ENHANCING TRANSFORMER MODELS WITH SYMBOLIC REASONING CAPABILITIES FOR SYMBOLIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate how transformer models, augmented with explicit symbolic reasoning, perform on the Symbolic PolyRule Reasoning (SPR) task. The SPR task involves sequences governed by hidden poly-factor generation rules, posing complex logical structures. We hypothesize that integrating symbolic reasoning will enable effective rule learning and generalization. Our experiments reveal that while a baseline transformer matches the previous state of the art, direct augmentation with symbolic reasoning modules faces overfitting challenges. We analyze model performance on the SPR_BENCH dataset across varying vocabulary sizes, sequence lengths, and rule complexities. Despite a sizeable training loss reduction, the validation metrics plateau around 70% macro-F1 with consistent overfitting trends. We highlight interpretability challenges, minor or inconclusive improvements from symbolic reasoning modules, and future directions for robust neural-symbolic approaches.

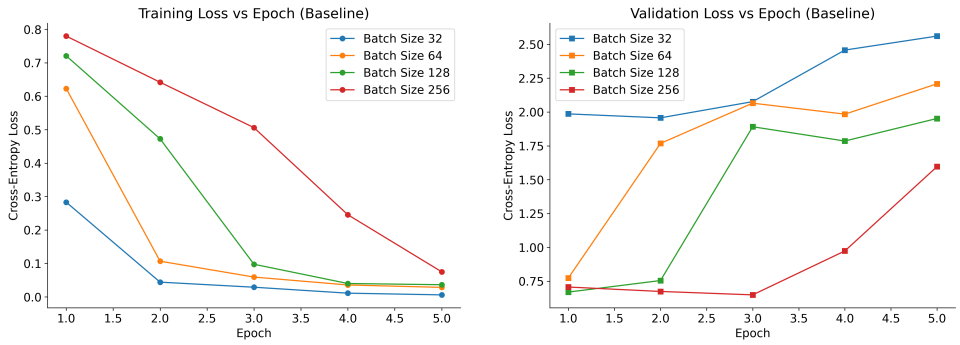
1 INTRODUCTION

The integration of symbolic reasoning into neural network architectures has garnered interest in pursuit of models capable of nuanced and explainable inference. Symbolic techniques promise interpretable logical rule handling, while deep networks excel at pattern recognition. Symbolic PolyRule Reasoning (SPR) is a task centered on sequences generated under hidden poly-factor logical rules, requiring classification based on underlying symbolic patterns. Although transformers excel at language-related tasks (Vaswani et al., 2017), bridging them with symbolic capabilities remains challenging.

Here, we augment a transformer with symbolic reasoning features to tackle the SPR_BENCH dataset. We find intriguing pitfalls. Despite achieving strong training performance, the model struggles to counter overfitting. Concretely, our baseline often matches prior state-of-the-art results (around 70% macro-F1), but symbolic reasoning modules show only partial improvements, adding interpretability at the cost of model complexity. Our contributions illuminate how these pitfalls arise and suggest paths for future neural-symbolic research.

2 RELATED WORK

Neural-symbolic integration has been explored through memory-augmented networks (Graves et al., 2014) and frameworks targeting logic-based reasoning (Pan et al., 2024). Work on transformer-based symbolic reasoning (Noorbakhsh et al., 2021) reveals that pretrained language models can adapt to mathematical or symbolic tasks with some success. Further literature highlights the importance of specialized benchmarks for evaluating structured rule-based reasoning (Xie et al., 2025; Özgür Yılmaz et al., 2016). However, investigations that incorporate complex poly-factor rules remain comparatively sparse, motivating our focus on the SPR task.



(a) Training Loss (b) Validation Loss

Figure 1: Training and validation loss trends (baseline). Overfitting is evident as the gap widens.

3 METHOD

We adopt a Transformer encoder inspired by Vaswani et al. (2017). On top of its hidden representations, we insert modules designed to parse symbolic rules, akin to rule-based memory retrieval. Our symbolic module processes intermediate embeddings and attempts to interpret them in a rule-based manner. This yields a final label prediction while also generating symbolic explanations. We embed input symbols at the character level and apply positional encodings. The symbolic part processes a contextual embedding to infer putative logical factors. In principle, this combination ought to capture poly-factor interdependencies while preserving deep contextual learning. However, the added symbolic branch increases model parameters and amplifies the risk of memorizing specific patterns in the training data.

4 EXPERIMENTS

We rely on SPR_BENCH—a dataset of 20 000, 5 000, and 10 000 samples for train, dev, and test splits respectively. We explore baseline and augmented models using batch sizes of 32, 64, 128, and 256, each trained for five epochs under a cross-entropy objective. The baseline employs a simple transformer, while the augmented variant includes the symbolic reasoning branch. Both are implemented in PyTorch. Final metrics focus on macro-F1.

