Computer Science and Application 计算机科学与应用, 2020, 10(9), 1625-1630
Published Online September 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.109171

Hans汉斯

深度神经网络知识蒸馏综述

韩　宇

中国公安部第一研究所，北京
Email: 2863428104@qq.com

收稿日期：2020年9月3日；录用日期：2020年9月17日；发布日期：2020年9月24日

摘　要

深度神经网络在计算机视觉、自然语言处理、语音识别等多个领域取得了巨大成功，但是随着网络结构的复杂化，神经网络模型需要消耗大量的计算资源和存储空间，严重制约了深度神经网络在资源有限的应用环境和实时在线处理的应用上的发展。因此，需要在尽量不损失模型性能的前提下，对深度神经网络进行压缩。本文介绍了基于知识蒸馏的神经网络模型压缩方法，对深度神经网络知识蒸馏领域的相关代表性工作进行了详细的梳理与总结，并对知识蒸馏未来发展趋势进行展望。

关键词

神经网络，深度学习，知识蒸馏

A Review of Knowledge Distillation in Deep Neural Networks

Yu Han

The First Research Institute, The Ministry of Public Security of PRC, Beijing
Email: 2863428104@qq.com

Received: Sep. 3rd, 2020; accepted: Sep. 17th, 2020; published: Sep. 24th, 2020

Abstract

Deep neural networks have achieved great success in computer vision, natural language processing, speech recognition and other fields. However, with the complexity of network structure, the neural network model needs to consume a lot of computing resources and storage space, which seriously restricts the development of deep neural network in the resource limited application environment and real-time online processing application. Therefore, it is necessary to compress the deep neural network without losing the performance of the model as much as possible. This article introduces

文章引用: 韩宇. 深度神经网络知识蒸馏综述[J]. 计算机科学与应用, 2020, 10(9): 1625-1630.
DOI: 10.12677/csa.2020.109171

the neural network model compression method based on knowledge distillation, combs and summarizes the relevant representative works in the field of deep neural network knowledge distillation in detail, and prospects the future development trend of knowledge distillation.

Keywords

Neural Network, Deep Learning, Knowledge Distillation

1. 引言

近年来，随着人工智能的不断兴起，深度神经网络已经被广泛应用在计算机视觉、自然语言处理、语音识别等多个领域，并取得了巨大的成功。然而随着深度学习模型性能增加，网络结构越来越深，模型参数越来越多，导致模型需要消耗大量的计算资源和存储空间，这给模型的训练和使用带来了很大的困难。网络结构加深使得模型训练周期变长，且需要大量的数据和强大性能的机器进行支撑；在模型的使用过程中，许多实际应用场景(如自动驾驶、智能对话等)对实时性有较高的要求，并且许多设备(如移动终端)不具备很高的存储条件，这严重制约了深度神经网络在资源有限的应用环境和实时在线处理的应用上的发展。因此，如何在尽量不损失复杂深度神经网络模型的性能的情况下，对模型进行压缩与加速从而有效减小模型的计算量和存储空间，成为了深度神经网络模型有效利用的一个重要问题。主流的深度神经网络压缩与加速的方法主要分为三类[1] [2]：1) 在已有的网络结构基础上进行参数的剪枝、共享、和低秩分解等操作来压缩模型的大小[3]-[9]；2) 通过设计更加紧密的网络结构来进行模型压缩[10] [11] [12]；3)使用知识迁移的方式，将大模型中的知识蒸馏到小模型中，从而提升小模型的性能[13]-[28]。剪枝、量化等参数压缩方法应用在硬件上时往往达不到很好的效果，而基于知识蒸馏的方法能够有效地对模型进行压缩，同时不显著地改变模型的性能。目前，基于知识蒸馏的压缩方法已经被广泛应用于复杂深度学习模型的压缩与加速。本文主要对基于知识蒸馏的神经网络模型压缩方法进行详细地介绍。

