Towards Interpretable Deep Learning Models for Knowledge Tracing

Yu Lu[1,2], Deliang Wang[2], Qinggang Meng[1], and Penghe Chen[1(✉)]

[1] Advanced Innovation Center for Future Education, Beijing Normal University, Beijing, China
{luyu,chenpenghe}@bnu.edu.cn
[2] School of Educational Technology, Beijing Normal University, Beijing, China

Abstract. Driven by the fast advancements of deep learning techniques, deep neural network has been recently adopted to design knowledge tracing (KT) models for achieving better prediction performance. However, the lack of interpretability of these models has painfully impeded their practical applications, as their outputs and working mechanisms suffer from the intransparent decision process and complex inner structures. We thus propose to adopt the post-hoc method to tackle the interpretability issue for deep learning based knowledge tracing (DLKT) models. Specifically, we focus on applying the layer-wise relevance propagation (LRP) method to interpret RNN-based DLKT model by backpropagating the relevance from the model's output layer to its input layer. The experiment results show the feasibility using the LRP method for interpreting the DLKT model's predictions, and partially validate the computed relevance scores. We believe it can be a solid step towards fully interpreting the DLKT models and promote their practical applications.

Keywords: Knowledge tracing · Interpretability · Deep learning

1 Introduction

The rapid development of ITS and MOOC platforms greatly facilitates building KT models by collecting a large size of learner's learning and exercise data in a rapid and inexpensive way. Yet, the collected massive and consecutive exercise questions are usually associated with multiple concepts, and the traditional KT models cannot well handle the questions without explicit labels and capture the relationships among a large size of concepts (e.g., 100 or more concepts). Accordingly, deep learning models are recently introduced into the KT domain because of their powerful representation capability [12]. Given the sequential and temporal characteristics of learner's exercise data, the recurrent neural network (RNN) [14] is frequently adopted for building the deep learning based knowledge tracing (DLKT) models. Since it is difficult to directly measure the actual knowledge state of a learner, the existing DLKT models often adopt an alternative solution that minimizes the difference between the predicted and the real

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 185–190, 2020.
https://doi.org/10.1007/978-3-030-52240-7_34

responses on exercise questions. Hence, the major output of DLKT models are the predicted performance on next questions. As a popular implementation variants of RNN, the long short-term memory (LSTM) unit [11] and GRU [7] are widely used in the DLKT models, and have achieved comparable or even better prediction performance in comparison to the traditional KT models [6,12].

Similar as the deep learning models operating as a "black-box" in many other domains [10], the existing DLKT models also suffer from the interpretability issue, which has painfully impeded the practical applications of DLKT models in the education domain. The main reason is that it is principally hard to map a deep learning model's abstract decision (e.g. predicting correct on next question) into the target domain that end-users could easily make sense of (e.g., enabling the ITS designers or users to understand why predicting correct on next question). In this work, we attempt to tackle the above issue by introducing the proper interpreting method for the DLKT models. In particular, we adopt a post-hoc interpreting method as the tool to understand and explain the RNN-based DLKT models, and the experiment results validate its feasibility.

2 Related Work

As indicated earlier, deep learning models are recently introduced into the KT domain, as they have enough capacity to automatically learn the inherent relationships and do not require explicit labels on the concept level. Deep knowledge tracing (DKT) [12] that utilizes LSTM can be regarded as the pioneer work, while some limitations have been reported [15]. Subsequently, other DLKT models [5,6,17,18] are proposed to improve KT performance.

The interpretability can be categorized into *ante-hoc* and *post-hoc* interpretabilities. Among different methods for *post-hoc* interpretability, the LRP method [3] can be regarded as a typical one, where the share of model output received by each neuron is properly redistributed by its predecessors to achieve the relevance conservation, and the injection of negative relevance is controlled by its hyperparameters. LRP method is applicable and empirically scales to general deep learning models. It has been adopted for image classification [1], machine translation [8] and text analysis [2]. In the education domain, researchers have started interpreting KT models [16], but most studies target on the traditional simple-structured Bayesian network-based ones [4,13]. In this work, we mainly focus on explaining the DLKT models by using the LRP interpretability method.

