How To Enhance The Generalization Capacity in Data Defficient Settings?

Yilu Guo, Shicai Yang, Weijie Chen, Liang Ma, Di Xie, and Shiliang Pu

Hikvision Research Institute, Hangzhou, China
gyl.luan0@gmail.com

Abstract. Convolutional neural networks (CNNs) have achieved great success in image classification by utilizing large-scale datasets. However, it is still of great challenge to learn from scratch on small-scale datasets. When the dataset is limited, the concept of category will be ambiguous and the over-parameterized CNNs tend to simply memorize the dataset. Therefore, it is crucial to study how to learn more discriminative representations while avoiding over-fitting. In this paper, we propose a new framework that consists of Contrastive Regularization and Auxiliary Classifier to learn representations efficiently, and Mean Teacher and Symmetric Cross Entropy to constrain the fitting spend and fitting balance. Together with other tricks, such as aggressive data augmentation, TenCrop inference and models ensembling, we achieve competitive performance in the VIPriors Image Classification Challenge.

1 Introduction

Convolutional neural networks (CNNs) have achieved tremendous success in image classification. However, it deeply depends on large-scale datasets, such as ImageNet [6] and OpenImage [13]. Generally, CNNs learn to generalize well with massive data. When trained on a small-scale dataset, they are required to be pre-trained on a large-scale dataset in a supervised or unsupervised manner. Herein, we can't help to ask, can we achieve comparable results on a small dataset by learning from scratch without any pre-training. This is an interesting and significant topic proposed in the VIPriors Image Classification Challenge.

The objective of VIPriors Image Classification Challenge is to increase the Top-1 Accuracy on ImageNet dataset by only using a small subset of ImageNet dataset. The data is divided into three splits, including a training set, a validation set, and a testing set which is unavailable during the model optimization. The training and validation splits are two subsets of the original training split. The test set is taken from the original validation split directly. Each split includes 1,000 classes which are the same as the original ImageNet and 50 images per class, resulting in 50,000 images in total.

When the training data is limited, especially when the image amount of each category is quite small (even less than the number of classes), the concept of category tends to be ambiguous. Hence, it is a challenging problem to extract discriminative representations by learning from scratch on a small dataset. Also,

it is crucial to alleviate over-fitting since the models with a great network capacity are easy to memorize the dataset, leading to a poor generalization ability. In this paper, we deal with the data defficient learning from two perspectives, i) to enhance the representational capacity of the model, ii) to mitigate the over-fitting problem.

Contrastive learning [1, 9] is usually utilized to learn discriminative representations in a supervised or unsupervised manner. We propose to use contrastive learning on the prediction of the model to strengthen the learning of representations, which is termed as contrastive regularization in this paper. An auxiliary classifier is added to the intermediate stage of the model to learn the representations more efficiently. For alleviating the over-fitting problem, sufficient augmentation strategies are necessary, such as random erasing [31], Mixup [27], CutMix [26], AutoAugment [3], RandAugment [4], etc. We further use Mean Teacher [23] to constrain the fitting spend and to learn more stable features. And Symmetric Cross Entropy (SCE) [24] is used to balance the fitting of different classes.

2 Method

Our framework consists of Contrastive Regularization and Auxiliary Classifier to learn representations efficiently, and Mean Teacher and Symmetric Cross Entropy to constrain the fitting spend and fitting balance. The entire pipeline is shown in Fig. 1.

2.1 Contrastive Regularization

Contrastive learning [1, 9, 12] is a framework to learn similar/dissimilar representations from data that are organized into similar/dissimilar pairs. Recently, contrastive learning has promoted the performance of various tasks, including semi-supervised learning [2, 14], learning with noisy label [30, 32] and so on. Supervised Contrastive Learning (SCL) [12] adapts contrastive learning to the fully supervised setting to learn more informative representations by effectively leveraging label information. SCL is an excellent representation learning method, but the model learned by SCL often needs to finetune on downstream tasks. We integrate SCL with the target task by applying SCL on the final prediction of the model, which is called Contrastive Regularization.

Within a multiview batch, let $i \in I \equiv \{1...2N\}$ be the index of an arbitrary augmented smaple, the original loss of SCL is as follow:

$$L_{SCL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}$$

Here, z_l is the feature representation, τ is a temperature hyper-parameter, $A(i) \equiv I \setminus \{i\}$, $P(i) \equiv \{p \in A(i) : \tilde{\boldsymbol{y}}_p = \tilde{\boldsymbol{y}}_i\}$ is the set of indices of all positives in the multiviewed batch distinct from i, and $|P(i)|$ is its cardinality.

