Does Large Dataset Matter? An Evaluation on the Interpreting Method for Knowledge Tracing

Yu LU[a, b], Deliang WANG [a*], Penghe CHEN [b] & Qinggang MENG [b]
[a]*School of Educational Techonology, Beijing Normal University, China*
[b]*Advanced Innovation Center for Future Education, Beijing Normal University, China*
*wangdeliang97@mail.bnu.edu.cn

Abstract: Deep learning has become a competitive method to build knowledge tracing (KT) models. Deep learning based knowledge tracing (DLKT) models adopt deep neural network but lack interpretability. The researchers have started working on interpreting the DLKT models by leveraging on methods in explainable artificial intelligence (xAI). However, the previous study was conducted on a relatively small dataset without comprehensive analysis. In this work, we perform the similar interpreting method on the largest public dataset and conduct the comprehensive experiments to fully evaluate its feasibility and effectiveness. The experiment results reveal that the interpreting method is feasible on the large-scale dataset, but its effectiveness declines with the larger size of learners and longer sequences of learner exercise.

Keywords: Knowledge tracing, deep learning, explainable artificial intelligence

1. Introduction

Knowledge tracing (KT) attempts to model learners' dynamic knowledge states on the skill level and predict their performance on the following exercises. With strong capacity to learn the inherent relationships from exercise data, deep learning has been adopted to build KT models. However, deep learning based knowledge tracing (DLKT) models have an untransparent decision process impeding their deployment. By leveraging on a technique called layer-wise relevance propagation (LRP) (Bach et al., 2015), we explored interpreting the DLKT model on a small dataset (Lu et al., 2020). It is still an open question whether the post-hoc interpreting method is feasible on large datasets.

In this work, we adopt the LRP method on one of the largest datasets, called EdNet (Choi et al., 2020). Specially, we clarify the technique of the LRP method in section 3, and perform the experiments to evaluate the feasibility and effectiveness of the method in section 4. The results reveal the LRP method is feasible on the large dataset, but its effectiveness declines with the larger size of learners and longer exercise sequence. By demonstrating the effectiveness issues of the current interpreting method, this work would be a solid step to build a fully transparent DLKT models.

2. Related Work

2.1 Knowledge Tracing

Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1995) can be regarded as the most prominent KT model, adopting the hidden Markov model (HMM) to estimate learner's mastery state on individual skill. Subsequent studies consider more factors to improve BKT, e.g., knowledge prior (Chen et al., 2017). Besides, logistic regression models have been deployed to build KT models. Recently, deep learning was introduced into KT domain. Deep knowledge tracing (DKT) (Piech et al., 2015) was the pioneer work. Then, the dynamic key-value memory network (DKVMN) (Zhang et al., 2017) and its variants (Chaudhry et al., 2018) were adopted to improve model performance. The attention network (Su et al., 2018) had been adopted to better represent question semantics. Besides, other information, e.g., prerequisite information (Chen et al., 2018), was utilized to design new DLKT models.

2.2 Explainable AI

The intransparent decision process of deep learning models is often hard to understand for human. To tackle this issue, researchers have proposed many explainable AI methods to interpret models' outputs and their inner working mechanism. The interpretability can be classified as ante-hoc and post-hoc: the ante-hoc interpretability focuses on training simple-structured machine learning models (Melis & Jaakkola, 2018), e.g., linear regression. The post-hoc interpretability focuses on interpreting the trained models. Among the post-hoc interpretability, the local methods such as backward propagation (Zeiler & Fergus, 2014) mainly aim to clarify the importance of the input features to model's predictions. In this work, we adopt a backpropagation method, namely LRP method, to interpret the DLKT models.