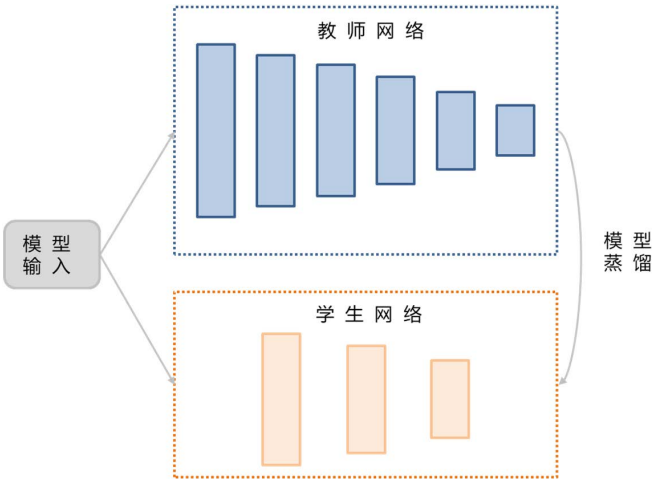

Figure 1. Typical deep neural network knowledge distillation framework
图 1. 典型的深度神经网络知识蒸馏框架

Hinton 等人[13]在 NIPS 2014 中提出了知识蒸馏(Knowledge Distillation, KD)的概念，知识蒸馏是一种常见的模型压缩方法，其将复杂模型或多个模型集成学习到的知识迁移到另一个轻量级的模型之中，使得模型变轻量的同时尽量不损失性能。典型的深度神经网络知识蒸馏框架如图 1 所示，将原始较大的或者集成的深度网络称为教师网络，用于获取知识；将轻量级的模型称为学生网络，用于接收教师网络的知识，并且训练后可用于前向预测。

知识蒸馏方法中的“知识”可以宽泛和抽象地理解成模型参数、网络提取的特征和模型输出等。现有的深度神经网络蒸馏方法根据学习位置的不同可分为基于最后输出层的蒸馏方法、基于中间结果层的蒸馏方法以及基于激活边界的蒸馏方法；根据学习方式的不同可分为基于自我学习的蒸馏方法和基于对抗学习的蒸馏方法。本文将对知识蒸馏各个类别的代表性研究成果进行详细介绍。

2. 基于最后输出层的蒸馏方法

基于最后输出层的模型蒸馏方法的主要思想是以教师模型的输出结果作为先验知识，结合样本真实类别标签来共同指导学生模型的训练。2014 年 Hinton 等人[13]在 NIPS 上提出了一种基于教师-学生网络的知识蒸馏框架，该文章是知识蒸馏的开山之作。Hinton 等人提出的知识蒸馏框架通过软化教师网络的输出来指导学生网络，将学生模型的优化目标分为两部分：1) 硬目标(Hard Target)：学生模型输出的类别概率与样本真实的类别标签(One-hot)之间的交叉熵；2) 软目标(Soft Target)：学生与教师模型软输出结果之间的交叉熵，软输出为经过带温度参数的 Softmax 的输出结果，带温度参数的 softmax 如公式(1)所示：

$$q_i = \frac{\exp\left(z_i/T\right)}{\sum_j \exp\left(z_j/T\right)} \tag{1}$$

其中 T 为温度参数，z_i 是神经网络得到的概率分布，q_i 为软输出。将这两个优化目标进行组合，使得学生模型能够模仿教师模型输出的概率分布，并具有与教师模型相近的拟合能力。

为了使学生模型能够更好地理解教师模型，Kim 等人[14]提出了一种相关因子迁移法(Factor Transfer, FT)来进行知识蒸馏，其主要思想为对模型的输出进行编码和解码。该模型利用卷积运算对教师模型的输出进行编码，并为学生模型添加一个卷积操作，用来学习翻译后的教师知识，最后通过 FT 损失函数来最小化教师和学生网络之间的因子差异。Passalis 等人[15]提出一种概率分布学习法，该方法让学生模型学习教师模型的概率分布。将教师模型中的知识使用概率分布进行表示，通过最小化教师和学生之间概率分布的散度指标来进行知识迁移，使得学习更加容易。

