Knowledge Distillation and its Computational Complexity

Teo Feliu
May 03, 2023

Abstract

Knowledge distillation is a promising technique in deep learning that allows for the transfer of knowledge from a large, complex teacher model to a smaller, more efficient student model while maintaining high levels of performance. In this paper, we provide a comprehensive overview of knowledge distillation, including explanations of key concepts, as well as the distinctions between ensemble distillation and distillation from a single large teacher. We examine various algorithms and their impact on computational complexity reduction and accuracy preservation in the distilled student models. A significant portion of the paper is dedicated to understanding the trade-offs between computational complexity and accuracy in knowledge distillation. We discuss the results from recent studies that demonstrate the potential of multi-teacher knowledge distillation approaches to reduce the computational complexity of an ensemble model while preserving much of its classification accuracy. In addition, we analyze the relationship between size and complexity in student and teacher models, highlighting the feasibility of achieving substantial size reductions in student models without sacrificing performance. This paper serves as a valuable resource for researchers and practitioners interested in harnessing the power of knowledge distillation to develop efficient deep learning models that maintain high levels of accuracy. By thoroughly examining the theoretical foundations and practical implications of knowledge distillation, we aim to provide insights that can guide the development and optimization of future deep learning models and applications.

1. Introduction

The rapid advancements in deep learning have led to the development of increasingly complex and powerful models. However, these models often come with a high computational cost, making them difficult to deploy in resource-constrained environments or real-time applications. Knowledge distillation, a technique that addresses this challenge, has emerged as a promising solution to efficiently transfer the knowledge from a complex, high-performing teacher model to a smaller, more efficient student model.

At its core, knowledge distillation involves training a smaller student model to mimic the output or behavior of a larger teacher model. The student model aims to capture the essential knowledge of the teacher model while maintaining high levels of performance. This process allows for the development of models that are more suitable for deployment in resource-limited settings, without sacrificing the performance gains achieved by the complex teacher models.

Through the analysis of computational complexity in the context of knowledge distillation, we hope to provide a comprehensive understanding of the challenges and opportunities in this field. Ultimately, our goal is to contribute to the ongoing efforts in creating more efficient and effective deep learning models that can be deployed across a wide range of applications and environments.

2. Key concepts in knowledge distillation

2.1 Teacher and student models

In the process of knowledge distillation, there are two fundamental components: the teacher model and the student model. The teacher model, which is pre-trained and often more complex, acts as a guide for the student model. The student model learns from the teacher's outputs, or predictions, which are in the form of logits. Logits are raw model outputs that represent the confidence level for each class prediction. A softmax function is then applied to these logits, converting them into a probability distribution known as "soft targets" (Gou, Yu, Maybank, & Tao, 2021). These soft targets are used to train the student model, which learns from the teacher model's class prediction probabilities. True class labels, on the other hand, are the actual ground truth labels for each input data point. We will see later how student models learn better from the teacher's soft targets than from the more accurate and confident true class labels.

2.2 Softmax function and temperature

Softmax and temperature are crucial components in the process of knowledge distillation. Softmax is a function that converts raw model outputs (logits) into a probability distribution. It essentially assigns a probability to each class based on the logits, making sure that the sum of probabilities for all classes equals 1.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Figure 1. Softmax function

$\sigma = softmax$
$\vec{z} = input\ vector$
$e^{z_i} = standard\ exponential\ function\ for\ input\ vector$
$e^{z_j} = standard\ exponential\ function\ for\ output\ vector$
$K = number\ of\ classes\ in\ the$
$multiclass\ classifier$

In a large model, we want it to be confident in its predictions, which means having a sharp probability distribution with high probabilities for the correct class and low probabilities for the others.

