Memory-Modular Classification: Novel-Class Generalization with Web-Crawled Memory

Dahyun Kang[1] Ahmet Iscen[2] Eunchan Jo[1] Sua Choi[1] Minsu Cho[1] Cordelia Schmid[2]
Pohang University of Science and Technology[1] Google Deepmind[2]

Abstract

We propose a memory-modular learner for zero-shot image classification that separates knowledge memorization from reasoning. Our model enables generalization to novel visual concepts 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 visual concepts by simply replacing the memory contents. The key differentiator is that our learner meta-learns to perform classification tasks with web-crawled data for classifying novel visual concepts. 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 [1, 2, 13, 25, 31] are trained on massive datasets using immense computational resources. They result 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 [5, 9] enhances image representations with their relevant data retrieved from an external knowledge source. Such 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.

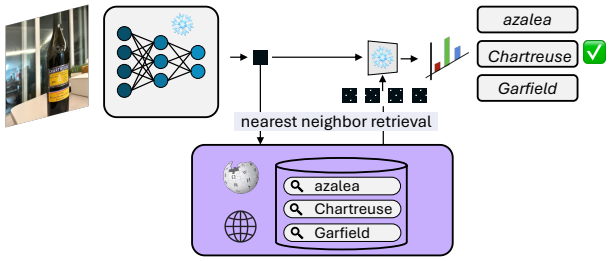

Figure 1. **Memory-modular learner (MML) for web-assisted zero-shot classification.** MML classifies novel classes by loading the relevant contents crawled from internet into the memory.

Inspired by this, we introduce a 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. 2). 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 that MML remains applicable as external knowledge evolves.

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 memory.
- We investigate the generalizability in adapting to new visual concepts by replacing the memory with related content, without tuning the model weights.
- 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

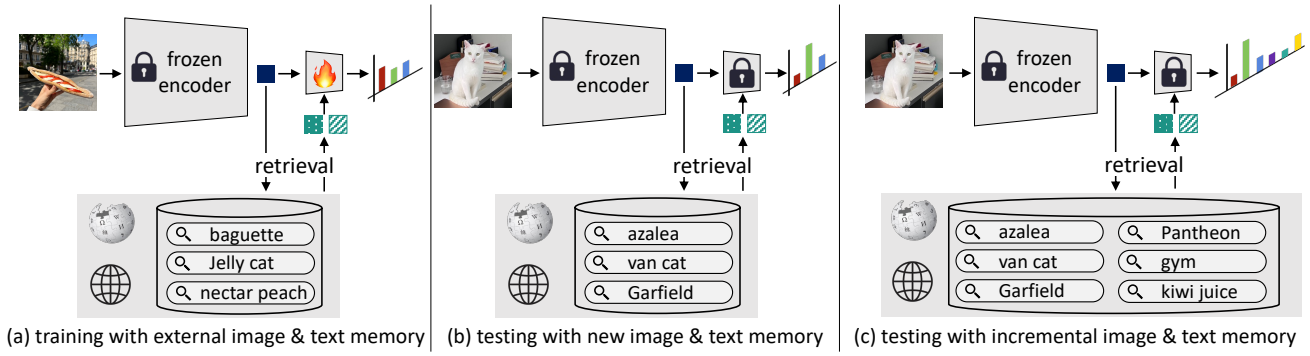

Figure 2. **Training and evaluation stages of MML.** MML constructs image/text memory with text keyword search on the internet given target classes. The memory provides relevant image/text features which are integrated via a trainable knowledge integration module (a). On evaluation, the memory can be 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.

knowledge collected from web.

2. Memory-modular learner

We address the problem of classifying an 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.

The memory-modular learner starts by loading the knowledge memory and generating class prototypes for target classes (Sec. 2.1). 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. 2.2). 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.

2.1. Memory and class prototype construction

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 [7, 12]. We follow a similar strategy for text memory. In this work, textual information relevant to each target class name is retrieved by querying Wikipedia [8, 17, 24]. 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 [23] 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:

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

where $\text{argmax}^{M}_{\boldsymbol{s}\in\mathcal{S}}(\cdot)$ denotes the top-M operator that returns the best M items from the set $\mathcal{S}$ maximizing the operand function. Likewise, the text prototype is obtained using the average text-to-image similarity.

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.

Table 1. Zero-shot cross-dataset transfer. MML is trained with 1 or 4 samples from ImageNet1K classes and tested on 10 fine-grained datasets with zero shot.

method	ImgNet1K	Caltech101	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	avg.
		objects	pets	cars	flowers	food	airplanes	scenes	textures	land	actions	
zero-shot CLIP [20]	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 [18]	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**

Table 2. Few-shot classification on ImageNet-S

method	4-shot	16-shot
linear-prob CLIP [20]	72.1	80.6
ProtoNet [23]	76.4	76.5
RAC [16]	66.8	78.1
kNN classifier [18]	77.2	77.2
MML	**82.8**	**83.5**

2.2. 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.

Retrieval. For an input image feature $\boldsymbol{f}$ extracted from the image encoder, its k-nearest-neighbor (kNN) image items are retrieved based on cosine similarity with all image memory items of all target classes. The text kNNs are 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 [10, 26] and then integrated with the input embedding $\boldsymbol{f}$. The cross-attention learns to integrate the nearest neighbor (NN) features into the input feature:

$$\boldsymbol{f}^{\text{img}} = \boldsymbol{f} + \sigma\left(\frac{\mathbf{Q}(\boldsymbol{f}) \cdot [\mathbf{K}(\boldsymbol{v}_k)]_{k=1}^K}{\sqrt{d}}\right)[\mathbf{V}(\boldsymbol{v}_k)]_{k=1}^K, \quad (2)$$

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 [18].

