Clustered ImageNet Labels for Training Production-friendly Image Classifier

Youngsoo Lee Seongjoo Moon Yongwoo Lee
KAIST
{youngsoo.lee, sj.moon, ywlee97}@kaist.ac.kr

Abstract

While using a pre-trained model is the easiest option for non-researchers when implementing a deep learning algorithm, most of the pre-trained models are trained with benchmark datasets which are far from the production applications. In this study, we present Clustered ImageNet Labels (CIL) as a dataset for production-friendly image classifiers. To reduce the management and computational cost, we merge labels of ImageNet while keeping the data in ImageNet unchanged, instead of composing a completely new dataset. CIL includes 488 reduced classes, where too finely divided classes in ImageNet are grouped into a general category. We demonstrate that a model trained with CIL obtains higher accuracy compared to the original ImageNet labels, and we show that creating a ready-to-hand image classifier can be done easily by finetuning the ImageNet pre-trained model with CIL.

1. Introduction

While deep learning spreads out the world, it is still not easy to utilize deep learning algorithms for non-researchers. One of the easiest ways for non-researchers to access deep learning algorithms is to use a pre-trained model, which is provided in Tensorflow Hub[1], GitHub[2], or other websites. However, most of the pre-trained models are trained with famous *benchmark* datasets, which are not designed for real-world applications.

As a representative example, ImageNet [3] has been widely used in the image classification, which contains 1,000 classes of images. However, non-expert humans have difficulties in classifying an image into the ImageNet classes since the classes in ImageNet are too fine-grained [7, 8]. For example, ImageNet contains 118 dog species as classes, yet it is difficult for non-experts to distinguish the exact dog species. We argue that most applications do not require fine-grained classification results, and ImageNet pre-trained models are not suitable for production.

[1] https://www.tensorflow.org/hub
[2] https://github.com

Instead of ImageNet, we might consider CIFAR-100 dataset, which contains 100 classes of images. Yet still, CIFAR-100 is not suitable for training real-world models because its resolution is too low (32×32). Creating a new dataset is also *not* considerable. First, building a new dataset requires huge resources and costs. Second, researchers are already too familiar with ImageNet, and it is not a good idea to have separate datasets from the research-purpose to production-friendly since it increases management and computational cost.

In this paper, we introduce *Clustered ImageNet Labels* (CIL) as a solution to convert the ImageNet pre-trained models to production-friendly image classifiers. Instead of composing a completely new dataset, we merge labels of ImageNet to be suitable for production, while keeping the data in ImageNet unchanged. With CIL, researchers can use ImageNet as a benchmark dataset as like before, yet they also can create a production-friendly trained model before deployment. CIL includes 488 classes, where the too finely divided classes (e.g., subclasses in dog) are grouped into a general category. We show that creating a ready-to-hand image classifier can be done easily by finetuning the ImageNet pre-trained model with CIL through this study. We further demonstrate that a model finetuned with CIL achieves better accuracy compared to the ImageNet pre-trained model.

Our contributions consist of the following:

- We propose clustering strategies for ImageNet classes. (§3)
- We present *Clustered ImageNet Labels* (CIL) as a dataset for training production-friendly image classifiers. (§4)
- We demonstrate that a model trained with CIL achieves higher accuracy compared to the original ImageNet pre-trained model. (§5.1)
- We show that creating a ready-to-hand image classifier can be done easily by finetuning a ImageNet pre-trained model with CIL at a low cost. (§5.2)

