CAKT: Coupling contrastive learning with attention networks for interpretable knowledge tracing

1st Shuaishuai Zu
Southwest University
Chongqing, China
zushuaishuai@email.swu.edu.cn

2nd Li Li
Southwest University
Chongqing, China
lily@swu.edu.cn

3rd Jun Shen
University of Wollongong)
NSW, Australia
jshen@uow.edu.au

2023 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-8867-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCNN54540.2023.10191799

***Abstract*—In intelligent systems, knowledge tracing (KT) plays a vital role in providing personalized education. Existing KT methods often rely on students' learning interactions to trace their knowledge states by predicting future performance on the given questions. While deep learning-based KT models have achieved improved predictive performance compared with traditional KT models, they often lack interpretability into the captured knowledge states. Furthermore, previous works generally neglect the multiple semantic information contained in knowledge states and sparse learning interactions. In this paper, we propose a novel model named CAKT that couples contrastive learning with attention networks for interpretable knowledge tracing. Specifically, we use three attention-based encoders to model three dynamic factors of the Item Response Theory (IRT) model, based on designed learning sequences. Then, we identify two key properties related to the knowledge states and learning interactions: consistency and separability. We utilize contrastive learning to incorporate the semantic information of the above properties into the representations of knowledge states and learning interactions. With the training goal of contrastive learning, we can obtain more representative representations of them. Extensive experiments demonstrate the excellent predictive performance of CAKT and the positive effects of considering the two properties. Additionally, CAKT can exhibit high interpretability for captured knowledge states.**

***Index Terms*—knowledge tracing, consistency, separability, contrastive learning, attention networks, IRT**

I. Introduction

Recently, intelligent tutoring systems (ITS) have become a fundamental component of online learning, which can provide massive open courses, extensive assessments, and questions. Knowledge tracing (KT) is the task of discovering the students' knowledge states given the learning interactions consisting of responses to questions. Using estimated knowledge states, tutors can capture the students' mastery level of concepts contained in given questions, and then develop a comprehensive ability profile. Due to the difficulty of capturing changing knowledge states, knowledge tracing also predicts students' future performance based on their historical learning interactions as an alternative. In recent decades, massive efforts have been devoted to knowledge tracing [1], they are mainly divided into traditional structured models and deep learning-based models.

The first category generally attempts to attain psychologically meaningful parameters based on human-engineered features [2]–[4]. These models estimate student performance by handcrafted logistic functions and extracting static factors from students' learning interactions, which attempt to provide essential interpretability for captured students' knowledge states. However, since those models treat concepts contained in questions independently, they are unable to make learning transfer for unpractised concepts. With the advance of deep learning, the second category can capture the latent relations across concepts and achieve excellent predictive performance [5]–[8]. Nevertheless, due to the black-box nature of deep neural networks [9], these models generally provide very limited interpretability for students' knowledge states. Therefore, several researchers attempt to combine IRT with neural networks to assign the parameters psychological meaning [10], [11]. However, these models neglect the properties of students' knowledge states and then easily acquire over-fitted representations.

Generally, existing KT methods use sparse learning interactions to capture the knowledge states. They generally concatenate a question representation with a corresponding response representation to define a learning interaction. And, the seperate representations of questions and responses may hinder to improve the predictive performance [5]. Most of those methods ignore utilizing the semantic information contained in learning interactions and knowledge states to enhance corresponding representations. Inspired by previous works [11]–[13], we identify two properties related to students' learning interactions and knowledge states: consistency and separability. As for consistency, it refers to that students can acquire knowledge about the contained concept whether to solve the given question correctly (true-interaction) or not (false-interaction) [12]. And, two consecutive knowledge states of one student are more likely to be triggered in students' knowledge states with little change [14]. As for separability, it refers to that students can bolster their mastery of different concepts while practicing questions belonging to different concepts. Also, the knowledge states of students are distinguishing with each other at the same timestamp, because their historical learning interactions are diverse [15]. Previous research has pointed out that it is valuable to regularize the representations of knowledge states and learning interactions for an interpretable insight, instead of solely purse of high predictive performance [16]. So, we regularize the

Li Li is corresponding author.

two properties of the learning interactions and knowledge states for attaining their more representative representations. Consistency aims to assign semantic similar information to the representations. Separability favors to preserve the semantic dissimilar information in the representations.

