CONTRASTIVE LEARNING WITH BIDIRECTIONAL TRANSFORMERS FOR KNOWLEDGE TRACING

Huijing Zhan, Jung-Jae Kim, Guimei Liu

Institute for Infocomm Research (I2R), A*STAR, Singapore

ABSTRACT

Knowledge tracing aims to predict students' probability of correctly answering the next question based on their interaction history. Previous methods employ left-to-right unidirectional transformers to encode the historical behaviors into hidden representations, especially with contrastive learning methods. Using uni-directional models to model student behaviors can only learn the hidden representation from its previous items, which restricts the power of the representation capability. Inspired by the success of BERT in text understanding, we propose a novel **Bi**directional Transformer encoder guided **C**ontrastive **L**earning framework **f**or deep **K**nowledge **T**racing system, named as **Bi-CL4KT** to generate the accurate response predictions for the next question. We incorporate the Cloze task and carefully design data augmentation methods to generate high-quality positive and negative instances for contrastive learning. Extensive experiments conducted on three real-world education datasets show that the proposed method significantly outperforms the state-of-the-art methods.

Index Terms— Knowledge tracing, bidirectional transformers, contrastive learning.

1. INTRODUCTION

With the increasing use of online learning platform, which has led to exponential growth in learner data, accurately tracing a student's mastering of specific knowledge concepts or skills has become possible. This process is referred to as knowledge tracing (KT). Its goal is to predict the probability whether students can correctly answer given questions based on their historical practices. KT methods have been extensively investigated, and a variety of approaches have been proposed to address this problem.

Current works on knowledge tracing can be broadly categorized into two genres: 1) Bayesian-based methods with hidden Markov modeling such as Bayesian Knoweledge Tracing (BKT) [1, 2] which employ a binary variable to represent a student's knowledge state in relation to a specific skill and 2) deep knowledge tracing with sequential models such as Recurrent Neural Networks (RNN) [3], Long-Term Memory Networks (LSTM) [4] or Transformers [5] aiming to effectively capture the evolution of student's learning activities over time and model students' knowledge state with summarized hidden vectors.

Due to the sparsity of interactions between students and questions [6], contrastive learning approaches [7] with left-to-right unidirectional transformers are introduced to mitigate the over-fitting issue in knowledge tracing. However, such unidirectional models [8] can only consider information in a left-to-right manner, which is not sufficient to learn optimal student behavior representations [9]. In contrast, bidirectional transformers incorporate contextual information by modeling the attention on items from both directions, resulting in improved performance. Nevertheless, the integration of the BERT-based model [10] with contrastive learning to enhance the effectiveness of knowledge tracing remains relatively unexplored. Addressing this challenge necessitates the design of a contrastive learning framework using bidirectional transformers.

This motivation prompts us to carefully examine the following questions when considering the successful incorporation of the contrastive paradigm into the architecture of bidirectional transformers:: Q1) How to select the augmentation method to generate high-quality positive samples? Q2) How to appropriately select negative instances for contrastive learning? Q3) How to address the distinctive challenges posed by the knowledge tracing task, particularly in distinguishing discrepancies in the representation of questions and responses?

In this paper, we introduce a contrastive learning paradigm, Bi-CL4KT, designed for enhancing the performance of bidirectional transformers in the knowledge tracing task. Specifically, our approach involves masking questions and knowledge concepts to generate variant maskings of the original sequence. Subsequently, we replace responses with mask tokens to enhance the model's prediction of responses. Furthermore, we integrate the Cloze task and produce a set of positive samples, which serve as the basis for defining a multi-pair contrastive loss for parameter optimization. To

This research / project is supported by the Ministry of Education, Singapore, under its Science of Learning Grant (Award ID MOE-MOESOL2021-0006). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

