are limited to a specific target task. Moreover, they assume that the memory content is retained throughout training and testing; the generalizability of the learned models when faced with substantial memory updates or replacements remains unexplored.

In this paper, we introduce a novel learning architecture, the *memory-modular learner* (MML), for image classification. MML leverages an external memory to perform input-adaptive reasoning during the classification process. A key advantage of MML is its ability to generalize with memory replacement, *i.e.*, memory-modular generalization. By simply plugging in new-class content into memory, MML can adapt to novel classification tasks *without requiring any architectural modifications* (Fig. 1). The external memory used by MML is populated by web-crawled images and text obtained by keyword search of the target class names. This approach facilitates the incorporation of up-to-date world knowledge into the memory, ensuring

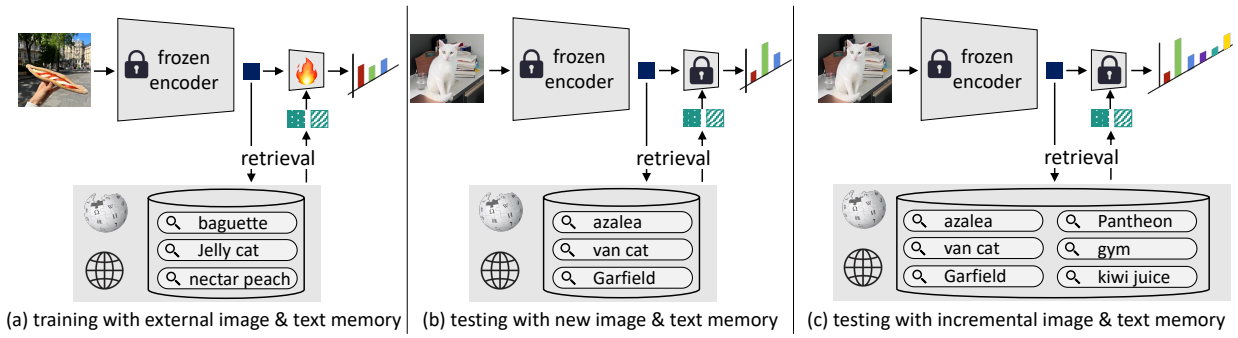

Figure 1: **Training and evaluation stages of MML for web-assisted zero-shot classification.** MML constructs image/text memory with text keyword search on the internet given target classes. The memory provides relevant image/text features which are integrated via a trainable knowledge integration module (a). On evaluation, the memory can be replaced or detached from the model such that MML joins the new knowledge as memory, while the rest of the model remains unchanged. Once trained, MML handles zero-shot classification on unseen classes with memory replacement (b) and incremental classes with memory expansion (c) using the new knowledge collected from web to solve zero-shot classification.

that MML remains applicable as external knowledge evolves. Despite the potential introduction of data noise from web crawling, MML demonstrates robust classification performance in practice. This remarkable robustness allows MML to effectively leverage the noisy memory contents for accurate image classification.

One critical observation of our work is that by representing classification as *metric learning* (Vinyals et al., 2016; Snell et al., 2017), MML becomes less susceptible to overfitting on the specific content of the memory. This allows it to learn more effectively how to perform classification reasoning with arbitrary memory contents. Specifically, we represent the classifier weight vectors, or the *class prototypes*, as the average of representative memory items rather than as learnable parameters. When a new set of classes is given, the class prototypes are immediately computed with the average of the memory items of the highest cross-modal similarity. The input query is classified by the class of the closest class prototype. This design choice allows us to update the memory and adapt to new classes without retraining the entire model. Due to its inherent flexibility, our meta-learned model can handle zero- to multi-shot samples, as well as a variable number of classes with the knowledge collected from web[1]. Experimental results in various scenarios, including zero-shot/few-shot classification of previously unseen classes, fine-grained classification, and class-incremental classification, demonstrate the promising performance of MML.

Our contributions can be summarized as follows.

- We introduce a memory-modular learner (MML) for image classification, that performs adaptive reasoning using external and replaceable memory.
- We investigate the generalizability in adapting to new classes by replacing the memory with related content, without tuning the model weights.
- We provide in-depth analyses on the memory-modular generalization to unseen classes in realistic setups, *i.e.*, using a noisy web-crawled memory.
- We show that MML achieves promising gains in various scenarios such as zero-shot, few-shot, fine-grained, and class-incremental classification by leveraging target-class knowledge collected from web.

[1]We clarify our zero-shot classification approach that accesses to unlabeled web data as *web-assisted zero-shot classification*.

2 Related work

2.1 Few-shot and zero-shot classification with the assistance of external web data

Few-shot image classification (Fei-Fei et al., 2006) aims to generalize to arbitrary unseen classes given a few support images from a target class set. The conventional experimental setup of few-shot classification (Vinyals et al., 2016; Allen et al., 2019; Triantafillou et al., 2020; Doersch et al., 2020; Zhang et al., 2020; Kang et al., 2021) assumes at least a few hundred labeled images used for (meta-)training before the actual few-shot inference stage. We, however, adopt a more label-efficient and realistic approach for this task; we train a model with even fewer labeled training samples *e.g.*, ≤ 16; instead we assume retrieval access to external unannotated data. **Zero-shot classification** (Larochelle et al., 2008; Yu & Aloimonos, 2010) aims for generalization beyond seen classes without the use of few-shot support images for the target classes. Instead, classification is conducted based on non-visual clues such as textual information of the images (Fu et al., 2015; Akata et al., 2016), yes-or-no attributes (Lampert et al., 2013) or the class name in text (Socher et al., 2013) of arbitrary classes. The conventional zero-shot tasks have assumed no use of images from *target classes* during training, but with the advent of web-driven pretrained models, recent "zero-shot" methods (Iscen et al., 2024; Liu et al., 2023) started to use the expression in a more relaxed way, meaning no use of *manually-annotated images* from target classes, thus allowing access to noisy web data. For example, a vision-and-language foundation model named CLIP (Radford et al., 2021) trains image and text encoders with 400 million image-and-caption pairs from internet which likely overlap with standard zero-shot classification benchmark categories. We follow this usage in our paper and leverage web data to leverage the external world knowledge for zero-shot classification. We thus clarify that our approach as **web-assisted zero-shot classification** with the terminology of "shot" denoting the number of *class-annotated images* for each target class.

