Project B: Knowledge Distillation for Building Lightweight Deep Learning Models in Visual Classification Tasks

Abstract—This paper is a report for Project A of ECE1512 2022W, University of Toronto. In this paper, we introduce two tasks assigned to us in detail, including the implementation and evaluation of KD based on two couples of Teacher-Student Models. For the

I. Task 1: Knowledge Distillation in MNIST dataset

In this task, we basically implement the load & preprocess of dataset, model construction, training and evaluation for teacher and student models using our own training functions. In addition, we implement the Early-Stopping Knowledge Distillation as the improving algorithm.

A. Question 1

In this section, we read the paper of Geoffrey Hinton [6], and answer the assigned questions.

1) SubQuestion a: The purpose of using Knowledge Distillation is to compress the knowledge in an ensemble to a single model which is much easier to deploy.

2) SubQuestion b: In the paper, what knowledge is transferred from the teacher model to the student model?

3) SubQuestion c: What is the temperature hyper parameter T? Why do we use it when transferring knowledge from one model to another? What effect does the temperature hyper parameter have in KD?

4) SubQuestion d: Explain in detail the loss functions on which the teacher and student model are trained in this paper. How does the task balance parameter affect student learning?

5) SubQuestion e: Can we look at the KD as a regularization technique, here? Explain your rationale.

II. Task 2: Knowledge Distillation in MHIST dataset

A. Question 1

1) SubQuestion a: How can we adapt these models for the MHIST dataset using transfer learning? Talk about the Feature Extraction and Fine-Tuning processes during transfer learning.

The core idea of Transfer learning is consists of taking features learned on one problem, and leveraging them on a new, similar problem. [18] Feature Extraction is
Fine tuning [19]

2) SubQuestion b: What is a residual block in ResNet architectures?

3) SubQuestion c: What are the differences between the ResNetV1 and ResNetV2 architectures?

4) SubQuestion d: What are the differences between the MobileNetV1 and MobileNetV2 architectures?

5) SubQuestion e: How can ResNet architectures, regardless of model depth, overcome the vanishing gradient problem?

6) SubQuestion f: Is MobileNetV2 a lightweight model? Why?

B. Question 2

Explain the effect of transfer learning and knowledge distillation in the performance of the student model. Do pre-trained weights help the teacher and student models perform well on the MHIST dataset? Does knowledge transfer from the teacher to the student model increase the student's performance?

References

[1] Cristian Bucilu˘a, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 535–541, 2006. https://dl.acm.org/doi/10.1145/1150402.1150464.

[2] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4794–4802, 2019. http://openaccess.thecvf.com/content_ICCV_2019/papers/Cho_On_the_Efficacy_of_Knowledge_Distillation_ICCV_2019_paper.pdf.

[3] Xiang Deng and Zhongfei Zhang. Comprehensive knowledge distillation with causal intervention. Advances in Neural Information Processing Systems, 34, 2021. https://openreview.net/forum?id=ch9qlCdrHD7.

[4] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. International Journal of Computer Vision, 129(6):1789–1819, 2021. https://link.springer.com/article/10.1007/s11263-021-01453-z.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In European conference on computer vision, pages 630–645. Springer, 2016. https://arxiv.org/abs/1603.05027.

[6] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2(7), 2015. https://arxiv.org/abs/1503.02531.

[7] Takumi Kobayashi. Extractive knowledge distillation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3511–3520, 2022. https: //openaccess.thecvf.com/content/WACV2022/papers/Kobayashi_Extractive_Knowledge_Distillation_WACV_2022_paper.pdf.

[8] Yann LeCun, L´eon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998. http://vision.stanford.edu/cs598_spring07/papers/Lecun98.pdf.

[9] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 5191–5198, 2020. https://ojs.aaai.org/index.php/AAAI/article/view/5963/5819.

[10] Rafael Mu ¨ller, Simon Kornblith, and Geoffrey Hinton. Subclass distillation. arXiv preprint arXiv:2002.03936, 2020. https://arxiv.org/abs/2002.03936.

[11] Dae Young Park, Moon-Hyun Cha, Daesin Kim, Bohyung Han, et al. Learning student-friendly teacher networks for knowledge distillation. Advances in Neural Information Processing Systems, 34, 2021. https://proceedings.neurips.cc/paper/2021/file/6e7d2da6d3953058db75714ac400b584-Paper.pdf.

[12] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3967–3976, 2019. https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

[13] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550, 2014. https://arxiv.org/abs/1412.6550.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015. https://arxiv.org/abs/1409.0575.

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4510–4520, 2018. https://openaccess.thecvf.com/content_cvpr_2018/papers/Sandler_MobileNetV2_ Inverted_Residuals_CVPR_2018_paper.pdf.

[16] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. arXiv preprint arXiv:1910.10699, 2019. https://openreview.net/pdf?id=SkgpBJrtvS.

[17] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In International Conference on Artificial Intelligence in Medicine, pages 11–24. Springer, 2021. https://arxiv.org/abs/2101.12355.

[18] Transfer learning & fine-tuning, Keras developer guide, 2020, https://keras.io/guides/transfer_learning.

[19] Anusua Trivedi, Deep Learning Part 2: Transfer Learning and Fine-tuning Deep Convolutional Neural Networks, Revolutions. https://blog.revolutionanalytics.com/2016/08/deep-learning-part-2.html.