Self-training with Noisy Student Improves ImageNet Classification

Abstract

- a semi-supervised learning approach
- extends the idea of self-training and distillation with the used of equal-or-larger student model
- noise added to the student during learning

Noisy Student Training

- train an EfficientNet model on a labeled images
- use it as a teacher to generate pseudo-labels for 300M unlabeled images
- train a larger EfficientNet as a student model on the combination of labeled and pseudo-labeled images
- iterate this process by putting back the student as the teacher
- during the learning of the student, inject noise such as dropout, stochastic depth, and data augmentation (RandAugment) to the student

Algorithm

Require: Labeled images

$\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

unlabeled images $\{\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_m\}$

1: learn teacher model θ_*^t which minimizes cross entropy loss on labeled images

$$\frac{1}{n}\sum_{i=1}^{n} l(y_i, f^{noised}(x_i, \theta^t))$$

2: use an unnoised teacher model to generate soft or hard pseudo labels for unlabeled images:

$$\tilde{y}_i = f(\tilde{x}_i, \theta_*^t), \forall i = 1, \cdots, m$$

---

3. learn equal-or-larger student model θ_*^s which minimizes the cross entropy loss on labeled images and unlabeled images with noise added to the student model

$$\frac{1}{n}\sum_{i=1}^{n} l(y_i, f^{noised}(x_i, \theta^s)) + \frac{1}{m}\sum_{i=1}^{M} l(\tilde{y}_i, f^{noised}(\tilde{x}_i, \theta^s))$$

4. Iterative Training: Use the student as a teacher and go back to step 2.

- the algorithm is an improved version of self-training, a method in semi-supervised learning and distillation

- key improvement lie in adding noise to the student

- and using student models that are equal to or larger than the teacher

Noise Injection

- input noise - data augmentation (RandAugment)
- model noise - dropout, stochastic depth
- when applied to unlabeled data, noise has a compound benefit of enforcing local smoothness in the decision function on both labeled and unlabeled data
- when dropout and stochastic depth function are used as noise, the teacher behaves like an ensemble at inference time, whereas the student behaves like a single model

Other Techniques

- data filtering, balancing
- filter images that the teacher model has low confidences on since they are usually out-of-domain images
- balance the number of unlabeled images for each class
- soft pseudo labels work slightly better for out-of-domain unlabeled data

Experiments

- labeled dataset - ImageNet 2012 ILSVRC challenge prediction task
- unlabeled dataset - JFT dataset (300M images)
- ignore the labels and treat them as unlabeled data
- perform data filtering and balancing
- run an EfficientNet-B0 trained on ImageNet over the JFT dataset to predict a label for each image

- select images that have confidence of the level higher than 0.3 for each class
- we select at most 130K images that have the highest confidence
- do not tune these hyper-parameters extensively since our method is highly robust to them
- architecture: EfficientNet as baseline model because they provide better capacity for more data

- training details - for labeled images, batch size 2048 by default

- train the student model for 350 epochs for models larger than EfficientNet-B4 (including EfficientNet-L2)

- train smaller student models for 700 epochs

- use a large batch size for unlabeled images, to make full use of large quantities of unlabeled images

- apply the recently proposed technique to fix train-test resolution discrepancy for EfficientNet-L2

- noise - stochastic depth, dropout, RandAugment
- survival probability in stochastic depth to 0.8 for the final layer and follow the linear decay rule for other layers with a dropout rate of 0.5
- For RandAugment, we apply two random operations with magnitude set to 27
- iterative training - the best model in our experiments is a result of 3 iterations of putting back the student as the new teacher