An Experimental Comparison of Knowledge Transfer Algorithms in Deep Neural Networks

Seán Quinn, Kevin McGuinness, Alessandra Mileo

Insight Centre for Data Analytics
Dublin City University

Abstract

Neural knowledge transfer methods aim to constrain the hidden representation of one neural network to be similar, or have similar properties, to another by applying specially designed loss functions between the two networks hidden layers. In this way the intangible knowledge encoded by the network's weights is transferred without having to replicate exact weight structures or alter the knowledge representation from its natural highly distributed form. Motivated by the need to enable greater transparency in evaluating such methods by bridging the gap between different experimental setups in the existing literature, the need to cast a wider net in comparing each method to a greater number of its peers and a desire to explore novel combinations of existing methods we conduct an experimental comparison of eight contemporary neural knowledge transfer algorithms and further explore the performance of some combinations. We conduct our experiments on an image classification task and measure relative performance gains over non-knowledge enhanced baseline neural networks in terms of classification accuracy. We observed (i) some interesting contradictions between our results and those reported in original papers, (ii) a general lack of correlation between any given methods standalone performance vs performance when used in combination with knowledge distillation, (iii) a general trend of older simpler methods outperforming newer ones and (iv) Contrastive Representation Distillation (CRD) achieving best performance.

Keywords: Knowledge Transfer, Knowledge Distillation, Deep Learning, Representation Learning, Machine Learning

1 Introduction

Modern deep neural networks solve difficult tasks by learning incredibly complex decision functions from raw training data and appropriate reward signals such as labels. The fields of artificial intelligence, machine learning and computer vision have benefited hugely from rapid advances in the theory of neural learning in recent years [LeCun et al., 2015]. However, as neural approaches begin to achieve a level of maturity and widespread adoption, attention has increasingly shifted from focusing on how best to exploit their well-established strengths to identifying and addressing their inherent and more fundamental limitations [Lake et al., 2017]. Parallels have been drawn between the somewhat opposing strengths and weaknesses of deep neural networks and "old-school" symbolic AI systems [Sun, 1999] – a key attribute of which is the re-use of sophisticated bodies of knowledge, albeit knowledge represented in a vastly different format. The knowledge learned by a deep neural network is encoded across a network of thousands to millions of neurons arranged in sequential layers, often hundreds deep. Individual neurons and their learned weights, the concrete entities which underpin such knowledge, are uninterpretable and of no utility when considered in isolation of the wider network structure. Therefore the knowledge contained within a neural network is an abstract function encoded in a distributed format across the entirety of the network's weights. In this sense the knowledge we seek to reuse is intangible and the existing knowledge transfer and reuse methodologies of the symbolic AI domain are not readily applicable in the deep learning domain. However, the lessons learned from knowledge representation in symbolic

AI along with the envisioned benefits it could bring to neural learning [Lake et al., 2017], has meant that the question remains – how do we emulate these knowledge transfer and reuse capabilities in the vastly different neural learning paradigm? Distillation based neural knowledge transfer approaches seek to do this by augmenting a neural networks learning process with inputs derived from another neural network (a teacher) where it is believed the teachers weights contain knowledge beneficial to the student network on a chosen task. This involves constraining the student networks' hidden representation to have similar properties to the teachers, thereby transferring knowledge from teacher to student in an abstract manner. This is achieved by applying specially designed loss functions between transformed representations of activations collected from hidden layers across both teacher and student networks for the same sample of data. By encouraging the student to represent individual pieces of data in a similar way to the teacher, the student's weights are also encouraged to encode a function similar (or with similar properties) to the teachers while appreciating that the student cannot replicate exact weight structures. In this paper we will contrast the performance of eight such contemporary neural knowledge transfer methods on a computer vision task. Comparisons such as this one are necessary to bridge the gap between the sometimes vastly different experimental setups that appear alongside original publications and to cast a wider net in comparing each method to a greater number of its contemporary peers. The comparison is intended to enable greater transparency and more reliable conclusions in establishing the state-of-the-art. Further to this, as an exploratory component, we will examine the combination of the Knowledge Distillation (KD) method [Hinton et al., 2015] with each of the other seven methods, some of these being previously unexplored combinations.