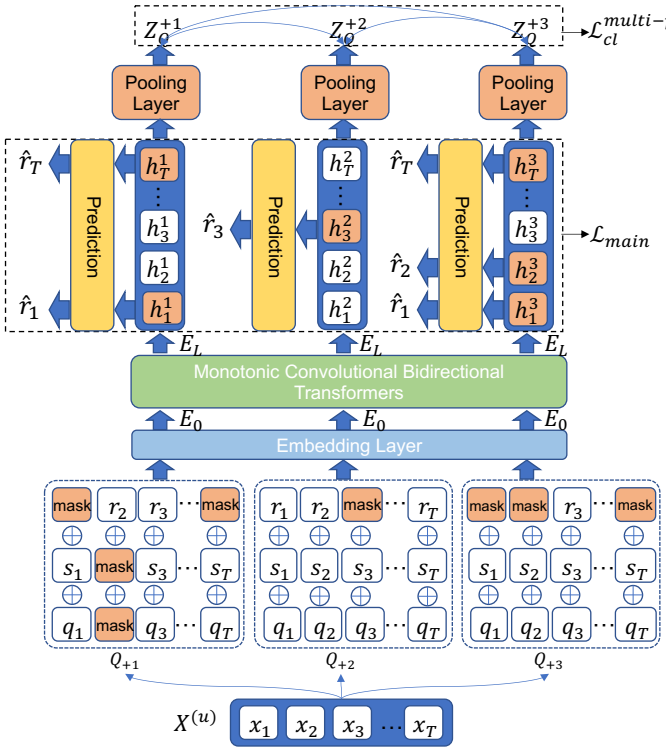

Fig. 1. The architecture of the Bi-CL4KT framework. $\oplus$ represents the addition operation.

curate high-quality negative samples, we leverage interaction histories from other users within the same batch and apply a probabilistic reversal to facilitate the contrastive learning process. Extensive experiments on several benchmarks validate the efficacy of the proposed methodology.

2. PROBLEM STATEMENT

We denote $X^{(u)} = \{x_1^{(u)}, x_2^{(u)}, \cdots, x_T^{(u)}\}$ as an interaction sequence of student u sorted in chronological order, where T is the length. To facilitate representation, we will omit the subscript (u) throughout the paper. At time t, the exercise is represented as $x_t = (q_t, s_t, r_t)$ with a question $q_t \in Q$, its skill $s_t \in S$ and the student's response $r_t \in \{0, 1\}$, respectively. $r_t \in \{0, 1\}$ represents whether the student answers the specified question correctly ($r_t = 1$) or not ($r_t = 0$). Given the exercise sequence $X_t = \{x_1, x_2, \cdots, x_t\}$ and the new question q_{t+1}, the goal of KT is to predict the probability of correctly answering the particular question, denoted as $P(r_{t+1} = 1 | X_t, q_{t+1}, s_{t+1})$.

3. PROPOSED FRAMEWORK

This section provides a detailed description of the proposed framework: Knowledge Tracing with Contrastive Learning using Bidirectional Transformers. The overall framework is illustrated in Fig. 1. Firstly, three branches are generated from the original user sequence X. In order to enhance the diversity of positive samples, we apply the differentiated strategy of masked language modeling during training: for one branch, all the questions, skills, and responses are randomly masked; for the other two branches, only responses are randomly masked. The first branch masking is to learn generic representation of interaction sequences. The last two branch masking is to focus on the goal of the knowledge tracing task, while the two branches may enrich the encoding like ensemble approach. Subsequently, these sequences pass through the embedding layer, resulting in transformed embedding vectors. These embedding vectors are then processed by the monotonic convolutional bidirectional transformers, which yields high-quality representations for the knowledge state's hidden features. To predict the values of masked responses, we incorporate the Cloze task as our primary training objective. Furthermore, to minimize the contrastive loss, we introduce a set of positive samples alongside carefully selected negative instances.

3.1. Base Model

3.1.1. *Embedding Layer*

In the embedding layer, we construct question embedding matrix $\mathbf{E}^Q \in \mathbb{R}^{|Q| \times d}$, skill embedding matrix $\mathbf{E}^S \in \mathbb{R}^{|S| \times d}$, response embedding matrix $\mathbf{E}^r \in \mathbb{R}^{|r| \times d}$ and a positional embedding matrix $\mathbf{E}^P \in \mathbb{R}^{T \times d}$, where d is the embedding size and T refers to the maximum sequence length. Given a question q_t, the representation which is the output of the embedding layer is denoted as below:

$$\mathbf{E}_0 = \mathbf{E}^Q_{q_t} + \mathbf{E}^S_{s_t} + \mathbf{E}^P_t + \mathbf{E}^r_{r_t}. \tag{1}$$

3.1.2. *Monotonic Convolutional Bidirectional Transformer*

The bidirectional transformer in Fig 1 consists of a stack of monotonic convolutional multihead layers and a feed-forward network. Following [11], we employ the monotonic convolutional layer, a mixed attention block that integrates the monotonic self-attention mechanism and span-based dynamic convolution to better model both global and local dependencies with reduced redundancy.