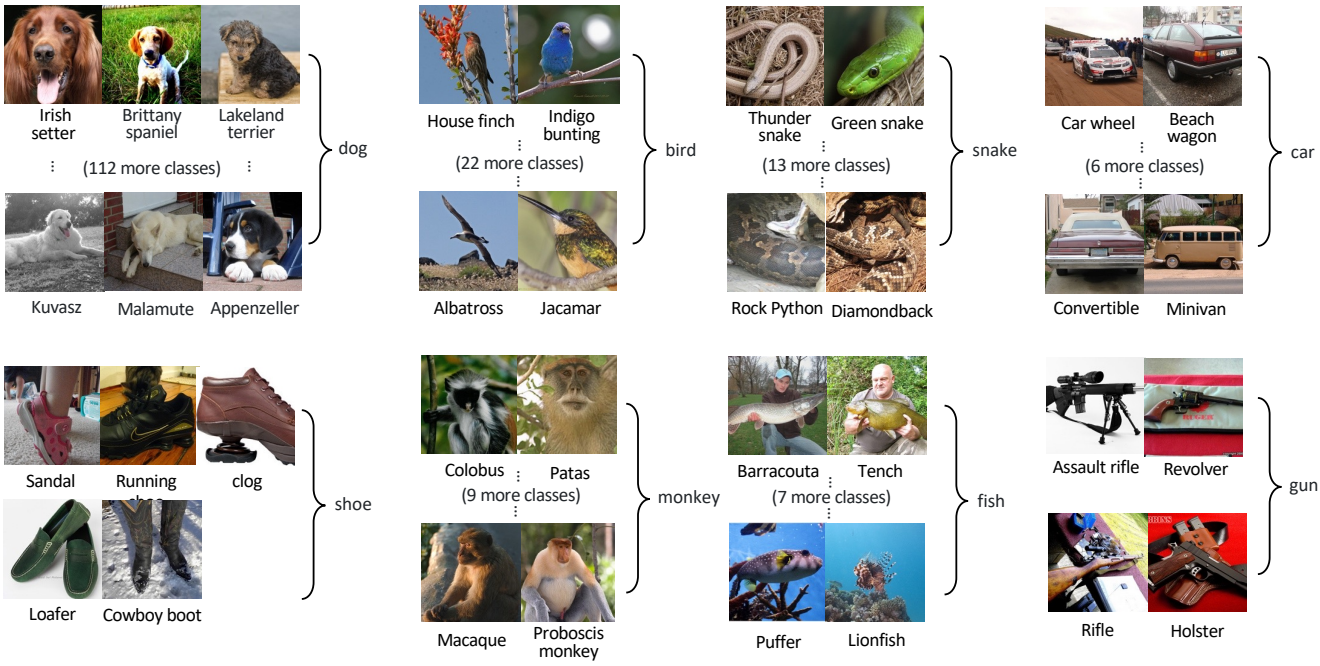

Figure 1: Sample mapping rules in *Clustered ImageNet Labels*, where too finely divided classes in original ImageNet are clustered into a general class.

2. Related Work

To the best of our knowledge, this is the first approach to propose a dataset for production-friendly image classifiers. Although there is no prior work covering production-friendly datasets, many studies have covered the limitations of ImageNet [3] and proposed to modify the data or labels of ImageNet to overcome the limitations.

Variation of ImageNet for faster training. Tiny ImageNet [9] is a variation of ImageNet, which contains fewer classes, fewer images, and lower resolution. Downsampled ImageNet [2] is another variation, where only the resolution is reduced while the number of images and classes are unchanged. Like [9], we reduce the number of classes in ImageNet, however, we carefully group the classes after analyzing the ImageNet hierarchy and user perception. Also, we do not reduce the resolution or number of the images; our goal is not to reduce the training time, but to provide ready-to-hand classifiers.

New ImageNet labels for better training/validation. Several studies [1, 6] have analyzed the limitations of ImageNet labels and propose new labels as alternative options for evaluating vision models. [10] further proposes new ImageNet labels to improve model accuracy on the training stage. While they make modifications on ImageNet labels, their considerations are to validate existing models with new labels or to train better a model with new labels. Our work has a similarity with these studies in that we modify labels of ImageNet, yet we merge classes of ImageNet to produce a production-friendly dataset.

3. Clustering Strategy for ImageNet Classes

Clustering the existing 1,000 labels into a smaller number is a quite subjective task. For example, a *dog expert* might be able to distinguish the exact category when a picture of a Yorkshire terrier is given, but a non-expert may only recognize it as a "dog." In other words, the clustering strategy can be different depending on the targeting domain. As a goal of our paper is to provide ready-to-hand classifiers that can be used generally, our clustering strategy aims to merge labels from an **average person's point of view** as much as possible.

3.1. Rule-based Hierarchical Clustering

As ImageNet is built upon WordNet [5], it is possible to merge labels from the WordNet hierarchy. For example, we could merge classes into one-level higher or two-levels higher ancestor. However, this approach fails on most classes in ImageNet since the depths of hierarchy are not evenly constructed. Figure 2 illustrates a counterexample of the rule-based hierarchical clustering approach.