2.2 Image recognition with memory retrieval

One of the earliest works of using an external memory in machine learning is the k-nearest neighbor (kNN) classifier (Hart, 1968), which retrieves k-nearest neighbors from memory for class prediction. Recent work constructs memory from large-scale pre-trained models and performs kNN retrieval for class prediction (Khandelwal et al., 2020; Nakata et al., 2022). This straightforward method revisits the potential of external memory for class reasoning, being decoupled from encoder learning (Graves et al., 2014). Image recognition models have also been trained using external image-text paired memory (Jia et al., 2021b; Long et al., 2022; Iscen et al., 2023). Our approach assumes a more weakly-supervised type of memory, collecting image and text memory contents separately. Other external memory-based image recognition work focuses on training multi-modal feature encoders (Wei et al., 2023; Hu et al., 2023b) or training CLIP models with external image-text paired data (Iscen et al., 2024; Liu et al., 2023). One common theme among the existing memory-based models is that they are either trained for a specific task (Long et al., 2022; Hu et al., 2023b; Iscen et al., 2023) or static memory (Iscen et al., 2024). Among them, REVEAL (Hu et al., 2023b) is perhaps most similar to ours. The memory-augmented learning architecture of REVEAL and MML is indeed similar in terms of architecture but different in terms of the role of memory. The memory of REVEAL serves as a general knowledge bank to assist VQA and captioning tasks. On the other hand, the memory of MML contains specifically related contents of the target classes for classification, crawled from web. Therefore, memory contents can be completely replaceable when the target classes are updated – the memory is *modular*. Note that any other previous models do not replace memory contents completely. Also, the difference of general and specific memory also leads to the size difference. REVEAL contains 20.3M memory items of general image-text pairs. MML requires only 0.7M image and 0.2M text memory items for 1K classes of ImageNet1K, (see Sec. 4.1) which is the 4.4 % size of REVEAL. In contrast to the previous related work, MML aims to generalize beyond a seen class set and modular memory that can be updated at any time. To the best of our knowledge, MML is the first to investigate the memory replacement with new memory contents to tackle unseen-class generalization.

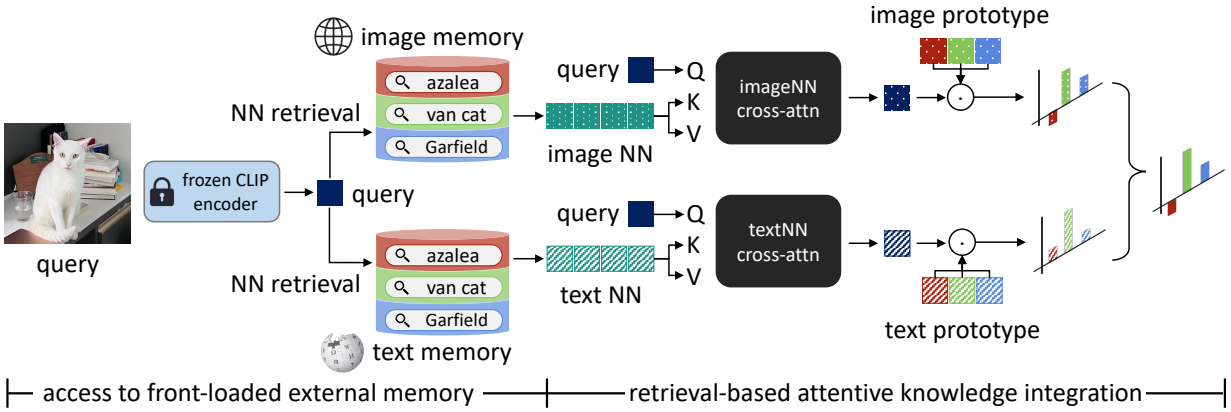

Figure 2: **Memory-modular learner (MML)** constructs image/text memory by web-crawling with text keyword search. Given a query image, its kNN features are retrieved from each memory and used for attentive knowledge integration. The class prototypes are constructed with the average of the memory elements of the highest cross-modal similarity. MML derives class reasoning with the nearest neighbors (NNs) from the external memory. This modular memory enables MML to perform web-assisted zero-/few-shot classification on unseen classes by memory replacement and class-incremental classification by memory expansion.

2.3 Class-incremental classification

Class-incremental classification (Rebuffi et al., 2017; Zhu et al., 2023) assumes that a set of unseen classes arrives at each stage and aims to classify the input into all known classes given limited access to the old class data. The most critical challenge of this task is catastrophic forgetting, *i.e.*, directly training neural networks with the new-class data leads to significant performance drops in old classes. To address this challenge, recent work (Yan et al., 2021; Wang et al., 2022a; Zhou et al., 2023b; Douillard et al., 2022; Wang et al., 2022b) introduces a memory to store data from previously seen classes as a training source to compile knowledge into a model. In contrast, the memory in MML plays the role of a replaceable and extensible world-knowledge reference. The purpose of memory in these two models are different: the memory of class-incremental learners helps not to forget the previously seen classes (Belouadah & Popescu, 2019; Iscen et al., 2020), however, the memory of MML assists the current classes of interest, which might have not been seen.

3 Memory-modular learner

We address the problem of classifying an input image into target classes that are represented by a class name in text, *i.e.*, zero-shot classification, or additional few support images, *i.e.*, few-shot classification. To this end, we introduce a *memory-modular learner* that performs adaptive reasoning using an external memory that is updatable and replaceable. Our memory-modular learner takes advantage of both vision and language modalities using the CLIP encoder (Radford et al., 2021) as a base feature extractor for image and text. Since our method is not restricted to CLIP, any other image-text model, *e.g.* ALIGN (Jia et al., 2021a) or LiT (Zhai et al., 2022), can also be adopted. Figure 2 illustrates the overall architecture of our approach.

