Authors:

Omer Ben Simon

Roei Aviv

Yuval Loria

Github Link : https://github.com/OmerBentzi/Knowledge_Distillation

Department of Intelligent Systems Engineering, Afeka College of Engineering, Tel Aviv, Israel

Abstract

Knowledge Distillation (KD) serves as a critical technique for compressing large-scale teacher models into efficient student networks, enabling deployment in resource-constrained environments. Traditional KD methods focus on soft output alignment, often neglecting the internal spatial reasoning and attention mechanisms that are vital to the teacher's superior performance. Recent research has explored Attention Transfer (AT) and Feature Alignment (FA) as individual enhancements to KD, yet the synergistic effect of their combination remains underexplored.

In this study, we introduce a comprehensive KD framework that integrates both Attention Transfer and Feature Alignment simultaneously, guided by causally influential regions derived from gradient-based attention masks. We hypothesize that this combined approach can significantly enhance the student model's ability to replicate the teacher's internal reasoning patterns, leading to superior accuracy and generalization.

Experiments conducted on the CIFAR-10 dataset validate this hypothesis. While standard KD improved validation accuracy from 80.56% (Baseline) to 83.64%, and individual enhancements (FA and AT) yielded 82.78% and 83.53% respectively, the combined method (FA + AT) achieved the highest validation accuracy of 83.77%. This demonstrates that the integration of attention and alignment mechanisms provides complementary benefits, facilitating more effective knowledge transfer than using either method in isolation.

Our findings highlight that synergistically combining attention-based and alignment-based distillation strategies is essential to maximizing student model performance. This approach narrows the performance gap to larger teacher models while maintaining model efficiency, paving the way for robust and deployable KD frameworks in real-world applications.

Introduction

Background and Motivation

Knowledge Distillation (KD) has emerged as a pivotal technique for compressing large, high-performance neural networks into compact student models suitable for deployment in environments with limited computational resources. By leveraging the soft target distributions of a well-trained teacher model, KD enables student networks to inherit knowledge beyond hard ground-truth labels, facilitating better generalization and efficiency.

However, traditional KD approaches are inherently limited in that they primarily focus on aligning the soft output logits, often neglecting the internal reasoning processes and spatial attention mechanisms that are critical to the teacher model's superior performance. The student model, in these cases, learns to imitate the final outputs but lacks the deeper understanding of where and how the teacher focuses within the input space, leading to sub-optimal feature learning, reduced robustness to corrupted data, and limited generalization.

Recent advancements have introduced mechanisms such as Attention Transfer (AT) and Feature Alignment (FA) to address this gap by encouraging student models to mimic the teacher's attention patterns and intermediate feature representations. However, most prior works have examined these enhancements in isolation, without fully exploring the synergistic effect of combining attention transfer with alignment mechanisms in a unified KD framework.

This research is motivated by the hypothesis that combining Attention Transfer and Feature Alignment, while focusing only on causally relevant regions of the input (using gradient-based causal masks), can significantly enhance the effectiveness of knowledge distillation. By guiding the student not only in output distribution alignment but also in spatial focus and feature-level reasoning, the distillation process becomes more comprehensive and structurally informative.

Thus, the primary objective of this study is to investigate whether a joint distillation strategy that integrates both masked attention transfer and masked alignment can lead to superior student performance, in terms of accuracy, robustness, and efficiency, compared to traditional KD methods and singular enhancement techniques. The CIFAR-10 dataset serves as the benchmark for evaluating this proposed methodology.

Research Questions and Objectives

1. Can the combination of Attention Transfer (AT) and Feature Alignment (FA) within a Knowledge Distillation (KD) framework significantly improve the accuracy, generalization, and robustness of compact student models compared to applying these techniques individually?

2. Does selectively transferring only the causally relevant attention regions from the teacher to the student enhance the efficiency of the knowledge transfer process?

3. How does the integration of AT and FA affect the trade-off between model performance and inference efficiency in real-world deployment scenarios?

Research Objectives:

- To develop a comprehensive KD framework that integrates both Attention Transfer and Feature Alignment simultaneously.

- To implement causal masking mechanisms that filter irrelevant attention regions, ensuring focused and effective knowledge transfer.

- To evaluate and compare the impact of individual techniques (AT and FA) versus their combination on student model performance using the CIFAR-10 benchmark.

