ATTENTION PROBE: VISION TRANSFORMER DISTILLATION IN THE WILD

Jiahao Wang[1] *Mingdeng Cao*[1] *Shuwei Shi*[1] *Baoyuan Wu*[2] *Yujiu Yang*[1*]

[1] Tsinghua Shenzhen International Graduate School, Shenzhen, China
[2] The Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

Vision transformers (ViTs) require intensive computational resources to achieve high performance, which usually makes them not suitable for mobile devices. A feasible strategy is to compress them using the original training data, which may be not accessible due to privacy limitations or transmission restrictions. In this case, utilizing the massive unlabeled data in the wild is an alternative paradigm, which has been proved effective for compressing convolutional neural networks (CNNs). However, due to the significant differences in model structure and computation mechanism between CNNs and ViTs, it is still an open issue that whether the similar paradigm is suitable for ViTs. In this work, we propose to effectively compress ViTs using the unlabeled data in the wild, consisting of two stages. First, we design an effective tool in selecting valuable data from the wild, dubbed *Attention Probe*. Second, based on the selected data, we develop a probe knowledge distillation algorithm to train a lightweight student transformer, through maximizing the similarities on both the outputs and intermediate features, between the heavy teacher and the lightweight student models. Extensive experimental results on several benchmarks demonstrate that the student transformer obtained by the proposed method can achieve comparable performance with the baseline that requires the original training data. Code is available at: `https://github.com/IIGROUP/AttentionProbe`.

***Index Terms*—** Transformer, data-free, distillation

1. INTRODUCTION

Vision transformer [1, 2] has prevailed in a series of computer vision tasks due to its superior capability in capturing long-distance dependencies based on self-attention mechanism [1, 3]. Modern vision transformers, such as ViT [4], DeiT [5], and IPT [6], are able to learn high-dimensional semantic information [4, 5, 7] or low-dimensional texture information [8, 9, 6] from video or images and achieve comparable performance compared to CNNs. However, the promising performance comes at a high computational cost, which blocks them from being deployed on mobile devices, *e.g.*, smartphones, cameras, and micro-robots. To this end, various attempts have been made to compress and accelerate the pre-trained heavy networks, including quantization [10, 11], pruning [12, 13] and knowledge distillation [14, 5]. Existing model compression techniques can obtain compressed models with low computation complexity, but their basic premise is that we can access the whole original training data to fine-tune the svelte networks to recover the performance. However, due to privacy constraints and transmission limitations, the original data is often unavailable, especially those related to personal privacy, such as face ID, voice and fingerprint. Therefore, recent trend of model compression technology is to pack the pre-trained heavy models in the absence of original data [15, 16, 17, 18, 19]. However, existing works focus mainly on CNNs, and distilling portable student transformers in the wild dataset remains an open issue. Since their calculation paradigm is completely different, directly transporting the existing success on CNNs to transformers will inevitably incur accuracy degradation.

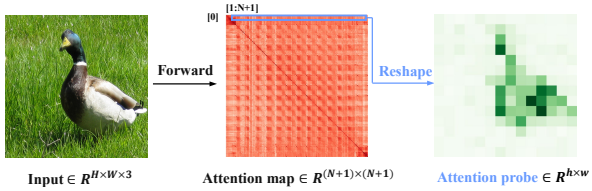

Fig. 1. Calculation paradigm of the attention probe, where $N = HW/p^2 = hw$ denotes patch number of the image.