传统的蒸馏学习方法直接最小化教师和学生模型输出值之间的相似性损失，使得学生模型的输出能够尽量接近教师模型的输出。文献[16]和[17]认为这些方法使得学生模型只能学习到教师模型的输出表现，无法真正学习到结构信息。因此，Park 等人[16]提出了一种关系型蒸馏学习法(RKD)，利用多个教师模型的输出构成结构单元，使用关系势函数从结构单元中提取关系信息，并将信息从教师模型传递给学生模型，从而更好的指导学生模型的训练。Peng 等人[17]认为传统的知识蒸馏只关注于教师和学术网络之间的实例一致性，他们提出了相关一致性知识蒸馏方法(CCKD)，该方法不仅考虑了实例一致性，还设计了一个样本间的相关性损失函数约束来实现多个实例之间的相关一致性。

目前以 BERT 为代表的一系列大规模的预训练语言模型成为了自然语言处理领域的流行方法，它们在许多自然语言处理任务上能够取得非常优异的成果，但是这些模型结构十分庞大，且需要大量的数据和设备资源来完成训练过程。2019 年 Tang 等人[18]提出了一种对 Bert 模型进行知识蒸馏的方法，将 Bert 模型蒸馏成 Bi-LSTM 模型。该方法与经典的蒸馏网络类似，其损失函数由两部分组成：Bi-LSTM 与真实

标签之间的交叉熵以及教师和学生网络的概率分布(Logits)之间的均方误差。

3. 基于中间结果层的蒸馏方法

在深度学习中，一般将网络隐藏层的输出看作模型所学习到的特征。基于中间结果层的模型蒸馏方法利用网络中间隐藏层学习到的特征作为知识，指导学生模型学习。

Romero 等人[19]首次提出了基于教师模型中间层进行知识蒸馏的方法 FitNets，该方法不仅让学生模型拟合教师模型的软输出(Soft Targets)，还关注于教师网络隐藏层所抽取的特征。FitNets 方法训练分成两个阶段，第一阶段利用中间层的监督信号指导学生网络，使得学生网络中间层输出拟合教师网络中间层输出；第二阶段使用教师网络的输出作为软目标(Soft Target)对学生网络整体进行蒸馏。

文献[20]提出了基于注意力迁移的蒸馏方法(Attention Transfer, AT)，其本质是指导中间结果的输出，使得学生网络能够模仿教师网络中间层的注意力特征图(Attention Maps)，从而显著提高学生模型的性能。

Huang 等人[21]提出了一种神经元选择性知识蒸馏方法，该方法训练学生网络使其中间层的激活分布与教师中间层的激活分布对齐，采用最大均值差异作为损失函数，衡量教师和学生模型特征之间的差异。

文献[22]认为学习的本质不是学习输出结果，而是学习层与层之间的关系，因此该文章将教师网络中层与层之间的关系映射作为学生网络的学习目标，通过优化教师与学生网络对应层之间的关系矩阵 FSP 之差的 L2 范数来进行蒸馏训练。

4. 基于激活边界的蒸馏方法

从分类的决策边界角度分析，一个关键的问题是分类边界样本的分类问题。简单的小模型真正的弱点在于对边界样本的分类困难，而大模型处理边界的能力优于小模型，因此可以利用学生模型学习教师模型的边界分布，从而实现对教师模型的蒸馏。

Heo 等人[23]提出一种 AB 激活边界学习法进行知识蒸馏，他们认为在模型蒸馏的过程中，不能只利用神经元的激活值来进行蒸馏约束，还应该使用神经元的激活区域进行约束。该方法通过最大化边界误差来引导学生模型获得更强的边界约束能力。此外，Heo 等人[24]还提出一种利用对抗样本进行边界激活学习的方法，该方法利用对抗攻击策略来发现位于决策边界的样本。为了能够更准确地传达关于决策边界的信息，该文献基于对抗边界样本训练学生分类器，以提升学生网络对决策边界的鉴别能力。