Monotonic Self Attention. Following [11], we use a modified monotonic version of the scaled dot-product attention merchanism. Considering the temporal nature of learning and the decreased significance of distant past interactions in comparison to recent practices, we incorporate an additional multiplicative exponential decay term denoted as below:

$$s_{t,\tau}(Q, K, V) = \text{softmax}\left(\frac{\exp(-\theta \cdot d(t, \tau)) \cdot \mathbf{Q}_t^T \mathbf{K}_\tau}{\sqrt{D_k}}\right), \tag{2}$$

where θ is a learnable decay parameter and $d(t,\tau)$ is a context-aware distance measure between the time steps t and τ, denoted as below:

$$d(t,\tau) = |t-\tau| \cdot \sum_{t'=\tau+1}^{t} \mathrm{softmax}(\frac{\mathbf{Q}_t^T \mathbf{K}_{t'}}{\sqrt{D_k}}). \quad (3)$$

Span-based Dynamic Convolution. A depth-wise separable convolution is utilized to gather the information of a span of tokens and then dynamically generates convolution kernel [12]. It can be formulated as:

$$\mathrm{SDC}(Q,K,V) = \mathrm{LConv}(V, \mathrm{softmax}(\mathbf{W}_f(Q \otimes K))), \quad (4)$$

where LConv denotes the light-weight convolution defined in [12] and $\otimes$ indicates the point-wise multiplication.

The mixed attention, corresponding to the monotonic convolutional multihead attention, is represented as below:

$$\text{Mixed-Attn}(Q,K,V) = \mathrm{Cat}(s_{t,\tau}(Q,K,V), SDC(Q,K,V)), \quad (5)$$

where Cat denotes the concatenation operation. The input representation $\mathbf{E}^0 = [\mathbf{E}_1^0, \mathbf{E}_2^0, \cdots, \mathbf{E}_T^0]$ progresses through L layers of bidirectional monotonic multi-convolutional transformers, and the representation for the l-th layer is as follows:

$$\mathbf{E}^l = \mathrm{MonoConvTrm}(\mathbf{E}^{l-1}), \quad (6)$$

where $\mathbf{E}^l$ is the output of l-th monotonic convolutional mixed attention layer (MonoConvTrm). The hidden representation $\mathbf{E}^L = [\mathbf{E}_1^L, \mathbf{E}_2^L, \cdots, \mathbf{E}_T^L]$ captures the final output of the bidirectional transformer encoder with the mixed attention mechanism. The output from the mix-attention stage is subsequently fed into to the feed forward network to generate the ultimate representation.

3.2. Data augmentation for contrastive learning

Due to the unique combination of interaction data, that is, tuples of question ID, skill ID and response, directly adopting the data augmentation methods from computer vision (CV), natural language processing (NLP) or even sequential recommendation poses challenges to the knowledge tracing task. A carefully designed data augmentation method is required to select high-quality positive and negative instances for the contrastive learning framework. Below we introduce two data augmentation operators on the interaction sequence to create various views of the same sequence.

Positive Samples. Inspired by the success of masked language models such as BERT [13], we adopt a question and response masking strategy to generate high-quality positive samples. We randomly replace question/skill IDs and responses with [MASK] token. During training, the objective is to predict the masked token response. The training data generator randomly selects 15% of the token positions. 80% of the selected positions are filled with [MASK] token, 10% are replaced with a random token. For the question or skill masking, a random token replaces the question or skill ID from the range $[1, \mathrm{Max}(Q)]$ or $[1, \mathrm{Max}(S)]$, respectively. For the response masking, the selected response is replaced with either 0 or 1. Properly masked learning histories, with question/skill or response masking, can be viewed as noisy representations of the original learning sequence. During testing, masking is applied at the last position of the sequence for prediction, utilizing the learning history to predict the correct response.

Negative Samples. The inclusion of appropriate negative samples is crucial for computing the contrastive loss. We randomly select interactions with probability γ_{neg} and reverse the response with $\tilde{r}_t = 1 - r_t$. The representations of the negative samples are integrated for computation as shown in Eq. 8. Given a batch of sequences $\{X^{(u)}\}_{u=1}^N$, two distinct views of the same original learning sequence serve as positive samples while the other $2(N-1)$ hidden representations from the other users within the same batch are treated as the negative samples [14].