In this paper, we propose a novel model named CAKT, that couples contrastive learning [17] with attention networks [18] for interpretable knowledge tracing. Specifically, we extract three different learning sequences from initial learning interactions and then feed them into three specialized attention-based encoders: ability encoder, difficulty encoder, and discrimination encoder. Each learning sequence is fed into one specialized encoder to capture one factor of IRT model. The attention-based encoders can summarize the latent information from historical learning sequences and characterize the changing factors over time. To enhance the representations of learning interactions and knowledge states, we utilize contrastive learning to regularize the two key properties: consistency and separability. As for each interaction, we select the true-interaction and its opposite false-interaction as positives mutually and the interactions contained other concepts as negatives. As for each student's knowledge states, we select its next knowledge states as positives and other students' knowledge states at the same timestamp as negatives. During the optimization of contrastive loss, CAKT can incorporate the semantic information of consistency and separability into the representations of learning interactions and knowledge states. In this way, CAKT can guarantee excellent predictive performance with explainable and meaningful knowledge states. Our main contributions are as follows:

- We propose three attention-based encoders to learn the dynamic factors of IRT model so that the parameters in the model CAKT can be assigned meaningful explanations while guaranteeing the predictive power of deep neural networks.
- We identify two key properties related to the students' learning interactions and knowledge states: consistency and separability. In order to utilize contrastive learning to regularize the representations, we design the selection way of positives and negatives. With the optimization of contrastive loss, CAKT can learn more representative representations of the knowledge states and learning interactions.
- Extensive experiments on four public datasets demonstrate that CAKT performs better than other baseline models. And, CAKT can capture students' changing knowledge states accurately. We also conduct studies to analyze the positive effect of each property.

II. Related Work

A. Knowledge Tracing

Knowledge tracing (KT) is the fundamental task in intelligent tutor systems, which aims to capture the students' knowledge states based on their learning interactions. Currently, existing KT models can be mainly classified into two categories: traditional structure models and deep learning-based models. Item Response Theory (IRT) [3] is a classic structured model that was inspired by educational psychology and has strong interpretability for acquired knowledge states. It extracts multiple static factors from students' learning interactions and estimates students' knowledge states based on a logistic function. Generally, researchers utilize feature engineering to extract the static factors: student ability (h^{ab}), concept difficulty (h^{diff}), and question discrimination (h^{dis}). Then, the probability of student i correctly answering question j is calculated by (1).

$$P(r_{i,j}|h_i^{ab}, h_j^{diff}, h_j^{dis}) = \sigma(h_j^{dis}(h_i^{ab} - h_j^{dif})) \qquad (1)$$

where $h^{ab}, h^{diff}, h^{dis}$ are all uni-dimensional parameters, σ is sigmoid activation. Several extensions of IRT have been proposed by incorporating question structures (HIRT) [19], guessing and slipping factors (TIRT) [20]. However, extracting the factors of IRT by humans is labor-intensive when facing large-scale education data. Besides, these models can't deal with unpracticed questions. And the learned factors are identical for every student and fixed during the whole learning process. Recently, many deep learning-based models are proposed for the KT task, which can address these issues [5], [7], [8], [21]. Self-Attentive Knowledge Tracing (SAKT) [7] is the first to use the attention mechanism [18] to capture context information for the KT task. And, attention mechanisms have demonstrated strong potential in the KT task. Subsequently, Context-Aware Attentive Knowledge Tracing (AKT) [8] modifies and supplements the attention mechanism, resulting in excellent performance in the KT task. Although these models have strong predictive power, they can only offer a limited meaningful explanation for their tens of thousands of parameters. Furthermore, researchers attempt to incorporate traditional structure models into deep neural networks [10], [11]. Neural Cognitive Diagnosis Model (NeuralCDM) [11] utilizes neural networks to learn a complex function to estimate students' knowledge states instead of a handcrafted logistic function. Deep-IRT [10] uses the Dynamic Key-Value Memory Network (DKVMN) to model students' profiles over time and the IRT model to predict the probability that students answer the next question correctly. Albeit explainable, they ignore the differences among questions containing the same concept and don't make full use of students' learning interactions to capture different factors of IRT.