3. Building and Interpreting DLKT Models

3.1 DLKT Models on EdNet and ASSISTments

We adopt the long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) to build the DLKT models in this work. Figure 1 illustrates the basic architecture of the DLKT model on EdNet. The interaction between learner and question can be denoted as the question-answer pair $x_t = \{(q_t, a_t) | \mathrm{t} = 1, \ldots, \mathrm{N}\}$, where q_t is the representation of question information, $a_t \in \{0,1\}$ is the binary representation of correct or false answer, and N > 0 is the sequence length. The LSTM accordingly maps the input sequence vectors $\{\ldots x_{t-1}, x_t, x_{t+1} \ldots\}$ to the output vector $\{\ldots y_{t-1}, y_t, y_{t+1} \ldots\}$. Given most of the individual questions in EdNet covering multiple skills, an additional layer is adopted, which simply sets the average probabilities of all the skills covered by the next question as the final prediction z_t as below:

$$z_t = \frac{y_t \cdot q_{t+1}}{\mathrm{m}}, \tag{1}$$

where the dot product operation is performed between y_t and q_{t+1}, and *m* is the number of skills in next question. For ASSISTments dataset (Feng et al., 2009), this additional layer is not necessary.

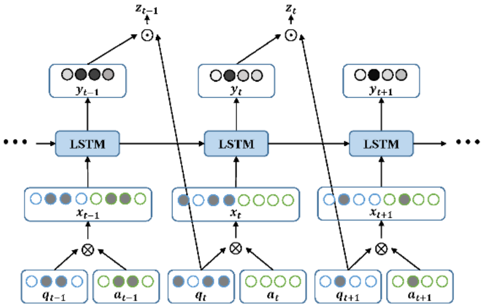


Figure 1. The Architecture of a RNN-based DLKT Model.

3.2 Interpreting Method

The LRP method interprets the DLKT models by analyzing the contribution of the individual input to the model's final prediction. Given a prediction made by the DLKT model, the LRP method would first sets the model's prediction value as the output layer neuron's relevance, and then backpropagate the relevance from the output layer to the input layer. During the backpropagating process, it needs to handle two different connections in the intermediate layers, namely *weighted linear connection* and *multiplicative connection*. The *weighted linear connection* can be written in a general form:

$$a_j^{(l+1)} = \sum_i w_{ij}\ a_i^{(l)} + b_j \tag{2}$$

where $a_j^{(l+1)}$ is the information the neuron *j* in layer *l+1* receives from the forward direction, w_{ij} and b_j are the weight and bias term. Given the relevance the neuron *i* in layer *l* receives from the neuron *j* in the layer *l+1* is $R_{i \leftarrow j}^{(l)}$, we have

$$R_{i\leftarrow j}^{(l)} = \frac{w_{ij}a_i^{(l)} + \frac{sign\left(a_j^{(l+1)}\right)\varepsilon + b_j}{N}\delta}{a_j^{(l+1)} + sign\left(a_j^{(l+1)}\right)\varepsilon} * R_j^{(l+1)} \tag{3}$$

where N is the number of neurons in layer l, and the item $sign\left(a_j^{(l+1)}\right) * \varepsilon$ prevents $R_{i\leftarrow j}^{(l)}$ becoming unbounded with $sign\left(a_j^{(l+1)}\right)$ as 1 or -1 and ε as a small positive value. We set δ as 0 to conserve relevance for the lower-level neurons. $R_j^{(l+1)}$ is the total relevance of neuron j in the layer $l+1$. For multiplicative connections, we define the neuron whose output ranging between 0 to 1 as "gate" neuron, and the remaining one as the "source" neuron. The multiplicative connection can be written as:

$$a_j^{(l+1)} = a_g^{(l)} \odot a_c^{(l)} \tag{4}$$

where $a_g^{(l)}$ and $a_c^{(l)}$ respectively are the message the "gate" neuron g and the "source" neuron s receive from layer l. During the forward propagating process, the "gate" neuron decides how much of the information should be retained in the upper-layer neurons and contributed to the model's decision (Arras et al., 2017). We set its relevance $R_{g\leftarrow j}^{(l)}$ as zero and give the full credit $R_j^{(l+1)}$ to the "source" gate.

4. Evaluation

4.1 Datasets and DLKT Models

We choose ASSISTment2009 and EdNet as the two datasets for the experiments. Table 1 summarizes the statistics of the preprocessed datasets. The built DLKT models adopt the LSTM network and RMSprop optimization for model training, with the iteration number and learning rate as 500 and 0.01. We set the hidden dimensionality, mini-batch size and the dropout rate to 200, 100 and 0.5 respectively. For both datasets, we utilize 64% data for training, 16% data for validating and the remaining ones for testing. After five-fold cross-validation, overall prediction accuracy (ACC) and AUC achieve 0.70 and 0.73 for the DLKT model on ASSISTment2009, and achieve 0.68 and 0.66 on EdNet.

