Effectiveness of Function Matching in Driving Scene Recognition

Shingo Yashima
Denso IT Laboratory, Inc.

Abstract. Knowledge distillation is an effective approach for training compact recognizers required in autonomous driving. Recent studies on image classification have shown that matching student and teacher on a wide range of data points is critical for improving performance in distillation. This concept (called function matching) is suitable for driving scene recognition, where generally an almost infinite amount of unlabeled data are available. In this study, we experimentally investigate the impact of using such a large amount of unlabeled data for distillation on the performance of student models in structured prediction tasks for autonomous driving. Through extensive experiments, we demonstrate that the performance of the compact student model can be improved dramatically and even match the performance of the large-scale teacher by knowledge distillation with massive unlabeled data.

Keywords: Knowledge distillation, Function matching, Model compression

1 Introduction

Recognition models in autonomous driving need to be lightweight and operate in real-time. On the other hand, there is a growing performance discrepancy between large-scale, state-of-the-art models and such lightweight models. This stems not only from the insufficient model capacity but also from the lack of the inductive biases in compact models: it has been well known that compact networks are harder to learn and generalize than large networks, even though they may have the capacity to represent the generalized solutions of large models. Therefore, establishing learning methods that exploit the full ability of compact models is crucial in this field.

Knowledge distillation [10,4] overcomes the lack of inductive biases by training a compact student model to emulate a large, generalized teacher model. Its superiority to standard supervised learning has been confirmed in numerous studies [5,18], and the state-of-the-art performance of a given architecture on typical vision tasks (e.g., ResNet-50 on ImageNet classification) is achieved by knowledge distillation [3,17]. In the literature, these decent performances of knowledge distillation have been mainly attributed to the following two factors: (1) Soft labels or features maps produced by the large-scale teacher model contain rich semantic information (i.e., dark knowledge [10]). This leads the student to capture representations that generalize well [16].
(2) Student’s output can be matched to the teacher’s output on data point far from the original training data by using intensely augmented data or auxiliary unlabeled data. This allows the student to be more thoroughly signaled for generalization than supervised learning only with the labeled training data [3].

Most of the existing studies mainly focused on the former factor and developed various distillation methods under a setting where same training data or augmentations as the supervised learning were applied for distillation [8,14]. However, recent studies [3,17,1] reveal that the latter factor is essential to achieving strong performance and simple output matching [10] on such massive data points exhibits state-of-the-art performance in image classification. This explanation on how knowledge distillation works is called function matching [3] and its effectiveness is also discussed theoretically [11,6].

When considering learning compact recognizers for autonomous driving, the concept of function matching seems attractive: we can easily access an almost infinite amount of unlabeled driving scene data. From this viewpoint, matching the student’s output to the teacher’s output at such massive data points is expected to yield well-performed compact models. However, existing studies on knowledge distillation in structured scene recognition mainly focused on designing task-specific distillation losses [12,13] rather than quantitatively investigating the impact of increasing data points on which distillation is performed.

In this study, we examine the effectiveness of function matching in driving scene recognition by conducting extensive experiments on two fundamental tasks in this field: semantic segmentation and object detection. We found that the performance of the student distilled with a large unlabeled dataset not only exceeds the one by standard supervised learning but also matches the performance of the teacher. This result indicates performance gap between teacher and student in supervised learning comes not from the difference in representational capacity but from that in inductive bias. In addition, we show that such a significant improvement is not observed when distilling only with original training data, indicating the number of distilling data points is crucial in function matching.

**Remark.** Although many studies have investigated knowledge distillation methods suitable for structured prediction tasks [12,13] that utilize spatial relation of predictions or features, we focus on the plain output matching as we are interested in how well the student can match the teacher as a function. Somewhat surprisingly, the student can mimic the teacher sufficiently well with such a simple distillation when large unlabeled dataset is available.

In addition, we note that the use of an unlabeled dataset is slightly unusual, but is not novel, having been used effectively in previous studies [10,13]. However, their usage is rather an additional option, and the size of their unlabeled dataset is not substantially larger than the original training dataset. We consider the setting where the number of unlabeled data is almost infinite and we can take (nearly) a new sample in each iteration, which is realistic in driving scene recognition and expected to be more effective.
2 Related Work

Knowledge distillation. Knowledge distillation [10,14] transfers knowledge from a large-scale model to a compact model so as to improve the performance of compact networks. It has been applied to image classification by using the class probabilities produced from the large model as targets for training the compact model [10,12] or transferring the intermediate features [18,21]. More recently, the concept of function matching was introduced, and it was shown that the original output matching [10] on various data points achieves state-of-the-art performance in ImageNet classification [3,17].