2 Experimental Comparison

2.1 Models & Data

We use ResNet models [He et al., 2016] in our experimental setup as they are one of the few widely known models that are viable for use with all of the methods examined here. Our teacher, the ResNet-56, is 2.7 times deeper and has 3.1 times as many parameters as the ResNet-20 we use as student; thus it has a much greater representational capacity and the ability to encode a much more powerful decision function. In selecting the parameters for methods reproduced here, we endeavoured to stay as close as possible to those reported in the original papers. However, many of the original papers do not report essential training parameters. In these cases we make choices based on (i) achieving functional gradient descent (ii) achieving parameter consistency across methods (where possible) and (iii) maximising classification accuracy on the validation dataset. We use the CIFAR-10 and CIFAR-100 image classification datasets [Krizhevsky, 2009] in this experimental evaluation. All details necessary for reproduction of these experiments including data loading parameters and exact training parameters for all models trained are available on the GitHub page at `github.com/squinn95/KD_IMVIP_21`.

2.2 Discussion of Results

The results of our experimental comparison are shown in Table 1. We also report the performance of a non-knowledge enhanced baseline student network as a benchmark with which to measure relative performance gains. We observe that **KD** [Hinton et al., 2015] performs strongly in the evaluation, showing best performance on CIFAR-10 and second best on CIFAR-100, this is broadly in line with the strong results reported for KD in the papers which reproduce it [Passalis and Tefas, 2018, Huang and Wang, 2017, Tian et al., 2020]. In our paper it outperforms all methods except CRD. The original **FitNet** evaluation does not report a baseline student figure [Romero et al., 2014], so we cannot contrast relative performance gains, it does however report that the method outperforms its teacher despite the student being a much lower capacity model. We did not observe performance this strong in our evaluation for FitNet or for any of the other methods or combination methods. Despite this we observed stronger gains over the baseline for FitNet than expected, as it is reproduced in several other papers [Yim et al., 2017, Passalis and Tefas, 2018, Huang and Wang, 2017]. The original **FSP** evaluation [Yim et al., 2017] compares with only one other method, FitNet, which it is shown to outperform.

Table 1: Top-1 classification accuracy (%) on the CIFAR datasets [Krizhevsky, 2009] – knowledge transfer from ResNet56 to ResNet20 [He et al., 2016]. Relative improvement over baseline student shown in brackets.

Model	**CIFAR-10**	**CIFAR-100**
Teacher	93.68	72.46
Baseline Student	91.74	68.3
KD [Hinton et al., 2015]	**92.66** (+0.92)	69.78 (+1.48)
FitNet [Romero et al., 2014]	92.22 (+0.48)	69.26 (+0.96)
FSP [Yim et al., 2017]	92.18 (+0.44)	68.76 (+0.46)
PKT [Passalis and Tefas, 2018]	91.88 (+0.14)	69.26 (+0.96)
MMD-Linear [Huang and Wang, 2017]	91.86 (+0.12)	68.9 (+0.6)
MMD-Polynomial [Huang and Wang, 2017]	92.06 (+0.32)	68.46 (+0.16)
MMD-Gaussian [Huang and Wang, 2017]	92.62 (+0.88)	69.38 (+1.08)
CRD [Tian et al., 2020]	91.88 (+0.14)	**70.66** (+2.36)
FitNet+KD	92.36 (+0.62)	69.98 (+1.68)
FSP+KD	92.34 (+0.6)	70.42 (+2.12)
PKT+KD	92.56 (+0.82)	69.88 (+1.58)
MMD-Linear+KD	92.04 (+0.3)	69.32 (+1.02)
MMD-Polynomial+KD	92.52 (+0.78)	69.9 (+1.6)
MMD-Gaussian+KD	92.76 (+1.02)	69.08 (+0.78)
CRD+KD	**92.96** (+1.22)	**70.9** (+2.6)

