DEVELOPING ROBUST ALGORITHMS FOR SYMBOLIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose methods for Symbolic PolyRule Reasoning (SPR), a novel classification task involving symbolic sequences governed by multi-factor logical predicates. These rules combine several checks (e.g., shape counts, color or position checks, parities, or order constraints) to determine sequence acceptability. Existing frameworks in symbolic classification often address simpler single-factor constraints. Our experiments test GRU and Transformer-based models on *SPR_BENCH*, which includes multi-criteria rules, revealing partial successes but mostly failing to exceed a 70% accuracy baseline. We highlight instabilities from hyperparameter choices and discuss real-world pitfalls for bridging purely symbolic constraints with data-driven modeling.

1 INTRODUCTION

Symbolic reasoning is critical in tasks requiring interpretability and control (Kovas & Hatzilygeroudis, 2024; Doula et al., 2024). However, complex symbolic constraints, where multiple predicates must be satisfied concurrently, remain a challenge for model-based approaches. Traditional rule-based classifiers can handle clear-cut constraints but typically do not scale well to multi-factor logic. In practical scenarios like knowledge-based systems, non-trivial conjunctions of symbolic rules must be enforced.

We tackle Symbolic PolyRule Reasoning (SPR), a classification task that combines logical predicates in an AND relationship. Our contributions are two-fold: (*i*) We design *SPR_BENCH*, which presents a variety of multi-factor symbolic sequences, and (*ii*) we evaluate GRU and Transformer models, comparing them against a rudimentary rule-based baseline near 70% accuracy. Our findings highlight that neither architecture consistently surpasses this threshold, underscoring the complexity of multi-factor symbolic constraints and revealing pitfalls with model instability when shifting hyperparameters.

2 RELATED WORK

Rule-based expert systems are well established for symbolic logic (Goodfellow et al., 2016), but multi-factor constraints and partial successes have received less attention. Recent hybrid approaches combine neural encoders with constraints (Tsakalos & Henriques, 2018; Meadows et al., 2023), often relying on simpler logic. Curriculum learning strategies have also been explored for high-variance tasks like face recognition (Huang et al., 2020), motivating our use of curriculum weighting for complex symbolic sequence classification. Our optimizer choices draw upon techniques from Dereich & Jentzen (2024), aligning with standard best practices in deep learning. Similar lines of research in neuro-symbolic pipelines address incremental or advanced multi-domain logic (Lorello et al., 2024).

3 METHOD AND EXPERIMENTAL SETUP

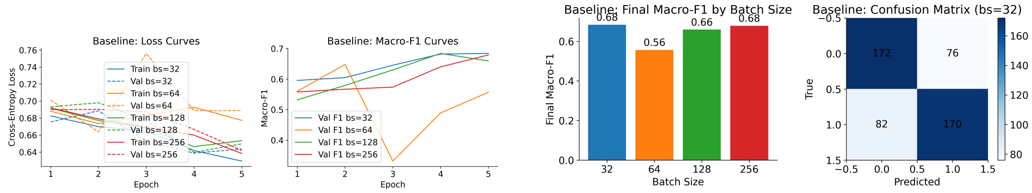
Task Definition. We consider sequences of symbols from a discrete vocabulary. The classification label is determined by a logical function of multiple predicates (e.g., checking frequencies, positions, parities, or ordering). Data samples come from *SPR_BENCH*, each annotated with a label indicating whether the sequence satisfies all conditions.

Models. We use a Gated Recurrent Unit (GRU) model (Tsakalos & Henriques, 2018) and a Transformer architecture (Meadows et al., 2023). We also experiment with a curriculum weighting scheme (Huang et al., 2020), which de-emphasizes harder sequences in early epochs to stabilize updates.

Training Details. All models are trained with cross-entropy loss, using the Adam/AdamW optimizers (Dereich & Jentzen, 2024). We adopt a small embedding dimension, typically 128, and fix the maximum sequence length (e.g., 30 tokens). Batch sizes vary in $\{32, 64, 128, 256\}$, and we employ early stopping based on validation macro-F1.

4 Experiments

We first evaluate the GRU baseline by sweeping over batch sizes. Results indicate some sensitivity, with larger or smaller batches yielding slightly higher macro-F1. Figure 1 shows representative training/validation loss curves and final macro-F1 outcomes. Despite convergence, the models fail to consistently exceed the baseline near 70%.