Table 1. *Statistics of the Preprocessed Two Datasets ASSISTment2009 and EdNet*

Dataset	Learners	Skills	Questions	Interactions
ASSISTment 2009	3,091	110	16,850	320,582
EdNet	442,030	188	12,372	93,359,825

4.2 Feasibility Evaluation

4.2.1 Consistency Experiment

Given the calculated relevance for each question-answer pair, we investigate whether the sign of the relevance is consistent with the correctness of the answer. We define correctly-answered questions with positive relevance or falsely-answered questions with negative relevance as *consistent questions* and define the percentage of the *consistent question* in each sequence as *consistent rate*. A high consistent rate reflects that the LRP method could properly differentiate the correctly-answered and falsely-answered questions. Specifically, we utilize 7,143 and 1,187,377 sequences with a length of 15 in the two datasets as the test data. For each sequence, the first 14 question-answer pairs are the input, and the last pair is to validate the model's prediction. We obtain 4,972 correctly-predicted sequences in ASSISTment2009 and 799,857 correctly-predicted sequences in EdNet.

Figure 2 gives the histogram of the consistent rate on the two datasets. Nearly 80% sequences achieve a high consistent rate (i.e., 90% or above) in ASSISTment2009, while only around 50% sequences achieve a high consistent rate (i.e., 90% or above) in EdNet. Both distributions clearly show that the majority of sequences in both datasets receive 70% consistent rate or above, which demonstrate the sign of the calculated relevance values on both the small and large datasets.

4.2.2 Deletion Experiment

We further quantitatively investigate the relevance for both datasets by performing the deletion experiment. Specifically, for each correctly-predicted sequence, we delete the question-answer pair in a decreasing order of their relevance values for positive predictions or in an increasing order for negative predictions, and then record the predictions accuracy after each deletion. We also delete the question-answer pairs at random for comparison. Figure 3 illustrates the results on ASSISTment2009 and EdNet. For both datasets, all the accuracy lines drop down from 1.0 with an increasing number of the question-answer pair deletions, but the LRP lines drop much faster than the random lines.

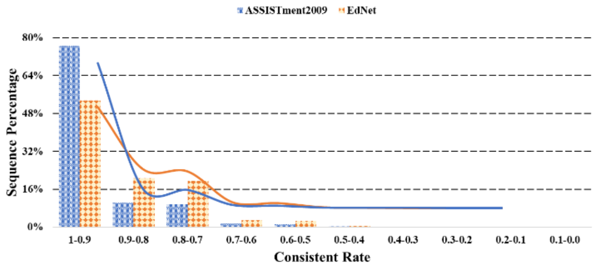
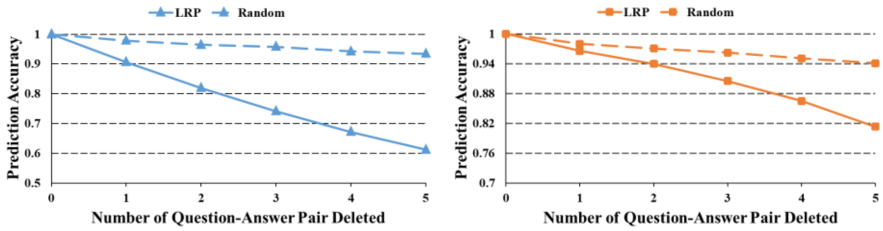
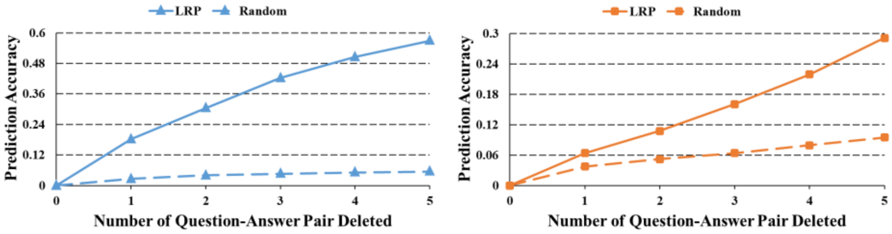


Figure 2. Histogram of the Consistent Rate on ASSISTment2009 and EdNet.



