Distillation for High-Quality Knowledge Extraction via Explainable Oracle Approach

Myunghak Lee1, Wooseong Cho1, Sungsik Kim1, Jinkyu Kim1(jinkyukim@korea.ac.kr), Jaekoo Lee1(jaekoo@kookmin.ac.kr)
1College of Computer Science Kookmin University Seoul, Korea
2Department of Computer Science and Engineering Korea University, Seoul, Korea

1. Introduction: Knowledge Distillation and Its Problem

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

2. Oracle Teacher, Trained with Relevance-Reinforced Inputs

Step A: Generating Relevance-Reinforced Inputs
- 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.

2-a. Generating Relevance-Reinforced Inputs

- Original Data

- Reinforced Data

- Student Model

- Model Output

2-b. Generating Relevance-Reinforced Inputs

- Original Data

- Reinforced Data

- Student Model

- Model Output

3. Oracle Teacher Model

- Our Oracle Teacher Model Good Enough?

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

- r-SNE (t-distributed Stochastic Neighbor Embedding)

- Silhouette Score

- Entropy

- ECE (Expected Calibration Error)

3-b. Knowledge Distillation Performance Comparison (Accuracy)

- Accuracy

- Scratch (Baseline)

- Our Oracle Teacher Model

4. 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
- Achieve SOTA in knowledge distillation field
- Demonstrate the validity and usefulness of it with ECE, t-SNE, and silhouette score.