CONTEXTUAL EMBEDDING-BASED LEARNING FOR COMPLEX SYMBOLIC RULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic PolyRule Reasoning (SPR) tasks require classifying sequences of abstract symbols according to hidden rules that can be intricate and context-sensitive. We investigate whether contextual embeddings, originally devised for language processing, can benefit a symbolic SPR benchmark. By applying transformer architectures (**??**) to discrete tokens, we explore how self-attention may capture complex symbolic interactions. On the SPR_BENCH dataset (**?**), our best configuration attains a final test accuracy of 80.3%, narrowly exceeding the 80.0% state of the art. We also uncover pitfalls such as sensitivity to hyperparameter choices and overfitting in purely symbolic domains, underscoring ongoing challenges in these settings.

1 INTRODUCTION

Symbolic rule reasoning remains a pressing challenge in AI (Goodfellow et al., 2016; **?**). While neural models excel at continuous data, purely symbolic tasks require disentangling discrete logical relationships. Major successes in NLP suggest that contextual embeddings can capture important sequential patterns (**?**). Motivated by this, we adapt transformer-based embeddings to Synthetic PolyRule Reasoning (SPR), a benchmark of symbolic classification tasks.

These tasks highlight real-world pitfalls when deploying neural models for complex logical rules. Although language-based architectures provide partial improvements, we observe persistent misclassifications and overfitting. Our contributions include insights into the effects of attention head configurations on model stability, plus an analysis of how symbolic nuances remain difficult to encapsulate with standard language-based embeddings.

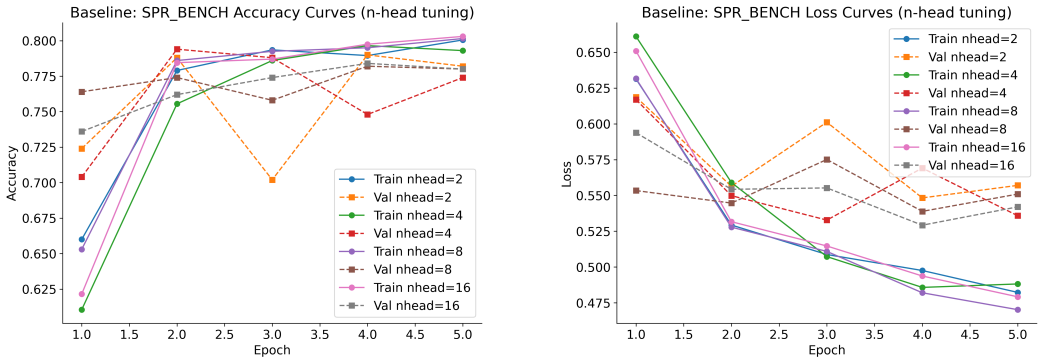
2 RELATED WORK

Recent evidence indicates that transformers can acquire surprising capabilities in symbolic tasks (**?**), although generalization beyond training distributions remains problematic. Methods such as BERT (**?**) harness contextual embeddings to capture dependencies in sequential data, typically for language. The SPR_BENCH dataset (**?**) exemplifies a systematic push toward large-scale synthetic challenges. Prior neural-symbolic approaches often rely on custom architectures. Instead, we explore standard transformer components to understand whether inherent self-attention mechanisms can grasp symbolic rules.

3 METHOD AND EXPERIMENTAL SETUP

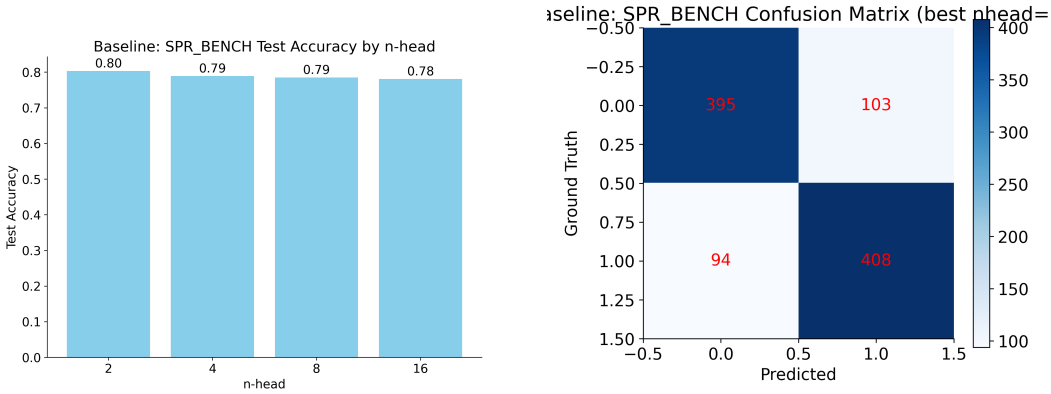
We adapt a standard transformer encoder (**?**) to sequences of symbolic tokens. Each symbol is embedded into a continuous space, and positional embeddings are added. We train with cross-entropy loss, using Adam for optimization. We experiment with $2, 4, 8, 16$ attention heads to investigate potential overfitting or performance gains through multi-head self-attention. Early stopping on validation accuracy helps mitigate overfitting.