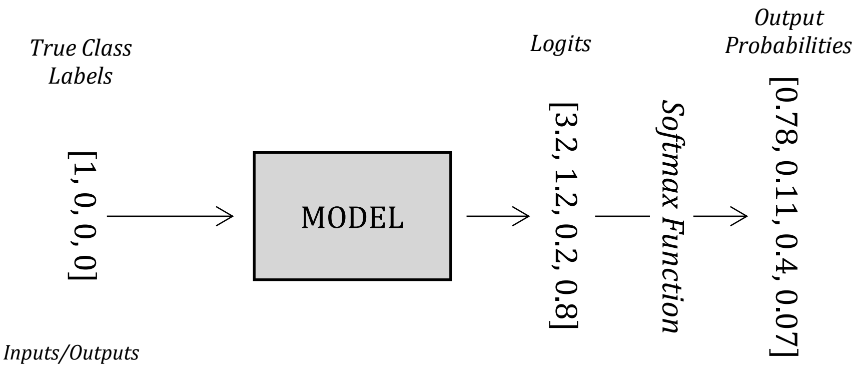

Figure 2. Model Inputs/Outputs

However, when we want a student model to learn from a teacher model's outputs, a sharp probability distribution may not be the most effective way to transfer knowledge. That's where the temperature comes into play. Temperature (T) is a hyperparameter that is used to control the "smoothness" or "sharpness" of the probability distribution produced by the softmax function (Hinton, Vinyals, & Dean, 2015). T>1 results in a smoother distribution, while a lower temperature leads to a sharper distribution. Here is the softmax formula with the added temperature parameter (T):

$$\text{softmax}(x)_i = \frac{e^{\frac{y_i}{T}}}{\Sigma_j^N e^{\frac{y_j}{T}}}$$

Figure 3. Softmax with Temperature

In the context of knowledge distillation, increasing the temperature makes the teacher model's probability distribution smoother, which provides more information about the relationships between different classes. What is meant by smoother is that the differences between the elements of the probability distribution vector get smaller, decreasing the contrast. Instead of just learning the correct class, the student model can now learn about the teacher's evaluation process and how it differentiates between various classes. By adjusting the temperature, we can fine-tune the balance between confidence in the predictions and the richness of information shared between the teacher and student models. In this way, the student model can effectively learn from the teacher model's insights and thought process, resulting in a more accurate and well-informed smaller model.

3. General overview of knowledge distillation

3.1 The Knowledge Distillation process

Train the Teacher Network: The complex teacher network is first trained using the complete dataset. This step requires high computational performance and is typically done offline on high-performing GPUs. This is usually done no differently than when training any other large AI model.

Establish Correspondence: When designing a student network, a correspondence needs to be established between the intermediate outputs of the student and teacher networks. The method of correspondence varies depending on the type of distillation, but it can can involve directly passing the output of a layer in the teacher network to the student network or performing some data augmentation before passing it to the student network.

Forward Pass through the Teacher Network: The input data is passed through the teacher network to obtain all intermediate outputs and then apply data augmentation (if any).

Backpropagation through the Student Network: The outputs from the teacher network and the correspondence relation are used to backpropagate the error in the student network, allowing it to learn to replicate the behavior of the teacher network. Using gradient descent, the student's outputs are compared to those of the teacher model and to the ground-truth labels, or the actual true labels given in the corpus dataset. Frequently, the knowledge distillation process involves observing not only the final layer outputs but also the outputs of the hidden layers, allowing the student model to learn intermediate representations from the teacher model. Using gradient descent and backpropagation, the model's parameters are iteratively updated to minimize the loss. Backpropagation calculates the gradients of the loss with respect to each parameter, and gradient descent adjusts the parameters in the direction of the negative gradient. This process continues until the model's performance converges or is satisfactory.

3.2 Importance and relevance in machine learning

The need for knowledge distillation arises from the fact that state-of-the-art deep learning models, which have achieved remarkable results in various domains such as computer vision, speech recognition, and natural language processing, often require vast computational resources that may not be available on edge devices. Behind the smart models that have been so frequently used by the masses, like ChatGPT, there is an immense amount of computational power needed that is not available to many. By compressing the knowledge from large, cumbersome models into smaller, more efficient models, knowledge distillation enables the deployment of deep learning models on devices with limited computational capabilities, such as mobile phones and embedded systems.