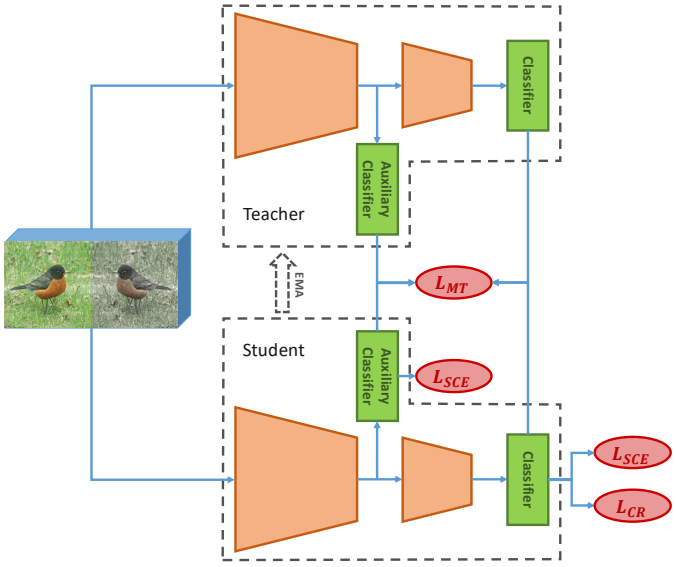

Fig. 1. The pipeline of our method. The two views of images by different augmentation are mainly used for Contrastive Regularization (CR). And they will also be utilized in the consistency loss of Mean Teacher (MT) and Symmetric Cross Entropy (SCE).

The proposed Contrastive Regularization changes the feature z_l to the softmax output o_l of the model:

$$L_{CR} = \sum_{i \in I} \frac{-1}{|P^{'}(i)|} \sum_{p \in P^{'}(i)} \log \frac{\exp(o_i \cdot o_p / \tau)}{\sum_{a \in A(i)} \exp(o_i \cdot o_a / \tau)}$$

Here, $P^{'}(i) \equiv \{p \in A(i) : dis(\tilde{\boldsymbol{y}}_p, \tilde{\boldsymbol{y}}_i) < \delta\}$, since we use the cutmix and mixup during the training stage and the hard-label are changed to soft-label. We use Jensen-Shannon divergence as the dis, and use δ to adjust the attention level to individual-wise or class-wise contrasting.

The small *delta* is used to strengthen the individual information to improve the standard classification. And the individual information also can avoid the model to memorise the noisy class.

2.2 Mean Teacher

Mean Teacher [23] is proposed for semi-supervised learning. Here, we adapt it to stabilize model learning. Mean teacher maintains the exponential moving average (EMA) weights of the model and uses the EMA weights as a teacher model. More formally, we define $\theta_t^{'}$ at training step t as the EMA of successive θ_t weights.

$$\theta_t^{'} = \alpha \theta_{t-1}^{'} + (1 - \alpha)\theta_t$$

And we define the consistency loss L_{MT} as the expected Kullback-Leibler divergence between the prediction of the student model and the prediction of the teacher model.

$$L_{MT} = \mathbb{E}_x[KL(f(x,\theta)||f(x,\theta^{'}))]$$

Averaging model weights over training steps tends to produce a more stabilized model and can provide more accurate soft-labels. Thus, the model can avoid learning some inaccurate information via the consistency loss. The learning process will be more robust.

2.3 Symmetric Cross Entropy

Symmetric Cross Entropy (SCE) [24] is a simple yet effective loss for learning with noisy label. It aims to simultaneously address the hard class learning problem and the noisy label overfitting problem of Cross Entropy.

The label of ImageNet dataset is well-known to contain errors [18, 17]. And there are many similar category concepts that will be more ambiguous when the image amount of each category is limited. So, there may be some noisy labels. And we find that accuracy varies greatly among different classes. Thus, we employ Symmetric Cross Entropy [24] to balance the fitting of different classes. The Symmetric Cross Entropy is easily constituted by standard cross entropy and reverse cross entropy.

$$L_{SCE} = H(p,q) + H(q,p)$$

Here, $H(p,q)$ is the standard cross entropy.