The memory-modular learner starts by loading the knowledge memory and generating class prototypes for target classes (Sec. 3.1). These front-loaded memory items and prototypes are all stored as frozen features from a pre-trained image-text encoder. They are replaceable whenever the target classes change or the external knowledge sources are updated. Given an input image, the memory-modular learner accesses the knowledge memory, retrieves k-nearest-neighbor (kNN) items, and predicts the corresponding class via cosine-similarity with class prototypes (Sec. 3.2). Since class prototypes are generated immediately from the memory items, the prototype-based classifier can adapt to new target classes of updated memory contents without additional training.

3.1 Memory construction and prototype generation

Given target class names or descriptions, we construct the knowledge memory based on available image and text data and generate class prototypes using the memory. As the world knowledge is updated, these memory items can be added or deleted, and even completely replaced, without updating the model weights.

Knowledge memory

The image memory is constructed using images obtained from keyword searches on the internet. For each target class c, images are collected using the class name as the search keyword on a search engine, *e.g.*, Google or Flickr (Kim et al., 2023; Hou et al., 2018). We follow a similar strategy for text memory. In this work, textual information relevant to each target class name is retrieved by querying Wikipedia (Tian et al., 2022; Hu et al., 2023a; Naeem et al., 2023). These web-crawled images and texts may be noisy, but consist of scalable memory contents that reflect the world knowledge. After collecting the relevant images and texts for each target class c, we extract their d-dimensional features with the image-text encoder, and then store them in the image and text memory: $\mathcal{M}_c^{\text{img}} = \{\boldsymbol{v}_i\}_{i=1}^{N_c^{\text{img}}}$ and $\mathcal{M}_c^{\text{txt}} = \{\boldsymbol{t}_j\}_{j=1}^{N_c^{\text{txt}}}$, respectively.

Class prototypes

For zero-shot classification, we construct class prototypes based on cross-modal consensus between image and text memory items. For each target class c, we first compute the cross-modal cosine similarity $\cos(\cdot,\cdot)$ from each image to all text items of the same class and then select the top-M images with the highest similarity to the texts, *i.e.*, images with high cross-modal consensus. The image prototype for class c is then set to be the average of the M features:

$$\boldsymbol{p}_c^{\text{img}} = \frac{1}{|\mathcal{T}|}\sum_{\boldsymbol{v}\in\mathcal{T}} \boldsymbol{v},\ \mathcal{T} = \text{argmax}_{\boldsymbol{v}'\in\mathcal{M}_c^{\text{img}}}^{M}\left(\sum_{\boldsymbol{t}\in\mathcal{M}_c^{\text{txt}}} \cos(\boldsymbol{v}',\boldsymbol{t})\right), \tag{1}$$

where $\text{argmax}_{\boldsymbol{s}\in\mathcal{S}}^{M}(\cdot)$ denotes the top-M operator that returns the best M items from the set $\mathcal{S}$ maximizing the operand function. Based on image-text consensus, this process constructs robust and representative class prototypes from noisy data in the absence of human annotation. Likewise, the text prototype is obtained using the average text-to-image similarity. This zero-shot prototype construction resembles Prototypical Networks (Snell et al., 2017), which averages the M image examples for each class. On the other hand, we build multi-modal prototypes by averaging the representative M samples collected without given annotations. For few-shot classification, *i.e.*, when a few support image samples are available for the target class name, we simply construct class prototypes by averaging the given samples as done in Snell et al. (2017).

Memory update for adapting to unseen classes

The knowledge memory contents and class prototypes are modular and replaceable. When target classes are updated, *e.g.*, classification of unseen classes or incremental classes, new memory contents are collected to pertain to the new classes. Subsequently, the prototypes for the classes are updated accordingly using Eq. 1.

3.2 Reasoning with memory access

Given an input image for classification, we incorporate memory knowledge into reasoning. Items relevant to the input are retrieved from image/text memory and integrated with the input feature through cross-attention. The input is then correlated with the image and text class prototypes. Finally, the predictions from the image and text branches are merged at the logit level for class prediction.

Memory retrieval

For an input image feature $\boldsymbol{f}$ extracted from the image encoder, its k-nearest-neighbor image items are retrieved based on cosine similarity with all image memory items of all target classes:

$$\mathcal{N}^{\text{img}} = \text{argmax}_{\boldsymbol{v}\in\mathcal{M}^{\text{img}}}^{K}\left(\frac{\boldsymbol{f}\cdot\boldsymbol{v}}{|\boldsymbol{f}||\,||\boldsymbol{v}||}\right), \tag{2}$$

Table 1: Zero-shot cross-dataset transfer. MML is trained with 1 or 4 samples from ImageNet1K coarse-grained classes and tested on 10 fine-grained datasets with zero shot, which is roughly a domain shift scenario.

method	ImgNet1K	Caltech101	OxfordPets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	avg.
		objects	pets	cars	flowers	food	airplanes	scenes	textures	land	actions	
zero-shot CLIP (Radford et al., 2021)	66.7	75.9	63.6	62.9	54.7	74.5	18.2	55.3	33.3	43.0	58.7	55.2
kNN classifier (Nakata et al., 2022)	55.7	87.6	72.7	68.6	75.2	75.6	**29.6**	56.2	33.2	37.3	63.2	59.5
MML (ImageNet1K-1)	48.3	92.6	86.4	68.1	76.2	81.8	26.2	60.0	41.6	45.6	64.2	62.8
MML (ImageNet1K-4)	**69.0**	**93.5**	**86.7**	**68.9**	**77.5**	**84.2**	26.3	**64.7**	**42.8**	**48.2**	**66.5**	**66.2**

where $\mathcal{M}^{\text{img}} = \cup_c \mathcal{M}_c^{\text{img}}$. The text kNNs are also retrieved by querying the image feature to the text memory.