B. Contrastive Learning

Contrastive Learning (CL) [22] aims to provide high-quality representations of all examples in the embedding space, in which similar examples are gathered together, and dissimilar examples are taken apart. It is widely applied in both supervised [23] and unsupervised [24] settings, and it has been the most powerful approach to providing effective representations. CL generally uses the examples generated by data augmentations as positives and the other examples within the same batch as negatives. During the training process, the distance between each example and its positives will be minimized,

while the distance between each example and its negatives will be maximized in the embedding space. For example, SimCSE [24] uses entailment pairs as positives and contradiction pairs as negatives for incorporating more meaningful information into representations, which can significantly improve the performance of downstream tasks. Although CL has become increasingly popular, with great success in a few areas such as computer vision [25] and natural language processing [24], it is still in its infancy for the KT task. It is worth noting that the optimization goal of CL is consistent with the proposed two key properties: consistency and separability. Consistency favors to keep semantic similar information, while separability favors to keep semantic dissimilar information. Therefore, we propose to use CL to regularize the two properties by carefully selecting positives and negatives without using any data augmentations, and hence enhance the representations of learning interactions and knowledge states.

III. Preliminary

In this section, we will present some preliminary foundations of our work, including the problem definition, and designed learning sequences.

A. Problem Definition

Generally, suppose there are K student, M questions, and N concepts. Students attempt to master each concept by practicing some corresponding questions and giving their responses. At timestamp t, the question q_j is given to a student, where q_j is affiliated with a concept c_k. Then the student's response r_t will be recorded into interaction $i_t = \{q_j, c_k, r_t\}$ during the learning process. However, the number of questions is far greater than the number of concepts, and many questions are given to a few students. To alleviate the sparse problem of question representations, we propose the discrimination parameter $\alpha = \{a_1, a_2, a_3, \cdots, a_{M-1}, a_M\}$. In this way, the q_t is denoted as $\{a_j, c_k\}_t$ instead of a independent question representation. If the student answer the question correctly, we set $r_t = r^1$, otherwise $r_t = r^0$. Based on historical learning interactions $I_{1:t} = \{i_1, i_2, i_3, \cdots, i_{t-1}, i_t\}$, the knowledge tracing task is to capture the student's knowledge states h_t and then predicts its responses for the next given question.

B. Learning Sequences

We expand the learning interactions $I_{1:t}$ into three different learning sequences: $X_{1:t}, Y_{i:t}, Z_{1:t}$. Each learning sequence is designed for learning one corresponding dynamic factor of the IRT model. They are defined as follows:

- Sequence $X = \{x_i\}_{i=1}^{T}$: The student ability H_t^{ab} is to measure the students' mastery level of each concept at timestamp t. And, we record $x_t = \{c_k, r_t\}$ to model the multi-dimensional factor H_t^{ab}.
- Sequence $Y = \{y_i\}_{i=1}^{T}$: In order to capture the concept difficulty H_t^{dif} at timestamp t, we record $y_t = \{c_k\}$ to model the multi-dimensional factor H_t^{dif} without using the students' responses $\{r_i\}_{i=1}^{t}$.
- Sequence $Z = \{z_i\}_{i=1}^{T}$: The question discrimination H_t^{dis} is to measure the difference among questions containing the same concept at timestamp t. And we record $z_t = \{a_j, c_k, r_t\}$ to denote acquired conceptual knowledge varying among questions and utilize $Z_{1:t}$ to model the multi-dimensional factor H_t^{dis}.

IV. The CAKT Model

In this paper, we propose a novel model named CAKT, which uses attention networks to capture dynamic factors of the IRT model and uses contrastive learning to learn more representative representations of learning interactions and knowledge states. The overall structure of CAKT is shown in Fig. 1. Specifically, we first regularize the two properties of learning interactions with contrastive learning. After obtaining representations of learning interactions, we design three attention-based encoders and feed them with different learning sequences to capture three dynamic factors. Then, we use these factors to capture students' knowledge states. Simultaneously, we regularize the representations of knowledge states by using contrastive learning again. Finally, the prediction will be given by a mapping function, which selects the corresponding element as the probability to answer the next question correctly. As we can see, the training goal of CAKT consists of three components: $\mathcal{L}_I^{CL}$, $\mathcal{L}_K^{CL}$, and $\mathcal{L}_{pre}$. In this section, we will give detailed descriptions of each component of our training goal.