5. 基于自我学习的蒸馏方法

根据学习方式的不同，深度神经网络蒸馏方法又可分为基于自我学习的蒸馏方法和基于对抗学习的蒸馏方法。基于自我学习的蒸馏方法不同于传统的教师－学生形式的拟合教师蒸馏框架，自蒸馏不是训练学生网络去拟合教师网络的输出，而是在网络内部对知识进行蒸馏。

Zhang 等人[25]提出了一种自我学习的方法用于知识蒸馏，它通过缩小卷积神经网络的规模来显著提高卷积神经网络的性能(准确性)。该方法首先将网络划分为几个部分，然后将网络较深部分的知识压缩到较浅部分，即蒸馏策略中教师模型与学生模型来自于同一个模型，该方法可以降低蒸馏学习的复杂度。

与自学习的蒸馏方法类似，相互学习策略中也不存在教师模型，而是多个模型之间的学习。[26]提出一种深度互学习方法，该方法的主要目的是为了提升模型的效果，而不过多的考虑模型压缩。该方法对多个学生网络同时进行训练，并利用多个网络的输出结果来相互借鉴、共同学习。各个模型的损失函数包括两部分：1) 经典的真值与输出值之间的交叉熵；2) 不同模型输出分布之间的 KL 散度(Kullback-Leibler Divergence)。

Meng 等人[27]提出一种新的知识嫁接的方法，旨在提高深度神经网络的表示能力。该方法并行地训练多个网络，并对网络的参数进行重要性排序，利用并行网络中的有效权重来替换当前网络中不重要的

权重。为了更好地执行嫁接过程，他们提出了一种基于熵的准则来测量滤波器的信息量，并提出了一种自适应加权策略来平衡网络间的嫁接信息。

6. 基于对抗学习的蒸馏方法

基于对抗学习的蒸馏方法利用生成对抗网络(GAN)来进行模型蒸馏。Xu 等人[28]提出一种使用生成对抗网络进行知识蒸馏的方法，利用条件对抗网络来学习损失函数，以实现知识从教师网络到学生网络的转移。该方法将学生网络作为生成器，生成概率分布结果(Logits)，使用预训练好的教师网络获取输入数据的输出概率分布，然后使用判别器来判别学生网络的输出和教师网络的输出，最终的学习目标为判别器无法区分学生网络和教师网络的输出。

7. 总结与展望

随着深度学习的快速发展，深度神经网络在一些任务中已经超越了人类识别的准确率，但是由于模型结构复杂，参数量巨大等问题的存在，导致模型对存储空间和计算力有较大的要求，因此如何对深度神经网络模型进行压缩成为了一个重要问题。知识蒸馏是一种被广泛应用的模型压缩方法，本文对深度神经网络知识蒸馏相关代表方法进行了详细的梳理与总结。

现阶段深度神经网络知识蒸馏方法大多针对模型结构进行蒸馏，考虑在许多实际应用场景中，不仅有计算资源和存储空间的限制，还存在着在线和离线特征不一致的问题。知识蒸馏方向未来的发展不仅要对复杂深度学习模型的结构进行蒸馏，还应该考虑对模型的输入特征进行蒸馏，利用教师模型将在线计算阶段难以获取到的特征直接蒸馏到学生模型中，使得学生模型在线预测时不需要加入离线特征，从而减少学生模型的特征输入并提升模型性能。

参考文献

[1] 纪荣嵘, 林绍辉, 晁飞, 吴永坚, 黄飞跃. 深度神经网络压缩与加速综述[J]. 计算机研究与发展, 2018, 55(9): 1871.

[2] 张弛, 田锦, 王永森, 刘宏哲. 神经网络模型压缩方法综述[C]//中国计算机用户协会网络应用分会 2018 年第二十二届网络新技术与应用年会论文集. 苏州, 2018: 5.

[3] Han, S., Mao, H. and Dally, W.J. (2015) Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding.

[4] Yoon, J. and Hwang, S.J. (2017) Combined Group and Exclusive Sparsity for Deep Neural Networks. *International Conference on Machine Learning*, Sydney, 6 August 2017, 3958-3966.