2.4 Auxiliary Classifier

Supervisions to the intermediate output are usually used in deep learning to reduce the difficulty of optimizing the deep network [21, 29] or to enhance the information from different scales [25]. For better extracting the image information, we add an Auxiliary Classifier to the intermediate output of the model. And during inference, the prediction is computed by the weighted average of the intermediate output and the final output, which is termed as Auxiliary Fusion.

3 Experiments

3.1 Dataset

We train models on the subset of the ImageNet [6] which was given by the VIPrior Image Classification Challenge without any pre-trained models. In the early phase of the competition, we study different methods by training models on the training split and verifying the effectiveness on the validation split. And in the final stage, we combine the training and validation splits for training and randomly split a few samples for validation.

Table 1. Model architecture comparison on validation split.

Model	top-1 acc. (%)
EfficientNet-b2	41.11
EfficientNet-b4	49.40
ResNet50	40.91
ResNeXt101_32x4d	45.32
Swin Transformer	26.43

3.2 Implementation Details

We use the RMSprop [5] optimizer with alpha set to 0.9 and momentum set to 0.9. Models are trained with 8 GPUs and 16 samples per GPU. Due to the small batchsize per GPU, we have used the Synchronized Batch Normalization (syncbn). Our learning rates are adjusted according to a cosine decaying policy [8] and the initial learning rate is set to 0.005. The warm-up [8] strategy is applied over the first 3 epochs, gradually increasing the learning rate linearly from 1e-6 to the initial value for the cosine schedule. The weight decay is set to 1e-5. The default image resolution is 320x320 during the training.

3.3 Ablation Study

We have trained some models with different structures on the training split with only some simple regularizations like dropout.

Table 1 shows the performances for different model architectures. We can see that the larger models perform the better. Specifically, EfficientNet [22] surpasses ResNet [10] in this task. While Swin Transformer [15] behaves badly. As the name of the challenge, say "Visual Inductive Priors", it needs more priors to learn from scratch on a small dataset, while Transformer, as is well known, lacks some of the inductive biases inherent to CNNs [7].

We use some common data augmentation methods and regularization methods that are as follows: AutoAugment [3], random erasing [31], dropout [19] with probability of 0.3, label smoothing [21], mixup [27] with alpha of 0.5 and cutmix [26] with alpha of 1.0. In addition, prolonged training epochs as 460 epochs are used to improve performance. The above methods enhance the generalization capacity of the model, leading to a quite good performance, and we pick it as our strong baseline. In Table 2, we compare some different augmentations with Strong Baseline.

Table 3 shows the ablation results for EfficientNet-b2 [22] which are trained on the training split and verified on the validation split. The techniques we use in Fig. 1 further improve the top-1 accuracy from 58.09% to 61.18%. The performance is even close to some single models trained with the training and validation data last year.

Table 2. Compare different augmentations with Strong Baseline on validation split.

EfficientNet-b2	top-1 acc. (%)
Strong Baseline	58.09
AutoAugment → RandAugment	57.25
Mixup+Cutmix → Mixup	57.75
Mixup+Cutmix → Cutmix	57.22

Table 3. Ablation study on validation split.

EfficientNet-b2	top-1 acc. (%)
Strong Baseline	58.09
+SCE	59.10
+Mean Teacher	60.07
+Contrastive Regularization	60.84
+Auxiliary Classifier	60.98
+Auxiliary Fusion	61.18

3.4 Final Results

We find that a larger resolution can further boost the performance both on the training and the inference. We use a larger resolution as 448x448 during training. We only train a few larger resolution models entirely and finetune a few epochs with a larger resolution on the default resolution models due to the high resource-consuming.

During the inference, the 448x448 resolution and TenCrop are utilized. After using these, we get an excellent performance (top-1 accuracy of 72.14%) by single model (EfficientNet-b7). Table 4 shows the comparison with last year's single model.

Experimental evidence shows that the ensemble method is usually much more accurate than a single model. We average the predictions of above methods in total of 16 models including EfficientNet-b5 [22], EfficientNet-b6, EfficientNet-b7, DSK-ResNeXt101 [20], ResNet-152 [10], SEResNet-152 [11]. Finally, we got the top-1 accuracy of 74.49% on the testing set. We also compare our result after ensembling with last year's final results in Table 5.

Table 4. Compare with last year's single model.