A. Representations of Learning Interactions

For each question, students will acquire knowledge of its corresponding concept regard the response is correct or not [12]. So, we consider all possible learning interactions and use contrastive learning to enhance the representations of them. Specifically, $I^+ = \{(c_1, r^1), (c_2, r^1), (c_3, r^1), \cdots, (c_N, r^1)\}$ denotes all true-interactions, $I^- = \{(c_1, r^0), (c_2, r^0), (c_3, r^0), \cdots, (c_N, r^0)\}$ denotes all false-interactions. To regularize the properties of consistency and separability, we pull close the representations of interactions containing the same concept and pull apart the representations of learning interactions containing different concepts in the embedding space. Then, we select the opposite learning interaction containing the same concept as positive mutually and other $2N-2$ learning interactions containing different concepts as negatives. For interaction (c_i, r^1), we select (c_i, r^0) as its positive and select $\{(c_j, r^1), (c_j, r^0)\}_{j=1}^{N} (j \neq i)$ as its negatives. The contrastive loss l_i^{CL} is defined as follows:

$$l_i^{CL} = -\log \frac{\exp\left(\frac{I_i^+ \cdot I_i^-}{\tau}\right)}{\sum_{k=1}^{N}\left(\exp\left(\frac{I_i^+ \cdot I_k^+}{\tau}\right) + \exp\left(\frac{I_i^+ \cdot I_k^-}{\tau}\right)\right)} \tag{2}$$

where τ is a temperature parameter to control the strength of separating the semantic information with negatives. The " $\cdot$ " symbol denotes the inner (dot) product, I_i^+, I_i^- denote i-th elements of I^+, I^- respectively. Then, the holistic contrastive

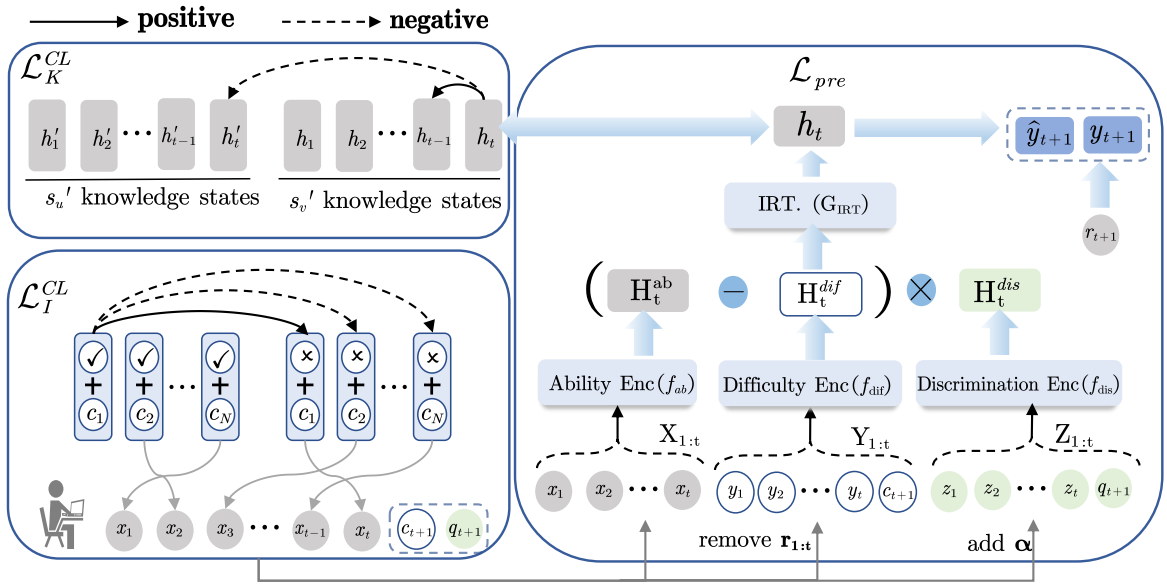

Fig. 1. The overall architecture of the CAKT model. "✓" denotes the response is correct, while "✗" denotes the response is not correct.

loss $\mathcal{L}_I^{CL}$ of learning representations of learning interactions is defined as follows:

$$\mathcal{L}_I^{CL} = \sum_{i=1}^{N} l_i^{CL} \tag{3}$$

B. Attention-based Encoders

To model the dynamic factors of the IRT model, we extract three different learning sequences and design three attention-based encoders. The attention mechanism can summarize the learning gains from historical learning sequences. Besides, the forgetting behaviors of students can't be ignored [26]. As Fig. 2 shows, the input sequence is mapping into corresponding query Q, key K, and value V by multipling embedding matrixes W^Q, W^K, W^V respectively. Then, the attention values $A_{t,t'}$ are calculated as follows:

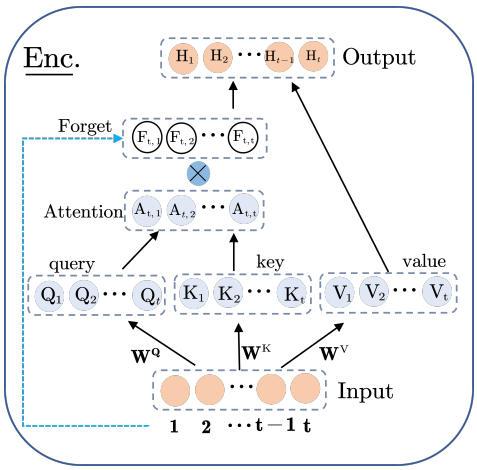

Fig. 2. The architecture of attention-based encoder.

$$A_{t,t'} = \frac{\exp\left(\frac{\mathbf{Q}_t^\top \mathbf{K}_{t'}}{\sqrt{d}}\right)}{\sum_{i=1}^{t} \exp\left(\frac{\mathbf{Q}_t^\top \mathbf{K}_i}{\sqrt{d}}\right)} \tag{4}$$

where $t' < t$ denotes historical timestamp, d is the dimension of Q_t, K_t. To model the forgetting behaviors of students, we add multiplicative exponential decay term to the attention values. The forget term $F_{t,t'}$ is calculated as follows:

$$F_{t,t'} = \frac{\exp\left(\frac{t-t'}{t}\right)}{\sum_{i=1}^{t} \exp\left(\frac{t-i}{t}\right)} \tag{5}$$

Then, the dynamic factor H_t is calculated as follows:

$$H_t = \sum_{t'<t} (A_{t,t'} \cdot F_{t,t'}) \mathbf{V}_{t'} \tag{6}$$

In this paper, we propose three different attention-based encoders: student ability encoder f_{ab}, concept difficulty encoder f_{dif}, and question discrimination encoder f_{dis}. They own the same structure but don't share the parameters. Specifically, the formal calculation process of the proposed three factors is defined as follows:

$$H_t^{ab} = f_{ab}(Q = x_t, K = \{x_i\}_{i=1}^{t}, V = X_{1:t}) \tag{7}$$

$$H_t^{dif} = f_{dif}(Q = c_{t+1}, K = \{c_i\}_{i=1}^{t}, V = Y_{1:t}) \tag{8}$$

$$H_t^{dis} = f_{dis}(Q = q_{t+1}, K = \{q_i\}_{i=1}^{t}, V = Z_{1:t}) \tag{9}$$

where $H_t^{ab}, H_t^{dif}, H_t^{dis} \in \mathbb{R}^d$ are learned dynamic factors of IRT at timestamp t. In this way, the current factors are not only related to historical learning sequences but also the relative distance with past timestamps.

C. Representations of Knowledge States

We incorporate the semantic information of consistency and separability into the representations of knowledge states by using CL to regularize them again. The knowledge state $h_t \in \mathbb{R}^N$ at timestamp t of student s_v is calculated as follows:

$$h_t = G_{IRT}(H_t^{dis} \cdot (H_t^{ab} - H_t^{dif})) \tag{10}$$

where G_{IRT} is a two-layer fully connected networks. For h_t of the student, we select his next knowledge state h_{t+1} as positive, and other students' knowledge states at timestamp t within the batch size M as negatives. CL aims to increase the similarity of h_t, h_{t+1} while keeping difference with h_t and other students' latent states $h_t^{'}$. The contrastive loss l_t^{CL} at timestamp t is defined as follows:

$$l_t^{CL} = -s\left(h_{t+1}^i, h_t^i\right) + \delta \sum_{k=1, k\neq i}^{M} s\left(h_t^i, h_t^k\right) \tag{11}$$

where h_t^i denotes the knowledge state of student s_i at timestamp t. $s(\cdot)$ is a cosine similarity function to measure the distance between two representations. δ denotes the weight of separating the representations of different students' latent knowledge states. The holistic contrastive loss $\mathcal{L}_K^{CL}$ of learning the representations of knowledge states during the whole learning period is defined as follows:

$$\mathcal{L}_K^{CL} = \sum_{t'=1}^{T} l_{t'}^{CL} \tag{12}$$

where T represents the all timestamps.