To this end, we propose an *Attention Probe* method to utilize a large amount of unlabeled data in the wild to complete the data-free distillation task. Images most relevant to the original dataset will be selected from the massive wild data (e.g., Flickr [20]) to conduct the distillation task. Different from CNN case [17] that samples with higher confidence scores provided by the teacher output, we tend to use attention probe to select the appropriate training data for vision transformer. The attention probe is defined as a special vector in the attention map of the transformer, as a probe of it. It is a special representation of the image, representing the key area the class token tends to focus on when the model completes the classification task. To further improve the performance, an attention probe-based knowledge distillation is exploited for extracting feature information from the teacher. The portable student model is optimized by making its output and interme-

*Corresponding author. Email: yang.yujiu@sz.tsinghua.edu.cn

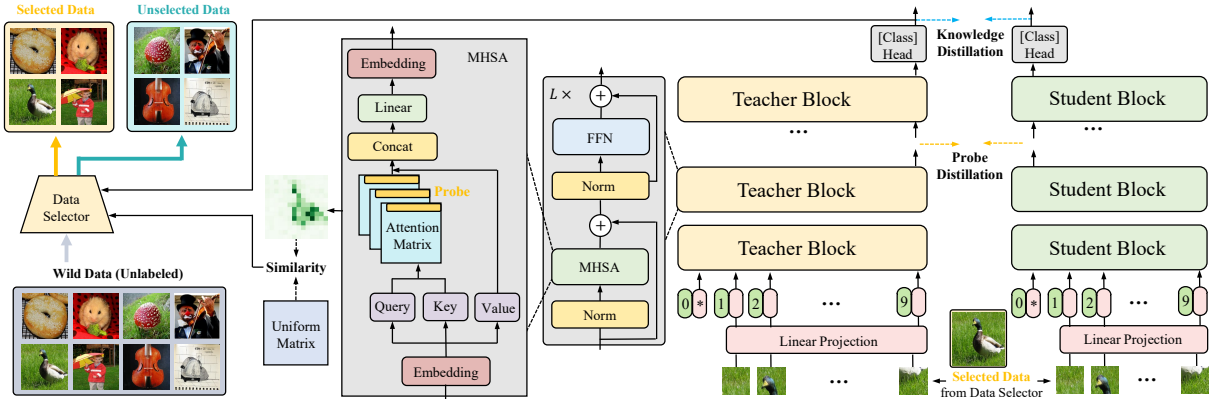

Fig. 2. Illustration of the proposed method for learning student vision transformer in the wild. The data selector utilizes the output of the teacher network and the attention probe to pick out useful data from the wild data and use it for the next distillation process. Probe distillation is adopted for exploiting the information of the intermediate features of the pre-trained heavy transformer.

diate attention probe close to the teacher, as shown in Fig. 2. Experiments demonstrate the proposed framework can surpass all data-free distillation methods designed for CNNs, and the accuracy of the learned portable student is comparable to or even surpass that of the student trained using original data.

2. METHOD

2.1. Attention Probe

In a standard transformer module, the input sequence $X \in \mathbb{R}^{(N+1)\times D}$ are first applied with three different linear transformations and output queries, keys and values $Q, K, V \in \mathbb{R}^{M\times(N+1)\times d_h}$, where N is the number of tokens, concatenated with the `[class]` token. M is head number, d_h is the dimension of queries, keys and values for each head. Generally, $D = Md_h$ and the multi-head self-attention is based on the pairwise similarity between two elements of the sequence:

$$\mathrm{A}(Q, K, V) = \mathrm{softmax}\Big(\frac{QK^T}{\sqrt{d_h}}\Big)V, \tag{1}$$

where $\mathrm{A} \in \mathbb{R}^{M\times(N+1)\times(N+1)}$ represents the relationship of each patch to all the other N patches in the input sequence. Given random input x^U from the wild data, we define the *attention probe* as a special section of the attention map:

$$\mathrm{P}(x^U) = \gamma(\mathrm{A}(x^U)[0, 1 : N + 1]), \tag{2}$$

where $\mathrm{P} \in \mathbb{R}^{h\times w}$, $N = \frac{HW}{p^2} = hw$, p is patch size and γ means the reshape operation which convert $\mathbb{R}^{1\times N}$ to $\mathbb{R}^{h\times w}$. As shown in Fig. 1, the attention map of each layer $\mathrm{A} \in \mathbb{R}^{M\times(N+1)\times(N+1)}$ is first obtained by forward propagation of the input image. After that, the last N elements of the first row of the attention map are extracted and reshaped. The attention probe represents how much attention the model's `[class]` token pays to the other N image patches. It will be utilized to collect data from the massive unlabeled wild data.

