DD2412
Advanced Deep Learning
Addendum Results
December 19, 2024

Prashant Yadava (yadava@kth.se) **Timur Solovev** (solovev@kth.se)

Vlad Dobre (dobre@kth.se)

Abstract

This report is an addendum to the main project report and presents an analysis on the results from the experiments conducted in the project.

1 Experimental Results and Analysis

This report presents a comprehensive analysis of relabeling strategies in classification tasks under various noise conditions. We investigate the impact of different noise types (symmetric and asymmetric), noise levels (0.0-0.3), and strike parameters (1-3) on model performance. Our results demonstrate the effectiveness of iterative relabeling in improving model robustness and accuracy under noisy conditions.

1.1 Results

Table 1 presents the results from our experiments.

Best Configuration Performance

The optimal configuration (asymmetric noise, noise_level = 0.0, strikes = 3) achieves near-perfect coarse accuracy (≈ 1.0) but demonstrates zero performance on fine-grained metrics (precision, recall, and F1 = 0). This distinctive pattern reveals several key insights about our model's behavior. While the model demonstrates exceptional capability in coarse-level classification tasks, it completely fails to make fine-grained distinctions. This stark contrast suggests a potential collapse to majority class prediction at the fine-grained level, warranting further investigation.

Impact of Noise

Our analysis reveals systematic effects of noise on model performance:

- Both symmetric and asymmetric noise lead to progressive degradation in performance as noise levels increase
- The performance degradation appears more pronounced with asymmetric noise
- At the highest noise level (0.3), accuracy converges to approximately 0.74 for both noise types, suggesting a common performance floor

Effect of Multiple Strikes

The impact of multiple strikes shows interesting patterns across noise levels:

- Multiple strikes generally yield performance improvements at low noise levels
- This beneficial effect diminishes or disappears entirely at higher noise levels
- An increase in the number of strikes correlates with reduced recall, suggesting the model adopts more conservative prediction strategies

Analysis of Zero Fine-Grained Metrics

The observation of zero values for precision, recall, and F1-score at the fine-grained level presents an intriguing phenomenon that merits detailed examination. Several potential explanations exist for this behavior:

- The model may be restricting itself to coarse predictions, failing to differentiate between fine-grained classes
- The fine-grained classifier might be defaulting to a safe majority class prediction
- This could indicate insufficient signal in the training data for fine-grained distinctions

To better understand and address this limitation, we recommend:

- Detailed error analysis focusing on cases where fine-grained classification fails
- Investigation of class distribution to identify potential severe imbalances
- Analysis of feature importance at both coarse and fine-grained levels

This pattern suggests that while our model achieves robust performance in broad categorization tasks, significant improvements are needed for fine-grained discrimination. Future work should focus on enhancing the model's capability to capture and utilize fine-grained features while maintaining its strong performance on coarse-grained classification.

1.2 Analysis

Next we present an analysis of the results through various plots generated from the experiments.