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:

$$\boldsymbol{z}_c = \cos(\boldsymbol{p}_c^{\text{txt}}, \boldsymbol{f}^{\text{txt}}) + \cos(\boldsymbol{p}_c^{\text{img}}, \boldsymbol{f}^{\text{img}}). \quad (3)$$

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

2.3. Training

MML is trained with cross-entropy loss with one-hot ground-truth class label $\boldsymbol{y}$ and class probability. Note that we freeze the pre-trained image-text encoder and train the remaining parameters only, *i.e.*, those of attention layers on the image and text branches. The number of training parameters and the frozen CLIP is 6.3M and 151M, respectively.

3. Experiments

3.1. Experimental setup

Training details. For the image/text feature extractor, we use the pre-trained CLIP [20]. Unless specified, CLIP-B/32 is used. For training, we use a batch size of 256 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.

Memory and data. To construct the external image memory for ImageNet derivatives, we employ a readily available web-crawled image dataset, WebVision ver. 2 [14]. WebVision is collected from Google and Flickr by the keyword search of the 1000 class names of ImageNet1K [22]. 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 [19] built-in `topK` module.

3.2. 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 for web crawling.

Baselines: The kNN classifier [18] retrieves kNN of the input from memory[1] and immediately predict the class by majority voting. Zero-shot CLIP 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 [3, 29, 30] are also compared, which are trained with the total 8885 annotated images and text attributes of CUB.

Results: Table 1 compares zero-shot baselines and MML on cross-dataset transfer. MML is trained with a few Im-

[1]The original work [18] 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
(a)			80.1
(b)	✓		76.4
(c)		✓	75.6
MML	✓	✓	**83.0**

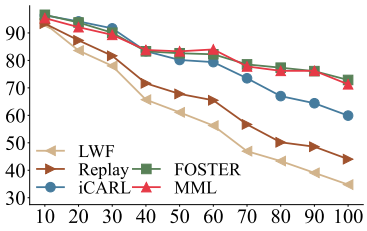

Figure 3. Class-incremental classification.

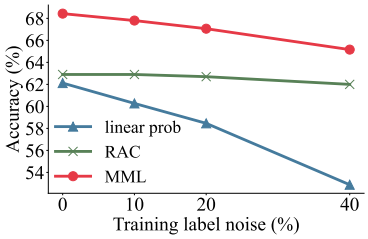

Figure 4. Result with training label noise.

ageNet samples but exhibits great performance on other datasets with extreme domain shifts, *e.g.*, from classifying general objects [22] to land [6], *by simply replacing memory* with the web-crawled memory. 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 visual concepts.

3.3. Few-shot/class-incremental classification

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

Few-shot classification. Few-shot classification [4, 27] represents target classes with few-shot image samples during testing. Table 2 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.

Class-incremental classification. 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. As seen in Fig. 3, MML outperforms or performs on par with the class-incremental learning specialists, without using specific techniques for the task such as distillation of old class knowledge in model weights [21] or storing the heavy model weights to the model memory [28].

3.4. Ablation study

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 [18]. 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.

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 that reasoning from the relevant external knowledge is more effective than reasoning from the memorized parameters. Figure 4 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, 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. Conclusion

We have presented the memory-modular learner and demonstrated its generalization ability for comprehending novel visual concepts. The experiments show that MML generalizes to unseen classes with memory replacement and exhibits robustness to noisy data. We frame the retrieval-based zero-shot classification as **web-assisted zero-shot classification**, which is believed to be promising with the advancements of web-trained foundation models. MML can benefit various areas beyond classification such as dense prediction tasks requiring visual concept discovery [11, 15], leaving them for future work.

Acknolwedgements. This work was supported by Samsung Electronics Co., Ltd. (IO201208-07822-01) and the IITP grants (RS-2022-II220290: Visual Intelligence for Space-time Understanding (30%), RS-2024-00457882: National AI Research Lab Project (20%), RS-2022-II220113: Sustainable Collaborative Multi-modal Lifelong Learning (50%)) funded by Ministry of Science and ICT, Korea.

References

[1] 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. 1

[2] 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. 1

[3] 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. 3

[4] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2006. 4

[5] 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. 1

[6] 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. 4

[7] Qibin Hou, Ming-Ming Cheng, Jiangjiang Liu, and Philip HS Torr. Webseg: Learning semantic segmentation from web searches. *arXiv preprint arXiv:1803.09859*, 2018. 2

[8] 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*, 2023. 2

[9] 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)*, 2023. 1

[10] 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. 3

[11] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*. Springer, 2024. 4

[12] 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)*, pages 9281–9288. IEEE, 2023. 2

[13] 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. 1

[14] 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. 3

[15] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proc. European Conference on Computer Vision (ECCV)*. Springer, 2024. 4

[16] 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. 3

[17] 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. 2

[18] 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. 3, 4

[19] 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. 3

[20] 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. 3

[21] 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. 4

[22] 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. 3, 4

[23] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3

[24] 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. 2

[25] 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. 1

[26] 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. 3

[27] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 4

[28] 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)*, 2022. 4

[29] 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. 3

[30] 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. 3

[31] 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. 1