MEAL V2: Boosting Vanilla ResNet-50 to 80%+ Top-1 Accuracy on ImageNet without Tricks*

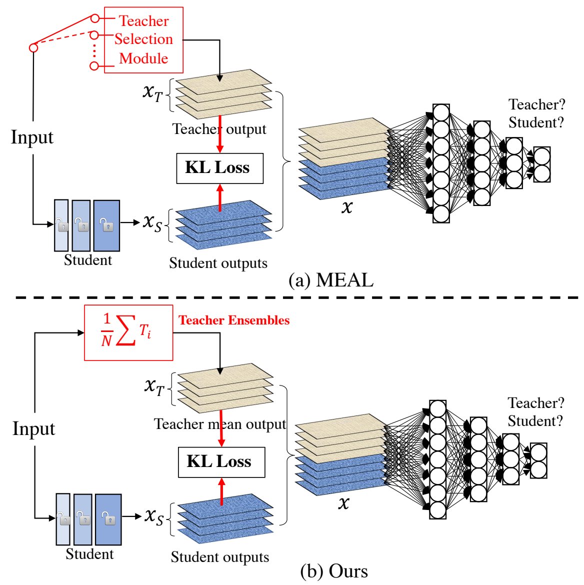
Zhiqiang Shen Marios Savvides
Carnegie Mellon University
{zhiqians,marioss}@andrew.cmu.edu

Abstract

*In this paper, we introduce a simple yet effective approach that can boost the vanilla ResNet-50 to 80%+ Top-1 accuracy on ImageNet without any tricks. Generally, our method is based on the recently proposed MEAL [18], i.e., ensemble knowledge distillation via discriminators. We further simplify it through 1) adopting the similarity loss and discriminator only on the final outputs and 2) using the average of softmax probabilities from all teacher ensembles as the stronger supervision for distillation. One crucial perspective of our method is that the one-hot/hard label should not be used in the distillation process. We show that such a simple framework can achieve state-of-the-art results without involving any commonly-used techniques, such as 1) architecture modification; 2) outside training data beyond ImageNet; 3) autoaug/randaug; 4) cosine learning rate; 5) mixup/cutmix training; 6) label smoothing; etc. On ImageNet, our method obtains **80.67%** top-1 accuracy using a single crop-size of 224×224 on the vanilla ResNet-50, outperforming the previous state-of-the-arts by a remarkable margin under the same network structure. Our result can be regarded as a new strong baseline on ResNet-50 using knowledge distillation. To our best knowledge, this is the first work that is able to boost vanilla ResNet-50 to surpass 80% on ImageNet without architecture modification or additional training data. Our code and models are available at: https://github.com/szq0214/MEAL-V2.*

1. Introduction

Convolutional Neural Networks (CNNs) [14] have been proven useful in many visual tasks, such as image classification [13, 8], object detection [6, 17], semantic segmentation [15], as well as some particular scenarios, like transferring feature representation [23], learning detectors from scratch [19], etc. In order to achieve highest possible accuracy, many training techniques and data augmentation methods have been proposed, such as mixup [25], cutmix [24], autoaug [1], randaug [2], fix resolution discrepancy [22], etc. Some works also focus on modifying the network structures, e.g., SENet [11], ResNeSt [26]. Our goal of this paper is to similarly obtain the best possible performance of a network, but our proposed method is orthogonal to the above techniques. In general, our method only relies on a *teacher-student* paradigm with a powerful ensemble of teachers and a good initialization of the student. It is simple, straight-forward, but effective and can achieve state-of-the-art performance on the large-scale dataset. The advantages of our method are: 1) no architecture modification is needed; 2) no outside training data beyond ImageNet; 3) no cosine learning rate; 4) no extra data augmentation like mixup, autoaug, etc; 5) no label smoothing.



(a) MEAL

(b) Ours

Figure 1. An illustration of the comparison between MEAL [18] and our method. We use an ensemble of all teacher networks instead of the teacher selection module as adopted in MEAL.

用教师网络输出平均作为监督

We also have a few interesting discoveries in our training process, for example, among them we would like to

[*] A short version of technical report.

