Interpretable neural networks vs. expert-defined models for learner behavior detection

Juan D. Pinto, Luc Paquette, Nigel Bosch
University of Illinois Urbana-Champaign
jdpinto2@illinois.edu, lpaq@illinois.edu, pnb@illinois.edu

ABSTRACT: Despite their powerful predictive abilities, neural networks' lack of transparency makes it difficult to interpret how they make decisions. Recent work has proposed approaches for increasing the interpretability of neural networks, including convolutional neural networks (CNNs). However, analysis of the meaning of discovered convolutional filters is yet to be done. Consequently, the present study explores interpretable CNNs to measure a well-studied learner behavior—that of gaming the system. Phase 1 of our research, presented in this paper, aims to answer: (1) How does the accuracy of a CNN model compare to an expert-created model for gaming-the-system classification? (2) What is the impact on performance when making CNN filters more interpretable via regularization? Our findings suggest that there is good reason to pursue interpretable machine learning models for detecting student behaviors. Future work will consist of an in-depth analysis of the specific patterns the CNN has identified in student actions.

Keywords: Interpretability, pattern mining, convolutional neural networks, gaming the system, learner behavior detection.

1 INTRODUCTION

Neural networks are increasingly used in educational contexts for tasks such as knowledge tracing (Sarsa et al., 2022) and learner behavior detection (Botelho et al., 2019). However, despite their powerful predictive abilities, neural networks' lack of transparency makes it difficult to interpret how they make decisions. In an effort to address this issue, Jiang & Bosch (2021) proposed an approach for increasing the interpretability of a specific type of neural network—convolutional neural networks (CNN)—by implementing regularization in the loss function that constrains the weights of the convolutional filters to be binary rather than continuous. Since each learned filter serves as a different pattern that the network is seeking in the data, making filter weights binary makes it far easier to understand which combinations of student actions the model considers important. This approach follows a philosophy of interpretability, which is fundamentally different than explainability methods such as SHapley Additive exPlanations (SHAP; Lundberg & Lee, 2017). Rather than conducting *a posteriori* analyses of inputs and outputs, Jiang & Bosch (2021) alter the inner workings of the model itself to make it easier for humans to understand.

While Jiang & Bosch (2021) demonstrated the effectiveness of their technique for discovering useful predictive patterns, they did not analyze the meaning of the discovered filters. To fill this gap, the present study explores interpretable CNNs to measure a well-studied learner behavior—that of gaming the system, or "attempting to succeed in an interactive learning environment by exploiting properties of the system rather than by learning the material" (Baker & de Carvalho, 2008). Models of gaming the system have been created using different approaches, including machine learning,

knowledge engineering, and a hybrid of the two (Paquette & Baker, 2019). Paquette et al. (2014) elicited details about the specific learner action patterns that experts consider when labeling gaming behavior. This provides an interesting baseline against which to compare an interpretable CNN, both in terms of accuracy and the specific patterns identified.

Phase 1 of our research, presented in this paper, aims to answer the following research questions: (1) How does the accuracy of a CNN model compare to an expert-created model for gaming-the-system classification? (2) What is the impact on performance when making CNN filters more interpretable via regularization? This project is about more than gaming the system or marginal performance improvements. Rather, we explore whether interpretable neural networks can provide valuable predictive information, while allowing us to understand the patterns they recognize in learner behavior. Future work will more specifically compare the CNN's discovered patterns to patterns identified through expert knowledge elicitation.

2 METHODS

To more accurately compare our results with prior studies, we used data previously reported in Paquette et al. (2014). It consists of sequences of actions from 59 students using the Cognitive Tutor Algebra system during an entire school year. Cognitive Tutor tasks students with solving multi-step mathematical problems and can provide on-demand hints. A total of 10,397 clips (i.e., student action subsequences) were previously labeled by an expert (Baker & de Carvalho, 2008) and contained 708 examples of gaming the system behavior (6.8%). Each clip contained a minimum of 5 actions and needed to last at least 20 seconds. If a clip was shorter than 20 seconds, additional blocks of 5 actions were added until the duration was at least 20 seconds. Overall, 84% of all clips contained 5 actions, while 12% contained 10 and the rest were various lengths. We used the same held-out test set as Paquette et al. (2014), consisting of 25% of the total data. The same study identified *constituents*, which are designed to capture elements of the students' problem-solving behavior that experts pay attention to when looking for gaming the system behavior: for example, whether a student reuses the same answer in a different part of the problem interface or the student enters consecutive similar answers. Each constituent is a binary feature (i.e., the constituent is either observed or not for a given action) and is designed to have a meaningful interpretation with regards to how students solve problems. We used these same constituents as our inputs when training the CNN.