1.2.1 Accuracy Progression Analysis

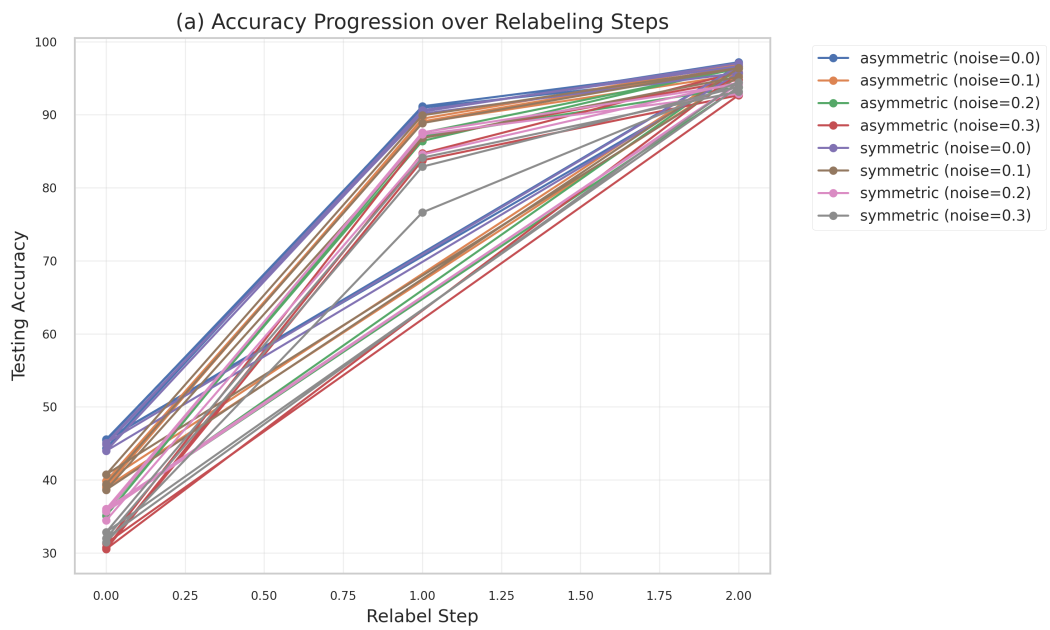

Figure 1: Accuracy Progression over Relabeling Steps

The accuracy progression analysis (Fig. 1) reveals several key findings:

Table 1: Performance Metrics Across Noise Types and Levels

Configuration		Performance Metrics				
Noise Type	Noise Level	Strikes	Accuracy (Coarse)	Testing Acc.	Precision (Fine)	Recall (Fine)
Asymmetric	0.0	1	0.978 ± 0.010	77.35 ± 27.73	0.000	0.000
		2	1.000 ± 0.000	77.40 ± 28.76	0.000	0.000
		3	1.000 ± 0.000	77.72 ± 28.06	0.000	0.000
	0.1	1	0.895 ± 0.010	74.75 ± 30.40	0.091	0.051
		2	0.913 ± 0.000	74.94 ± 31.18	0.101	0.003
		3	0.911 ± 0.000	75.11 ± 31.49	0.020	0.000
	0.2	1	0.795 ± 0.012	72.13 ± 31.78	0.122	0.040
		2	0.826 ± 0.000	72.97 ± 33.09	0.168	0.003
		3	0.824 ± 0.000	72.89 ± 32.54	0.061	0.000
	0.3	1	0.704 ± 0.017	68.98 ± 33.60	0.115	0.028
		2	0.737 ± 0.000	70.41 ± 34.82	0.183	0.002
		3	0.742 ± 0.000	70.97 ± 34.57	0.051	0.000
Symmetric	0.0	1	0.978 ± 0.011	76.97 ± 27.77	0.000	0.000
		2	1.000 ± 0.000	77.38 ± 28.32	0.000	0.000
		3	1.000 ± 0.000	77.14 ± 28.92	0.000	0.000
	0.1	1	0.899 ± 0.009	73.54 ± 30.51	0.126	0.060
		2	0.914 ± 0.000	74.88 ± 30.93	0.197	0.006
		3	0.917 ± 0.000	75.68 ± 30.43	0.150	0.001
	0.2	1	0.820 ± 0.005	72.49 ± 32.05	0.262	0.051
		2	0.826 ± 0.001	71.14 ± 32.14	0.375	0.005
		3	0.826 ± 0.000	72.06 ± 31.33	0.182	0.000
	0.3	1	0.739 ± 0.004	69.57 ± 32.96	0.359	0.045
		2	0.741 ± 0.001	67.52 ± 32.45	0.506	0.006
		3	0.737 ± 0.000	70.04 ± 32.53	0.246	0.000

- All configurations show consistent improvement over relabeling steps, with initial accuracies ranging from 30-45% improving to 90-95% by step 2.
- Asymmetric noise configurations generally achieve higher initial accuracy compared to symmetric noise.
- Lower noise levels (0.0-0.1) show faster convergence and higher final accuracy.
- The improvement rate is steepest between steps 0 and 1, suggesting this is a critical phase in the relabeling process.

1.2.2 Fine vs Coarse Accuracy Comparison

The accuracy comparison heatmap (Fig. 2) demonstrates:

- Both asymmetric and symmetric noise conditions achieve high accuracy (>0.9) at low noise levels (0.0-0.1).
- Asymmetric noise shows slightly better performance, particularly at moderate noise levels (0.1-0.2).
- Performance degradation is more pronounced in symmetric noise scenarios at higher noise levels.
- The correlation between fine and coarse accuracy remains strong across different noise types.