[5] Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S. and Zhang, C. (2017) Learning Efficient Convolutional Networks through Network Slimming. 2017 *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2736-2744. https://doi.org/10.1109/ICCV.2017.298

[6] He, Y., Zhang, X. and Sun, J. (2017) Channel Pruning for Accelerating Very Deep Neural Networks. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 1389-1397. https://doi.org/10.1109/ICCV.2017.155

[7] Sun, X., Ren, X., Ma, S. and Wang, H. (2017) Meprop: Sparsified Back Propagation for Accelerated Deep Learning with Reduced Overfitting.

[8] Denton, E., Zaremba, W., Bruna, J., Lecun, Y. and Fergus, R. (2014) Exploiting Linear Structure within Convolutional Networks for Efficient Evaluation. *Advances in Neural Information Processing Systems* 27 (*NIPS* 2014), Montréal, 8 December 2014, 1269-1277.

[9] Lebedev, V., Ganin, Y., Rakhuba, M., Oseledets, I. and Lempitsky, V. (2014) Speeding-Up Convolutional Neural Networks Using Fine-Tuned CP-Decomposition.

[10] Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Adam, H., *et al.* (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[11] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L. (2018) MobileNetV2: Inverted Residuals and Linear

Bottlenecks. *Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

[12] Zhang, X., Zhou, X., Lin, M. and Sun, J. (2018) ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6848-6856. https://doi.org/10.1109/CVPR.2018.00716

[13] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network.

[14] Kim, J., Park, S. and Kwak, N. (2018) Paraphrasing Complex Network: Network Compression via Factor Transfer. *Advances in Neural Information Processing Systems*, Montréal, 3 December 2018, 2760-2769.

[15] Passalis, N. and Tefas, A. (2018) Learning Deep Representations with Probabilistic Knowledge Transfer. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 283-299. https://doi.org/10.1007/978-3-030-01252-6_17

[16] Park, W., Kim, D., Lu, Y. and Cho, M. (2019) Relational Knowledge Distillation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 3967-3976. https://doi.org/10.1109/CVPR.2019.00409

[17] Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhang, Z., *et al*. (2019) Correlation Congruence for Knowledge Distillation. *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 5007-5016. https://doi.org/10.1109/ICCV.2019.00511

[18] Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O. and Lin, J. (2019) Distilling Task-Specific Knowledge from Bert into Simple Neural Networks.

[19] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C. and Bengio, Y. (2014) Fitnets: Hints for Thin Deep Nets.

[20] Zagoruyko, S. and Komodakis, N. (2016) Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer.

[21] Huang, Z. and Wang, N. (2017) Like What You Like: Knowledge Distill via Neuron Selectivity Transfer.

[22] Yim, J., Joo, D., Bae, J. and Kim, J. (2017) A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 4133-4141. https://doi.org/10.1109/CVPR.2017.754

[23] Heo, B., Lee, M., Yun, S. and Choi, J.Y. (2019) Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 3779-3787. https://doi.org/10.1609/aaai.v33i01.33013779

[24] Heo, B., Lee, M., Yun, S. and Choi, J. Y. (2019) Knowledge Distillation with Adversarial Samples Supporting Decision Boundary. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 3771-3778. https://doi.org/10.1609/aaai.v33i01.33013771

[25] Zhang, L., Song, J., Gao, A., Chen, J., Bao, C. and Ma, K. (2019) Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 3713-3722. https://doi.org/10.1109/ICCV.2019.00381

[26] Zhang, Y., Xiang, T., Hospedales, T.M. and Lu, H. (2018) Deep Mutual Learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4320-4328. https://doi.org/10.1109/CVPR.2018.00454

[27] Meng, F., Cheng, H., Li, K., Xu, Z., Ji, R., Sun, X. and Lu, G. (2020) Filter Grafting for Deep Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 6599-6607. https://doi.org/10.1109/CVPR42600.2020.00663

[28] Xu, Z., Hsu, Y.C. and Huang, J. (2017) Training Shallow and Thin Networks for Acceleration via Knowledge Distillation with Conditional Adversarial Networks.