Self-Supervised Neural Architecture Search for Imbalanced Datasets

Aleksandr Timofeev[1], Grigorios G. Chrysos[1], Volkan Cevher[1]

[1] École Polytechnique Fédérale de Lausanne (EPFL)

Motivation

Designing architectures with supervised labels is costly and time-consuming. Thus, we learn the architecture of a neural network assuming no labels. Since in reality the labels are often imbalanced, we use self-supervised learning for imbalanced datasets.

Neural Architecture Search

Neural Architecture Search is separated into three components:

- *Search space*: defines the set of architectures to be explored;
- *Search strategy*: determines how to explore the search space;
- *Performance estimation*: designed to estimate performance in each step.

- **DARTS.** The final architecture is constructed by stacking cells. Each cell is a directed acyclic graph (DAG) with N nodes. Each edge represents a candidate operation o_{ij} with input x_i and output x_j, where $x_j = \sum_{i<j} o_{ij}(x_i)$. The *softmax relaxation* between the candidate operations $\mathcal{O} = \{o_{ij}^1, o_{ij}^2, ..., o_{ij}^M\}$ is used. The following bi-level optimization problem describes the objective:

$$\min_{\alpha} \ell_{val}\left(\omega^*(\alpha), \alpha\right) \quad \text{s.t. } \omega^*(\alpha) = \arg\min_{\omega} \ell_{train}(\omega, \alpha).$$

  where ω and α denote normal and architecture weights, $\ell_{val}\left(\omega^*(\alpha), \alpha\right) \approx \ell_{val}\left(\omega - \nabla_\omega \ell_{train}\left(\omega, \alpha\right), \alpha\right)$.

- **FairDARTS.** DARTS has several disadvantages:
  - Skip connections are frequently selected as the operation o_{ij};
  - Lack of weights that significantly outperform others.

  The recent work of FairDARTS [2] mitigates those issues using two modifications:
  - It replaces the softmax operation with a sigmoid function to avoid the competition with skip connections as a candidate operation;
  - Sparsity in architecture weights is encouraged by adding the zero-one loss:

$$\ell_{0-1} = -\frac{1}{N}\sum_{i=1}^{N}(\sigma(\alpha_i) - 0.5)^2.$$

Self-supervised learning

The idea is to devise one task that the ”target label” is known, and use losses developed for supervised learning.

- **Barlow Twins.** Barlow Twins [11] creates a pair of images for every original image by applying two randomly sampled transformations. The model extracts the representations z^A and z^B of the two corresponding distorted versions of the original mini-batch. The objective function is

$$\ell_{\mathcal{BT}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2,$$

  where λ is a positive coefficient, $\mathcal{C}$ is a cross-correlation matrix of the outputs size computed between two outputs:

$$\mathcal{C}_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b \left(z_{b,i}^A\right)^2}\sqrt{\sum_b \left(z_{b,j}^B\right)^2}}.$$

  where b indexes batch samples and i, j index the vector dimension of the outputs.

Handling imbalanced datasets

- **Focal loss.** The idea behind the focal loss [6] is to give a lower weight to easily classified samples. In a binary case, we introduce $p_t = \mathbb{1}[y = 1]p + \mathbb{1}[y = -1](1 - p)$,where $y \in \{-1, 1\}$ are labels, p is model's estimated probability, and $\mathbb{1}[\cdot]$ is an indicator function. Then, the focal loss is defined as $\mathrm{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log p_t$, where α_t are loss weights which can be inverse class frequencies, γ is a tunable parameter.

- **Logit adjustment.** In [7], the authors propose to handle imbalance by correcting the output of the model before softmax operation. Specifically, they introduce the logit adjusted softmax cross-entropy loss:

$$\ell\left(y, f(x)\right) = -\log \frac{e^{f_y(x) + \tau \log \pi_y}}{\sum_{y' \in [L]} e^{f_{y'}(x) + \tau \log \pi_{y'}}},$$

  where L is a number of classes, $f_y(x)$ is a logit of the given class, π_y is empirical frequencies of classes. Therefore, we induce the label-dependent prior offset which requires a larger margin for rare classes.