3 Interpreting RNN-Based KT Model

3.1 RNN-Based DLKT Model

A number of DLKT models, such as DKT [12], adopt LSTM or similar architectures (e.g., GRU) to accomplish the KT task. As a typical RNN architecture, the model maps an input sequence vectors $\{\mathbf{x_0}, ..., \mathbf{x_{t-1}}, \mathbf{x_t}, ...\}$ to an output sequence vectors $\{\mathbf{y_0}, ..., \mathbf{y_{t-1}}, \mathbf{y_t}, ...\}$, where $\mathbf{x_t}$ represents the interaction

between learners and exercises, and $\mathbf{y_t}$ refers to the predicted probability vectors on mastering the concepts. The standard LSTM unit is usually implemented in the DLKT models as follows:

$$\mathrm{f_t} = \sigma\left(\mathbf{W_{fh}}\mathrm{h_{t-1}} + \mathbf{W_{fx}}\mathbf{x_t} + \mathrm{b_f}\right) \tag{1}$$

$$\mathrm{i_t} = \sigma\left(\mathbf{W_{ih}}\mathrm{h_{t-1}} + \mathbf{W_{ix}}\mathbf{x_t} + \mathrm{b_i}\right) \tag{2}$$

$$\widetilde{\mathrm{C_t}} = \tanh\left(\mathbf{W_{ch}}\mathrm{h_{t-1}} + \mathbf{W_{cx}}\mathbf{x_t} + \mathrm{b_c}\right) \tag{3}$$

$$\mathrm{C_t} = \mathrm{f_t} \odot \mathrm{C_{t-1}} + \mathrm{i_t} \odot \widetilde{\mathrm{C_t}} \tag{4}$$

$$\mathrm{o_t} = \sigma\left(\mathbf{W_{oh}}\mathrm{h_{t-1}} + \mathbf{W_{ox}}\mathbf{x_t} + \mathrm{b_o}\right) \tag{5}$$

$$\mathrm{h_t} = \mathrm{o_t} \odot \tanh\left(\mathrm{C_t}\right). \tag{6}$$

After getting the LSTM output h_t, the DLKT models usually further adopt an additional layer to output the final predicted results y_t as below:

$$\mathbf{y_t} = \sigma\left(\mathbf{W_{yh}}\mathrm{h_t} + \mathrm{b_y}\right) \tag{7}$$

From the above implementations, we see that the RNN-based DLKT models usually consist of two types of connections: ***weighted linear connection***, i.e., Eq. (1), (2), (3), (5), (7), and ***multiplicative connection***, i.e., Eq. (4) and (6). The two types would be interpreted by LRP in different ways.

3.2 Interpreting DLKT Models Using LRP Method

Considering the RNN-based DLKT model given in Eq. (1) to (7) and the LRP method, interpreting can be accomplished by computing the relevance as below:

$$R_{\mathrm{h_t}} = \frac{\mathbf{W_{yh}}\mathrm{h_t}}{\mathbf{W_{yh}}\mathrm{h_t} + \mathrm{b_y} + \varepsilon * \mathrm{sign}(\mathbf{W_{yh}}\mathrm{h_t} + \mathrm{b_y})} * R^d_{y_t} \tag{8}$$

$$R_{C_t} = R_{h_t} \tag{9}$$

$$R_{f_t C_{t-1}} = \frac{f_t C_{t-1}}{C_t + \varepsilon * sign(C_t)} * R_{C_t} \tag{10}$$

$$R_{C_{t-1}} = R_{f_t C_{t-1}} \tag{11}$$

$$R_{i_t \tilde{C}_t} = \frac{i_t \tilde{C}_t}{C_t + \varepsilon * sign(C_t)} * R_{C_t} \tag{12}$$

$$R_{\tilde{C}_t} = R_{i_t \tilde{C}_t} \tag{13}$$

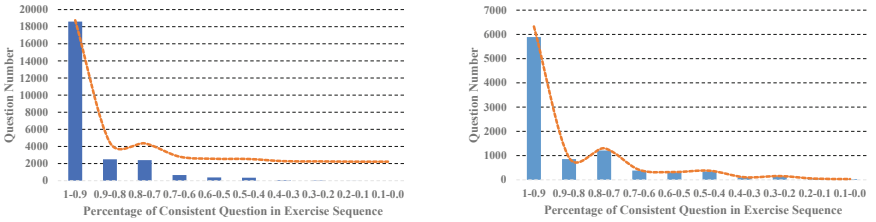
where $R^d_{y_t}$ is the value of the d^{th} dimension of the prediction output y_t, and the item $\varepsilon * sign()$ is a stabilizer. Finally, the calculated relevance value R_{x_t} for the input x_t can be derived as

$$R_{x_t} = \frac{\mathbf{W_{cx}}\mathbf{x_t}}{\mathbf{W_{ch}}\mathrm{h_{t-1}} + \mathbf{W_{cx}}\mathbf{x_t} + \mathrm{b_c} + \varepsilon * \mathrm{sign}(\mathbf{W_{ch}}\mathrm{h_{t-1}} + \mathbf{W_{cx}}\mathbf{x_t} + \mathrm{b_c})} * R_{\tilde{C}_t}$$

Note that the above process is applicable to computing the relevance of the model inputs (e.g., x_{t-1}), while computing $R_{C_{t-1}}$ might be slightly different.