Table 1. Item-by-item comparison of techniques that we use and do not use in our distillation training.

What we do not use		What we use	
architecture modification	✘	an ensemble of giant pre-trained teachers	✔
outside training data	✘	KL divergence loss	✔
hard/one-hot labels during distillation	✘	a good initialization for the student	✔
cosine/linear decay learning rate	✘	step decay learning rate (0.01-0.001)	✔
weight-decay	✘		
cutout [4]/mixup [25]/cutmix [24] training	✘		
label smoothing [20]	✘		
autoaug [1]/randaug [2], etc.	✘		
warmup [7]	✘		

emphasize that the one-hot/hard label[1] is not necessary and could not be used in the distillation process, which is important and critical for the distillation framework [9, 18]. Some discussions about this perspective are provided in Sec. 4. While some previous studies deem that structure might be more important and crucial than pre-trained parameters on some downstream tasks like object detection [19], segmentation [12], etc., we still believe that boosting the performance of standard and classical network structures is interesting and useful, especially the network is already tiny and compact, like MobileNet V3, EfficientNet-B0, as the proposed method is toilless to be generalized to other well-designed or searched architectures. That is to say, our proposed framework is a general design, literally easy to use and can be considered as a post-process to distill small and compact models for further boosting their performance, meanwhile, no modification is required.

2. Our Approach

We begin by introducing each component in our proposed framework, including: 1) teacher ensemble; 2) KL-divergence loss; 3) the discriminator. Then, we present the training details and techniques that we used and did not use in our distillation training.

Teachers Ensemble is used to generate more accurate predictions for guiding the student training. Different from MEAL [18] that selected one teacher through a teacher selection module in each training iteration, we adopt the average of softmax probabilities from multiple pre-trained teachers as an ensemble. Let $\mathcal{T}_\theta$ be the teacher network, the output ensemble probability $\hat{\mathbf{p}}_{\mathbf{e}}^{\mathcal{T}_\theta}$ can be described as:

均值

$$\hat{\mathbf{p}}_e^{\mathcal{T}_\theta}(X) = \frac{1}{K}\sum_{\mathbf{t}=1}^{K}\mathbf{p_t}^{\mathcal{T}_\theta}(X) \tag{1}$$

where $\mathbf{p}_{\mathbf{t}}^{\mathcal{T}_\theta}$ is the $\mathbf{t}$-th teacher's softmax prediction. X is the inout image and K is the number of total teachers.

KL-divergence is a measure metric of how one probability distribution is different from another reference distribution.

描述两个分布的相似性

In our approach, we train the student network $\mathcal{S}_\theta$ by minimizing the KL-divergence between its output $\mathbf{p}^{\mathcal{S}_\theta}(x_i)$ and the ensembled soft labels $\hat{\mathbf{p}}^{\mathcal{T}_\theta}(x_i)$ generated by the teacher ensemble. The loss function of KL-divergence can be formulated as:

$$\begin{aligned}\mathcal{L}_{KL}(\mathcal{S}_\theta) &= -\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{p}}_e^{\mathcal{T}_\theta}(x_i)\log(\frac{\mathbf{p}^{\mathcal{S}_\theta}(x_i)}{\hat{\mathbf{p}}_e^{\mathcal{T}_\theta}(x_i)}) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{p}}_e^{\mathcal{T}_\theta}(x_i)\log\mathbf{p}^{\mathcal{S}_\theta}(x_i) \\ &+\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{p}}_e^{\mathcal{T}_\theta}(x_i)\log\hat{\mathbf{p}}_e^{\mathcal{T}_\theta}(x_i)\end{aligned} \tag{2}$$

where N is the number of samples. The second term is the entropy of ensembled labels from teacher ensemble and is constant with respect to $\mathcal{T}_\theta$. We can remove it and simply minimize the rest cross-entropy loss as follows:

实际就是CE

$$\mathcal{L}_{CE}(\mathcal{S}_\theta) = -\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{p}}_e^{\mathcal{T}_\theta}(x_i)\log\mathbf{p}^{\mathcal{S}_\theta}(x_i) \tag{3}$$