2.2. Data Collection for Knowledge Distillation

Direct data distillation using the whole wild data is not only time-consuming, but also cannot guarantee the performance of the student network. Since the wild data contains many noisy data that are not related to the original data distribution, using them to optimize the student model can lead to inappropriate optimization direction. To this end, we aim to collect valuable data $\hat{x}^U$ whose distribution is close to the original distribution from the wild X^U to guide the distillation process. Our general goal can be formulated as follows [17]:

$$\hat{x}^U = \arg\min_{\hat{x}^U \in X^U} D_{KL}(\mathcal{N}_S(\hat{x}^U), \mathcal{N}_T(x^O)). \tag{3}$$

where $\mathcal{N}_S$ denotes the student transformer trained by the selected data $\hat{x}^U$. However, the data collection principle is intractable since the original data distribution x^O is unaccessible. Thus, we propose the surrogate principle to collect data based on the attention probe.

Given a random sample x^U from the unlabeled wild dataset, the attention probe of the pre-trained transformer is $\mathrm{P}(x^U)$. The attention value of x^U can be expressed as:

$$V_a(x^U) = \mathrm{sim}(\mathrm{P}(x^U), S_N), \tag{4}$$

where sim denotes cosine similarity, and $S_N \in \mathbb{R}^{h\times w}$ is a uniform matrix whose values are uniformly $\frac{1}{N}$. A useful sample x^U is expected to have a small attention value $V_a(x^U)$.

The intuitions behind the attention value are straightforward. We have a teacher transformer that is well trained on the original data, which shows better recognition capacity for data close to the original distribution. In other words, the teacher's `[class]` token knows which area of the input image should be paid attention to in order to complete the recognition task. Thus, the variance of the values in the attention probe is usually large and much less similar to a uniform matrix S_N. In contrast, when the sampled data has a lower probability of being from the original distribution, the pre-trained teacher is less sensitive to the key region of the input image, so the

variance of the attention probe is lower and is more close to S_N. Thus the approach can hinder many out-of-distribution jamming data in the wild from being collected.

Besides, samples with higher confidence scored by the teacher are more likely to ensure the student network achieve a good performance [17], which can be formulated as:

$$V_n(x^U) = D_{KL}(\mathcal{N}_T(x^U), \hat{\boldsymbol{y}}^{\boldsymbol{U}}), \tag{5}$$

where $\hat{\boldsymbol{y}}^{\boldsymbol{U}}$ is the one-hot pseudo label of x^U whose valid encoding is $\arg\max_i \mathcal{N}_T(y = y_i|x^U)$, which is predicted by the teacher transformer $\mathcal{N}_T$. By combining Eq. 4 and Eq. 5, we obtain the final value function as:

$$V(x^U) = \lambda_a V_a(x^U) + \lambda_n V_n(x^U), \tag{6}$$

In practice, we select samples with low value defined in Eq. 6, which means that samples with a fluctuating attention probe and a high output confidence score are preferred.

2.3. Probe Distillation using Intermediate Features

In this section, we discuss how to distill the student using intermediate features of the teacher model. We are motivated by the lack of supervised information in the data-free scenario since the collected data is unlabeled. As a result, the lack of the cross entropy between the student output with corresponding ground-truth labels will inevitably incur information loss and accuracy degradation. To compensate for the loss of supervised information, we embrace the intermediate information of the pre-trained teacher transformer. The intermediate layers contain more embedding, supplement richer features, thus allow the student transformer to acquire more information in addition to outputs. A straightforward method to utilize the intermediate information is to transfer the embedding features between teacher and student directly. However, there exists a limitation that their feature dimension must be the same.

