Literature review:
Knowledge distillation

Raouf Bahsoun
gusbahsra@student.gu.se
DIT892

Advisors

Yinan Yu
yinan@chalmers.se

Charlotte von Numers
charlotte.vonnumers@astrazeneca.com

Luis Arevalo
imarevost@gmail.com

Introduction

Deep learning has been a subject of great advancements over the past years, which comes with the need for efficient model compression techniques. One method that has received a considerable amount of attention is knowledge distillation. Knowledge distillation is a process where a smaller student model learns from a larger teacher model. This technique has shown to increase model performance and reduce model size. In this literature review, three articles that contribute to the advancement of knowledge distillation and model compression techniques are explored. The article “Knowledge Distillation: A Good Teacher is Patient and Consistent” by Lucas Beyer et al., explores knowledge distillation techniques, with focus on the importance of teacher model behavior in knowledge distillation. The second article, “CHEX: Channel Exploration for CNN Model Compression” by Zejiang Hou et al., discusses channel exploration for Convolutional Neural Networks (CNN). The third article, “Densely Distilled Flow-Based Knowledge transfer in Teacher-Student Framework for Image Classification” by Ji-Hoon Bae et al., presents a flow-based knowledge transfer approach within the teacher-student architecture.

Methodology

The authors from the articles use different methods and experiments. Beyer et al. introduce temperature scaling as a mechanism for controlling the level of similarity between teacher and student predictions. They use various datasets and CNN architectures in order to conduct their experiments. Empirical evidence demonstrates the effectiveness of the proposed knowledge distillation approach.

Hou et al. present a systematic approach to channel exploration, using L1 (Lasso) regularization for score calculation. Experiments are conducted on various datasets and CNN architectures. The results show the practicality and effectiveness of channel exploration in achieving substantial model compression while preserving performance.

Bae et al. investigate dense connections and knowledge distillation in the form of flows.

Their experiments, mainly on image classification tasks, confirm the effectiveness of their approach.

Findings

Beyer et al. discuss the critical role of teacher model behavior in knowledge distillation, emphasizing the significance of teacher patience and consistency. Their research shows the importance of patient and consistent teacher models, which contribute significantly to improving the student model's generalization and accuracy of student models. This demonstrates the exchange between teacher and student in the knowledge transfer process. The authors also emphasize the role of temperature scaling in knowledge distillation. Adjusting the temperature parameter can control the level of similarity between teacher and student predictions. Various loss functions for knowledge distillation, including mean squared error (MSE), Kullback-Leibler (KL) divergence, and cross-entropy loss are explored in order to provide insights into when each loss function is most effective for different knowledge transfer scenarios.

Hou et al. introduce channel exploration as an innovative approach to CNN model compression, with a focus on channel pruning. Channel pruning, guided by importance scores learned through L1 regularization, significantly reduces model size and computational complexity while maintaining accuracy. The article explores the impact of channel exploration on different CNN architectures, revealing promising outcomes for efficient deployment.

Bae et al. propose a flow-based knowledge transfer approach within the teacher-student framework for image classification. Their research introduces a dense distillation process that enhances knowledge transfer from teacher to student models, resulting in improved student performance. The flow-based knowledge transfer technique presents a unique perspective on model compression.

Future Developments

The findings of Beyer et al. hold substantial implications for the development of effective knowledge distillation strategies. In future developments, one may explore the extension of these techniques to specific domains and further optimization for real-world applications. Development in the dynamic of the teacher-student interaction mechanisms that adaptively adjust the teacher's behavior during training based on the student's learning progress. This could lead to more effective knowledge transfer.

Hou et al. findings have great implications for deep learning and computer vision, with opportunities to open up for more efficient CNNs in constrained environments. Future research may investigate the application of these techniques to other neural network architectures and the optimization of trade-offs between model size and performance. This could also include developing more efficient algorithms for channel exploration, such as techniques that identify and prune redundant channels with reduced computational complexity, making channel exploration more practical for large-scale networks.

Bae et al. extend the possibilities of knowledge transfer in teacher-student frameworks. Exploration of the adaption of flow-based knowledge transfer to other computer vision tasks and its application can be further researched, for example investigating how this approach can be optimized for natural language processing, speech recognition, or other tasks. Furthermore, optimizing the architecture of student models to further enhance performance could be investigated.

Conclusion

To conclude, these articles contribute important insights to the subject of knowledge distillation and model compression in deep learning. “Knowledge Distillation: A Good Teacher is Patient and Consistent” highlights the importance of teacher-student dynamics, “CHEX: Channel Exploration for CNN Model Compression” introduces a promising channel exploration technique, and “Densely Distilled Flow-Based Knowledge transfer in Teacher-Student Framework for Image Classification” presents a flow-based knowledge transfer approach. These articles aim to simplify the hunt for efficient and effective deep learning models. They provide technical insights and potential space for further research and future developments.

References

Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., & Kolesnikov, A. (2022). Knowledge distillation: A good teacher is patient and consistent. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Hou, Z., Qin, M., Sun, F., Ma, X., Yuan, K., Xu, Y., Chen, Y., Jin, R., Xie, Y., & Kung, S. (2022). CHEX: Channel Exploration for CNN Model Compression. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Bae, J., Yeo, D., Yim, J., Kim, N., Pyo, C. S., & Kim, J. (2020). Densely Distilled Flow-Based Knowledge Transfer in Teacher-Student Framework for Image Classification. *IEEE Transactions on Image Processing, 29,* 5698-5710.