Discriminator is a binary classifier to distinguish the input features are from teacher ensemble or student network. It consists of a *sigmoid* function following the *binary cross-entropy* loss. The loss can be formulated as:

$$\mathcal{L}_{\mathcal{D}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{y}_i\cdot\log\mathbf{p}^{\mathcal{D}}{}_i + (1-\mathbf{y}_i)\cdot\log(1-\mathbf{p}^{\mathcal{D}}{}_i)\right] \tag{4}$$

where $\mathbf{y}_i$ is the binary label for the input features x_i, $\mathbf{y}\in\{0,1\}$, and $\mathbf{p}^{\mathcal{D}}{}_i$ is the corresponding probability vector.

We define a *sigmoid* function to model the individual teacher or student probability:

$$\mathbf{p}^{\mathcal{D}}(x;\theta) = \sigma(f_\theta(\{x_{\mathcal{T}}, x_{\mathcal{S}}\})) \tag{5}$$

where f_θ is a three-fc-layer subnetwork and θ is its parameters, $\sigma(x) = 1/(1+\exp(-x))$ is the logistic function. In our model, we use the last output layer before softmax as the representation for the discriminator input.

判别器由三层全连接和sigmoid函数组成，输入教师和学生网络的softmax之前的特征

[1] the ground-truth labels.

Table 2. Comparison of validation accuracy on ImageNet dataset for ResNet-50 architecture under single crop evaluation. (*) indicates that they used horizontal flip, shifted center crop and color jittering for training.

Network	Resolution	#Params	**Top-1 (%)**	**Top-5 (%)**
ResNet-50	224	25.6M	76.51	93.20
ResNet-50 + DropBlock, (kp=0.9) [5]	224	25.6M	78.13	94.02
ResNet-50 + DropBlock (kp=0.9) [5] + label smoothing (0.1)	224	25.6M	78.35	94.15
ResNet-50 + MEAL [18]	224	25.6M	78.21	94.01
ResNet-50 + Ours (MEAL V2)	224	25.6M	**80.67**	**95.09**
ResNet-50 + FixRes [22]	384	25.6M	79.0	94.6
ResNet-50 + FixRes (*) [22]	384	25.6M	79.1	94.6
ResNet-50 + Ours (MEAL V2)	380	25.6M	**81.72**	**95.81**
ResNet-50 + FixRes [22] + CutMix	320	25.6M	79.7	94.9
ResNet-50 + FixRes [22] + CutMix (*)	320	25.6M	79.8	94.9
ResNet-50 + Ours (MEAL V2) + CutMix	**224**	25.6M	**80.98**	**95.35**

Table 3. Comparison of validation accuracy on ImageNet for MobileNet V3-Small 0.75/1.0/Large 1.0 and EfficientNet-B0 architectures.

Network	Resolution	#Params	**Top-1 (%)**	**Top-5 (%)**
MobileNet V3-Small 0.75 [10]	224	2.04M	65.40	–
+ Ours (MEAL V2)	224	2.04M	**67.60**	**87.23**
MobileNet V3-Small 1.0 [10]	224	2.54M	67.40	–
+ Ours (MEAL V2)	224	2.54M	**69.65**	**88.71**
MobileNet V3-Large 1.0 [10]	224	5.48M	75.20	–
+ Ours (MEAL V2)	224	5.48M	**76.92**	**93.32**
EfficientNet-B0 [21]	224	5.29M	77.3 (76.8)	93.5 (93.2)
+ Ours (MEAL V2)	224	5.29M	**78.29**	**93.95**

Consider that our teacher supervision is an ensemble of multiple networks, it is not convenient to obtain the intermediate outputs. Also, to make the whole framework neater, we only adopt the similarity loss and discriminator on the final outputs of networks for distillation. We show from our experimental results that supervision from the last layer of teacher ensemble is competent to distill a strong student.