Attentive knowledge integration

The knowledge of the retrieved memory items $\mathcal{N}^{\text{img}} = [\boldsymbol{v}_k]_{k=1}^K$ is aggregated by cross-attention (Vaswani et al., 2017; Jaegle et al., 2021) and then integrated with the input embedding $\boldsymbol{f}$. The cross-attention learns to integrate the nearest neighbor (NN) features into the input feature:

$$\boldsymbol{f}^{\text{img}} = \boldsymbol{f} + \sigma\left(\frac{\mathbf{Q}(\boldsymbol{f}) \cdot [\mathbf{K}(\boldsymbol{v}_k)]_{k=1}^K}{\sqrt{d}}\right)[\mathbf{V}(\boldsymbol{v}_k)]_{k=1}^K, \tag{3}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are projection layers with non-linearity, σ softmax over k items, and $[\cdot]$ concatenation. Similarly, the same step with the text NN features is performed in parallel. This process can be viewed as a learnable soft NN integration in contrast to the hard majority voting with NNs (Nakata et al., 2022).

Classification inference

The resulting embedding is matched against the multi-modal prototypes for all C target classes with cosine similarity $\cos(\cdot,\cdot)$ to produce classification score. The c-th class logit $\boldsymbol{z}_c$ is obtained with:

$$\boldsymbol{z}_c = \cos(\boldsymbol{p}_c^{\text{txt}}, \boldsymbol{f}^{\text{txt}}) + \cos(\boldsymbol{p}_c^{\text{img}}, \boldsymbol{f}^{\text{img}}). \tag{4}$$

Final class prediction is conducted simply by taking the class with the highest score.

3.3 Training

Our model is trained with cross-entropy loss with one-hot ground-truth class label $\boldsymbol{y}$ and the logit $\boldsymbol{z}$:

$$\mathcal{L} = -\sum_{c=1}^{C} \boldsymbol{y}_c \log \frac{\exp(\boldsymbol{z}_c/\tau)}{\sum_{c'}^{C} \exp(\boldsymbol{z}_{c'}/\tau)}, \tag{5}$$

where τ is a temperature for scaling. Note that we freeze the pre-trained image-text encoder backbone and train the remaining parameters only, *i.e.*, those of attention layers on the image and text branches. This helps to retain general knowledge in the pre-trained encoder, avoid overfitting, and reuse pre-computed features of memory items. In our case, the number of training parameters and the frozen backbone parameters amounts to 6.3M and 151M in total when using the CLIP-B/32 backbone. The small trainable part effectively learns memory-based reasoning for classification, as experimentally shown in the following section.

4 Experiments

4.1 Experimental setup

Training details

For the image/text feature extractor, we use the pre-trained CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021a). Unless specified, CLIP-B/32 is used. For training, we use a batch size of 256, a learning rate

Table 2: Comparison on zero-shot classification on CUB (Wah et al., 2011) with different backbones. Note that RECO (Iscen et al., 2024) is trained with CC12M.

method	backbone	accuracy (%)
CLIP (Radford et al., 2021)	CLIP-B/32	70.3
RECO* (Iscen et al., 2024)	CLIP-B/32	75.2
MML	CLIP-B/32	**76.7**

method	backbone	accuracy (%)
CLIP (Radford et al., 2021)	ResNet-101	68.8
Yu et al. (2020)	ResNet-101	72.4
Xu et al. (2020)	ResNet-101	73.8
Chen et al. (2022)	ResNet-101	76.1
MML	ResNet-101	**78.8**

of $1e^{-6}$ and weight decay of $5e^{-4}$ on a single 2080 Ti or an RTX 3090 GPU for all training and testing. We retrieve 32 NNs from both the image and text memory. We use $M = 16$ for prototype construction and set the logit temperature $\tau = 16$, which is chosen via hyperparameter search. We use three random seeds for drawing few-shot samples randomly and report the average.

Memory and data

To construct the external image memory for ImageNet derivatives, we employ a readily available web-crawled image dataset, WebVision ver. 2 (Li et al., 2017). WebVision is collected from Google and Flickr by the keyword search of the 1000 class names of ImageNet1K (Russakovsky et al., 2015). We use the image subset crawled from Google unless otherwise specified. To construct image memory for the other 10 datasets used in Table 1, as no public web-crawled datasets for the corresponding classes are available, we crawl a maximum of 100 images per class from Google with an auto crawler. For text memory, we query Wikipedia for each class name and retrieve the corresponding article text by web crawling. In such a way, the modest length of memory is obtained, *e.g.*, 0.7M images and 0.2M texts for the 1K classes of ImageNet1K, of which kNN search is feasible with the PyTorch (Paszke et al., 2017) built-in `topK` module. The dataset details used for zero-/few-shot, fine-grained, and class incremental classification are specified in the corresponding paragraph.

4.2 Web-assisted zero-shot classification

First of all, we evaluate our method on zero-shot classification setup, where no labeled images are provided for the target classes. The only information given for the task is a phrased class label for each class, *e.g.*, "van cat", which is used as the search keyword to collect the web-crawled memory.

Datasets: MML is evaluated on single-dataset and cross-dataset zero-shot classification benchmarks. For single-dataset zero-shot classification, ImageNet-S and CUB are used, where the classes of each dataset are split into disjoint sets for few-shot training and zero-shot testing. We adopt the existing zero-shot classification CUB benchmark (Wah et al., 2011; Akata et al., 2013) of which classes are split into 150/50 bird species classes for train/validation. Similarly, we introduce an ImageNet (Russakovsky et al., 2015) split such that it comprises 600/200/200 classes for train/validation/test and call it ImageNet-S (S stands for class *split*). We use 16 images per class for training. For testing on target classes, either zero or a few shots are used for zero- or few-shot classification scenarios. For the cross-dataset setting, we adopt a cross-dataset zero-shot transfer scenario (Zhou et al., 2022), where a model is trained with a few samples from ImageNet1K and transferred to 10 fine-grained datasets: Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FgvcAircraft, SUN397, DTD, EuroSAT, and UCF101. The total classes of these datasets amount to 1,310 classes and their details including the references are found in Table 10.