(a) ASSISTment2009 (b) Ednet

Figure 3. Accuracy Changes of Correctly-Predicted Sequences on the Two Datasets.



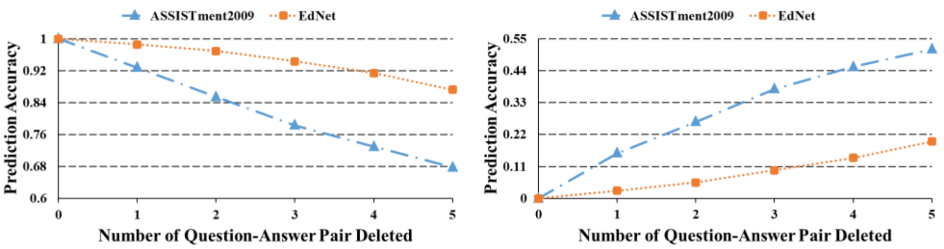
(a) ASSISTment2009 (b) Ednet

Figure 4. Accuracy Changes of Falsely-Predicted Sequences on the Two Datasets.

For each falsely-predicted sequence, we conduct similar experiments. Figure 4 shows that the accuracy lines rise from 0.0 with an increasing number of deletions for both datasets, but the LRP lines rise much faster than the random lines. All the deletion experiment results illustrate that the quantity of the relevance computed by the LRP method is possible to infer the question-level contribution to the final prediction result, and the LRP method is feasible on the large dataset EdNet.

4.3 Effectiveness Evaluation

To evaluate the effectiveness of the LRP method, we compare the accuracy changes between ASSISTment2009 and EdNet directly, deducting the random deletion effect. Figure 5 shows that the accuracy changes on ASSISTment2009 are much larger than on EdNet, which indicates the LRP method is less effective on the large dataset. This might be due to the larger number of learners. Another feature of the EdNet is the length of its sequences, which are larger than the ones in ASSISTment2009. We design the experiments on EdNet to evaluate whether the length of sequences (i.e., the number of question-answer pairs in one sequence) affect the effectiveness of the LRP method. Specifically, we divide the test data in EdNet into the sequences at a length of 15, 50, 100 and 200, and conduct the consistency and deletion experiments Table 2 summarizes sequence number at different lengths.



(a) Correctly-Predicted Sequences (b) Falsely-Predicted Sequences

Figure 5. Comparison of the Accuracy Changes between ASSISTment2009 and EdNet.

Table 2. *Number of Sequences at a Length of 15, 50, 100 and 200 in EdNet*

		Len15	Len50	Len100	Len200
Correctly Predicted	Positive Prediction	751,737	218,173	102,603	47,110
	Negative Prediction	48,120	11,698	5,384	2,467
Falsely Predicted	Positive Prediction	354,901	96,491	45,097	20,289
	Negative Prediction	32,619	8,540	3,908	1,745
Total		1,187,377	334,902	156,992	71,611

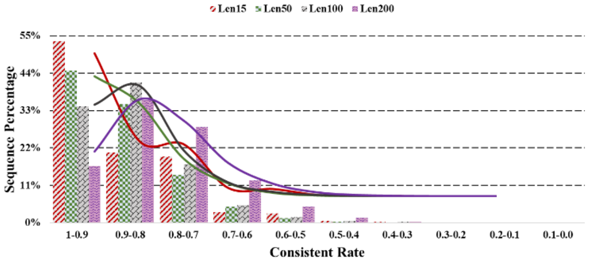
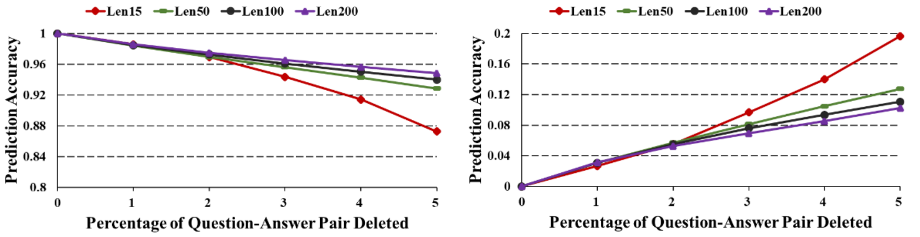


Figure 6. Histogram of the Consistent Rate at Different Lengths of the Sequences.