3.3. Multi-task learning with Cloze and contrastive losses

Learning with Cloze Task. To facilitate efficient training of bidirectional transformers, we introduce the Cloze task. For each learning sequence $X^{(u)}$, we generate m different masked sequence with distinct random seeds, denoted as Q_{+1}, Q_{+2}, Q_{+3} as shown in Fig. 1. In this paper, we set $m = 3$ as explained in Section 3. The objective of the task is to predict the masked responses' content. The loss function for the Cloze task is defined as follows:

$$\mathcal{L}_{main} = -\frac{1}{m}\sum_{j=1}^{m} \frac{1}{|I_j^u|} \sum_{t \in I_j^u} \log P(v_t | Q_{+j}), \quad (7)$$

where I_j^u denotes the position index of the masked question response. Note that we only consider the masked response for calculating the loss of the Cloze task.

Multi-Pair Contrastive Loss. The goal of contrastive learning is to bring the positive samples closer and push away the negative instances. The contrastive learning loss for one-pair positive instances based on InfoNCE [15] is formulated as below:

$$\mathcal{L}_{cl}(Q_{+x}, Q_{+y}) = -\log \frac{e^{\mathrm{sim}(z_Q^{+x}, z_Q^{+y})}}{e^{\mathrm{sim}(z_Q^{+x}, z_Q^{+y})} + \sum_{z_Q^-} e^{\mathrm{sim}(z_Q^{+x}, z_Q^-)}}, \quad (8)$$

where $Q_{+x}(z_Q^{+x})$ and Q_{+y} (z_Q^{+y}) indicate the one-pair positive samples with their representations for the entire sequence after a pooling layer as illustrated in Fig. 1. z_Q^- refers to the representation of the negative samples. Here we employ the cosine similarity as the measure for sim. The multi-pair contrastive loss is defined as:

$$\mathcal{L}_{cl}^{multi-pair} = \sum_{+x=1}^{m} \sum_{+y=1}^{m} 1_{(+x \neq +y)} \mathcal{L}_{cl}(Q_{+x}, Q_{+y}), \quad (9)$$

Table 1. Performance comparison of the baseline methods on several datasets in predicting the learner responses.

Dataset	DKT	DKVMN	SAKT	CL4KT	MCB	Proposed
ASSISTments2009	0.7285	0.7271	0.7179	0.7600	0.8002	**0.8073**
Algebra2005	0.8088	0.8146	0.8162	0.7871	0.8190	**0.8279**
Algebra2006	0.7939	0.7961	0.7927	0.7789	0.7997	**0.8050**

Table 2. Ablation studies on Algebra2005.

L_{main}	$+L_{cl}^{onepair}$	$+L_{cl}^{multi-pair}$
0.8208	0.8232	0.8279

where $1_{(\cdot)}$ is the indicator function and the value is equal to 1 if $+x \neq +y$ otherwise 0. The overall objective loss is shown as follows:

$$\mathcal{L} = \min_{\theta}(\mathcal{L}_{main} + \lambda_c \mathcal{L}_{cl}^{multi-pair}), \tag{10}$$

where λ_c denotes trade-off weights and θ refer to the trainable parameters of Bi-CL4KT.

4. EXPERIMENTS

4.1. Datasets

We conduct extensive experiments on three publicly available datasets. During the preprocessing stage, student data with fewer than five interactions are excluded. When an interaction instance has multiple skills (or concepts) in a single interaction, we treat the combination of concepts as a unique concept. The Algebra05 and Algebra06 datasets provided by the KDD Cup 2010 EDM Challenge [16]. The assistments 2009 dataset [1] is collected from the ASSISTment intelligent tutoring system.

4.2. Baselines

To evaluate the effectiveness of our model, we compare it with several state-of-the-art baselines including DKT [17], DKVMN [18], SAKT [19], CL4KT [7] and MCB [11].

4.3. Implementation Details

Following the experimental settings in [11], each dataset is divided into training, validation, and testing sets. The experimental results for the baseline methods are obtained from [11]. For the monotonic convolutional multihead attention, we employ eight attention heads for calculation, and the maximum sequence length is set as 100. The batch size is fixed at 32, the embedding size is set to 512 and λ is set to 0.01. The learning rate is chosen as 0.001. For hard negative samples, we tune γ_{neg} among the set of $\{0.1, 0.5, 1.0\}$. Parameter updating utilizes the Adam optimizer. Area under the curve (AUC) is employed as the evaluation metrics for performance comparison.

