Critique Report on the Paper: "Self-training with Noisy Student improves ImageNet classification [10]"

Sumaiya Tabassum Nimi

Background Information and Purpose of the Research

Deep learning models for computer vision can be made to achieve stunning, sometimes to the extent of human-level classification performance. The limitation is that this gain in performance can be gleaned only through supervised learning using datasets consisting of a limited number of labeled images. Whereas a larger collection of unlabeled images can be obtained with much less associated cost of labeling. The goal of the work [10] is to develop more general and robust deep learning models for classification of images using these inexpensive unlabeled images.

Methods Proposed

For developing more general and robust deep learning models by making use of unlabeled images, a semi-supervised training paradigm had been adopted in this paper [10], akin to knowledge distillation [6]. The proposed training algorithm can be summarised as follows.

- Step-1: Training teacher model on labeled dataset
- Step-2: Using teacher model to generate labels on unlabeled data, without adding noise to model or data
- Step-3: Training student model, equal or larger than the trained teacher model on both sets of labeled and unlabeled images, incorporating adversity in terms of adding noise to both data and model
- Step-4: Trained student became teacher and the process repeated iteratively from Step-2.

Strength of the Work

- The proposed training algorithm modified the traditional knowledge distillation method to generate student model that was larger than the teacher and trained with added adversity that came in the form of added noise to both model (using Dropout [9] and Stochastic Depth Function [7]) and data (using RandAugment [1]). Training in this fashion made the student model learn with more capacity compared to the teacher in a more adverse environment. The student model was forced to generate similar labels on images that were diversely modified. Also the added model noise made the teacher model behave like a more powerful ensemble that the student mimicked. This expansion of model capacity and injection of noise for obtaining more powerful and robust student model amounted to an exemplary out-of-the-box thinking and a major contribution of the paper.
- Starting with the model EfficientNet-B7 trained on ImageNet [2] dataset, three iterations of swapping the roles of students and teachers in the proposed algorithm, making use of 300M unlabeled images, led to the development of model called EfficientNet-L2 that obtained State-Of-The-Art

(SOTA) top-1 accuracy of 88.4% on ImageNet dataset. Also the size (in terms of number of parameters) of the model EfficientNet-L2 was half of the size of the model that had reported the previous SOTA accuracy on ImageNet dataset.

- Not only on clean ImageNet dataset, the best model EfficientNet-L2 achieved SOTA performance on datasets ImageNet-A [4], ImageNet-C [5] and ImageNet-P [3] containing distorted and corrupted ImageNet images also, proving that the model indeed was robust.
- Attention was paid to details like ensuring that the set of unlabeled images was class-balanced and also potential out-of-domain images were filtered based on class-confidence score.
- Also resolution of the train-test images was tuned towards best classification performance.
- Extensive ablation studies were done and the key findings were reported.

Limitations of the Work

- 300M unlabeled images were used for training in the work [10]. Naturally, as noted in the paper, many of the images will be Out-Of-Domain (OOD) for the ImageNet dataset. These images were filtered using just the class-confidence scores, which was not enough as it had been studied previously that OOD images are frequently classified with high confidence also [8]. So the student models would be trained on many OOD images. So the resulting model would wrongly classify many OOD input images, rendering the model *unreliable*.
- It seemed that in order to perform well, the proposed training algorithm required *everything* to be *large*. The teacher model should be sufficiently large, the student models would have to be even larger and there should be a large number of labeled as well as unlabeled images. The best model developed in the work called EfficientNet-L2 was so large that its training needed 6 days to converge even after using 2048 TPU cores. Large enough models trained on sufficiently large number of data to avoid overfitting are usually destined to perform well. The proposed training algorithm did not have room for low-budget cases. Especially when hardware for training bigger student models are unavailable, no scaling approach was designed.
- Continuous expansion of the student models will result in models that would perform well but will be suitable neither for deploying in memory-constrained edge devices nor for real-time inference.
- The effectiveness of the proposed training algorithm should have been tested on at least some of the smaller datasets like CIFAR-10, CIFAR-100 etc. using teacher models having limited capacity like ResNet, Densenet, VGG-19, VGG-16 etc. Because the real datasets will never be huge like ImageNet and the edge devices would not afford models as large as the EfficientNet models that have been reported in the paper.

Questions Unanswered

- If the proposed algorithm was used for training on smaller datasets consisting of fewer classes of images compared to ImageNet, then a larger proportion of the unlabeled images used for training would be OOD. The paper did not discuss how adversely it would have affected the classification performance in those cases and also how unreliable then the resulting model would have been.
- There was no discussion on incorporating zero-shot or few-shot learning in the proposed training algorithm, whereas these cases often appear in real-life learning tasks.

Suggested Future Studies

- OOD detection module can be incorporated into the training paradigm that will detect and filter out-of-domain images based on more sophisticated techniques that perform better than just checking class-confidence that has been exploited in the paper. If we introduce loss term in loss function for OOD detection, then using the proposed training algorithm, as student models develop iteratively, OOD detection can also become more accurate.
- An interesting future direction can be compressing models using knowledge distillation in the same way the models were expanded iteratively in the proposed algorithm. The best possible model with maximum capacity will be generated and then this model will be repeatedly pruned identifying the redundant parameters and even layers. There will be an obvious tradeoff between performance and model capacity. For performance equal or better than a fixed threshold, the most compact model can be obtained using this iterative pruning. In that way, models suitable for deploying in the edge devices can be obtained.

References

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Randaugment Le. Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2019.

[2] 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.

[3] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[4] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.

[5] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

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

[7] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016.

[8] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[9] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[10] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.