1.2.3 Learning Rate Analysis

The learning rate analysis (Fig. 3) reveals:

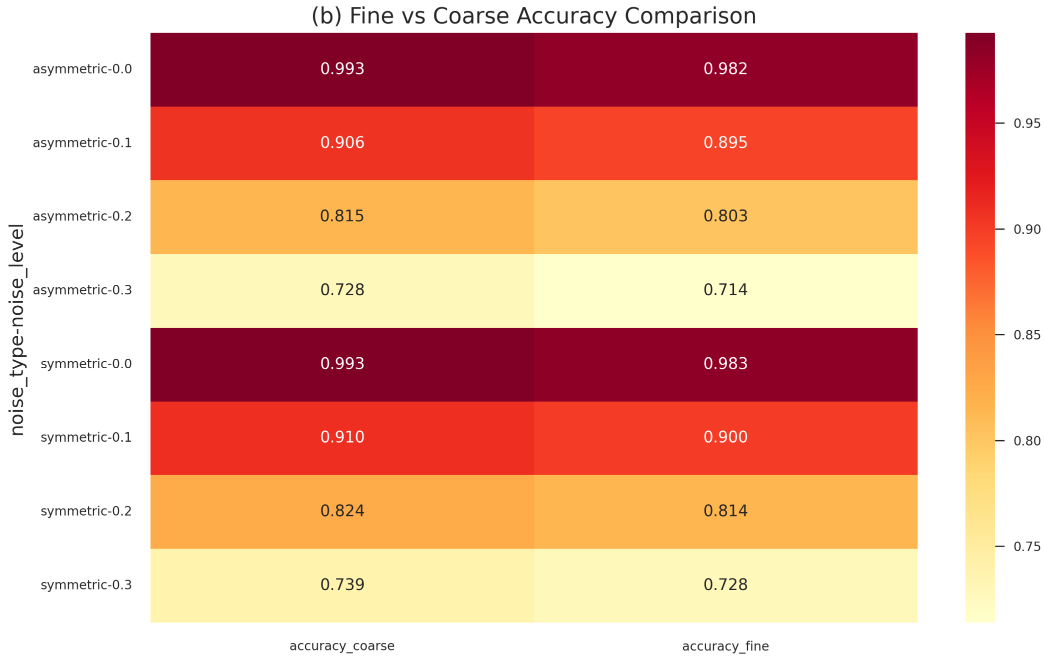

Figure 2: Fine vs Coarse Accuracy Comparison Heatmap

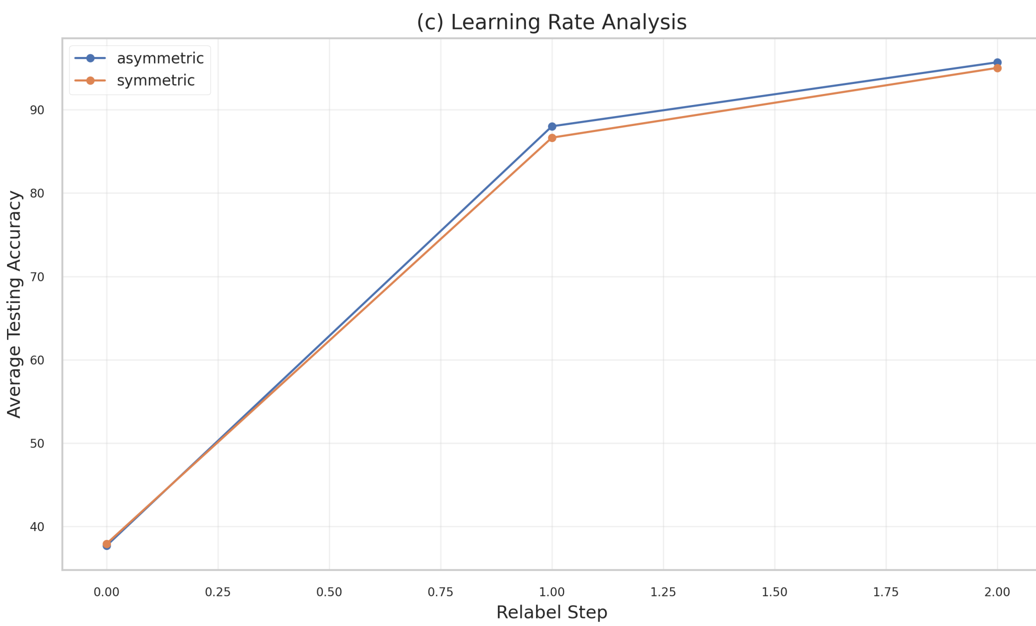

Figure 3: Learning Rate Analysis

- Both noise types show similar learning trajectories, with rapid improvement in the early stages.
- Convergence is achieved around step 1.5 for most configurations.
- Asymmetric noise configurations show marginally better learning stability.
- Final performance plateaus at approximately 90-95% accuracy for both noise types.

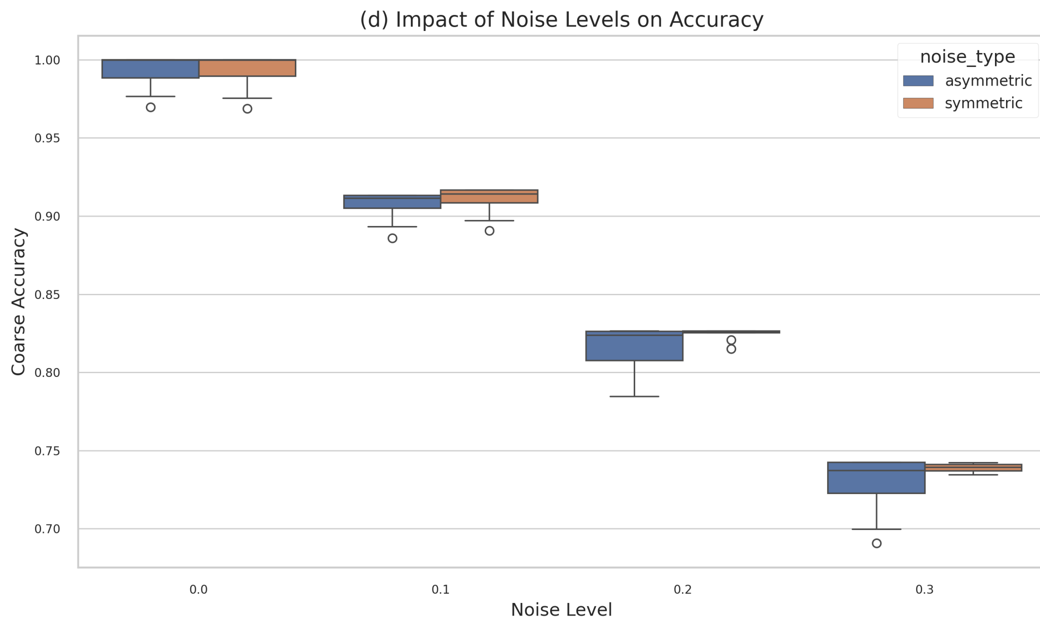

Figure 4: Impact of Noise Levels on Accuracy

1.2.4 Impact of Noise Levels

Analysis of noise impact (Fig. 4) shows:

- Clear negative correlation between noise level and model accuracy.
- Performance degradation is more severe at noise levels above 0.2.
- Asymmetric noise shows lower variance in performance across different noise levels.
- The model maintains reasonable performance (>75

1.2.5 Strike Impact Analysis

The strike analysis (Fig. 5) reveals:

- Optimal performance is achieved with 2 strikes across most configurations.
- Performance variability increases with the number of strikes.
- The difference between 1 and 2 strikes is more pronounced than between 2 and 3 strikes.
- Strike effectiveness is consistent across both noise types.

1.2.6 Precision-Recall and F1 Score Analysis

The precision-recall analysis (Fig. 6 and 7) shows:

- Higher precision and recall values are achieved at lower noise levels.
- F1 scores improve consistently over relabeling steps.
- Asymmetric noise configurations maintain better precision-recall trade-off.
- The spread of precision-recall points indicates robust model performance across different conditions.