- To analyze the efficiency and scalability of the proposed method in terms of model size, inference time, and accuracy.

- To provide empirical evidence that demonstrates the synergistic benefit of combining spatial attention mechanisms with feature alignment in KD.

Several key works laid the foundation for this research. Hinton et al. (2015) introduced the original concept of Knowledge Distillation (KD), demonstrating how soft target distributions from a large teacher model could guide a smaller student network towards better generalization. Zagoruyko and Komodakis (2017) expanded upon this by proposing Attention Transfer (AT), emphasizing the importance of aligning spatial attention maps between teacher and student to enhance knowledge transfer.

Literature Review

Knowledge Distillation

Knowledge Distillation, introduced by Hinton et al. (2015), is a technique aimed at transferring knowledge from a large, high-capacity teacher model to a smaller, compact student model. The classical KD approach focuses on aligning the soft output distributions (soft targets) of the teacher with those of the student, leveraging the additional information encoded in the soft probabilities to enhance the student's generalization. However, traditional KD methods are limited in that they do not account for the internal representations and spatial focus of the teacher, leading to a gap in the reasoning capability of the student network.

Feature Alignment

Feature Alignment strategies aim to bridge the gap in intermediate feature representations between teacher and student networks. These methods utilize contrastive losses or masked alignment techniques to align the spatial or semantic features of corresponding layers. Yet, aligning feature spaces is challenging due to inherent differences in architecture, depth, and capacity between teacher and student models. Without proper filtering, students may overfit to irrelevant teacher activations.

Attention Transfer

Attention mechanisms have been widely adopted in deep learning to improve model focus on salient regions of the input. Zagoruyko and Komodakis (2017) introduced Attention Transfer in KD, proposing that aligning spatial attention maps between teacher and student can enhance the distillation process. By transferring attention maps, the student learns where to focus, mimicking the teacher's internal reasoning. However, traditional AT methods often involve a blind copy of attention maps, without considering whether all attention regions are causally relevant or beneficial for the student.features, showing promise in complex tasks.

Challenges in Existing Works:

A significant challenge in both AT and FA is that not all attention regions or feature activations of the teacher are beneficial for the student. In some cases, the teacher may attend to regions irrelevant to the primary task, leading to noise in the distillation process. Additionally, balancing multiple loss functions (Cross-Entropy, KL Divergence, Attention Loss, Alignment Loss) introduces complexity in hyperparameter tuning. The interplay between these signals requires precise calibration to ensure no single component dominates the training dynamics.

Furthermore, most existing research has focused on applying AT or FA independently, without a structured investigation into the synergistic effect of combining them within a unified KD framework. There is a gap in literature regarding methods that selectively transfer only the causally influential attention regions, ensuring that the student inherits meaningful reasoning patterns from the teacher, while ignoring redundant information.

Research Gap and Motivation for This Study:
Despite the advancements in KD, AT, and FA, there is limited research addressing the integration of attention and alignment mechanisms with selective, causally guided filtering. This study aims to bridge this gap by proposing a KD framework that combines Attention Transfer and Feature Alignment simultaneously, guided by gradient-based causal masks to ensure that only the teacher's most impactful knowledge is transferred.

Methodology

This study proposes an enhanced Knowledge Distillation (KD) framework that integrates **Attention Transfer (AT)** and **Feature Alignment (FA)**, guided by **causal masking**, to optimize student model learning.

Experimental Setup

- **Dataset:** CIFAR-10, consisting of 60,000 RGB images (32x32 pixels) categorized into 10 balanced classes.
    - **Training Set:** 50,000 images.
    - **Validation Set:** 10,000 images.
    - **Preprocessing:** Random horizontal flips, random cropping with padding, and normalization using dataset-specific mean and standard deviation.
- **Model Architectures:**
    - **Teacher Model:** ResNet-50 augmented with Convolutional Block Attention Module (CBAM).
    - **Student Model:** Lightweight ResNet-18, selected for its balance between compactness and representational capacity.

Knowledge Distillation Configurations

We evaluated five distinct configurations to assess the contribution of each component:

1. **Baseline:** Student trained solely with Cross-Entropy (CE) loss on ground-truth labels.

2. **Standard KD:** Addition of KL Divergence (KD) loss to align the student's soft logits with those of the teacher.