We trained a neural network with a single convolutional layer. To make use of clips that were not exclusively 5 actions long, we included an adaptive max pooling layer that condensed the output from the convolutional filters into the same length, making it possible to feed them all into the same fully connected layer for prediction. We further divided our training data into randomly selected training/validation sets of 80%/20%. We separately trained three models to select an early stopping point before overfitting to the training data began. We then used the average number of epochs among these to train a final model on the combined training + validation set.

To answer RQ2, we conducted the process described above twice: once with a model without regularization, and once with a model using the regularization term described in Jiang & Bosch (2021). This regularization is minimized when the parameters of convolutional filters are either 0 or 1, which enables interpretation of a filter as a sequential pattern that matches action presence and absence.

3 FINDINGS AND FUTURE WORK

Based on the validation stage, we trained the model with regularization (the more interpretable one) for 300 epochs on the entire training set. It obtained a Cohen's kappa of .322 (accuracy=93.1%, precision=.490, recall=.278) on the held-out testing set. These results are comparable to the performance of the expert model described in Paquette et al. (2014) (kappa=.331, accuracy=88.7%, precision=.307, and recall=.528 on the held-out test set). It outperformed the machine learning model created by Baker & de Carvalho (2008), which achieved a kappa of .24 on held-out data (as reported in Paquette et al., 2014).

We also trained the model without regularization for 90 epochs, which obtained a kappa of .333 (accuracy=93.7%, precision=.582, recall=.261) on the held-out testing set. The accuracy improvement over the model with regularization was therefore very minimal. Given our goal of generating interpretable models, and the potential of our model with regularization to be more interpretable, these results are promising.

Our findings so far suggest that there is good reason to pursue interpretable machine learning models for detecting student behaviors such as gaming the system. CNNs in particular provide a powerful approach for analyzing sequential data that can be designed with both accuracy and interpretability in mind. To evaluate the claim of interpretability, the next phase of our project will consist of an in-depth analysis of the specific patterns the convolutional layer has identified in the student actions. As part of this analysis, we plan to compare these patterns with those defined as relevant to gaming-the-system behavior by a domain expert.

REFERENCES

Baker, R. S., & de Carvalho, A. M. J. A. (2008). Labeling student behavior faster and more precisely with text replays. *Proceedings of the 1st International Conference on Educational Data Mining (EDM)*, 38–47.

Botelho, A. F., Varatharaj, A., Patikorn, T., Doherty, D., Adjei, S. A., & Beck, J. E. (2019). Developing early detectors of student attrition and wheel spinning using deep learning. *IEEE Transactions on Learning Technologies*, *12*(2), 158–170. https://doi.org/10.1109/TLT.2019.2912162

Jiang, L., & Bosch, N. (2021). Predictive sequential pattern mining via interpretable convolutional neural networks. *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, 761–766.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Paquette, L., & Baker, R. S. (2019). Comparing machine learning to knowledge engineering for student behavior modeling: A case study in gaming the system. *Interactive Learning Environments*, *27*(5–6), 585–597. https://doi.org/10/gg2h95

Paquette, L., de Carvalho, A. M. J. A., & Baker, R. S. (2014). Towards understanding expert coding of student disengagement in online learning. *Proceedings of the 36th Annual Cognitive Science Conference*, 1126–1131.

Sarsa, S., Leinonen, J., & Hellas, A. (2022). Empirical evaluation of deep learning models for knowledge tracing: Of hyperparameters and metrics on performance and replicability. *Journal of Educational Data Mining*, *14*(2), Article 2. https://doi.org/10.5281/zenodo.7086179