We observed the opposite result, with FitNet achieving marginally higher performance on CIFAR-10 and significantly higher performance on CIFAR-100. We examined the previously unexplored combination of FSP+KD which achieved surprisingly good performance, second best overall on CIFAR-100, it is noteworthy that FSP seems to offer much more when combined with KD than as a standalone method. The original **PKT** evaluation [Passalis and Tefas, 2018] reports the methods outperforming FitNet and KD in a content-based retrieval experimental setup, we observe the opposite in our classification setup, with both methods outperforming PKT. The original **MMD** paper [Huang and Wang, 2017] reports MMD-Polynomial as the strongest performing of the three MMD kernel variants where in our experiments we observe MMD-Gaussian to be the stronger. Due to this finding the authors did not include results for the combinations MMD-Linear+KD and MMD-Gaussian+KD. While we did find MMD-Polynomial+KD to be the stronger combination on CIFAR-100, MMD-Gaussian+KD prevailed on CIFAR-10, making the extra combinations a worthwhile inclusion in our analysis. They further report MMD-Polynomial as surpassing both FitNet and KD on CIFAR-10 and FitNet on CIFAR-100 where we observe MMD-Gaussian surpassing FitNet only on both datasets and do not observe either of the other two kernel variants outperforming KD or FitNet. We see near identical performance for the **CRD** method on CIFAR-100 in our evaluation and the original paper [Tian et al., 2020], confirming it as the strongest performing method among those examined. Curiously it did not perform as strongly when used without KD on CIFAR-10, but held a clear advantage over other methods in the other three scenarios examined.

3 Conclusion

We find all methods examined improved performance over a non-knowledge enhanced baseline on both the CIFAR-10 and CIFAR-100 datasets. We achieved results roughly in line with expectations for three of the

methods examined (KD, MMD-Linear, CRD), weaker performance in three methods (MMD-Polynomial, PKT, FSP) and slightly stronger than expected performance in two methods (MMD-Gaussian and FitNet). We suspect that these some of these disparities may be due to a tendency for evaluation setups to be slightly biased towards the method they report. We found little correlation between the performance of methods when evaluated standalone vs when combined with KD – one does not offer any reliable insight into the other. This is illustrated by the fact that two of the three new knowledge transfer scenarios explored here (FSP+KD, MMD-Gaussian+KD) achieved surprisingly strong results. From this we conclude that there is value in exploring method combinations even when their constituent parts do not appear especially promising. We observed a general trend of older simpler methods (KD, FitNet) outperforming some more recent methods (FSP, PKT, MMD) in contradiction to the initial evaluations which accompanied these newer methods. Finally, we confirm CRD to be the best performing standalone method and CRD+KD to be the overall strongest knowledge transfer regime.

Acknowledgments

Funded by the Irish Research Council GOIPG/2018/2501 and partially by Science Foundation Ireland SFI/12/RC/2289_P2. Supported by an Nvidia Corporation research hardware grant.

References

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Hinton et al., 2015] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[Huang and Wang, 2017] Huang, Z. and Wang, N. (2017). Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

[Krizhevsky, 2009] Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report.

[Lake et al., 2017] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

[Passalis and Tefas, 2018] Passalis, N. and Tefas, A. (2018). Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284.

[Romero et al., 2014] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2014). Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

[Sun, 1999] Sun, R. (1999). Artificial intelligence: Connectionist and symbolic approaches.

[Tian et al., 2020] Tian, Y., Krishnan, D., and Isola, P. (2020). Contrastive representation distillation. In *International Conference on Learning Representations*.

[Yim et al., 2017] Yim, J., Joo, D., Bae, J., and Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141.