Method

The approach consists of two steps:

- *Architecture search:* Our method is build on top of FairDARTS. We also replace the supervised loss with Barlow Twins loss which does not require labels. Additionally, we use the supernet with only three cells for all steps. This is beneficial for three reasons:
  - The training process is efficient and affordable for slow GPUs, while it produces small but powerful architectures;
  - The designed architecture is appropriate for the final model as its depth is unchanged;
  - The learned weights in the first step are fully utilized in the second step.
- *Fine-tuning:* To fine-tune the designed model, we add on the top another layer which projects the output matrix into the output classes and train it with the focal loss and the logit adjustment in a supervised manner to improve learning of rare classes.

The main advantages of our method (in comparison to similar works) are:

- FairDARTS avoids aforementioned drawbacks and robust to initialization(requires only run);
- Barlow Twins can be executed with a small batch (non-typical for other methods);
- We use a smaller supernet which improves training time and produces more efficient architectures;
- We skip the self-supervised pretraining step without occurring a loss in performance (≈ 30% overhead in the running time);
- Our method is specifically developed for imbalanced datasets.

Results

The first experiments are conducted on artificially long-tailed CIFAR-10 [5] with imbalance factor of 10. The number of parameters is reported in millions ($\times 10^6$). A supernet has 3 cells with 4 nodes each. We use SGD with learning rate 0.025, momentum 0.9, and weight decay 3×10^{-4} with cosine annealing learning rate scheduler. A batch size of 32 is used, while we train the architecture for 100 epochs. The experiments are performed on NVIDIA Tesla K40c. The fine-tuning is run for 600 epochs.

Method	# Params	Error (↓)
ResNet-32 + Focal	21.80	13.34
ResNet-32 + SGM [3]	21.80	12.97
ResNet-32 + BSGM [3]	21.80	12.51
LDAM-DRW [1]	21.80	11.84
smDragon [9]	21.80+	12.17
VE2 + smDragon [9]	21.80+	11.84
SSNAS [4]	0.83	18.84
Our method	**0.81**	**10.91**

Then, we test the method on a naturally imbalanced ChestMNIST [10] dataset from medical imaging. The resolution of input images are indicated in the parenthesis.

Method	# Params	Error (↓)
ResNet-18 (28)	11.68	94.7
ResNet-18 (224)	11.68	**94.8**
ResNet-50 (28)	25.56	94.7
ResNet-50 (224)	25.56	94.7
auto-sklearn (28)	-	64.7
AutoKeras (28)	-	93.9
Google Auto ML (28)	-	94.7
SSNAS (28) [4]	**0.57**	94.7
Our method (28)	0.82	**94.8**

We also verify transferability of the architecture obtained with ChestMNIST by training it on COVID-19 X-ray dataset [8].

Method	# Params	Error (↓)
DarkCovidNet (224) [8]	1.16	98.08
SSNAS (224) [4]	**0.57**	**98.40**
SSNAS (28) [4]	**0.57**	98.08
Our method (224)	0.82	**98.40**
Our method (28)	0.82	**98.40**

References

[1] Kaidi Cao et al. “Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss”. In: *Advances in neural information processing systems (NeurIPS)*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[2] Xiangxiang Chu et al. “Fair DARTS: Eliminating Unfair Advantages in Differentiable Architecture Search”. In: *European Conference on Computer Vision (ECCV)*. 2020.

[3] Yin Cui et al. “Class-balanced loss based on effective number of samples”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 9268–9277.

[4] Sapir Kaplan and Raja Giryes. “Self-supervised Neural Architecture Search”. In: *arXiv preprint arXiv:2007.01500* (2020).

[5] Alex Krizhevsky et al. “Learning multiple layers of features from tiny images”. In: (2009).

[6] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 2980–2988.

[7] Aditya Krishna Menon et al. “Long-tail learning via logit adjustment”. In: *International Conference on Learning Representations (ICLR)*. 2021.

[8] Tulin Ozturk et al. “Automated detection of COVID-19 cases using deep neural networks with X-ray images”. In: *Computers in Biology and Medicine* 121 (2020), p. 103792. ISSN: 0010-4825.

[9] Dvir Samuel, Yuval Atzmon, and Gal Chechik. “From generalized zero-shot learning to long-tail with class descriptors”. In: *Winter Conference on Applications of Computer Vision (WACV)*. 2021.

[10] Xiaosong Wang et al. “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[11] Jure Zbontar et al. “Barlow twins: Self-supervised learning via redundancy reduction”. In: *arXiv preprint arXiv:2103.03230* (2021).