Distillation for High-Quality Knowledge Extraction via Explainable Oracle Approach

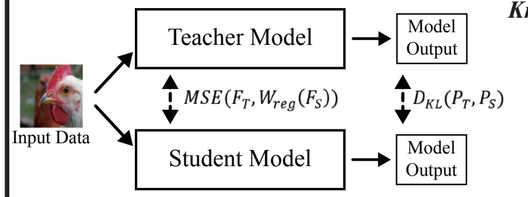
Myunghak Lee[1], Wooseong Cho[1], Sungsik Kim[1], Jinkyu Kim[2](jinkyukim@korea.ac.kr), Jaekoo Lee[1](jaekoo@kookmin.ac.kr)

[1]College of Computer Science Kookmin University Seoul, Korea

[2]Department of Computer Science and Engineering Korea University, Seoul, Korea

BMVC 2023

1. Introduction: Knowledge Distillation and Its Problem



Knowledge Distillation

- Technique used to transfer knolwedge from teacher to student
- Goal of student: Achieving similar performance as the teacher model with fewer parameters
- However, if the teacher model is larger than the student model, several problems arised.

Problem 1: Capacity Gap Problem

- Failing for the student model to appropriately receive the knowledge from the teacher model

Problem 2: Distillation Link Problem

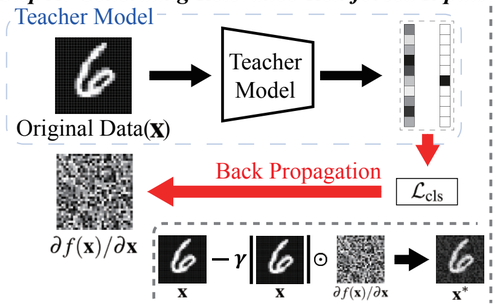
- Establishing links of the knowledge between the teacher and the student models

Solution: Explainable Oralce Approach

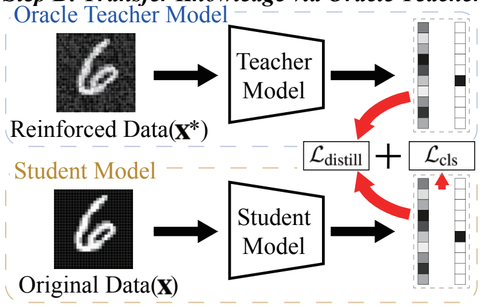
- We make relevance-reinforced data that contain ground-truth information.
- If we use relevance-reinforced data, the student model can be trained well even with a teacher model of the same size as the student model.

2. Oracle Teacher, Trained with Relevance-Reinforced Inputs

Step A: Generating Relevance-Reinforced Inputs



Step B: Transfer Knowledge via Oracle Teacher



Relevance-Reinforced Data

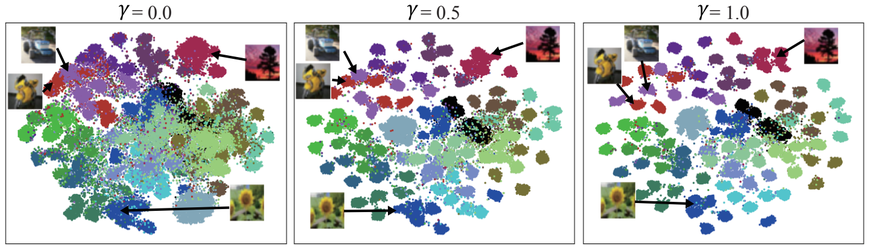
- Contrary to the adversarial example, we created relevance-reinforced data by subtracting the gradient from the original data.

Oracle teacher model

- We put the relevance-reinforced data into the teacher model once again, creating a model with extremely high accuracy.
- At this time, since the input data includes ground-truth information, the teacher model that inputs relevance-reinforced data was defined as the oracle teacher model.

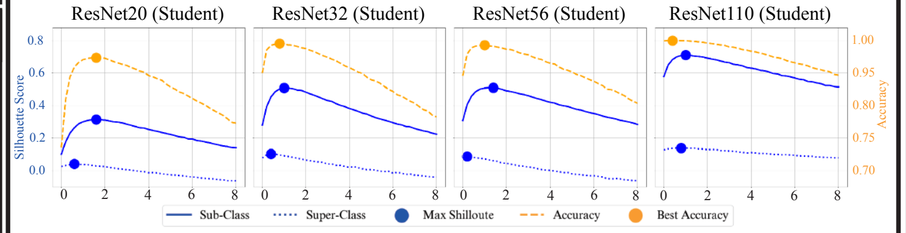
3-a. Is the Knowledge Obtained from Our Oracle Teacher Model Good Enough?

t-SNE (t-distributed Stochastic Neighbor Embedding)



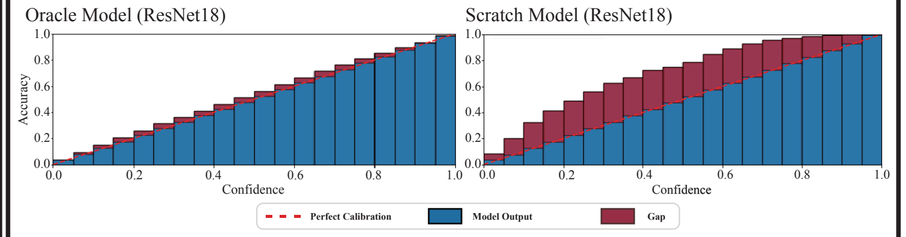
- The larger the γ value, the better the clusters between similar classes were expressed.
- It means a higher γ value makes similar classes cluster together better.

Silhouette Score



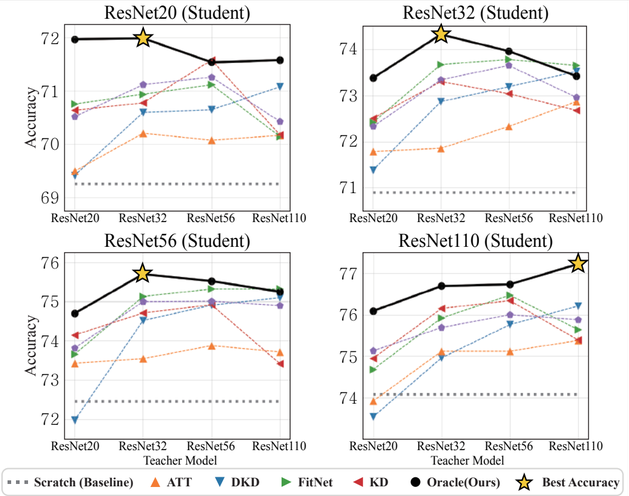
- Variation of silhouette score and top-1 accuracy with different γ
- The silhouette score is largest at an appropriate gamma value.

ECE (Expected Calibration Error)



- The Orcal model is much better calibrated than the scratch.
- It means the Oracle model's response knowledge is more reliable than the scratch model's.

3-b. Knowledge Distillation Performance Comparison (Accuracy)



- We test 16 different pairs of ResNet-based backbones on CIFAR-100 dataset.
- Interestingly, despite setting teacher and student models to be same, ours outperforms all others.
- This demonstrates the effectiveness of our method in extracting and transferring high-quality knowledge, even with a smaller teacher model.

5. Conclusion

- We propose using gradient-based explainable AI techniques to improve the model performance and compression effect of knowledge distillation techniques.
- Our approach effectively extract high quality knowledge using reinforced data.
- Reduce the commonly observed KD problems with small teacher model
- Acheive SOTA in knowledge distillation field
- Demonstrate the validity and usefulness of it with ECE, t-SNE, and silhouette score.