3. **KD + Feature Alignment:** Incorporating a masked Mean Squared Error (MSE) loss to align attention maps of teacher and student, applied only to causally relevant regions identified via gradient-based masks.

4. **KD + Attention Transfer:** Guiding the student's focus through masked attention transfer, using squared activation maps filtered by causal masks.

5. **Combined Method (KD + FA + AT):** A comprehensive approach combining CE loss, KD loss, masked attention alignment, and masked attention transfer into a unified training objective.

Training Strategy

- **Optimizer:** Stochastic Gradient Descent (SGD) with momentum of 0.9.

- **Learning Rate:** Initialized at 0.01.

- **Weight Decay:** 5e-4.

- **Batch Size:** 64.

- **Early Stopping:** Implemented to prevent overfitting, based on validation loss plateauing.

- **Epochs:** Maximum of 30 epochs per experiment.

The teacher model's parameters were kept frozen throughout training, and only the student model was updated via backpropagation. For configurations involving attention alignment and transfer, attention maps were extracted from intermediate convolutional layers of both teacher and student models. Gradient-based causal masks were computed to filter out irrelevant regions, ensuring that alignment and transfer losses were applied selectively.

Evaluation Metrics

- **Top-1 Accuracy** on the validation set.
- **Model Efficiency:** Number of parameters and inference time per sample.
- **Robustness:** Qualitative assessment of model performance on corrupted or perturbed inputs.

Results

The proposed Knowledge Distillation (KD) framework was evaluated through a systematic comparison of five configurations, aiming to quantify the contribution of Attention Transfer (AT), Feature Alignment (FA), and their combination. All experiments were conducted on the CIFAR-10 dataset, utilizing a ResNet-50 teacher model with CBAM and a lightweight ResNet-18 student model.

Key Observations:

1. **Standard KD** improved the baseline student model by **+3.08%**, demonstrating the effectiveness of soft target alignment in enhancing generalization.

2. Introducing **Feature Alignment (FA)** alone resulted in a +2.22% gain over the baseline, yet slightly underperformed compared to Standard KD, suggesting that feature alignment without soft logit supervision might not fully capture the teacher's reasoning.

3. **Attention Transfer (AT)** alone achieved +2.97% improvement over the baseline, approaching the performance of Standard KD, indicating that guiding the student's spatial focus can be nearly as effective as soft label alignment.

4. The **Combined Method (FA + AT)** yielded the best performance, reaching 83.77%, outperforming all other configurations. This result highlights the synergistic effect of integrating both attention mechanisms and feature alignment, leading to superior knowledge transfer.

Model Efficiency and Inference Time:

Despite the performance gains, the proposed methods maintained the compactness and efficiency of the student model:

- **Student Model Size:** Remained consistent across all configurations (~11M parameters).
- **Inference Time:** Negligible differences were observed between configurations, ensuring the enhancements did not introduce computational overhead during deployment.

Robustness to Input Corruption:

Qualitative assessments on corrupted CIFAR-10 samples (e.g., Gaussian noise, occlusions) revealed that configurations involving attention mechanisms (AT and FA) exhibited better focus on salient image regions, resulting in more stable predictions under perturbations. The combined method further amplified this robustness by ensuring spatial and feature-level alignment.

Performance Metrics

	Group	Best Validation Accuracy
0	Group 1: Baseline	0.8056
1	Group 2: KD	0.8364
2	Group 3: Feature Alignment	0.8278
3	Group 4: Attention Transfer	0.8353
4	Group 5: Combined (Align + Attention)	0.8377

Discussion

The results of this study clearly demonstrate the effectiveness of integrating Attention Transfer (AT) and Feature Alignment (FA) within a unified Knowledge Distillation (KD) framework. By analyzing the comparative performance across all evaluated configurations, several key insights emerge regarding the strengths and limitations of existing KD approaches and the impact of the proposed method.

Comparison to Baseline

The Baseline student model, trained solely using cross-entropy loss with ground-truth labels, achieved a validation accuracy of 80.56%, serving as a lower-bound reference. This result reflects the limited generalization capacity of lightweight models when trained without any form of teacher supervision.

Applying Standard KD, which incorporates KL divergence loss for soft target alignment, resulted in a substantial improvement to 83.64%, a +3.08% gain over the baseline. This highlights the effectiveness of soft label supervision in transferring class similarity information from the teacher to the student. However, this method does not account for the teacher's internal spatial reasoning, limiting the student's ability to replicate more complex decision-making patterns.