3. Experiments

3.1. Dataset

We conduct experiments on ILSVRC 2012 classification dataset [3] that consists of 1,000 classes, with a number of 1.2 million training images and 50,000 validation images. We adopt the basic data augmentation scheme following [16], i.e., *RandomResizedCrop* and *RandomHorizontalFlip*, and apply the single-crop operation at test time.

3.2. Experimental Settings

We use a mini-batch size of 512 with 8 GPUs for training our models. SGD optimizer is adopted with a step learning rate decay scheduler. The initial learning rate is set to 0.01. We train with a total number of 180 epochs and the learning rate multiplied by 0.1 at 100 epoch. The weight decay is not used (set to 0) in our training. We apply this strategy to all our experiments regardless of what kind of teacher and student architectures we choose. We use the models in timm[2]. If the input size of a student network is 224 × 224, we choose senet154 and resnet152_v1s as teachers according to the input size of the pre-trained models. For 380 × 380, we use efficientnet_b4_ns and efficientnet_b4 as teachers. Our code and all trained models are available at: https://github.com/szq0214/MEAL-V2.

3.3. Results

On ResNet-50. Our results on ResNet-50 are shown in Table 2. Under 224 × 224 input size, our method achieves 80.67% Top-1 accuracy, outperforming the previous state-of-the-art method MEAL [18] by 2.46%. Furthermore, our results are even better than ResNeSt-50 [26] (fast) that requires to modify the network architecture and is learned with many training tricks. After enlarging the input size to 380×380, our performance is further improved to 81.72%, outperforming FixRes (*) [22] by 2.62% with slightly smaller input.

On Small Networks. We choose MobileNet V3 Small-0.75/1.0/Large-1.0 and EfficientNet-B0 networks which are already compact models to verify the effectiveness of our proposed method. Our results are shown in Table 3, on MobileNet V3-Small 0.75 and 1.0, our method improves the original models by 2.20% and 2.25% accuracies without any architecture modification. Such huge increases are fairly surprising since the models are already compact, more importantly, the gains are totally free during inference stage.

[2] https://github.com/rwightman/pytorch-image-models/.

On MobileNet V3-Large 1.0 and EfficientNet-B0, although the improvement is not as enormous as Small 0.75 and 1.0, we still obtain 1.72% and 1.49% increases on ImageNet. Note that for EfficientNet-B0, 77.3/93.5 accuracy is from their paper [21] and 76.8/93.2 is the accuracy from their pre-trained models in timm.

With more data augmentation. We'd like to further explore whether our models have been saturated on the target data by injecting more data augmentation like CutMix in training. The results are shown in Table 2, we involve CutMix and keep other settings the same as our basic experiments, we obtain Top-1/5 80.98%/95.35%, which outperforms the baseline MEAL V2 by 0.31%/0.26%. While the improvement is not so large, it indicates that our model is not yet over-fitting and still has room to boost. Moreover, our results are 1.18%/0.45% better than FixRes+CutMix (*) under smaller input resolution (224 vs. 320). Intriguingly, the results on ResNet-50 are very close to the teachers we used in distillation (81.22%/95.36% and 81.01%/95.42%), since the scale of our student is much smaller than the teacher architectures, it's surprising that the student can catch up the teachers without additional training data.

3.4. Analysis

We know there are many factors in knowledge distillation to determine and affect the performance of a student. Since we use the same teacher ensemble for all ResNet-50, MobileNet V3 and EfficientNet-B0 under 224×224 input, the results indicate that the student architecture or capacity itself is a crucial indicator. If we compare MEAL V1 and V2 we can further derive the conclusion that teacher's performance, i.e. the quality of supervision, is another factor for the student, generally, the stronger teachers can consistently distill stronger students. To verify whether the initialization of a student has a big impact, we conduct the ablation study through adopting *tf_efficientnet_b0* (Top-1/5: 76.85%/93.25%) and *efficientnet_b0* (77.70%/93.53%) as the student initialization in timm, respectively. They have the same architecture but the training settings and performance are different. Interestingly, we got Top-1 78.29% and 78.23% respectively for the two initializations with the same teacher ensembles and training hyper-parameters. It seems that a good initialization of a student only helps to speed up the converge of training, but it has no big impact on the final performance of the student.