We address the problem by introducing a patch-level relationship into knowledge distillation. By considering the transformer as a feature extractor, we can obtain the patch embedding $\mathcal{F} \in \mathbb{R}^{(N+1)\times D}$ of each intermediate layer. The inner patch relationship contains much information, which helps the student network to learn the relationship between patch embedding of the pre-trained teacher. Specifically, for a certain intermediate layer of the student network, we define the correlation between patches as follows:

$$\mathcal{R}_S = \frac{\mathcal{F}_S}{\|\mathcal{F}_S\|_2} \cdot \left(\frac{\mathcal{F}_S}{\|\mathcal{F}_S\|_2}\right)^\top, \tag{7}$$

where $\mathcal{F}_S$ denotes the intermediate feature of a specific layer of the student, and $\mathcal{R}_S \in \mathbb{R}^{(N+1)\times(N+1)}$ represents the inter-relationship between image patches. Our goal is to transfer the relationship between the `[class]` token and the other N patches of the pre-trained teacher model, *i.e.* , the attention probe is to enrich the learning of the lightweight transformer. Thus, we derive our objective for probe distillation as:

$$\mathcal{L}_P(\mathcal{N}_S) = \|\mathcal{R}_T(x)[0, 1:N+1] - \mathcal{R}_S(x)[0, 1:N+1]\|_2^2, \tag{8}$$

Note that in Eq. 8 instead of having the inner patch relationship of each intermediate feature of the student network exactly mimic the teacher network, we simply require the student network to maintain the attention weight on the other N patches of the `[class]` token that completes the recognition task.

We also follow [17] to establish a noisy adaption matrix Q after the softmax layer of the student network to transform the probability of true label to the noisy probability. Considering the common knowledge distillation term $\mathcal{L}_{KD}$, the student transformer is optimized through the selected data:

$$\mathcal{L}_{KD}(\mathcal{N}_S) = D_{KL}(\mathcal{N}_S(x^U), \mathcal{N}_T(x^U)). \tag{9}$$

Thus, the objective for probe distillation can be formulated as:

$$\mathcal{L}_{PD}(\mathcal{N}_S) = \mathcal{H}_{CE}(Q(\mathcal{N}_S(x)), \hat{\boldsymbol{y}}) + \alpha\mathcal{L}_P(\mathcal{N}_S) + \beta\mathcal{L}_{KD}(\mathcal{N}_S), \tag{10}$$

where $\hat{\boldsymbol{y}}$ is teacher predicted one-hot pseudo label with valid encoding $\arg\max_i \mathcal{N}_T(y = y_i|x)$, α, β are hyper-parameters.

3. EXPERIMENTS

3.1. Experiments on MNIST

The widely used MNIST benchmark is first selected as the original dataset. We use EMNIST [22] as the unlabeled dataset, which contains 814K images in the same data format as MNIST. 500,000 images are selected from the wild dataset by the attention probe method. We take DeiT-XSmall with an embedding dimension of 384 and 3 heads as the teacher, and take DeiT-XTiny with an embedding dimension of 128 and 2 heads as the student. The patch size of the teacher and student transformer is set as 4. Both are optimized using AdamW [23] optimizer for 100 epochs with a 256 batch size. The initial learning rate and weight decay are 7×10^{-4} and 0.025. We used the attention probe in layers 4 and 8 to help collect useful data with their hyper-parameter λ_a in Eq. 6 are both set as 0.05. λ_n is set as 0.9. α and β in Eq. 10 are 0.8 and 1, respectively. Detailed results are shown in Table 2. The teacher transformer achieves a 99.39% accuracy and the student achieves a 99.06% accuracy when trained from scratch or a 99.04% accuracy when directly performing distillation on the original MNIST dataset. We also compare with the method without using the original data. DFND [17] develops an output confidence-based data selection method and a noisy knowledge distillation algorithm to improve student performance. Our method performs better than DFND and achieves better results than training with the original dataset, demonstrating the effectiveness of data selection and knowledge distillation based on the attention probe for vision transformer.