D. Response Prediction

Consequently, the prediction $\hat{y}_{t+1}$ of the next question q_{t+1} is given by selecting the i-th element of h_t, in which i is the index of corresponding concept contained in q_{t+1}. The prediction loss $\mathcal{L}_{pre}$ is the cross entropy between y_{t+1} and $\hat{y_{t+1}}$, and is calculated as follows:

$$\mathcal{L}_{pre} = \sum_{i=1}^{M} \sum_{t=1}^{T} \left(\hat{y}_t^i \log y_t^i + \left(1 - \hat{y}_t^i\right) \log\left(1 - y_t^i\right)\right) \tag{13}$$

where $\hat{y}_t^i$ denotes the prediction of student s_i at timestamp t, y_t^i is the ground truth. Eventually, the overall training goal $\mathcal{L}$ is consist of $\mathcal{L}_{pre}$ and CL loss $\mathcal{L}_{CL}$, $\mathcal{L}_{CL} = \mathcal{L}_K^{CL} + \mathcal{L}_I^{CL}$. $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{pre} + \mathcal{L}_{CL} \tag{14}$$

The training loss $\mathcal{L}$ is minimized using the Adam optimizer. And more details will be shown in the part of experimental settings.

V. Experiments

In this section, we first describe the datasets used in the experiments. Then, we conduct extensive experiments to demonstrate the effectiveness of our model CAKT from the following aspects: (1) the predictive performance of our proposed model CAKT. (2) the positive effect of regularizing the two properties by using the designed contrastive losses, $\mathcal{L}_I^{CL}$ and $\mathcal{L}_K^{CL}$. (3) the case study to demonstrate the interpretability of acquired knowledge states.

A. Datasets

There are four widely-used datasets evaluating our model, and the details of their statistical information are shown in Table I. ASSIST and Span are online tutoring systems that teach and assess students, in which we can obtain the datasets ASSIST2009, ASSIST2015, ASSISTs2017, and Span.

TABLE I
STATISTICS OF DATASETS AFTER PREPROCESSING. "—" REPRESENTS THE DATASET DOES NOT CONTAIN THE RELEVANT INFORMATION.

	ASSIST2009	ASSIST2015	ASSIST20017	Span
#Student	3852	19840	1709	182
#Concept	123	100	102	221
#Question	16891	—	3162	409
#Interaction	325637	683801	942816	578726

B. Baseline Models

We compare our model with several baseline models. They are listed as follows:

- **IRT** [3] is a classical model and the simplest model for factor analysis, which generally uses two static factors to make predictions.
- **PFA** [4] is a traditional knowledge tracing model based on IRT, which considers the students' successful and unsuccessful attempts.
- **DKT** [5] uses a recurrent neural network (RNN) [27] to model students' learning interactions, in which the knowledge state is represented by latent state of RNN.
- **DKVMN** [21] uses two matrices to get richer interpretable student knowledge state, in which one key matrix stores latent knowledge state and one dynamic value matrix updates the corresponding knowledge state through operations of reading and writing.
- **Deep-IRT** [10] uses the DKVMN model to process the student's learning interactions and estimate the concept difficulty and the student ability over time.
- **SAKT** [7] is the first model to utilize self-attention mechanism to capture historical information for prediction.
- **AKT** [8] proposes a novel monotonic attention mechanism to weight historical learning interactions and utilize a Rasch model to regularize the representations of questions.

C. Experimental Settings

In this section, we demonstrate the parameter settings and training settings. Firstly, the embedding dimension d of concepts, responses, and three dynamic factors are fixed to 128. The embedding dimension of the knowledge state is equal to the number of concepts N. We use the Adam optimizer with a learning rate of 10^{-4}. Furthermore, we truncate learning sequences that are longer than 200 for computational efficiency reasons. If a student has more than 200 interactions, we break

them up into multiple shorter interactions. In order to evaluate the performance of all models more accurately, we use 5-fold cross-validation for all models and report the average results. For each fold, the partition ratio of the training set, validation set, and test set is set to 6: 2: 2. In addition, all models are implemented by PyTorch and trained on a Linux server with a GTX 3090.

D. Experiments

In this paper, we use the Area Under Curve (AUC) metric to measure the prediction performance of each method. Table II shows the AUC results of all models. For dataset ASSIST2015, the question discrimination encoder isn't used, because the dataset doesn't provide any question information. As for other datasets, we use all three encoders. There are some observations from the results:

TABLE II
AUC RESULTS OF ALL MODELS ON FOUR DATASETS (THE HIGHER, THE BETTER). THE BEST RESULTS ARE MARKED IN BOLDFACE FONTS.