Besides image classification, several distillation methods have been proposed for structured prediction task, such as semantic segmentation [24,13] or object detection [12,23,7]. They reported improved performance by exploiting spatial relationship of predictions or feature maps. Our primary interest differs from these studies in that we focus on the impact of the dataset size used in distillation rather than the specific design of distillation loss.

Semi-supervised learning. The use of both labeled and unlabeled data can be seen as semi-supervised learning. Typical setting of semi-supervised learning is few-shot classification, where unlabeled data are utilized because labeled data are insufficient to learn a meaningful classifier [15,19]. In contrast, we assume the size of labeled data is large enough to learn the strong teacher model and leverage unlabeled data for further exploiting the ability of compact student models (overcoming the lack of inductive biases [6]).

3 Experiments

We use BDD100K [25] as a driving scene dataset. It consists of 100,000 driving videos, each 40 seconds long at 720p. Annotations are provided for each reference frame of 100,000 videos for object detection and 10,000 subset videos for semantic segmentation. We adapt these annotated data as an original training dataset for each task. For an auxiliary unlabeled dataset, we use frames cut from original video clips at 6 fps that result in 16,125,000 images, which is more than 1,000 times larger than the labeled dataset for segmentation. Note that train/val/test splits are common in original video clips and annotated datasets, so auxiliary unlabeled dataset does not contain val/test data of annotated datasets.

Throughout the experiments, we compare the performance of the following three training strategies:

(1) Directly train the compact model by supervised learning with the labeled dataset (Supervised).
(2) First train the large model by supervised learning with the labeled dataset, then distill it to the compact model with the same labeled dataset (Distill).
(3) First train the large model by supervised learning with the labeled dataset, then distill it to the compact model with the auxiliary unlabeled dataset (Distill-Aux).
Roughly speaking, we can see the effect of using soft targets instead of hard labels by comparing the performance of (1) and (2). Furthermore, we can see the impact of increasing matching data points by comparing the performance of (2) and (3) as the main interest in this study.

3.1 Semantic Segmentation

We use the segmentation architecture of PSPNet \[27\] with a ResNet-101 and ResNet-18 \[9\] as the teacher and the student network, respectively. For supervised learning, we adopt the standard pixel-wise cross entropy.

**Distillation loss.** For distillation in Distill and Distill-Aux, we consider a simple pixel-wise KL divergence loss:

\[
L_{\text{distill}} = \frac{1}{|I|} \sum_{(x,y) \in I} \sum_{c \in C} \left( -p_{x,y,c}^T \log p_{x,y,c}^S + p_{x,y,c}^T \log p_{x,y,c}^T \right)
\] (1)

where \( p_{x,y,c}^T \) and \( p_{x,y,c}^S \) represent the probability of class \( c \) on the pixel coordinate \((x, y)\) produced by the teacher and the student, respectively. \( I \) is a set of pixel coordinates in prediction, and \( C \) is a set of classes. We do not use any additional loss term with respect to the original dataset’s hard labels as in \[3\].

**Training setup.** In all settings, we train segmentation networks from ImageNet pre-trained weights using stochastic gradient descent (SGD) with the momentum (0.9) and the weight decay (0.0001) with batch size 8. The learning rate is initialized to be 0.01 and is multiplied by \((1 - \frac{\text{iter}}{\text{max.\,iter}})^{0.9}\). For the number of training steps, we sweep over \{80K, 160K, 320K, 640K, 1280K\} as the optimal iteration number depends on training strategies. We apply standard augmentations, including random scaling (from 0.5 to 2.0) and random horizontal flipping, except for the Distill-Aux setting, where we have massive data and do not need to care about overfitting. In all settings, patches of size \(512 \times 1024\) are randomly cropped from (possibly augmented) images.

**Results.** The result in Figure 1 shows that the compact model trained by Distill-Aux performs significantly better than the one trained by Supervised or Distill. This indicates that increasing the number of matching data points is the critical component of the distillation, while the number of labeled data is the same for all training strategies. Table 1 presents class-wise IoU of the best performing model among all training steps on each training strategy. While the improvement in major classes such as road or sky is relatively small, the significant performance gain is observed in more structured classes such as motorcycle or truck. This indicates that the semantic information in the teacher is transferred well to the student in Distill-Aux, thanks to the massive amount of matching data points.