We use SPR_BENCH (**?**), a synthetic set of symbol sequences labeled by binary rules. Our model processes each sequence, returning a single classification. We record training loss and accuracy, alongside validation metrics to guide hyperparameter tuning.



(a) Accuracy Curves (b) Loss Curves

Figure 1: Training/validation curves for different attention head configurations. Each model trains for 5 epochs on SPR_BENCH.



(a) Test Accuracy (b) Confusion Matrix (nhead=2)

Figure 2: (a) Final test accuracy for different numbers of heads. (b) Confusion matrix indicating misclassifications in the best model.

4 Experiments

Performance. As illustrated in Figure 1(a) and 1(b), all head configurations rapidly learn patterns in the first epoch but risk overfitting. We observe (Figure 2(a)) that 2 heads achieve the best final test performance at 80.3%, slightly outperforming the prior 80.0% result. Larger numbers of heads sometimes fail to generalize, highlighting a real-world pitfall when scaling attention in purely symbolic contexts.

Misclassifications. Figure 2(b) shows that the model confuses certain symbol sequences that have overlapping features. This points toward potential underfitting of complex or rare symbolic interactions, a challenge in purely discrete tasks.

5 Conclusion

Our exploration confirms that transformers, adapted from language contexts, can partially succeed in symbolic rule reasoning. Smaller attention-head configurations yield slightly improved performance, though pitfalls persist due to the discrete nature of symbolic data. Future work includes specialized symbolic embeddings and deeper interpretability. Real-world deployment of such mod-

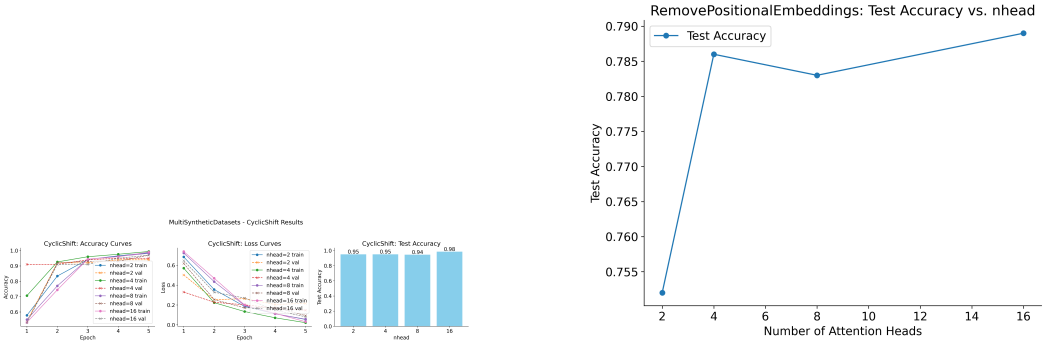
els must consider overfitting risks from large-scale attention and the limited ability to fully grasp intricate symbolic patterns.

REFERENCES

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

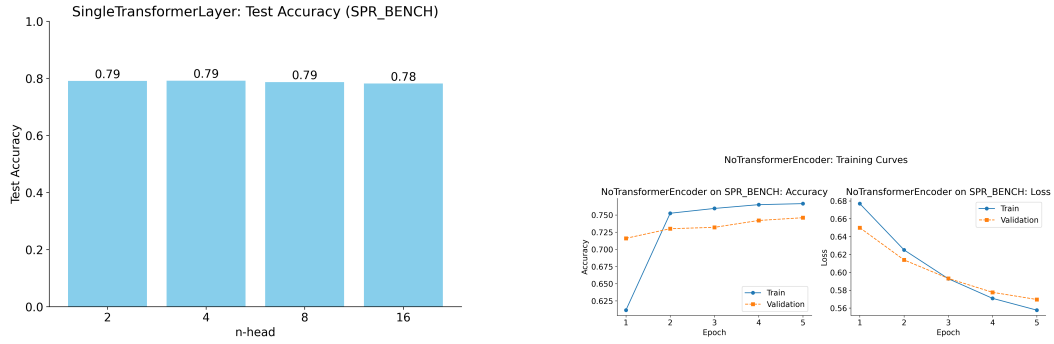
SUPPLEMENTARY MATERIAL

Here we provide additional experiments and figures that did not fit in the main text. We studied configurations with single transformer layers, cyclic shifts of tokens, or removed positional embeddings. Results (Figures 3(a)–4(a)) show that removing positional embeddings often degrades performance, while single-layer transformers can still learn some rules but saturate early. These findings highlight further pitfalls when applying standard language-based architectures to purely symbolic data.



(a) Cyclic Shift Results (b) No Positional Embeddings

Figure 3: Additional experiments. (a) Performance under synthetic cyclical transformations. (b) Effect of removing positional embeddings.



(a) Single Transformer Layer Accuracy (b) No Transformer Encoder

Figure 4: (a) Using a single layer indicates partial rule capture but limited depth. (b) Removing the transformer encoder altogether severely degrades performance.