3.3 Knowledge Distillation versus independently training a smaller model

While the theory of creating a reduced model by learning to mimic a more complex model is conceptually appealing, a question that may arise is whether this method of training an efficient model—through knowledge distillation—truly yields superior accuracy compared to independently training a new model on the original corpus of data, with the explicit goal of having fewer parameters, layers, and overall smaller size. After all, the input data is consistently confident in its right answer, and the teacher outputs are a probability distribution with an emphasis on the most likely class.

Creating a smaller model on the same corpus of data as a larger, more complex model may not yield the same level of accuracy, primarily due to the differences in their ability to explore and find optimal solutions given the large span of data that it's given (search space) and their tendencies to converge to different final solutions (convergence behavior). A complex model, having a larger search space, can potentially find better solutions compared to a smaller model. However, simply training a smaller model on the same dataset doesn't guarantee that it will arrive at the same optimal solution (convergence) as the larger model. A smaller, less complex model will likely not know how to use its input data as well as a more complex model. This could result in a student model whose final solution significantly deviates from the optimal solution found by its teacher model.

To address this issue, a teacher model is employed to guide the student model to replicate its behavior. This enables the student model to have its convergence space overlap with the original teacher network's convergence space. By leveraging the teacher model's knowledge representation and soft output distribution, the student model can learn a more concise representation of the knowledge encoded in the larger model. The better performance of a smaller student model over an independently-trained small model will be apparent after discussing results later in the paper.

In essence, the process of knowledge distillation capitalizes on the large model's superior ability to learn concise knowledge representations from data, and transfers this knowledge to the smaller model, which may not have been able to learn it on its own. It is in this context that the accuracy of the naming for "student" and "teacher" models is apparent, as a student is much more likely to learn effectively through a teacher who not only provides it with data but also with insight on reasoning and the thought process, than from just a textbook, even if the textbook is the most explicit and concise way of sharing the content. This allows for the deployment of an efficient and accurate model that is both computationally economical and suitable for resource-constrained environments.

4. Contextualizing knowledge distillation

4.1 Ensemble learning and model compression

The concept of knowledge distillation has grown to be used in areas beyond just student-teacher reduced models. Ensemble learning, which involves training multiple models on the same data and averaging their predictions, has been proven to improve the performance of various machine learning algorithms (Hinton, Vinyals, & Dean, 2015). However, using an ensemble of models, while linearly complex, can still be computationally expensive and cumbersome, particularly for deployment in resource-constrained environments and when the number of models reach the thousands. In their paper, Caruana and his collaborators demonstrated that it is possible to compress the knowledge contained in an ensemble into a single model, called Model Compression, which is easier to deploy (Buciluă, Caruana, & Niculescu-Mizil, 2006). This approach was further developed using a different compression technique, resulting in significant improvements in various applications, including the acoustic model of a heavily used commercial system. The main idea behind model compression is to use a fast and compact model to approximate the function of several equally-sized models working together. The slower, complex ensemble model can be compressed into the fast, compact student model with little loss in performance. When treating the ensemble of models like one, complex model, it is apparent how reducing an ensemble of models into a singular model is the same concept as reducing a cumbersome teacher model into a more efficient student model (Allen-Zhu & Li, 2023).

5. Overview of knowledge distillation techniques

This section provides an overview of various knowledge distillation techniques that have been developed to effectively transfer the knowledge from a teacher model to a student model. These techniques aim to capture different aspects of the teacher model's knowledge, including output probabilities, attention maps, and intermediate representations, in order to help the student model learn more effectively and efficiently. By understanding and combining these approaches, researchers and practitioners can develop more accurate and efficient student models that are suitable for a wide range of real-world applications.