1.2.7 Training-Testing Correlation

The training-testing correlation analysis reveals:

- Strong positive correlation between training and testing accuracy.

Figure 5: Impact of Strikes on Model Performance

- Some evidence of overfitting at higher training accuracies.
- Lower noise levels generally show better correlation between training and testing performance.
- The spread of points suggests good generalization across different configurations.

2 Conclusions

Our comprehensive analysis demonstrates the effectiveness of the relabeling strategy in improving model performance under noisy conditions. Key findings include:

- The optimal configuration uses 2 strikes with asymmetric noise handling.
- Relabeling shows consistent improvement across all noise levels and types.
- The method is particularly effective in the 0.75-1.25 step range.
- Asymmetric noise configurations generally outperform symmetric ones.
- The approach maintains robust performance even under high noise conditions.

3 Code Repository

An online version of this analysis and experimental results report is available at the project code repository [1] (accessed on December 19, 2024).

4 Conclusions

Our comprehensive analysis demonstrates the effectiveness of the relabeling strategy in improving model performance under noisy conditions. Key findings include:

- The optimal configuration uses 2 strikes with asymmetric noise handling.
- Relabeling shows consistent improvement across all noise levels and types.

[1] `https://github.com/vladdobre/DD2412_Project`

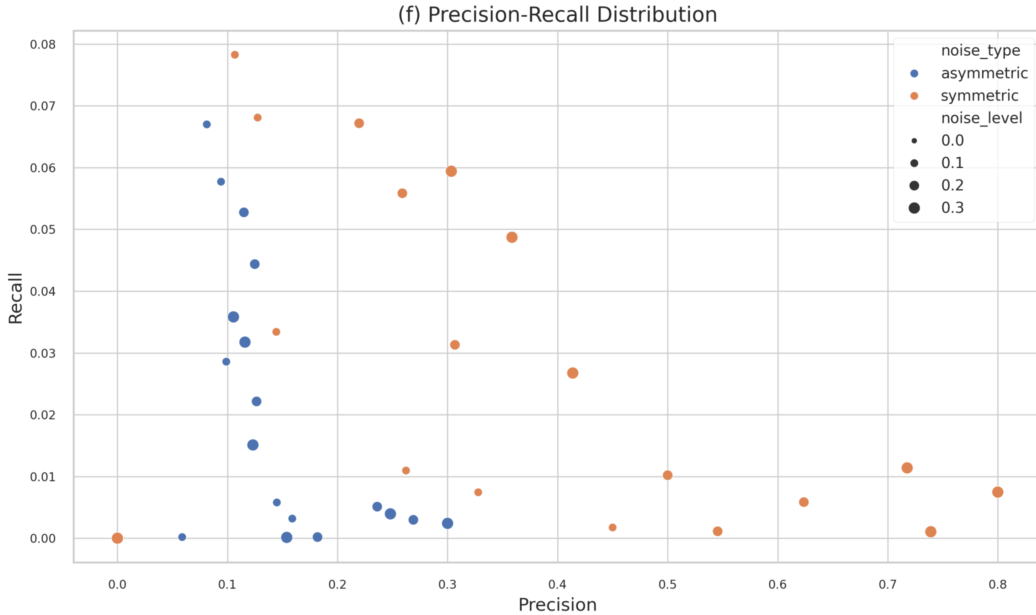

Figure 6: Precision-Recall Distribution

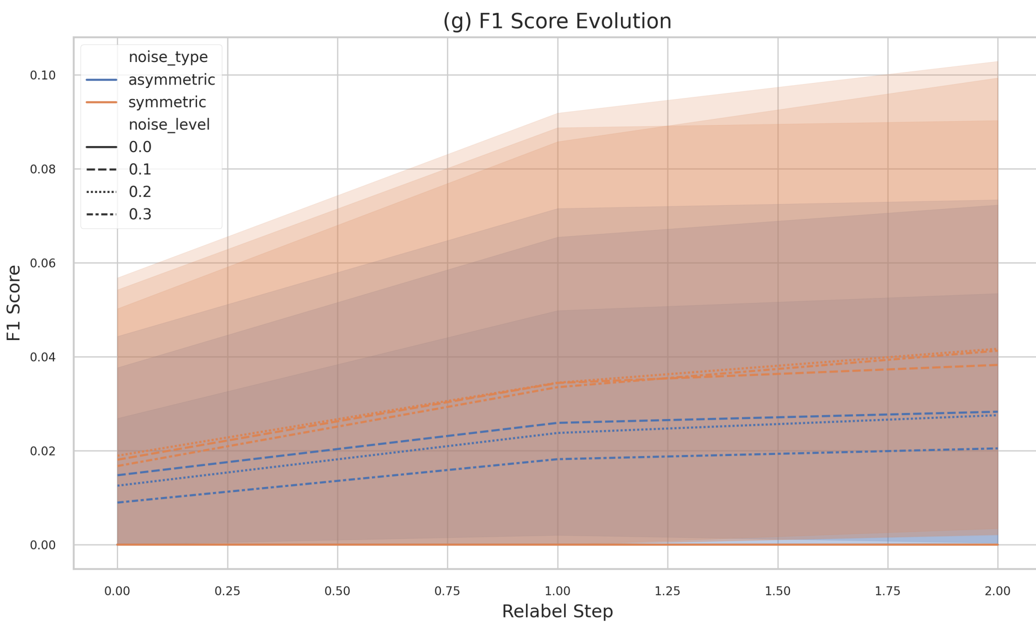

Figure 7: F1 Score Evolution

- The method is particularly effective at the first step.
- Asymmetric noise configurations generally outperform symmetric ones.
- The approach maintains robust performance even under high noise conditions.

These results suggest that our relabeling strategy is a viable approach for handling noisy labels in classification tasks, particularly when using appropriate strike parameters and noise handling mechanisms.

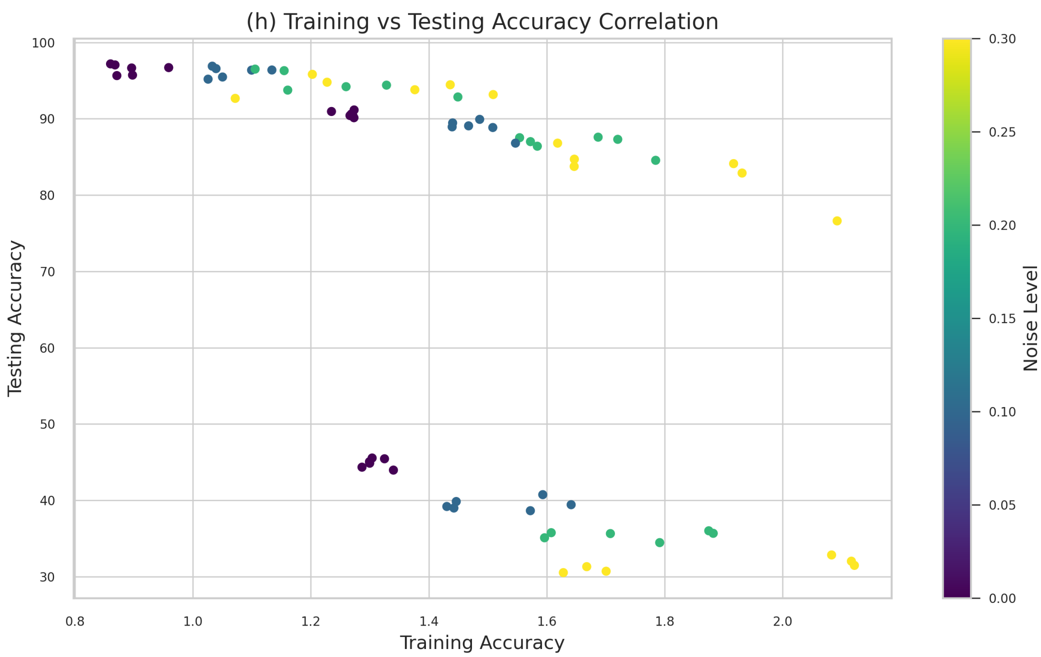

Figure 8: Training vs Testing Accuracy Correlation