Method	top-1 acc. (%)
Ours	72.14
Sun's [20]	69.59
Luo's [16]	66.36
Zhao's [28]	66.20

Table 5. Compare with last year's final results.

Method	top-1 acc. (%)
Ours	74.49
Sun's [20]	73.08
Luo's [16]	70.15
Zhao's [28]	68.80

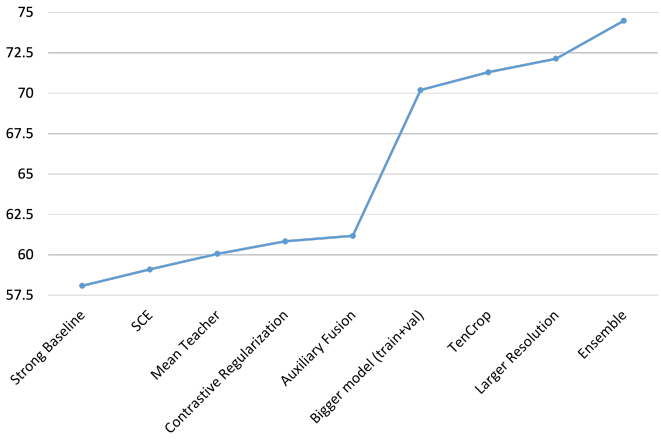

Fig. 2. Perfromance overview.

Fig. 2 shows an overview of methods and appearances. No external data or pre-trained models were used throughout the competition.

4 Conclusions

In this paper, we discuss and explore how to enhance the generalization capacity in data defficient dettings. We deal with this problem by learning more discriminative representations while avoiding over-fitting. And We propose a new framework which consists of Contrastive Regularization, Auxiliary Classifier, Mean Teacher and Symmetric Cross Entropy. Ablation studies show that our framework is effective in data defficient learning. Finally we achieve competitive performance in the VIPriors Image Classification Challenge together with other tricks.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML 2020: 37th International Conference on Machine Learning. vol. 1, pp. 1597–1607 (2020)
2. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: Advances in Neural Information Processing Systems. vol. 33, pp. 22243–22255 (2020)
3. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 113–123 (2019)
4. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: Randaugment: Practical automated data augmentation with a reduced search space. In: Advances in Neural Information Processing Systems. vol. 33, pp. 18613–18624 (2020)
5. Dauphin, Y.N., de Vries, H., Chung, J., Bengio, Y.: Rmsprop and equilibrated adaptive learning rates for non-convex optimization. arXiv: Learning (2015)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR 2021: The Ninth International Conference on Learning Representations (2021)
8. Goyal, P., Dollár, P., Girshick, R.B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9729–9738 (2020)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
11. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. vol. 42, pp. 2011–2023 (2018)
12. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 18661–18673 (2020)
13. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982 (2018)
14. Li, J., Xiong, C., Hoi, S.C.H.: Comatch: Semi-supervised learning with contrastive graph regularization. arXiv preprint arXiv:2011.11183 (2020)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
16. Luo, Z., Li, G., Zhang, Z.: A technical report for vipriors image classification challenge. arXiv: Computer Vision and Pattern Recognition (2020)

17. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021)
18. Northcutt, C.G., Jiang, L., Chuang, I.L.: Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research **70**, 1373–1411 (2021)
19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
20. Sun, P., Jin, X., Su, W., He, Y., Xue, H., Lu, Q.: A visual inductive priors framework for data-efficient image classification. In: European Conference on Computer Vision. pp. 511–520 (2020)
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
22. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114 (2019)
23. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: ICLR (Workshop) (2017)
24. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 322–330 (2019)
25. Xie, S., Tu, Z.: Holistically-nested edge detection. International Journal of Computer Vision **125**(1), 3–18 (2017)
26. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6022–6031 (2019)
27. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2017)
28. Zhao, B., Wen, X.: Distilling visual priors from self-supervised learning. In: European Conference on Computer Vision. pp. 422–429 (2020)
29. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6230–6239 (2017)
30. Zheltonozhskii, E., Baskin, C., Mendelson, A., Bronstein, A.M., Litany, O.: Contrast to divide: self-supervised pre-training for learning with noisy labels. In: arXiv e-prints (2021)
31. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13001–13008 (2020)
32. Zhou, Y., Ge, Y., Wu, J.: Friends and foes in learning from noisy labels. CoRR **abs/2103.15055** (2021), https://arxiv.org/abs/2103.15055