3.2. Experiments on CIFAR

We then validate our method on the CIFAR-10 and CIFAR-100 datasets. The teacher and student transformer are the same as the experiments on MNIST. ImageNet dataset [24] is chosen as

Table 1. Classification result on the CIFAR dataset.

Algorithm	Required data	FLOPS	#params	CIFAR-10	CIFAR-100
Teacher	Original data	∼1.38G	∼21.3M	96.65%	76.30%
Student	Original data	∼153M	∼2.38M	86.31%	65.46%
Knowledge Distillation [21]	Original data	∼153M	∼2.38M	87.66%	67.38%
DAFL [15]	No data	∼153M	∼2.38M	56.12%	40.49%
DFND [17]	Unlabeled data	∼153M	∼2.38M	93.14%	71.24%
Attention Probe	**Unlabeled data**	**∼153M**	**∼2.38M**	**93.95%**	**71.82%**

Table 2. Classification result on the MNIST dataset.

Algorithm	FLOPS	#params	Accuracy
Teacher	∼1.06G	∼21.3M	99.39%
Student	∼118M	∼2.38M	99.06%
KD [21]	∼118M	∼2.38M	99.04%
DFND [17]	∼118M	∼2.38M	98.84%
Attention Probe	**∼118M**	**∼2.38M**	**99.07%**

the unlabeled dataset, and all images are resized to $32 \times 32 \times 3$ to fit the input size of the transformer. We use the AdamW [23] and cosine learning rate decay with an initial learning rate of 7.5×10^{-4}. We select 650,000 images from the wild data and train 300 epochs. Hyper-parameters in Eq. 6 and Eq. 10 are the same as experiments on MNIST. The experimental results are shown in Table 1. For CIFAR-10 results, the teacher transformer achieves a 96.65% accuracy and the student achieves a 86.31% accuracy training from scratch or a 87.66% accuracy with direct knowledge distillation on the original CIFAR-10 dataset. Chen *et.al.*proposed DAFL [15] to generate fake original training data through the pre-trained network. However, we show that DAFL only achieves a 56.12% accuracy, which demonstrates that DAFL does not show good transferability to ViTs. Our method performs better than DFND and achieves a 93.95% accuracy, surpassing distillation on the original dataset. Besides CIFAR-10, we further validate the effectiveness of the proposed method on the CIFAR-100. The student transformer learned by exploiting the attention probe obtained a 71.82% accuracy with the original data unavailable.

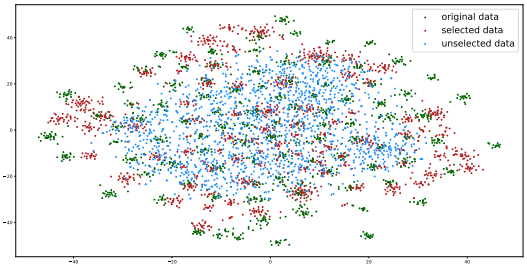

Fig. 3. T-SNE of teacher's output feature of the data selected by the attention probe, the original and unselected data.

3.3. Experiments on ImageNet

We then conduct experiments on the ImageNet dataset. We choose Flicker1M as the unlabeled dataset. The experimental setting is the same as those in the CIFAR-10 experiments, except only 10% of the original ImageNet is used. The teacher transformer achieves a 95.6% Top-5 accuracy and the student transformer achieves a 91.1% Top-5 accuracy when trained from scratch or a 91.9% accuracy with direct knowledge distillation on the original ImageNet. For the data-free setting, our method achieves a 78.8% Top-5 accuracy. Note that the size of the wild dataset we use is still much smaller than the original ImageNet dataset. Our method still achieves a promising result when using only a tiny part of the original dataset.

Table 3. Ablation study of the proposed attention probe.