4. Discussions

Why is the hard/one-hot label not necessary in knowledge distillation? The one-hot labels in ImageNet are annotated by humans, thus there are definitely some incorrect or missing annotations into them. Also, a non-negligible proportion of images in ImageNet contain more than one object within a single image, the one-hot label is determined by the annotators among multiple objects which cannot represent the complete content of this image precisely. We argue that if the teacher ensembles are strong enough, which can provide high-quality predictions for the input image, involving such inaccurate hard labels will mislead the student to a wrong optimum and incur inferior performance.

标准错误，无标签，以及多目标只有一个标签

How does the discriminator help the optimization? The discriminator is used to prevent the student from being overfitting on the training data. It can slow down the moving of a student to mimic the teachers' output, which can be regarded as a regularization effect. In training, a very small learning rate is adopted to tune the discriminator's parameters to ensure that it will not converge too fast, which is discrepant from the backbone network. In the scenario of MEAL V2, our teacher ensembles are usually powerful and strong, meanwhile, the student architectures are always smaller and more compact than the teachers, it means that the student's capability and learning ability are also much worse than the pretrained teacher networks, even we force the student to produce the same predictions as strong teachers, the outputs between student and teacher ensembles still have inevitable gaps which cannot be eradicated through the KL-divergence loss. That is to say, the discriminator is very easy to distinguish that the feature is from a student or teacher ensemble and the regularization effect will be weakened. Nevertheless, in MEAL V2 we still see slight improvement on performance by using the discriminator.

防止过拟合，减缓学生模仿教师视为正则效果

训练中判别器采用较小的学习率，以免收敛过快

How about the generalization ability of our method on large students? We tried to use some large models like ResNeXt-101 32×48d for the students as used in teacher networks, meaning that the student has similar capability with teachers. As expected, the improvement is not as considerable as those of small students, we still see some increase on performance. Generally, the soft supervision from teacher ensembles is better than the human-annotated hard labels. Especially when the scale and performance gap between teachers and students are enormous, the improvement will be more effective and notable. That is to say, in most of our experimental cases, the stronger teachers can consistently produce and distill stronger students.

大模型蒸馏小模型效果更好

Is there still room to improve the performance of vanilla ResNet-50? It's definitely *Yes*. Replacing the teacher ensembles we used with more and stronger networks could be helpful, but the training cost will be increased accordingly. Also, some of the common tricks like cosine decay learning rate might be useful for the performance but we did not have enough resources to test all of them. The current choices are just the compromise and a trade-off under the considerations of training efficiency, computational resources, etc. Our purpose of this paper is mainly to verify the effectiveness of our proposed perspective, rather than the high accuracy. Still, it will be very interesting to explore the upper bound performance of a fixed-structure network,

采用更大网络改善学生大模型

such as ResNet-50.

5. Conclusion

We have presented a new paradigm of knowledge distillation based on a teacher ensemble and a discriminator. We show that such a simple framework can achieve promising results without tricks on a variety of network structures including the extremely tiny and compact models. On ImageNet dataset, our method achieves **80.67%** top-1 accuracy using a single crop of 224×224 on the vanilla ResNet-50. Our results show that existing networks' potential has not been fully exploited and there is still room to boost and enhance through our framework. We hope the proposed method can inspire more studies along this direction of boosting tiny and compact models through knowledge distillation.

References

[1] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[2] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[5] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*, pages 10727–10737, 2018.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[7] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[10] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1314–1324, 2019.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[12] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 11–19, 2017.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[18] Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4886–4893, 2019.

[19] Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue. Dsod: Learning deeply supervised object detectors from scratch. In *Proceedings of the IEEE international conference on computer vision*, pages 1919–1927, 2017.

[20] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[21] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019.

[22] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. In *Advances in Neural Information Processing Systems*, pages 8252–8262, 2019.

[23] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In

Advances in neural information processing systems, pages 3320–3328, 2014.

[24] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019.

[25] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[26] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.