(a) Correctly-Predicted Sequences (b) Falsely-Predicted Sequences

Figure 7. Comparison of the Accuracy Changes at Different Lengths of the Sequences.

Figure 6 gives the histogram of the consistent rate at different lengths. Less than 5% sequences at different lengths have a consistent rate below 0.6, showing the feasibility of the LRP method. Different lengths exhibit distinct distributions: for the shorter ones, e.g., length of 15 and 50, the highest sequence percentage bars appear in the consistent rate range 1-0.9, and then they sharply drop down with the decreasing consistent rate. For the longer sequences, e.g., length of 100 and 200, the highest sequence percentage bars appear in the consistent rate range 0.9-0.8, and then drop down smoothly. In other words, the distributions tend to display their non-monotonic and long-tail patterns, which indicates that the long sequences might affect the sign of the relevance calculated by the interpreting method.

Figure 7 presents the deletion experiment results at different lengths. For the correctly-predicted sequences, Figure 7(a) shows all the accuracy lines drop down with an increasing number of deletions. However, the lines of shorter sequences (e.g., length of 15) drop much faster than the longer ones (e.g., length of 200). For the falsely-predicted sequences, in Figure 7(b) all the accuracy lines rise up with an increasing number of deletions. However, the lines of shorter sequences (e.g., length of 15) rise much faster than the longer ones (e.g., length of 200). Both experiment results indicate that the relevance for longer sequences are more difficult to reflect the question-level contributions to the prediction. It is more difficult for the LRP method to capture important question-answer pairs from longer sequences.

5. Conclusion

In this work, we first build the RNN-based DLKT models on both ASSISTment2009 and EdNet, and then perform the LRP methods on both the small-scale and large-scale models. Both the consistency and deletion experiments validate the feasibility of the interpreting method on the large dataset EdNet. However, the current interpreting method performs less effective on EdNet, which might be mainly due to its bigger size of learners and longer sequence of learner exercise. On a broader canvas, this work validates the basic interpreting method for explaining the DLKT model's predictions, but suggests the new studies to improve the current interpreting methods due to the large-scale educational datasets.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 62077006, 61807003), and the Fundamental Research Funds for the Central Universities.

References

Arras, L., Montavon, G., Müller, K. R., & Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. *EMNLP 2017,*page 159.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *Plos One*, 10(7):0130140.

Chaudhry, R., Singh, H., Dogga, P., & Saini, S. K. (2018). Modeling hint-taking behavior and knowledge state of students with multi-task learning. In *Proceedings of Educational Data Mining*.

Chen, P., Lu, Y., Zheng, V. W., & Pian, Y. (2018). Prerequisite-driven deep knowledge tracing. In *2018 IEEE International Conference on Data Mining (ICDM),* pages 39–48. IEEE.

Chen, Y., Liu, Q., Huang, Z., Wu, L., Chen, E., Wu, R., Su, Y.,& Hu, G. (2017). Tracking knowledge proficiency of students with educational priors. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management,* pages 989–998. ACM.

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., & Heo, J. (2020). Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education*, pages 69–73. Springer.

Corbett, A. T. & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, 4(4):253–278.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modelingand User-Adapted Interaction*, 19(3):243–266.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

Lu, Y., Wang, D., Meng, Q., & Chen, P. (2020). Towards interpretable deep learning models for knowledge tracing. In *International Conference on Artificial Intelligence in Education*, pages 185–190. Springer.

Melis, D. A. & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. In *Proceedings of Advances inNeural Information Processing Systems(NIPS),* pages 7786–7795.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Advances in NeuralInformation Processing Systems*, pages 505–513.

Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., Ding, C., Wei, S., & Hu, G. (2018). Exercise-enhanced sequential modeling for student performance prediction. In *32nd AAAI Conference on Artificial Intelligence*.

Zeiler, M. D. & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of European Conference on Computer Vision*, pages 818–833.

Zhang, J., Shi, X., King, I., & Yeung, D. Y. (2017). Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web*, pages 765–774.