In addition to the performance with respect to the ground truth, we show the prediction agreement between teacher and student in Figure 2. Although it
is natural that the agreement is low in Supervised, where we do not leverage the teacher’s prediction, we can see Distill-Aux achieves much higher fidelity between the teacher and the student than Distill. This high fidelity to the teacher possibly leads to the large performance gain of Distill-Aux from Distill with respect to the ground truth in Figure 1.

Table 1. Class-wise IoUs for semantic segmentation. We pick the best-performing model among all training steps on each training strategy.

| Model          | Method   | mIoU road | swalk | build. | wall | fence | pole | tlight | tsign | veg. |
|----------------|----------|-----------|-------|--------|------|-------|------|--------|-------|------|
| Teacher (R101) |          | 64.83     | 95.68 | 70.68  | 87.94| 28.97 | 54.00| 58.30  | 64.73 | 63.83| 87.49|
|                | Supervised| 61.48     | 94.84 | 66.76  | 86.42| 30.84 | 48.52| 52.04  | 59.99 | 58.67| 86.43|
|                | Distill   | 61.70     | 94.96 | 66.60  | 86.44| 25.75 | 48.89| 51.66  | 58.82 | 58.89| 86.55|
|                | Distill-Aux| 64.74    | 95.59 | 71.11  | 87.92| 28.80 | 54.00| 54.66  | 62.59 | 61.43| 87.82|

| Model          | Method   | terrain | sky    | person | rider | car    | truck | bus   | train | mbike | bike  |
|----------------|----------|---------|--------|--------|-------|--------|-------|-------|-------|-------|-------|
| Teacher (R101) |          | 51.51   | 95.59  | 69.00  | 36.33 | 91.55  | 58.12 | 83.60 | 0.00  | 70.80 | 63.57 |
|                | Supervised| 47.63   | 94.91  | 65.10  | 35.54 | 90.68  | 54.17 | 80.54 | 0.00  | 56.38 | 58.63 |
|                | Distill   | 48.02   | 94.92  | 64.56  | 41.61 | 90.82  | 56.45 | 81.16 | 0.00  | 57.40 | 58.82 |
|                | Distill-Aux| 51.96  | 95.71  | 68.76  | 44.95 | 92.04  | 62.75 | 83.34 | 0.00  | 63.23 | 63.32 |

### 3.2 Object Detection

As single stage detectors are easy to implement output matching, we use the object detection architecture of FCOS [22] with a ResNet-101 and ResNet-18 [9] as the teacher and the student network, respectively. The channel in the detector towers is set to 128 for the student and 256 for the teacher. For supervised learning, we adapt the original loss of FCOS.
Distillation loss. We use the following loss function for output matching, which is a natural extension of the original loss [22] for soft targets:

\[
L_{\text{cls}} = \frac{1}{N_{\text{pos}}} \sum_{(x,y) \in I} \sum_{c \in C} \left[ -p_{x,y,c}^T \log p_{x,y,c}^S - (1 - p_{x,y,c}^T) \log(1 - p_{x,y,c}^S) \right], \tag{2}
\]

\[
L_{\text{reg}} = \sum_{(x,y) \in I} \frac{1}{N_{\text{pos}}} \sum_{c \in C} \frac{p_{x,y,c}^T}{p_{x,y,c}^T} \text{IoU}(t_{x,y}^T, t_{x,y}^S), \tag{3}
\]

\[
L_{\text{center}} = \sum_{(x,y) \in I} \left[ -q_{x,y}^T \log q_{x,y}^S - (1 - q_{x,y}^T) \log(1 - q_{x,y}^S) \right], \tag{4}
\]

\[
L_{\text{distill}} = L_{\text{cls}} + L_{\text{reg}} + L_{\text{center}} \tag{5}
\]

where \(p_{x,y,c}, t_{x,y} \), and \(q_{x,y} \) are the outputs for class probability, regression coordinate, and centerness, respectively. \(I\) is a set of pixel coordinates in the prediction of all feature levels, and \(C\) is a set of classes. \(N_{\text{pos}} = \sum_{(x,y) \in I} \sum_{c \in C} p_{x,y,c}^T\) and IoU is the GIoU loss as in [26]. When replacing the teacher’s output with hard label, we recover the original FCOS loss.