Baselines: The kNN classifier (Nakata et al., 2022) retrieves kNN of the input from memory[2] and immediately predict the class by majority voting. Zero-shot CLIP/ALIGN extracts text embeddings of the text class names in the predefined templates, *e.g.*, a photo of a van cat, and matches them against the input image embedding. Three state-of-the-art zero-shot models (Yu et al., 2020; Xu et al., 2020; Chen et al., 2022) are also compared, which are trained with the total 8885 annotated images and text attributes of CUB.

[2]The original work (Nakata et al., 2022) leverages annotated datasets such as ImageNet1K as image memory, which is expensive to be used as memory. We thus replace it with the noisy web-crawled memory for reproduction.

Table 3: Ablation study on MML components

model	kNN retrieval	learnable integration	ImageNet-S	CUB
(a)			80.1	61.9
(b)	✓		76.4	67.9
(c)		✓	75.6	60.0
MML	✓	✓	**83.0**	**75.6**

Table 4: Effect of different class prototypes

prototype	ImageNet-S	CUB
avg of all memory items	81.2	74.4
avg of random memory items	81.9	70.4
text only	82.7	69.1
image only	82.3	**76.8**
MML (image & text)	**83.0**	75.6

Results: Table 1 compares zero-shot baselines and MML on cross-dataset transfer. MML is trained with a few ImageNet samples but exhibits great performance on other datasets with extreme domain shifts, *e.g.*, from classifying general objects (Russakovsky et al., 2015) to land (Helber et al., 2019), *by simply replacing the memory* with the web-crawled domain-related knowledge. In particular, compared with the kNN classifier, which is uni-modal and non-learnable, our method meta-learns to integrate the multi-modal kNNs and effectively transfers to unseen classes. Table 2 compares MML and other zero-shot models on the zero-shot CUB benchmark (Akata et al., 2013) with 150 training classes and 50 test classes, where MML demonstrates its outstanding effectiveness compared to other models. While the existing models train with full training images and ground-truth attribute annotations (*e.g.*, eye colors), MML learns with *zero* human-annotated attributes, but shows great performance based on integrating the retrieved knowledge from the external memory. Comparing CLIP and ours examines the *significant advantage of external memory access and knowledge integration* where ours obtains a 7.6-11.0 % point accuracy improvement. Plus, we examine the efficacy of different backbones in Table 15 in Appendix, where MML consistently outperforms the others. The following paragraphs continue with more analyses and ablation studies on zero-shot classification.

4.3 Model analyses

We present the analyses of the model components. All experiments are based on CLIP ViT-B unless specified.

Ablation study on model components

Table 3 presents the ablation study of the main model components of MML. The first model (a) is a zero-shot prototype classifier. When the kNN retrieval is added without the learnable kNN integration, the model (b) corresponds to the kNN classifier (Nakata et al., 2022), which is beneficial on CUB compared to the model (a). The model (c) examines the learnable integration of the cross-attention module without kNN retrieval, thus transforming the input feature with the learnable self-attention. The worst result of (c) implies that the additional cross-attention is even harmful without the proper source of kNN knowledge integration. The last row with the two components (MML) achieves the highest performance on the two datasets.

Effect of different class prototypes

Table 4 compares different methods to build class prototypes. We first try to naïvely average all the contents in each memory to obtain class prototypes without using the top-M operator (Eq. 1). This average aggregation is likely to include plenty of unfiltered data noise, resulting in poor performance on both datasets. Next, "avg of random memory items" randomly selects the same number of items with the proposed cross-modal prototype method and averages them per class. This noisy and unrepresentative prototype leads poor classification. We also attempt to use the single-modality class prototype. The image prototype is more helpful than the text prototype on CUB and the reverse on ImageNet-S, suggesting that the efficacy of the image and text prototype can be dependent on target dataset characteristics. All these methods do not use image-text consensus while our method carefully selects memory items that exhibits the high cross-modal similarity for constructing class prototype. In this way, the prototype is comprised of the representative class data and also avoids potential data noise. Using the multi-modal prototypes, our model achieves robust performance.

Effect of different memory types

Table 5 validates our dual-branch image and text memory. The "no kNN" baseline has the same architecture as the proposed model, but instead, it feeds the input feature for the key and value inputs in replace of the

Table 5: Effect of different memory types

retrieval from	ImageNet-S	CUB
no kNN	75.6	60.0
text memory	82.3	69.4
image memory	76.2	71.2
unified memory	76.8	71.3
MML (separate memory)	**83.0**	**75.6**

Figure 3: Effect of memory size on ImageNet-S

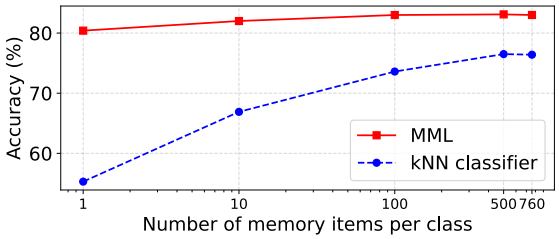

Table 6: Memory content robustness of MML. Memory replacement at testing time from one memory content to another.

image mem at train image mem at test	WebVision (WV) Google WV Google → WV Flickr	WebVision (WV) Flickr WV Flickr → WV Google
ImageNet-S	83.0 → 83.1 (+0.1)	83.2 → 83.2 (−0.0)
text mem at train text mem at test	**Wikipedia Wikipedia → text thumbnails**	**text thumbnails text thumbnails → Wikipedia**
ImageNet-S	83.0 → 82.7 (−0.3)	83.2 → 83.0 (−0.2)