5.1 Soft target training

Soft target training is a knowledge distillation technique that involves training the student model to match the teacher model's probability distribution over the classes. This is done by minimizing the loss between the student's soft targets and the teacher's soft targets, as well as the True class labels. As this standard technique serves as the basis for many variations and extensions in the field of knowledge distillation, it is used to explain the overarching concept and principles of knowledge distillation as a whole in this paper. This technique is most closely related to response-based knowledge representation, as it focuses on transferring the teacher model's output probabilities (responses) to the student model. By minimizing the loss between the student's soft targets and the teacher's soft targets, as well as the True class labels, the student model learns the teacher model's way of making predictions.

5.2 Attention transfer

In order to accurately learn from the teacher model, the student model cannot just look at the teacher's outputs, but also at the hidden layers within the teacher model and, in the case of Attention transfer, the way the teacher looks at the input data. Attention transfer, introduced by Zagoruyko and Komodakis, focuses on transferring the spatial attention maps of the teacher model to the student model (Zagoruyko & Komodakis, 2017). These attention maps represent how the teacher model "pays attention" to specific regions within the input data to make its predictions.

In the attention transfer process, the student model is trained to mimic the attention maps generated by the teacher model. By learning to focus on the same regions as the teacher model, the student model develops a better understanding of the most important features within the input data. This helps the student model to make more accurate predictions, even with its smaller and less complex architecture.

To implement attention transfer, a loss function is utilized to minimize the difference between the teacher's and student's attention maps at multiple layers of their respective architectures. Attention transfer is associated with feature-based knowledge representation, as it focuses on transferring the important features learned by the teacher model to the student model by aligning their attention maps. This helps the student model to focus on the same regions as the teacher model, developing a better understanding of the most important features within the input data. By leveraging the strengths of both techniques, the student model can better replicate the teacher model's knowledge, leading to improved accuracy and efficiency in various applications.

5.3 Knowledge distillation with hint layers

Building upon the concepts of soft target training and attention transfer, which look at how the teacher mimics the output and looks at the input, knowledge distillation can also be enhanced by incorporating hint layers into the process. Hint layers provide additional guidance to the student model by transferring intermediate representations from the teacher model (Romero et al., 2014). This approach helps the student model capture not only the final output probabilities but also the underlying reasoning of the teacher model at various stages of the learning process.

Selected intermediate layers from the teacher model, referred to as "hint layers," are used to provide guidance to corresponding layers in the student model, called "guided layers." The guided layers are chosen to have a similar structure as the hint layers but with a smaller and more compact architecture. By minimizing the difference between the activations of the hint and guided layers, the student model is encouraged to learn the teacher model's intermediate representations, thereby gaining a deeper understanding of the data.

To implement this technique, a hint loss function is employed to minimize the discrepancy between the activations of the hint and guided layers. This hint loss is then combined with the standard cross-entropy loss, or even with distillation loss from soft target training and attention loss from attention transfer, to create a joint loss function. By optimizing this joint loss function, the student model effectively learns from the teacher model's knowledge at multiple levels of abstraction, ultimately improving its overall performance.

This approach is related to both feature-based and relation-based knowledge representation, as it encourages the student model to learn the teacher model's intermediate representations (features) and the relationships between them at various levels of abstraction. By combining this approach with soft target training and attention transfer, the student model can effectively replicate the teacher model's knowledge across different stages of the learning process, leading to more accurate and efficient models suitable for a wide range of applications.

6. Types of knowledge distillation

Knowledge distillation techniques can be broadly categorized into three types of distillation schemes: offline distillation, online distillation, and self-distillation. Each of these schemes has its unique approach to transferring knowledge from a teacher model to a student model, offering different benefits and drawbacks depending on the specific use case.