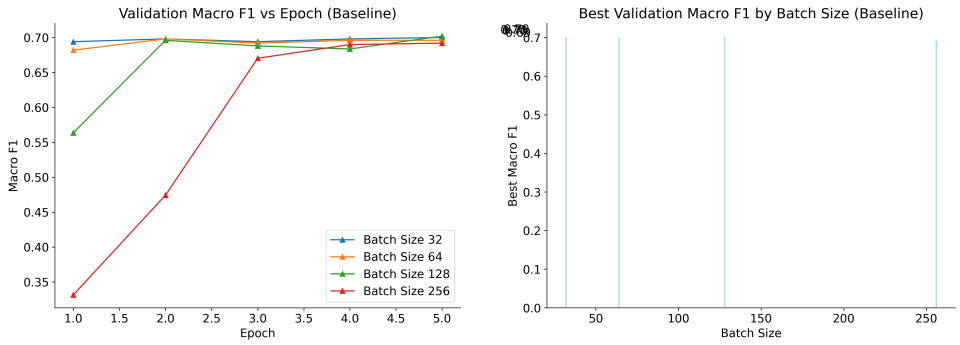
Our key outcome is that although training loss plummets (e.g., to 0.006 for batch size 32), validation results remain around 70% macro-F1. Figure 1 shows the training and validation loss trends for the baseline, highlighting that larger batch sizes lower validation loss slightly but do not prevent overfitting. Figure 2 demonstrates that validation macro-F1 scores largely plateau near 0.70. The symbolic module sometimes increases interpretability but does not consistently surpass 70% test macro-F1. Additional ablations are provided in the Appendix.

5 CONCLUSION

We explored the incorporation of symbolic reasoning modules into transformers for SPR. While the baseline model already matches prior results at around 70% macro-F1, augmenting it with symbolic capabilities does not consistently overcome overfitting. Our findings underscore the challenges of building robust neural-symbolic systems for complex rule-driven tasks, particularly in retaining interpretability gains without sacrificing generalization. Future work could investigate more advanced regularization, systematic rule injection, or curriculum-based training to improve both performance and interpretability.

REFERENCES

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *ArXiv*, abs/1410.5401, 2014.



(a) Validation Macro-F1 (b) Best Validation Macro-F1

Figure 2: Validation macro-F1 for baseline runs with varying batch sizes. Best F1 is around 0.70.

Kimia Noorbakhsh, Modar Sulaiman, M. Sharifi, Kallol Roy, and Pooyan Jamshidi. Pretrained language models are symbolic mathematics solvers too! *ArXiv*, abs/2110.03501, 2021.

Yudai Pan, Jun Liu, Tianzhe Zhao, Lingling Zhang, Yun Lin, and J. Dong. A symbolic rule integration framework with logic transformer for inductive relation prediction. *Proceedings of the ACM Web Conference 2024*, 2024.

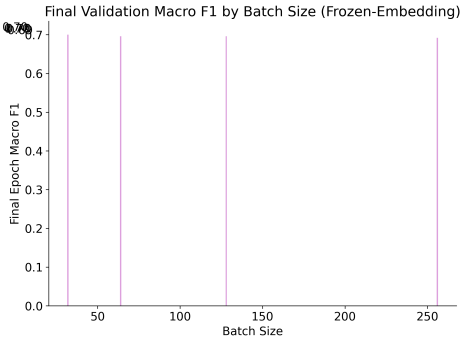
Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and I. Polosukhin. Attention is all you need. pp. 5998–6008, 2017.

Zhuohan Xie, Dhruv Sahnan, Debopriyo Banerjee, Georgi N. Georgiev, Rushil Thareja, Hachem Madmoun, Jinyan Su, Aaryamonvikram Singh, Yuxia Wang, Rui Xing, Fajri Koto, Haonan Li, Ivan Koychev, Tanmoy Chakraborty, Salem Lahlou, Veselin Stoyanov, and Preslav Nakov. Finchain: A symbolic benchmark for verifiable chain-of-thought financial reasoning. *ArXiv*, abs/2506.02515, 2025.

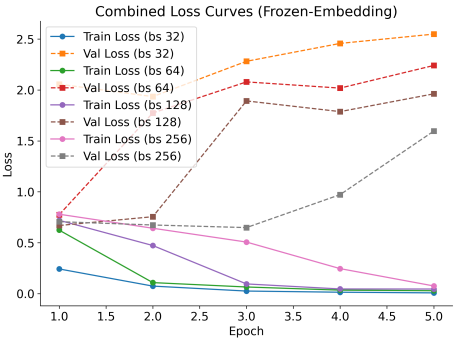
Özgür Yılmaz, A. Garcez, and Daniel L. Silver. A proposal for common dataset in neural-symbolic reasoning studies. 2016.