Table 7: Memory content robustness comparison with different noise levels on classification (class) and retrieval

memory noiseness	kNN retrieval rec@1	 rec@16	class acc.
no memory	-	-	75.6
noisy (WebVision)	65.5	90.3	83.0
clean (ImageNet1K)	66.4	93.5	84.7

kNNs, *i.e.*, the query feature is shared with Q, K, V in Figure 2. This baseline exhibits the lowest performance and signifies the importance of the kNN knowledge integration. Next, we ablate either image or text memory. It is noticed that the model using only the image memory is more effective than the one using the text memory on CUB, while this trend is reversed on ImageNet-S. The opposite trend suggests that the vast and detailed visual knowledge collected from the internet is beneficial for fine-grained image classification, on the other hand, textual information is useful for coarse-grained classification of general objects. Lastly, we merge the image and text memory contents and then retrieve the modality-agnostic kNN features, which are then passed to a single knowledge integration branch. We observe that the majority of kNNs are from the image memory, thus closely matching the performance of the image-memory model. To effectively interact with multi-modal kNNs, we choose to separate the image/text memories. This result signifies that the dual-branch multi-modal knowledge integration is crucial in zero-shot unseen class generalization.

Effect of memory size

To validate the size of the memory, we set the memory size per class from 1 to full for two retrieval-based models, the kNN majority voting classifier (Nakata et al., 2022) and MML, and verify the performance growth in Figure 3 and Table 12. While the larger memory is more helpful, the performance reaches plateau with the abundant memory size. We thus claim that MML requires a moderated size of the retrieval pool as the retrieval and knowledge integration effectively incorporate the useful data from the noisy data source.

Robustness of memory

Table 6 shows that MML does not overfit to certain memory contents and performs robustly to different memory contents with little loss of performance. Note that MML already makes unseen class predictions with completely new memory contents of the new classes at the zero-shot testing phase (Tables 1-2). This experiment further examines whether the model performs robustly when *different instances of the same classes* are plugged into the memory. We equip a pair of image/text data collection from two different sources of the same target classes. For image memory, we use the two disjoint sets from the WebVision (WV) dataset; one set is obtained through Google crawling and the other from Flickr, where both of them are from ImageNet-S classes. For text memory, we use the text articles from Wikipedia and the text thumbnails of Google text keyword search. Once MML is trained with one source, we test it by plugging the two different memory sources. The results show that the model exhibits a marginal performance gap when replacing the test memory from one to another and also proves the modular property of the memory.

Table 7 presents the robustness of the memory when compared with the clean (human-annotated) and the noisy (web-crawled) memory. The use of web data inherently introduces the trade-off between avoiding

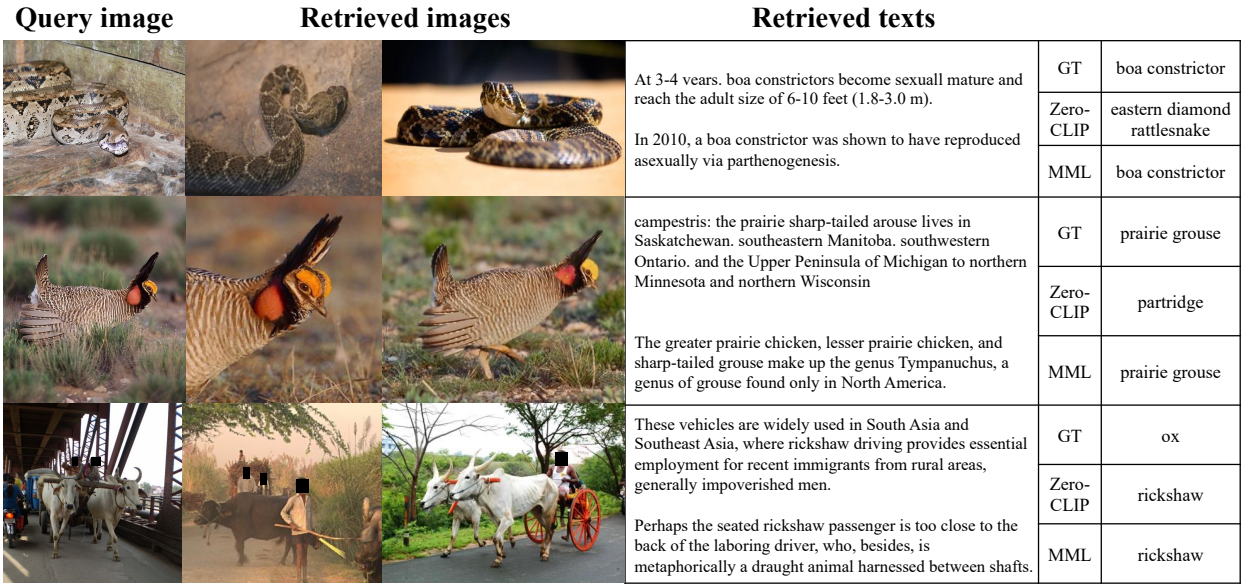

Figure 4: Examples of a query, image 2NNs and text 2NNs. Human faces are anonymized for visualization.

additional human annotation and data noise. Retrieval-based learning of MML is also for noise reduction. The retrieval-based knowledge integration remedies such problem by selecting the nearest, *i.e.*, the most relevant, samples to integrate them to establish the data and classifier prototype representation. To quantify the noisiness of the memory data pool, we also present the recall of retrieved items with the standard retrieval metric of recall@K (rec@K) (Jégou et al., 2011). The recall@K returns 1 if any instances from the ground-truth class are included in the kNN and 0 otherwise. Although the clean image memory enhances mid-level retrieval and end-task classification accuracy, the noisy memory model achieves comparable results to the upper bound. This experiment supports our modeling choice — utilizing web-crawled images as external memory — is reasonably effective and label-efficient compared to fully annotated memory.