4 Evaluation

We choose the public educational dataset ASSISTment 2009–2010 [9], and the dataset used for training the DLKT model consists of 325,637 answering records on 26,688 questions associated with 110 concepts from 4,151 students. The built DLKT model adopts the LSTM unit with the hidden dimensionality of 256. During the training process, the mini-batch size and the dropout are set to 20 and 0.5 respectively. Considering KT as a classification problem and the exercise results as binary variables, namely 1 representing correct and 0 representing incorrect answers, the overall prediction accuracy achieves 0.75.



(a) Positive Prediction Group (b) Negative Prediction Group

Fig. 1. Histogram of the consistent rate on both positive and negative prediction groups

We conduct the experiment to understand the relationship between the LRP interpreting results and the model prediction results. Specifically, we choose 48,673 exercise sequences with a length of 15, i.e., each sequence consisting of 15 individual questions, as the test dataset for the interpreting tasks. For each sequence, we take its first 14 questions as the input to the built DLKT model, and the last one to validate the model's prediction on the 15th question. As the result, the DKLT model correctly predicts the last question for 34,311 sequences, where the positive and negative results are 25,005 and 9,306 respectively. Based on the correctly predicted sequences, we adopt the LRP method to calculate the relevance values of the first 14 questions, and then investigate whether the sign of relevance values is consistent with the correctness of learner's answer. Specifically, we define *consistent question* among the previous exercise questions as "either the correctly-answered questions with a positive relevance value" or "the falsely-answered questions with a negative relevance value". Accordingly, we compute the percentage of such consistent questions in each sequence, and name it as *consistent rate*. Intuitively, a high *consistent rate* reflects that most correctly-answered questions have a positive contribution and most falsely-answered questions have a negative contribution to the predicted mastery probability on the given concept. Figure 1 shows the histogram of the consistent rate on both groups of positive prediction (i.e., the mastery probability above 50%) and negative prediction (i.e., the mastery probability below 50%). Clearly, we see that the majority of the exercise sequences achieve 90%

(or above) consistent rate, which partially validates the question-level feasibility of using LRP method to interpret DLKT model's prediction results.

5 Conclusion

We have introduced a post-hoc interpretability method into KT domain, which is applicable to general RNN-based DLKT models. We demonstrated the promise of this approach via using its LRP method to explain DLKT models. We conducted the preliminary experiments to validate the proposed method.

Acknowledgment. This research is partially supported by the National Natural Science Foundation of China (No. 61702039 and No. 61807003), the Fundamental Research Funds for the Central Universities, and CCF-Tencent Open Fund.

References

1. Arbabzadah, F., Montavon, G., Müller, K.-R., Samek, W.: Identifying individual facial expressions by deconstructing a neural network. In: Rosenhahn, B., Andres, B. (eds.) GCPR 2016. LNCS, vol. 9796, pp. 344–354. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45886-1_28
2. Arras, L., Horn, F., Montavon, G., Muller, K., Samek, W.: What is relevant in a text document? An interpretable machine learning approach. PLoS ONE **12**(8), 0181142 (2017)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), 0130140 (2015)
4. Baker, R.S.J., Corbett, A.T., Aleven, V.: More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 406–415. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-69132-7_44
5. Chaudhry, R., Singh, H., Dogga, P., Saini, S.K.: Modeling hint-taking behavior and knowledge state of students with multi-task learning. In: Proceedings of Educational Data Mining (2018)
6. Chen, P., Lu, Y., Zheng, V.W., Pian, Y.: Prerequisite-driven deep knowledge tracing. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 39–48. IEEE (2018)
7. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
8. Ding, Y., Liu, Y., Luan, H., Sun, M.: Visualizing and understanding neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1150–1159 (2017)
9. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User Model. User-Adap. Inter. **19**(3), 243–266 (2009). https://doi.org/10.1007/s11257-009-9063-7
10. Grégoire, M., Wojciech, S., Klaus-Robert, M.: Methods for interpreting and understanding deep neural networks. Digit. Signal Process. **73**, 1–15 (2018)

11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
12. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, pp. 505–513 (2015)
13. Qiu, Y., Qi, Y., Lu, H., Pardos, Z.A., Heffernan, N.T.: Does time matter? Modeling the effect of time with bayesian knowledge tracing. In: Proceedings of Educational Data Mining Workshop at the 11th International Conference on User Modeling, pp. 139–148 (2011)
14. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997)
15. Xiong, X., Zhao, S., Van Inwegen, E., Beck, J.: Going deeper with deep knowledge tracing. In: EDM, pp. 545–550 (2016)
16. Yang, H., Cheung, L.P.: Implicit heterogeneous features embedding in deep knowledge tracing. Cogn. Comput. **10**(1), 3–14 (2018)
17. Yeung, C.: Deep-IRT: make deep learning based knowledge tracing explainable using item response theory. In: Proceedings of Educational Data Mining (2019)
18. Zhang, J., Shi, X., King, I., Yeung, D.Y.: Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th International Conference on World Wide Web, pp. 765–774 (2017)