Effectiveness of Individual Enhancements (FA and AT)

When Feature Alignment (FA) was applied independently, the student achieved 82.78%, a +2.22% improvement over the baseline. This indicates that aligning intermediate attention maps helps the student network focus on relevant spatial features, but lacks the complementary benefits provided by soft target distributions.

Attention Transfer (AT) alone achieved 83.53%, nearly matching the performance of Standard KD. This suggests that guiding the student's focus towards the teacher's salient regions is a highly effective mechanism for knowledge transfer. However, the marginal difference between AT and KD indicates that neither spatial focus nor soft targets alone are sufficient to fully bridge the gap to the teacher's reasoning capabilities.

Synergistic Benefit of Combining AT and FA

The combined method (FA + AT) outperformed all individual configurations, achieving a validation accuracy of 83.77%, representing the highest observed performance. This +3.21% improvement over the baseline and +0.13% gain over Standard KD—though seemingly incremental—demonstrates a crucial point: the integration of spatial attention guidance with feature-level alignment yields complementary benefits, enabling the student to inherit both high-level semantic understanding and precise spatial focus from the teacher.

While the numerical improvement over Standard KD is modest, the qualitative robustness to corrupted inputs and the consistent generalization across various perturbations indicate a more resilient student model.

Comparison to Existing Works

Unlike prior studies that explored Attention Transfer and Feature Alignment in isolation, this research emphasizes the importance of a unified, causally guided distillation strategy. Previous AT methods often applied global attention transfers, which risked copying irrelevant focus regions. Similarly, FA methods lacked the spatial precision required to guide the student's attention effectively. Our method addresses these limitations by:

1. Applying gradient-based causal masking, ensuring only influential attention regions are transferred.

2. Combining soft target supervision with spatial and feature-level alignment, creating a more holistic knowledge transfer pipeline.

Thus, this work advances the state-of-the-art by demonstrating that selective, structured, and aligned knowledge transfer is essential for maximizing the effectiveness of KD, especially when deploying compact student models in constrained environments.

Conclusion

This study presents a comprehensive Knowledge Distillation (KD) framework that synergistically combines Attention Transfer (AT) and Feature Alignment (FA), guided by causal masking, to enhance the performance of compact student models. Through systematic evaluation on the CIFAR-10 dataset, we demonstrated that while traditional KD methods and individual enhancements (AT or FA) independently contribute to improved student accuracy, their combination yields superior results in terms of accuracy, generalization, and robustness, all while preserving model efficiency.

The proposed method achieved a validation accuracy of 83.77%, outperforming the baseline by +3.21% and marginally exceeding the performance of Standard KD. Beyond numerical gains, the selective transfer of causally relevant attention regions enabled the student model to better replicate the teacher's internal reasoning processes, resulting in a more resilient and focused learning paradigm.

Our findings highlight a critical insight: effective knowledge distillation requires not only soft target alignment but also the structured and selective transfer of spatial reasoning patterns embedded within the teacher's architecture. The integration of AT and FA, when applied with causal filtering, provides a holistic pathway for transferring both semantic understanding and spatial attention, narrowing the performance gap between compact students and their larger teacher counterparts.

Future Work

Building upon these promising results, future research directions include:
Extending the framework to larger and more complex datasets (e.g., ImageNet) to validate scalability.
Investigating the impact of different causal masking techniques (e.g., attention rollout, LayerCAM) to refine selective attention transfer.

Exploring student architectures with varying capacities to assess the generality of the proposed method across diverse model families.

Evaluating the framework's performance in real-world deployment scenarios, such as edge devices and mobile applications, where efficiency and robustness are critical.

References

1. **Hinton, G., Vinyals, O., & Dean, J. (2015).** Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

2. **Zagoruyko, S., & Komodakis, N. (2017).** Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *International Conference on Learning Representations (ICLR).*

3. **Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015).** FitNets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

4. **Yim, J., Joo, D., Bae, J., & 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 (CVPR)*, 1063-6919.

5. **Heo, B., Kim, J., Yun, S., & Han, B. (2019).** A comprehensive overhaul of feature distillation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1921-1930.

6. **Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017).** Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618-626.