Visualization of multi-modal kNN

Figure 4 visualizes the retrieved two image nearest neighbors (NNs) and two text NNs of the given input as well as the zero-shot CLIP prediction. Note that the images and texts in the example are independently retrieved from each memory. We notice that the image NNs often contain the query's noticeable visual patterns. From the text NNs, we observe that retrieved texts often contain synonymous keywords, *e.g.*, the scientific names of animals. The last example with cows contains multiple objects hence ambiguously class-labeled. In this case, MML is able to retrieve semantically related images and predicts a reasonable class than the ground truth.

4.4 Application to other classification setups

Few-shot to many-shot image classification

Problem setup: Few-shot classification (Fei-Fei et al., 2006; Vinyals et al., 2016) represents unseen classes with few-shot image samples for each target class during testing. We reuse ImageNet-S to make it a few-shot classification scenario by allowing access to additional 4 or 16 labeled images during validation and testing.

Baselines: Linear prob is the simplest few-shot classification baseline (Chen et al., 2019), where we add a class-length linear layer on top of the frozen backbone and train it with the given target class few-shot examples. ProtoNet (Snell et al., 2017) is another few-shot classification baseline, where the few-shot samples are averaged and used as a class prototype. We also compare ours with another memory-based classification model, Retrieval-Augmented Classification (RAC) (Long et al., 2022). RAC first retrieves the nearest images from an image memory and feeds their corresponding class text labels to the subsequent text encoder to obtain an auxiliary textual feature, which is then added to the input image feature. RAC was originally

Table 8: Few-shot classification on ImageNet-S

method	4-shot	16-shot
linear-prob CLIP (Radford et al., 2021)	72.1	80.6
ProtoNet (Snell et al., 2017)	76.4	76.5
RAC (Long et al., 2022)	66.8	78.1
kNN classifier (Nakata et al., 2022)	77.2	77.2
MML	**82.8**	**83.5**

Table 9: Few-shot to many-shot classification. TTT stands for test-time training with the given data (shots).

methods	TTT	4	16	64	128	256
(a) linear prob	✓	72.1	80.6	85.5	86.9	87.3
(b) MML		82.8	83.5	85.7	85.7	85.8
(c) MML*	✓	**83.4**	**87.0**	**87.4**	**87.9**	**88.3**
(b) - (a) gap		+10.8	+2.9	+0.2	-1.2	-1.5
(c) - (a) gap		+11.4	+6.4	+1.9	+1.0	+1.0

designed to be trained with abundant training data for long-tailed classification (Huang et al., 2016). We adapt RAC for few-shot classification and keep the text encoder frozen; otherwise, few-shot training fails to converge. All methods use the CLIP-B/32 backbone.

Results on few-shot classification: Table 8 compares MML and the aforementioned baselines on few-shot classification. While our MML outperforms the other methods, we observe that the performance gap between MML and the linear prob CLIP is bigger with fewer shots. This result implies that *the knowledge retrieval from external memory is especially effective when limited supervised data are available* as the external memory access can compensate for the lack of supervised data.

Extended results on many-shot classification: In addition, we attempt to increase the few shots from the target classes to many shots and demonstrate the performance trend in Table 9. While the linear prob (a) is directly trained with 4 to 256 shots from the target classes, our method (b) is trained on the non-target classes and tested *without* additional training with the 4 to 256 shots. MML outperforms linear prob by a significant margin with 4 shots, and the gain gradually diminishes with increasing training data for linear prob. The results of MML with the test-time training (c) with the 4 to 256 shots show that our model recovers the diminished gap and further improves performance. Note that this work primarily focuses on leveraging external knowledge with zero or minimal supervision, in addition, we also show that additional test-time training benefits MML orthogonally to retrieval-based reasoning.

Class-incremental classification

Problem setup: A class-incremental learning model is assumed to receive a set of new class data sequentially and is asked to classify a test image into the accumulated classes. As the model is not assumed to access to the previously seen data, the key challenge is not to forget the old classes.

Details on reproduction: For a fair comparison, we reproduce existing class-incremental learning methods (Li & Hoiem, 2017; Ratcliff, 1990; Rebuffi et al., 2017; Wang et al., 2022a) with the CLIP-B/32 backbone as well as ours on a unified codebase (Zhou et al., 2023a) by following the standard constraints.

Benchmark: We adopt a public benchmark, ImageNet100-Base0-Inc10 (Rebuffi et al., 2017), where 10 unseen classes and their annotated samples are sequentially given for 10 consecutive stages. For each stage, a model is required to classify an image into all the known classes, resulting in the accumulation of 100-class classification at the end. Across all stages, the size of the memory for each stage is always restricted to 2000 elements for all methods. For each stage, MML manages the memory length by dropping some old-class data from the memory and storing the new-class data such that the remaining memory elements are closest to the average of the memory contents, following (Rebuffi et al., 2017). Accordingly, MML updates the class prototypes with the updated memory elements at each stage. For evaluation, input images are classified into all the seen classes without stage-specific information.

Results: As seen in Figure 5, MML outperforms or performs on par with the class-incremental learning specialist models, without using specific techniques for the task such as distillation of old class knowledge in model weights (Rebuffi et al., 2017) or storing the heavy model weights to the model memory (Wang et al., 2022a). For CIFAR100, please see Figure 7 in Appendix.

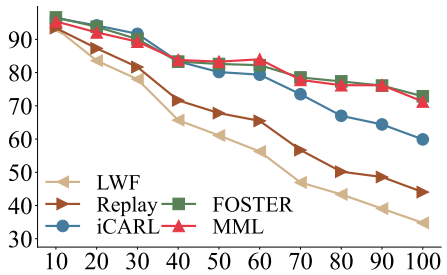

Figure 5: Class-incremental classification on ImageNet100

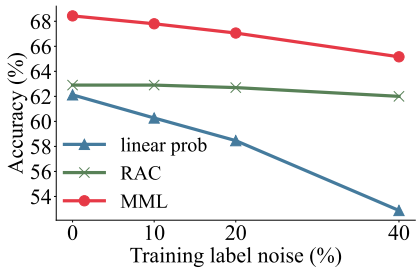

Figure 6: Result with training label noise

Training label noise robustness

We showcase that the reasoning procedure via memory retrieval is robust against the training data label noise. To simulate the label noise, we randomly permute from 10% to 40% of the class labels of training queries with a wrong class and train the architecture with the corrupted labels. This comparison validates the effectiveness of reasoning sources for classification: reasoning from the relevant external knowledge *vs.* reasoning from the memorized parameters. Figure 6 presents the comparison of the baselines and ours on ImageNet1K with the increasing portion of incorrect class labels. The memory-based models, RAC and MML, show robustness and powerful performance against training data noise. As MML predicts classes assisted by retrieving input-adaptive kNN from the frozen memory, MML can avoid directly compiling the wrong training data into parameters, particularly being more robust as the more incorrect label noise is injected in training. We hypothesize that retrieval-based reasoning encourages robust learning against the training label noise as the kNNs provide interactive reasoning with the neighborhood embeddings.