6.1 Offline Distillation

Offline Distillation is the most traditional and widely used approach to knowledge distillation, and is the main focus of this paper. In this scheme, the teacher network is pre-trained and then frozen, meaning it is not updated while the student network is being trained. Most existing knowledge distillation methods work in an offline manner. The primary focus in offline distillation research is on improving the knowledge transfer mechanism, with less attention given to the teacher network architecture.

6.2 Online Distillation

In cases where a large pre-trained teacher model is unavailable, as assumed in offline distillation methods, online distillation can be used (Gou, Yu, Maybank, & Tao, 2021). In this scheme, the teacher and student networks are trained simultaneously. For example, a recent paper proposed an online mutual knowledge distillation method that aims to fuse features from multiple sub-networks. These sub-networks, which are ensembled, and the fusion module are learned by mutually teaching each other using response-based knowledge.

A recent study by Harutyunyan et al. demonstrated the importance of online distillation in the cases where there is a large gap between teacher and student accuracy (Harutyunyan et al., 2023). The authors found that online distillation with proper temperature scaling can significantly decrease this gap, resulting in a more accurate student model. This finding highlights the potential of online distillation as an effective approach to improve the performance of student models while reducing computational complexity.

6.3 Self-Distillation

Conventional knowledge distillation faces two key challenges: the choice of teacher models significantly impacts student model accuracy, and the student models often cannot achieve the same level of accuracy as teacher models, leading to potential accuracy degradation during the inference period. Self-distillation addresses these problems by having the same network act as both the teacher and the student (Gou, Yu, Maybank, & Tao, 2021).

One method proposed in a paper addresses the weakly-supervised object detection problem using a comprehensive attention self-distillation (CASD) approach. This approach conducts self-distillation on the weakly-supervised object detection network itself to enforce consistent spatial supervision on objects, approximating the comprehensive attention simultaneously through multiple transformations and layers of the same image. Self-distillation enables instance-balanced and spatially consistent supervision, resulting in robust bounding box localization.

7. Computational Complexity

7.1 Ensemble Models

In a recent study by Konrad Zuchniak, a multi-teacher knowledge distillation approach was explored, demonstrating promising results in terms of accuracy and computational complexity (Zuchniak, 2023). The paper's teacher model is an ensemble of N teacher models, whose results are aggregated to perform much better than the results of one of the teacher models independently. The paper looks at reducing the ensemble model a singular student model the size and computational complexity of one of the teacher models in the ensemble. Out of the many techniques that were explored, the most successful involved having the student mimicking all of the teachers' results independently, and then aggregating them in the same way the ensemble model would. The student model's accuracy had a slightly lower classification accuracy than the ensemble of all teachers, but was found to be closer to the ensemble of teachers rather than that of a single teacher model, while maintaining the computational complexity equivalent to only one of the teachers.

Despite a slightly lower classification accuracy, these results should not be discarded as it is well known that a model can always be improved in accuracy by increasing the computational complexity and number of parameters. When sufficient computational power is available to run the ensemble of teacher models, it would indeed be beneficial to employ the ensemble, as increasing the number of teachers (N) in the ensemble would continuously improve the model's performance asymptotically. However, the findings of this study show that the multi-teacher knowledge distillation approach allows for the reduction of computational complexity to that of a single teacher model while essentially maintaining the entire classification accuracy of all N models.

These results are highly promising, especially in situations where the computational power required to run N teacher models is not available. The multi-teacher knowledge distillation approach presents an efficient and practical solution for real-world applications, allowing for the aggregation of knowledge from multiple teachers into a single student model without sacrificing much in terms of accuracy and requiring only the computational complexity of one base model for inference.

7.2 Complex Teacher Model Reduction to Student

In looking at results from a complex teacher model to a distilled student model, a study by Alkhulaifi et al. offers valuable insights into the relationship between size and complexity in student and teacher models (Alkhulaifi, Alsahli, & Ahmad, 2021). The authors conducted a comprehensive evaluation of various knowledge distillation algorithms and their impacts on size reduction and accuracy.