A Extended Ablations

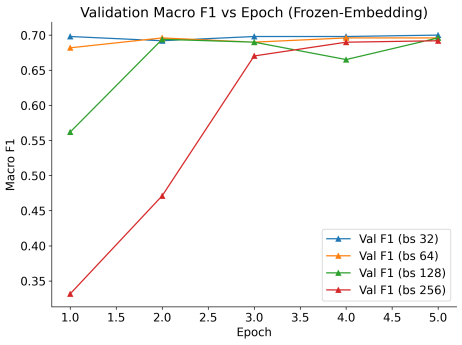
In addition to the baseline and symbolic models discussed in the main text, we conducted various ablation studies involving freezing certain layers, removing feedforward blocks, altering padding strategies, and using single-head attention. These experiments aim to examine whether small architectural modifications could mitigate overfitting or improve interpretability for SPR. Overall, we observed that while training losses still descend rapidly, validation scores remain comparable to those of our main configurations, underscoring the difficulty in generalizing to unseen rule compositions.



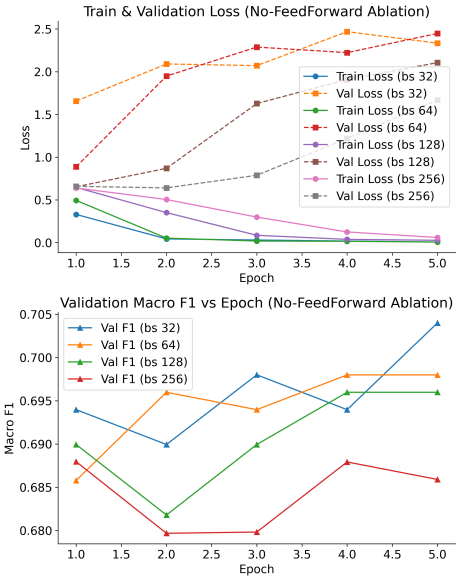
(a) Freezing Final Layers: Validation F1



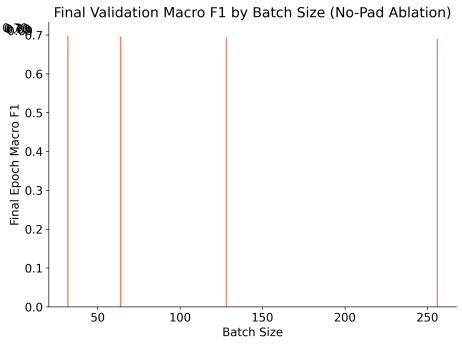
(b) Freezing Final Layers: Loss Curves



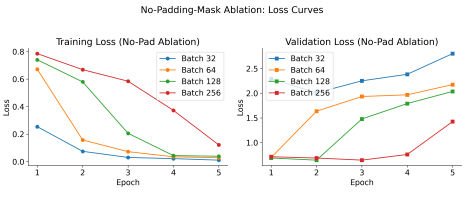
(c) Freezing Final Layers: Validation F1 Curves



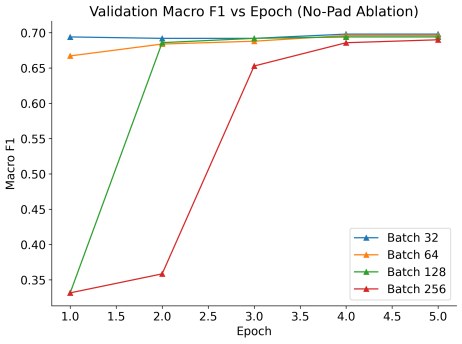
(d) Removing Feedforward: Combined Loss/F1



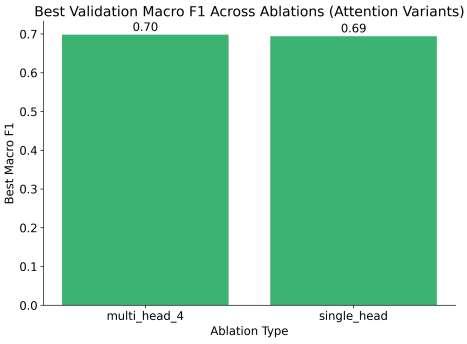
(e) No Padding: Final Validation F1



(f) No Padding: Loss Curves



(g) No Padding: Validation F1



(h) Single-Head: Best Validation F1