4.5 Discussion on scalability

MML is lightweight and introduces little computational overhead; feature extraction, kNN retrieval, and knowledge integration take 1129.9, 58.3, and 10.7 GFLOPs (94.2, 4.9, and 0.9 %), respectively. MML is efficient in that training with batch size 256 consumes only 2.2 GB GPU memory on a 2080Ti thanks to the frozen backbone and memory features. We verify that classification inference with MML scales up to 1000 classes at once which consumes only 3.4 GB memory on a single GPU. As MML is scalable with the increasing number of classes with the manageable size of memory, MML is expected to handle the dynamic number of classes for classification tasks in the real world.

5 Conclusion

We have presented the memory-modular learner and demonstrated its efficacy in various scenarios, investigating the memory-modular generalization for unseen classes. The experiments show that our memory-modular reasoning effortlessly generalizes to unseen classes with memory replacement and exhibits robustness to noisy memory data. We also frame our retrieval-based zero-shot classification as web-assisted zero-shot classification, which is believed to be more realistic in the future research with the growth of web-trained foundation models. Although the memory introduces space and time overhead than conventional models in the memory construction and retrieval, the overhead is greatly reduced thanks to our memory design; the memory consists of pre-computed features of the frozen encoder, which are reused for inference except for memory replacement. Although the memory retrieval brings a substantial gain, the use of the generalized vision encoder might block deploying on specialized target domains such as medical or industrial vision. These tasks necessitate specialized image-text encoders pre-trained on those domains to aim higher precision. We believe that memory-modular learning benefits various tasks in the areas of artificial intelligence beyond classification, leaving them for future work.

References

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. Infinite mixture prototypes for few-shot learning. In *Proc. International Conference on Machine Learning (ICML)*, 2019.

Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. Msdn: Mutually semantic distillation network for zero-shot learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2019.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 2004.

Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2006.

Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *Proc. International Conference on Machine Learning (ICML)*, 2020.

Peter Hart. The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory*, 14 (3):515–516, 1968.

Yangji He, Weihan Liang, Dongyang Zhao, Hong-Yu Zhou, Weifeng Ge, Yizhou Yu, and Wenqiang Zhang. Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Qibin Hou, Ming-Ming Cheng, Jiangjiang Liu, and Philip HS Torr. Webseg: Learning semantic segmentation from web searches. *arXiv preprint arXiv:1803.09859*, 2018.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *arXiv preprint arXiv:2302.11154*, 2023a.

Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Ahmet Iscen, Jeffrey Zhang, Svetlana Lazebnik, and Cordelia Schmid. Memory-efficient incremental learning through feature adaptation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

Ahmet Iscen, Alireza Fathi, and Cordelia Schmid. Improving image recognition by retrieving from web-scale image-text data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Retrieval-enhanced contrastive vision-text models. *Proc. International Conference on Learning Representations (ICLR)*, 2024.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proc. International Conference on Machine Learning (ICML)*. PMLR, 2021.

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 *Proc. International Conference on Machine Learning (ICML)*, 2021a.

Menglin Jia, Bor-Chun Chen, Zuxuan Wu, Claire Cardie, Serge Belongie, and Ser-Nam Lim. Rethinking nearest neighbors for visual classification. *arXiv preprint arXiv:2112.08459*, 2021b.

Herve Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2011.

Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *Proc. International Conference on Learning Representations (ICLR)*, 2020.

Namyup Kim, Taeyoung Son, Jaehyun Pahk, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Wedge: web-image assisted domain generalization for semantic segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9281–9288. IEEE, 2023.

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

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)*, 2013.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.

Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, 2008.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(12):2935–2947, 2017.

Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Kengo Nakata, Youyang Ng, Daisuke Miyashita, Asuka Maki, Yu-Chieh Lin, and Jun Deguchi. Revisiting a knn-based image classification system with high-capacity storage. In *Proc. European Conference on Computer Vision (ECCV)*, 2022.

M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop Autodiff*, 2017.

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 *Proc. International Conference on Machine Learning (ICML)*, 2021.

Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. International Conference on Machine Learning (ICML)*. PMLR, 2019.

Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *Proc. European Conference on Computer Vision (ECCV)*. Springer, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, 2011.

Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *Proc. European Conference on Computer Vision (ECCV)*, 2022a.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 2020.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Yixuan Wei, Yue Cao, Zheng Zhang, Houwen Peng, Zhuliang Yao, Zhenda Xie, Han Hu, and Baining Guo. iclip: Bridging image classification and contrastive language-image pre-training for visual recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4582–4591, 2017.

Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision (IJCV)*, 2016.

Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xiaodong Yu and Yiannis Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *Proc. European Conference on Computer Vision (ECCV)*, 2010.

Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. Episode-based prototype generating network for zero-shot learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. *arXiv preprint arXiv:2302.03648*, 2023a.

Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, and De-Chuan Zhan. A model or 603 exemplars: Towards memory-efficient class-incremental learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2023b.

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

Zhen Zhu, Weijie Lyu, Yao Xiao, and Derek Hoiem. Continual learning in open-vocabulary classification with complementary memory systems. *arXiv preprint arXiv:2307.01430*, 2023.