[1] https://sites.google.com/view/assistmentsdatamining

4.4. Performance Comparison and Ablation Study

Table 1 presents the results of the performance comparison with the baselines. Since the proposed framework's sequence encoder falls within the self-attention transformer family, we compare it with transformer architectures among the baselines. In terms of AUC score evaluation, we observe that our proposed method significantly outperforms the state-of-the-art methods. Our approach shows substantial improvement compared to another contrastive learning-based method CL4KT. Moreover, in comparison with MCB which solely utilizes prediction of the masked tokens, our incorporation of the multi-pair contrastive learning strategy underscores the effectiveness of carefully selected high-quality positive and negative samples. We also conduct ablation experiments on Algebra2005 to validate the significance of different modules, as shown in Table 2. Here $+$ is the sum of main task $\mathcal{L}_{main}$ and the auxiliary tasks with different number of pairs ($\mathcal{L}^{onepair}$ and $\mathcal{L}^{multi-pair}$). The improvement in accuracy achieved using the one-branch positive surpasses that utilizing $\mathcal{L}_{main}$, proving the effectiveness of the contrastive learning. And the increase from $+\mathcal{L}^{multi-pair}$ over $+\mathcal{L}^{onepair}$ confirms the efficacy of the multi-pair positive strategy. On the other perspective, the accuracy increase from employing one-pair positive to multi-pair positives once again demonstrates the superiority of using a collection of multiple positive samples.

5. CONCLUSION

In this paper, we introduce a novel framework named Contrastive Learning with Bidirectional Transformers for deep knowledge tracing. We propose a new strategy to generate high-quality positive and negative instances for calculating contrastive loss. Comparing to the traditional one-pair contrastive learning which doesn't fully leverage the advantage of bidirectional transformers, we incorporate more branches to create a multi-pair instances-based variant. Experimental results on several benchmark datasets demonstrate the superiority of Bi-CL4KT in predicting responses based on the historical interaction learning sequences.

6. REFERENCES

[1] Zachary A Pardos and Neil T Heffernan, "Modeling individualization in a bayesian networks implementation of knowledge tracing," in *User Modeling, Adaptation, and Personalization: 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20-24, 2010. Proceedings 18*. Springer, 2010, pp. 255–266.

[2] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon, "Individualized bayesian knowledge tracing models," in *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16*. Springer, 2013, pp. 171–180.

[3] Youngnam Lee, Youngduck Choi, Junghyun Cho, Alexander R Fabbri, Hyunbin Loh, Chanyou Hwang, Yongku Lee, Sang-Wook Kim, and Dragomir Radev, "Creating a neural pedagogical agent by jointly learning to review and assess," *arXiv preprint arXiv:1906.10910*, 2019.

[4] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu, "Ekt: Exercise-aware knowledge tracing for student performance prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 1, pp. 100–115, 2019.

[5] Aritra Ghosh, Neil Heffernan, and Andrew S Lan, "Context-aware attentive knowledge tracing," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 2330–2339.

[6] Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, AS Albahri, Bashar Sami Nayyef Aldabbagh, Mohammed A Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H Al-Timemy, et al., "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, pp. 46, 2023.

[7] Wonsung Lee, Jaeyoon Chun, Youngmin Lee, Kyoungsoo Park, and Sungrae Park, "Contrastive learning for knowledge tracing," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2330–2338.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1441–1450.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] Unggi Lee, Yonghyun Park, Yujin Kim, Seongyune Choi, and Hyeoncheol Kim, "Monacobert: Monotonic attention based convbert for knowledge tracing," *arXiv preprint arXiv:2208.12615*, 2022.

[12] Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan, "Convbert: Improving bert with span-based dynamic convolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12837–12848, 2020.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[15] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[16] J Stamper, A Niculescu-Mizil, S Ritter, G Gordon, and K Koedinger, "Algebra i 2005-2006 and bridge to algebra 2006-2007. development data sets from kdd cup 2010 educational data mining challenge," 2010.

[17] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein, "Deep knowledge tracing," *Advances in neural information processing systems*, vol. 28, 2015.

[18] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung, "Dynamic key-value memory networks for knowledge tracing," in *Proceedings of the 26th international conference on World Wide Web*, 2017, pp. 765–774.

[19] Shalini Pandey and George Karypis, "A self-attentive model for knowledge tracing," *arXiv preprint arXiv:1907.06837*, 2019.