Knowledge Distillation in Regression: A Comparative Study of Hard and Soft Distillation Techniques

Hira Sardar

Abstract

This study explores the application of knowledge distillation techniques in regression tasks. We compare hard and soft distillation methods, investigating their effectiveness in transferring knowledge from a complex teacher model to a simpler student model. Our experiments involve a dataset related to optical network transmission, comparing the performance of distilled models against the original teacher model using metrics such as Mean Squared Error (MSE) and R-squared (R2) scores.

1 Introduction to Knowledge Distillation and Literature Review

Knowledge distillation, introduced by Hinton et al. (2015), is a technique for model compression where a smaller model (student) is trained to mimic a larger, more complex model (teacher). While initially proposed for classification tasks, it has been extended to regression problems.

In the context of regression, Saputra et al. (2019) demonstrated the effectiveness of knowledge distillation in improving the performance of lightweight models for various regression tasks. Chen et al. (2017) proposed a method for distilling regression models using privileged information, showing improvements in prediction accuracy.

Our study focuses on comparing two main approaches to distillation in regression:

- Hard Distillation: The student model is trained on a combination of true labels and teacher predictions.

- Soft Distillation: The student model is trained on softened versions of the teacher's predictions, controlled by a temperature parameter.

2 Data Preparation

The dataset used in this study relates to optical network transmission. The preparation process involved:

1. Feature selection using SelectKBest with f-regression, choosing the top 10 most relevant features.
2. Normalization of input features and target variables using StandardScaler.
3. Splitting the data into training and testing sets.

3 Model Training

The teacher model in our experiment is a stacking ensemble, combining multiple base models for improved performance. For the student model, we used a Gradient Boosting Regressor with the following parameters:

```
GradientBoostingRegressor(
    n_estimators=100,
    learning_rate=0.1,
    max_depth=3,
    random_state=42
)
```

4 Hard Distillation

In hard distillation, we trained the student model using a combination of true labels and teacher predictions:

$$y_{combined} = (1 - \alpha) * y_{true} + \alpha * y_{teacher}$$

where $\alpha = 0.2$ was used as the mixing coefficient.

5 Soft Distillation

For soft distillation, we experimented with different temperature settings (T = 0.5, 1.0, 2.0, 5.0). The teacher's predictions were softened using:

$$y_{soft} = \frac{y_{teacher}}{T}$$

The student model was then trained on a combination of true labels and softened predictions:

$$y_{combined} = (1 - \alpha) * y_{true} + \alpha * y_{soft}$$

where $\alpha = 0.5$ was used as the mixing coefficient.

6 Analysis of Results

6.1 Performance Metrics

We evaluated the models using Mean Squared Error (MSE) and R-squared (R2) scores. Lower MSE and higher R2 indicate better performance.

6.2 Graphs

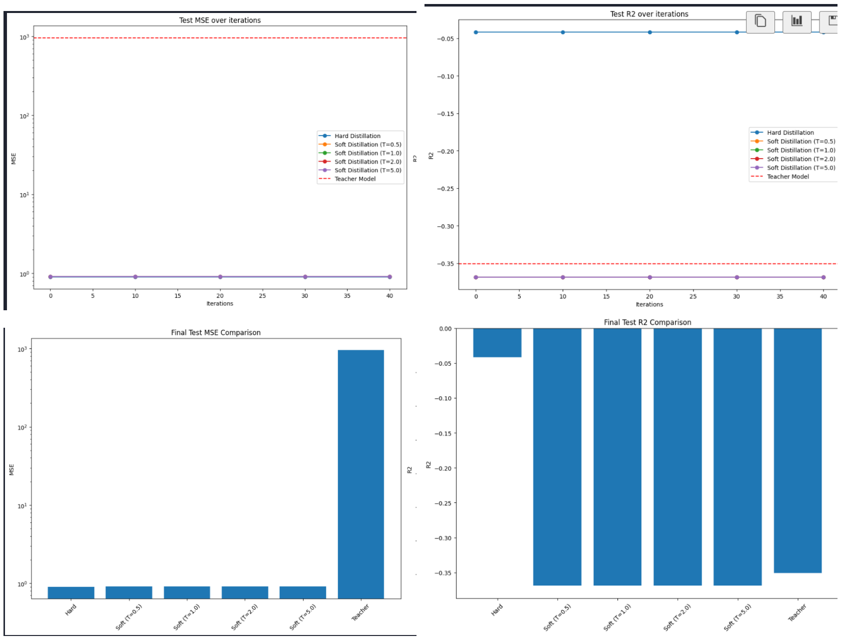

Figure 1: Four images inserted side by side.

6.3 Summarization of Findings

Our experiments revealed several key insights:

- Hard Distillation Performance: The hard distillation approach showed improvement over the teacher model, with a test MSE of 0.8997 and R2 of -0.0414, compared to the teacher's MSE of 957.1201 and R2 of -0.3508.

- Soft Distillation Performance: Soft distillation, regardless of the temperature setting, showed consistent performance with a test MSE of 0.9080 and R2 of -0.3686. This performance was better than the teacher model but slightly worse than hard distillation.

- Temperature Impact: In our experiments, varying the temperature in soft distillation did not significantly affect the results. This suggests that for this particular regression task, the choice of temperature may not be critical.

- Overall Improvement: Both distillation methods significantly outperformed the teacher model in terms of MSE, demonstrating the effectiveness of knowledge distillation in improving model efficiency for this regression task.

- R2 Scores: While the R2 scores improved compared to the teacher model, they remained negative, indicating that there's still room for improvement in capturing the variability of the target variable.

7 Conclusion

Our study demonstrates that knowledge distillation can be effectively applied to regression tasks in the domain of optical network transmission. Both hard and soft distillation methods showed significant improvements over the complex teacher model, with hard distillation slightly outperforming soft distillation in our experiments.

The consistently negative R2 scores across all models, including the teacher, suggest that the regression task at hand is challenging and that there might be underlying complexities in the data that are not fully captured by the current modeling approaches. Future work could explore more sophisticated distillation techniques, feature engineering, or alternative model architectures to further improve performance.

References

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

[2] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). Distilling knowledge from a deep pose regressor network. arXiv preprint arXiv:1512.00103.

[3] Li, Z., & Zhang, X. (2017). Learning efficient object detection models with knowledge distillation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.