The majority of the algorithms that performed better than average achieved more than 70% size reductions while maintaining less than a 5% reduction in accuracy. In contrast, algorithms that performed worse tended to have size reductions ranging from 50% to 60%, but some of them had less than a 1% reduction in accuracy. The performance of these algorithms is objective depending on the prioritization between computational complexity reduction and accuracy preservation.

To further evaluate the success of the distillation, the paper observed baseline models that were the same size as the student models but trained on the dataset and not through the teacher. The accuracy of the student model was almost always closer to that of teacher's than to the baseline model's accuracy, and in some cases the student's accuracy even surpassed the teacher's accuracy. The findings show that it is possible to achieve substantial size reductions in student models while maintaining or even improving their accuracy compared to their teacher counterparts. This highlights the potential of knowledge distillation in reducing computational complexity without sacrificing model performance.

8. Conclusion

Knowledge distillation is a powerful technique for model compression and optimization, offering significant reductions in computational complexity while maintaining or even improving the performance of the distilled student models. Throughout this paper, we have explored various aspects of knowledge distillation, from the underlying concepts such as softmax and temperature to the distinction between ensemble distillation and distillation from a large teacher. Furthermore, we discussed different algorithms and techniques and their impacts on size reduction and accuracy.

Recent studies showcase the promising potential of knowledge distillation in real-world applications where computational power and resources may be limited. The multi-teacher knowledge distillation approach, in particular, offers a practical solution for efficiently aggregating knowledge from multiple teachers into a single student model without sacrificing much in terms of accuracy.

As the field of deep learning continues to evolve, it is essential to develop strategies that can reduce computational complexity without compromising model performance. By letting billion-parameter models be the only forefront of AI model technology, we are constraining ourselves to let only those with large enough computational power to push the industry forward. Knowledge distillation, as demonstrated in this paper, is a promising technique that reduces industry-leading models into equally-accurate models that are accessible to everyone. Future research in this area should focus on exploring new distillation methods, optimizing the trade-off between computational complexity and accuracy, and investigating the applicability of knowledge distillation in different domains, such as natural language processing and reinforcement learning. By further refining and enhancing the knowledge distillation process, we can pave the way for the development of more efficient, accurate, and accessible deep learning models that cater to the diverse needs of real-world applications.

9. References

Alkhulaifi, A., Alsahli, F., & Ahmad, I. (2021). Knowledge Distillation in Deep Learning and its Applications. PeerJ Computer Science, 7, e474. doi:10.7717/peerj-cs.474

Allen-Zhu, Z., & Li, Y. (2023). Towards Understanding Ensemble, Knowledge Distillation, and Self-Distillation in Deep Learning. arXiv:2012.09816 [cs.LG]. Retrieved from https://arxiv.org/abs/2012.09816

Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model Compression. Retrieved from https://www.cs.cornell.edu/~caruana/compression.kdd06.pdf

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge Distillation: A Survey. International Journal of Computer Vision, 129(6), 1789-1819. doi:10.1007/s11263-021-01453-z

Harutyunyan, H., Rawat, A. S., Menon, A. K., Kim, S., & Kumar, S. (2023). Supervision Complexity and its Role in Knowledge Distillation. arXiv:2301.12245 [cs.LG]. Retrieved from https://arxiv.org/abs/2301.12245

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [stat.ML]. Retrieved from https://arxiv.org/abs/1503.02531

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). FitNets: Hints for Thin Deep Nets. arXiv:1412.6550 [cs.LG]. Retrieved from https://arxiv.org/abs/1412.6550

Zagoruyko, S., & Komodakis, N. (2017). Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. arXiv:1612.03928 [cs.CV]. Retrieved from https://arxiv.org/abs/1612.03928

Zuchniak, K. (2023). Multi-teacher knowledge distillation as an effective method for compressing ensembles of neural networks. ArXiv preprint arXiv:2302.07215.