Setting	AP	PD	CIFAR-10	CIFAR-100
M1			93.14%	71.24%
M2	✓		93.86%	70.74%
M3		✓	93.45%	70.25%
Ours	✓	✓	**93.95%**	**71.82%**

3.4. Ablation Study

Visualization of Features. To demonstrate the effectiveness of our attention probe-based data collection approach, we visualize the output softmax of the teacher transformer through the T-SNE method. We randomly select 1,500 images from the CIFAR-100 dataset and feed them into the pre-trained DeiT-XSmall model. As shown in Fig. 3, selected data through the attention probe shows a similar distribution to the original data. In contrast, the distribution of the unselected data is haphazard.

Ablation Study. We also conduct ablation study to illustrate the effectiveness of the attention probe. All experiments are done with the DeiT-XTiny network on the CIFAR dataset. The training settings are the same as those in Section 3.2. As is shown in Table 3, AP denotes data selection with the attention probe approach, and PD denotes probe distillation. M1 denotes none of the components are utilized and achieves a 93.14% accuracy on CIFAR-10. When further implementing probe distillation or collecting data with the attention probe, we observe a 0.31% or 0.72% accuracy improvement, which demonstrates the effectiveness of the proposed method.

4. CONCLUSION AND ACKNOWLEDGEMENTS

We propose the attention probe concept in order to select useful data from the wild, and propose a probe distillation approach to utilize the intermediate information of the pre-trained teacher. Extensive experiments demonstrate that we can learn portable transformers with the original data unavailable.

This research was supported by the Key Program of the National Natural Science Foundation of China under Grant No. U1903213, the Shenzhen Key Laboratory of Marine IntelliSense and Computation under Contract ZDSYS20200811142605016.

5. REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[2] Han Shu, Jiahao Wang, Hanting Chen, Lin Li, Yujiu Yang, and Yunhe Wang, "Adder attention for vision transformer," *NeurIPS*, 2021.

[3] Shuwei Shi, Qingyan Bai, Mingdeng Cao, Weihao Xia, Jiahao Wang, Yifan Chen, and Yujiu Yang, "Region-adaptive deformable network for image quality assessment," in *CVPRW*, 2021.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[5] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao, "Pre-trained image processing transformer," in *CVPR*, 2021.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.

[8] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Swinir: Image restoration using swin transformer," *arXiv preprint arXiv:2108.10257*, 2021.

[9] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu, "Uformer: A general u-shaped transformer for image restoration," *arXiv preprint arXiv:2106.03106*, 2021.

[10] Song Han, Huizi Mao, and William J Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," 2016.

[11] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *CVPR*, 2018.

[12] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf, "Pruning filters for efficient convnets," *ICLR*, 2016.

[13] Yunhe Wang, Chang Xu, Shan You, Dacheng Tao, and Chao Xu, "Cnnpack: packing convolutional neural networks in the frequency domain," in *NIPS*, 2016.

[14] Hanting Chen, Yunhe Wang, Han Shu, Changyuan Wen, Chunjing Xu, Boxin Shi, Chao Xu, and Chang Xu, "Distilling portable generative adversarial networks for image translation," in *AAAI*, 2020.

[15] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian, "Data-free learning of student networks," in *ICCV*, 2019.

[16] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *CVPR*, 2020.

[17] Hanting Chen, Tianyu Guo, Chang Xu, Wenshuo Li, Chunjing Xu, Chao Xu, and Yunhe Wang, "Learning student networks in the wild," in *CVPR*, 2021.

[18] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner, "Data-free knowledge distillation for deep neural networks," *arXiv preprint arXiv:1710.07535*, 2017.

[19] Suraj Srinivas and R Venkatesh Babu, "Data-free parameter pruning for deep neural networks," *arXiv preprint arXiv:1507.06149*, 2015.

[20] Mark J Huiskes and Michael S Lew, "The mir flickr retrieval evaluation," in *SIGMM*, 2008.

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[22] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik, "Emnist: Extending mnist to handwritten letters," in *IJCNN*, 2017.

[23] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *ICLR*, 2019.

[24] 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," *IJCV*, 2015.