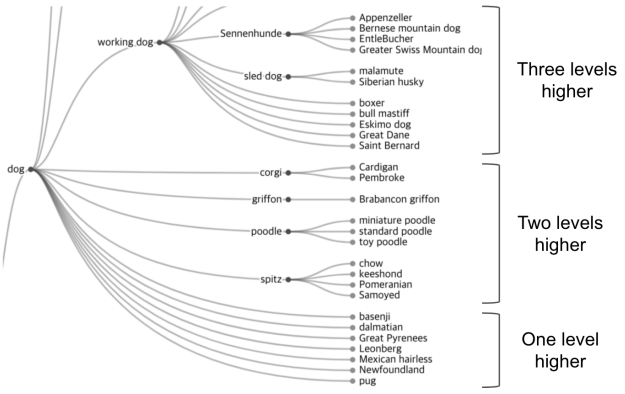

Figure 2: A counterexample of rule-based hierarchical clustering approach. Sub-classes under the class 'dog' have different depths from the ancestor 'dog'.

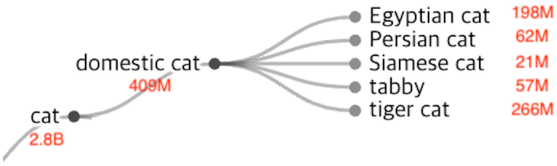

Figure 3: We could use the Google search count of each keyword in ImageNet hierarchy for clustering.

3.2. Using Search Count of Classes

Instead of the fixed rule-based hierarchical clustering algorithm, we could design a better algorithm using meta-information. Based on our goal that we merge labels from an average person's point of view, we crawl the number of Google search results of each class in the ImageNet hierarchy. Figure 3 shows an example part of the ImageNet hierarchy and the number of Google search results of each class. Since people are more familiar with the word "cat" than the word "Egyptian cat", we could cluster to the "cat" as it has more search results.

However, this approach still suffers from exceptional cases. First, there might be a keyword with a high search count at the top of the hierarchy tree. For example, animal has 2.9B search results, which is a greater number than dog (2.2B) and cat (2.8B), thus this approach will cause a too-coarsely clustered result. Second, there might be an incorrect search count due to the homonym. For example, python (409M) indicates a snake in ImageNet, but the search results are exaggerated from the programming language python.

3.3. Survey Application using Collective Intelligence

To overcome the limitations of previous approaches, we implement a survey application for generating clustered labels. The application is shown in Figure 4. To help users

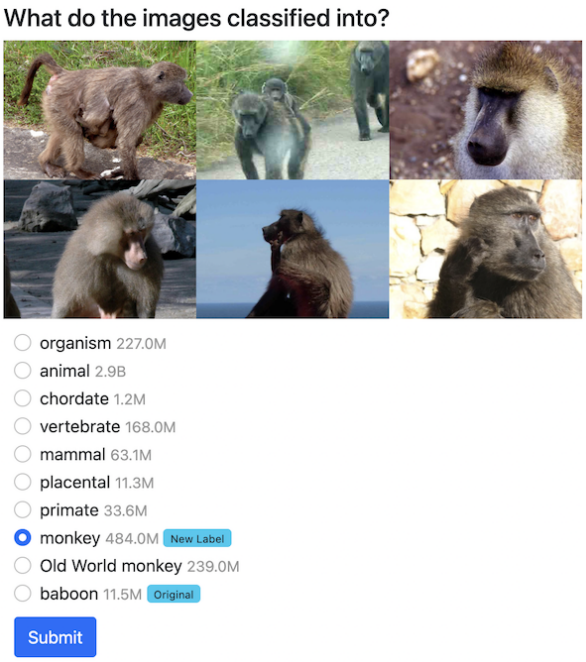

Figure 4: Survey application for generating clustered labels.

classify images more easily, candidates are provided based on the hierarchy, and the Google search count is displayed for each candidate. By using this application, we can collect data from multiple users and obtain clustered results from an average person's point of view while exceptions are also handled.

4. Clustered ImageNet Labels (CIL)

Since it was not possible to collect data from many users due to time and cost limitations in this project, we collected data from only three users using our survey application. Figure 1 illustrates representative samples of original labels and the new labels clustered by CIL. CIL includes 488 clustered classes, where 627 classes among 1,000 original classes are clustered with at least one another class. Among them, dog is clustered with 118 classes with the largest number, followed by bird which is clustered with 26 classes. The CIL file, and training code using CIL, and the survey application code are available at https://github.com/Prev/clustered-imagenet-labels.

