Task 13: L2 and Dropout Regularization Experiment

Objective

The objective of this experiment was to evaluate the impact of L2 weight regularization combined with dropout on the generalization performance of the InstruNet CNN model for musical instrument recognition. The goal was to reduce overfitting observed in the baseline and batch-normalized models while maintaining stable convergence and improving test-time robustness, especially for minority instrument classes.

Experimental Setup

To ensure a fair and controlled comparison, the following constraints were strictly maintained:

- Dataset: IRMAS mono spectrogram dataset (unchanged)
- Architecture depth: Same CNN backbone as baseline
- Optimizer: SGD with momentum (0.9) and Nesterov acceleration
- Learning rate: 0.01 (with ReduceLROnPlateau)
- Loss function: Binary Cross-Entropy
- Batch size: 128
- Epochs: 50 (with EarlyStopping)
- Class imbalance handling: Inverse frequency class weighting

Regularization Strategy

- L2 regularization ($\lambda = 1e\text{-}4$) applied to all convolutional and dense layers
- Dropout (0.4) applied only after Global Average Pooling
- No dropout inside convolutional blocks to preserve spatial and spectral locality

Training Behavior Analysis

Training vs Validation Curves

- Training accuracy increased steadily and converged smoothly.
- Validation loss showed a more stable trend compared to previous experiments, with reduced oscillations.
- The train–validation gap was visibly reduced, indicating effective regularization.
- EarlyStopping triggered later compared to non-regularized and dropout-heavy models, suggesting improved generalization capacity.

This behavior confirms that applying dropout only at the classification head, combined with L2 regularization, avoids excessive suppression of convolutional feature learning.

Quantitative Results (IRMAS Test Set)

Global Metrics

Metric	Value
Micro F1-score	0.6765
Macro F1-score	0.6573
Micro ROC–AUC	0.9432
Macro ROC–AUC	0.9357

Interpretation:

- The improvement in macro F1-score indicates better performance across all classes, including under-represented instruments.
- High ROC–AUC values confirm strong ranking ability and class separability.
- Compared to Task 11, both Micro and Macro F1 scores improved, validating the effectiveness of L2 + dropout regularization.

Per-Class Performance Analysis

Recall & Precision Highlights

- High recall achieved for voi (0.880), pia (0.769), org (0.777), and gac (0.758).
- Significant recall improvement observed for cel (0.638) and flu (0.500) compared to earlier experiments.
- Precision remained consistently high for most instruments, exceeding 0.70 for the majority of classes.

This indicates a better balance between false positives and false negatives, especially for previously weak classes.

Confusion Matrix Observations

- Reduction in false negatives for several string and wind instruments.
- Minority classes such as cel, flu, and vio showed improved true positive detection.
- Confusion remains for acoustically similar instruments (e.g., sax vs tru), which is expected in real-world polyphonic audio.

Polyphonic Test Set Evaluation

Sliding-Window Inference (Real-World Scenario)

Metric	Value
Micro F1-score	0.6365
Macro F1-score	0.4168

Observations:

- Performance on polyphonic data is lower than IRMAS test data, which is expected since the model was trained on single-label segments.
- Strong performance for dominant instruments such as voi, pia, sax, and gac.
- Lower recall for rare instruments (e.g., cel, flu) due to limited presence in the polyphonic test set.

Despite this, the regularized model outperformed the non-regularized model in both Micro and Macro F1 on real-world data.

Effectiveness of Regularization

What Changed

- Introduced L2 regularization to penalize large weights
- Restricted dropout to the dense-equivalent classification head

Why It Was Changed

- Previous experiments showed over-regularization when dropout was applied inside convolution blocks
- Batch Normalization and Global Average Pooling already provided implicit regularization

Impact on Overfitting

- Reduced train–validation gap
- Improved macro-level generalization
- More stable validation loss trends

Final Decision

The L2 + head-only dropout configuration should be retained, as it provides the best trade-off between bias and variance among all tested configurations.

Conclusion

The Task 13 experiment demonstrates that carefully balanced regularization significantly improves model generalization for musical instrument recognition. Applying L2 regularization across convolutional layers while restricting dropout to the classification head results in superior macro-level performance, stable convergence, and improved robustness on both curated and real-world polyphonic audio. This configuration represents the most reliable and production-ready version of the InstruNet model tested so far.