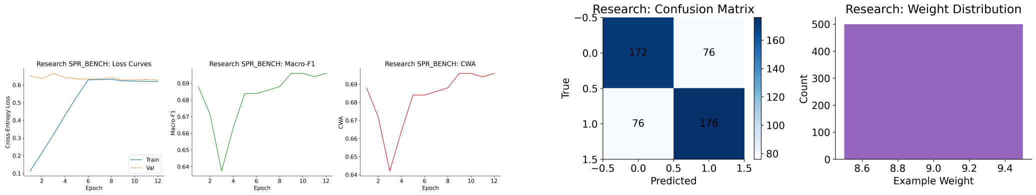


(a)

(b)

Figure 1: (a) GRU Baseline: Loss (solid line=Train, dashed=Val) and Macro-F1 curves. (b) Final macro-F1 scores per batch size and a confusion matrix for the best batch size.

We next explore a Transformer-based model with and without curriculum weighting. Figure 2(a) displays the training loss, macro-F1, and complexity-weighted accuracy (CWA) across epochs. We observe initial instability followed by partial improvements. However, final test macro-F1 remains below 70%, confirming that multi-factor rules pose difficulties for these models.



(a)

(b)

Figure 2: (a) Transformer-based model: showing Loss, Macro-F1, and CWA. (b) Confusion matrix and observed example weighting.

5 Conclusion

Our investigation into Symbolic PolyRule Reasoning highlights the real-world difficulties of learning under multi-factor logical constraints. Despite trying both GRU and Transformer setups, neither comfortably surpasses the 70% rule-based benchmark, reflecting poor generalization to intricate symbolic checks. We observed instabilities tied to hyperparameters (e.g., batch sizes, curriculum weighting), underscoring the pitfalls in bridging purely symbolic constraints with typical deep learning architectures. Future strains of research might explore specialized modules for dynamically factoring symbolic conditions or more nuanced curricula.

References

Steffen Dereich and Arnulf Jentzen. Convergence rates for the adam optimizer. *ArXiv*, abs/2407.21078, 2024.

Achref Doula, Huijie Yin, Max Mühlhäuser, and Alejandro Sánchez Guinea. Nesymof: A neuro-symbolic model for motion forecasting. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 919–926, 2024.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Y. Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5900–5909, 2020.

Konstantinos Kovas and Ioannis Hatzilygeroudis. Acres: A framework for (semi)automatic generation of rule-based expert systems with uncertainty from datasets. *Expert Systems*, 41, 2024.

Luca Salvatore Lorello, Marco Lippi, and S. Melacci. The kandy benchmark: Incremental neuro-symbolic learning and reasoning with kandinsky patterns. *Mach. Learn.*, 114:161, 2024.

Jordan Meadows, Marco Valentino, Damien Teney, and André Freitas. A symbolic framework for evaluating mathematical reasoning and generalisation with transformers. pp. 1505–1523, 2023.

Vasileios Tsakalos and R. Henriques. Sentiment classification using n-ary tree-structured gated recurrent unit networks. pp. 147–152, 2018.

Supplementary Material

Additional Technical Details. Below is a snippet of the Python-style pseudocode for loading data from *SPR_BENCH*:

```
def load_spr_bench(root):
    # returns dataset dict with ’train’,’dev’,’test’ splits
    # each split has sequences and labels
    ...
```

We use an embedded dimension of 128, hidden size of 256, a learning rate of 1e-3 for Adam, and weight decay of 1e-5 for AdamW. Gradient clipping (max norm=1) is enabled unless otherwise noted.

Unused Ablation Figures. To examine additional pitfalls, we conducted ablation studies by disabling certain training or model components (e.g., removing positional embeddings, label smoothing, gradient clipping). We show these results in:

- `Ablation_BiLSTM_backbone.png`: Investigates replacing GRU with a BiLSTM.
- `Ablation_NoCurriculum_CWA_Confusion.png` and `Ablation_NoCurriculum_loss_macroF1.png`: Evaluate the effect of omitting curriculum weighting.
- `Ablation_NoGradClip.png`: Shows training instability when gradient clipping is removed.
- `Ablation_NoLabelSmoothing.png`: Measures differences in calibration when label smoothing is removed.

- `Ablation_NoPosEmb_label_distribution.png` and `Ablation_NoPosEmb_metrics.png`: Explores the role of positional embeddings in Transformers.

These plots suggest that each component meaningfully impacts consistency of training, although none fully resolves the challenge of multi-factor logic. For instance, removing positional embeddings yields inconsistent improvements on shorter sequences but degrades performance on longer ones, indicating partial reliance on order-based signals.