5. Experiments

We present two experiments using CIL. We first show that classifying 488 classes is much easier than classifying 1,000 classes while clustered classes are much intuitive for the average human. Next, we show that creating a ready-to-hand image classifier can be done easily by finetuning an

ImageNet pre-trained model with CIL at a low cost.

Experimental Setup. Due the limited time and computational cost, we use 32×32 Downsampled ImageNet [2] instead of original ImageNet. ResNet-18 [4] is used as a baseline model in all experiments. The batch size, momentum, and weight decay are set to 256, 0.9, and 0.0001, respectively.

5.1. Accuracy Comparison

We compare accuracy on downsampled ImageNet using CIL against original ImageNet 1k labels. The models are trained with SGD for 40 epochs with the initial learning rate of 0.1, and the learning rate is divided by 2 at every 10 epochs. Results are given in Table 1, where we obtain higher accuracy on both top1 and top5 when using CIL.

Network	Label	top1 (%)	top5 (%)
ResNet-18	Original	36.3	61.0
ResNet-18	**CIL**	**46.8**	**70.4**

Table 1: Downsampled ImageNet classification accuracy with different labels.

Meanwhile, the results might be regarded to be expectable since classifying 488 classes is easier than classifying 1,000 classes. However, **difficulty** of the task is not just lowered in our experiment. Rather than, it is **adjusted to the level required by the average user**. We argue that this result is meaningful in that we demonstrate that CIL obtains accuracy gain in the applications that do not require too fine-grained classification.

5.2. Finetuning from ImageNet pretrained model

We compare the accuracy and training time of naive training and finetuning on downsampled ImageNet classification using CIL. The pre-trained model is borrowed from §5.1, and the fine-tuned model is trained with SGD for 10 epochs with the initial learning rate of 0.1, and the learning rate is divided by 2 at every 2 epochs.

Finetuning	Epochs	top1 (%)	top5 (%)
No	40	46.8	70.4
Yes	**10**	**48.8**	**72.0**

Table 2: Accuracy comparison between a model trained from the beginning with CIL and a model fine-tuned from an original ImageNet pre-trained model. ResNet-18 architecture is used in both models.

Table 2 and Figure 5 show the results of the finetuning experiment. We originally expected that finetuning would reduce the training time since the model could learn knowledge from the pre-trained model, yet we that finetuning not only increases the training speed but also increases the accuracy. We expect this result came out from the reason that the model trained with the original ImageNet learned more features.

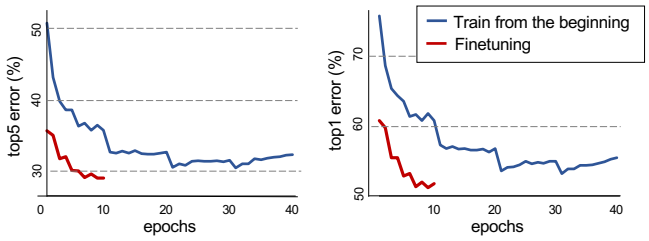

Figure 5: Training on CIL w/wo finetuning.

6. Conclusion

We proposed Clustered ImageNet Labels (CIL) as a compromise dataset for training production-friendly image classifiers. To compose CIL, we proposed three clustering strategies for ImageNet classes, where the survey application using collaborative intelligence was employed as the solution for generating the current version of CIL. By introducing CIL, we argue that non-researchers can implement deep learning algorithms in a much easy way. Our experiments show that a model trained with CIL achieves higher accuracy compared to the original ImageNet label, and we further demonstrated that creating a production-friendly image classifier can be done easily by finetuning an ImageNet pre-trained model with only 10 epochs.

References

[1] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[2] P. Chrabaszcz, I. Loshchilov, and F. Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. 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] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[6] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International*

Conference on Machine Learning, pages 5389–5400. PMLR, 2019.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[8] V. Shankar, R. Roelofs, H. Mania, A. Fang, B. Recht, and L. Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020.

[9] Tiny imagenet — kaggle. https://www.kaggle.com/c/tiny-imagenet.

[10] S. Yun, S. J. Oh, B. Heo, D. Han, J. Choe, and S. Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. *arXiv preprint arXiv:2101.05022*, 2021.