Training setup. Similar to semantic segmentation, we train detection networks from ImageNet pre-trained weights using stochastic gradient descent (SGD) with the momentum (0.9) and the weight decay (0.0001) with batch size 32. The learning rate is initialized to be 0.01 and is multiplied by \((1 - \frac{\text{iter}}{\text{max iter}})^{0.9}\). For the number of training steps, we sweep over \{80K, 160K, 320K, 640K\}. We apply standard augmentations, including random scaling (from 0.5 to 2.0) and random horizontal flipping, except for the Distill-Aux setting as in semantic segmentation.

Results. Similar to the results of semantic segmentation, Figure 3 shows that the compact model trained by Distill-Aux performs the best and almost reaches
Fig. 3. mAP results for object detection in all training steps.

Table 2. Detailed results for object detection. We pick the best-performing model among all training steps on each training strategy.

| Model         | Method     | mAP | AP50 | AP75 | APs | APm | API |
|---------------|------------|-----|------|------|-----|-----|-----|
| Teacher (R101)| Supervised | 31.4| 56.3 | 29.7 | 14.6| 37.9| 52.1|
| Student (R18) | Supervised | 29.5| 53.9 | 27.4 | 13.5| 34.4| 48.4|
|               | Distill    | 30.4| 55.1 | 28.3 | 13.7| 35.2| 51.0|
|               | Distill-Aux| **31.2**| **56.3**| **29.0**| **13.7**| **36.3**| **52.6**|

4 Conclusion

In this study, we experimentally investigated the influence of the amount of unlabeled data used in distillation for structured predictions. We confirmed that the use of large unlabeled data significantly improves compact models’ performance. We believe these results are promising for autonomous driving since collecting such a massive amount of unlabeled data is easy in this domain. Although the type of the backbone architecture used for teachers and students was identical (ResNet) in this study, it is an important issue to investigate the effectiveness of function matching in non-identical settings [20] (e.g., transformer teacher and ResNet student) because compatibility with hardware heavily depends on these architectures. This might give more principled strategies for learning real-time models for driving scene recognition.
References

1. Asano, Y.M., Saeed, A.: Extrapolating from a single image to a thousand classes using distillation. arXiv preprint arXiv:2112.00725 (2021)
2. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: Advances in Neural Information Processing Systems. vol. 27 (2014)
3. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., Kolesnikov, A.: Knowledge distillation: A good teacher is patient and consistent. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10925–10934 (2022)
4. Bucilă, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 535–541 (2006)
5. Cho, J.H., Hariharan, B.: On the efficacy of knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4794–4802 (2019)
6. Cotter, A., Menon, A.K., Narasimhan, H., Rawat, A.S., Reddi, S.J., Zhou, Y.: Distilling double descent. arXiv preprint arXiv:2102.06849 (2021)
7. Dai, X., Jiang, Z., Wu, Z., Bao, Y., Wang, Z., Liu, S., Zhou, E.: General instance distillation for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7842–7851 (2021)
8. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International Conference on Machine Learning. pp. 1607–1616. PMLR (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
11. Hsu, D., Ji, Z., Telgarsky, M., Wang, L.: Generalization bounds via distillation. In: International Conference on Learning Representations (2021)
12. Li, Q., Jin, S., Yan, J.: Mimicking very efficient network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6356–6364 (2017)
13. Liu, Y., Shu, C., Wang, J., Shen, C.: Structured knowledge distillation for dense prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
14. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. In: International Conference on Learning Representations (2015)
15. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(8), 1979–1993 (2018)
16. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? In: Advances in Neural Information Processing Systems. vol. 32, pp. 4694–4703 (2019)
17. Ridnik, T., Lawen, H., Ben-Baruch, E., Noy, A.: Solving ImageNet: a unified scheme for training any backbone to top results. arXiv preprint arXiv:2204.03475 (2022)
18. Romero, A., Kahou, S.E., Montréal, P., Bengio, Y., Montréal, U.D., Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (2015)
19. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems. vol. 33, pp. 596–608 (2020)
20. Somepalli, G., Fowl, L., Bansal, A., Yeh-Chiang, P., Dar, Y., Baraniuk, R., Goldblum, M., Goldstein, T.: Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13699–13708 (2022)
21. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: International Conference on Learning Representations (2020)
22. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: A simple and strong anchor-free object detector. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
23. Wang, T., Yuan, L., Zhang, X., Feng, J.: Distilling object detectors with fine-grained feature imitation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4933–4942 (2019)
24. Xie, Shuai, Hu, Lin, Zheng: Improving fast segmentation with teacher-student learning. In: Proceedings of British Machine Vision Conference (2018)
25. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2636–2645 (2020)
26. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM International Conference on Multimedia. pp. 516–520 (2016)
27. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2881–2890 (2017)