	ASSIST2009	ASSIST2015	ASSIST2017	Span
IRT	0.6950	0.6432	0.6820	0.6835
PFA	0.7219	0.6940	0.6256	0.7545
DKT	0.8152	0.7258	0.7636	0.8278
DKVMN	0.8044	0.7260	0.7042	0.8143
Deep-IRT	0.8165	0.7288	0.7156	0.8255
SAKT	0.7512	0.7275	0.7243	0.8113
AKT	0.8251	0.7310	0.7337	0.8320
CAKT	**0.8404**	**0.7514**	**0.7650**	**0.8437**

First, we can observe the noticeable performance differences on most datasets. The model IRT and PFA are inferior to all other models, which demonstrates the superiority of the deep learning-based methods. In addition, the performance of DKT is superior to SAKT on datasets ASSIST2009, ASSIST2017, and Span, indicating that using the attention mechanism directly does not work very well. Compared with SAKT, AKT shows remarkable improvements on most datasets, demonstrating that the modified attention mechanism can yield more performance improvements.

Second, the model CAKT outperforms other baseline models with varying degrees. On the datasets ASSIST2015, CAKT can perform significantly better than other models without using any question information. Furthermore, CAKT has achieved remarkable improvements to varying degrees on the datasets containing question information. This observation demonstrates that CAKT can further capture the question information and achieve better predictive performance.

E. Ablation Studies

In this section, we conduct ablation experiments on three variants of CAKT to verify the effectiveness of diffident implementations. The details of the three variants are defined as follows: (1) CAKT-I: Don't regularize the properties of learning interactions; (2) CAKT-K: Don't regularize the properties of knowledge states; (3) CAKT-IK: Don't regularize the consistency and separability of both learning interactions and knowledge states. From Table III, there are some observations:

TABLE III
PERFORMANCE COMPARISONS BETWEEN CAKT AND ITS VARIANTS. THE BEST RESULTS ARE MARKED IN BOLDFACE FONTS.

	ASSIST2009	ASSIST2015	ASSIST2017
CAKT-I	0.8313	0.7326	0.7543
CAKT-K	0.8247	0.7307	0.7419
CAKT-IK	0.8101	0.7267	0.7202
CAKT	**0.8404**	**0.7514**	**0.7650**

most importantly, CAKT outperforms other variants on all datasets by at least 1% over the closest variant. In addition, CAKT-I, CAKT-K outperform CAKT-IK on all datasets, which demonstrates that each component of our training loss $\mathcal{L}_{CL}$ can bring performance improvements. These results confirm our intuition that more representative representations of learning interactions and knowledge states can be beneficial to the KT task. Note that there are slight performance differences among all variant models on the ASSIST2015, implying that solely utilizing the concept information is still challenging for making promotion.

In addition, there are two noteworthy hyperparameters, namely τ and δ. The first one is τ which regulates the degree of attention to negatives, and the second one adjusts the weight to distinguish the knowledge states of different students within the batch size. We set $\delta = \{0.5, 1, 10\}$, $\tau = \{0.1, 0.5, 1, 10\}$ to demonstrate the performance in different implementations. From Table IV, if we set $\tau \leq 1$, the model's AUC performance is higher compared with $\tau = 10$. For datasets containing question information, when we set $\tau = 1, \delta = 1$, we can obtain the best AUC performance compared with other implementations. For the dataset ASSIST2015, when we set $\tau = 0.1, \delta = 1$, the performance can also beat other implementations. A smaller parameter of τ will make the model pay more attention to separating the negatives, which implies that the model needs to extract more information from the concept differences to achieve any improvements without the question information. When we fix the parameter τ, there are slight differences in performance under different δ settings. And, the improvement yielded by the δ is not as remarkable as the parameter of τ, indicating that regularizing the consistency of knowledge states is the key to make improvements.

F. Visualization

In this section, we conduct some visualization experiments to illustrate the interpretability of the CAKT model. We first visualize one student's 20 consecutive timestamps' knowledge states. And then, we show the clustering result of different students' knowledge states.

As Fig.3 shows, we visualize the evolution process of one student's knowledge states to demonstrate the interpretability of learned knowledge states. We visualize the student's knowledge state of 5 concepts for the first 20 timestamps. For better visualization, we conduct some necessary preprocessing. Firstly, we only select the first 20 questions that the student has answered. Secondly, we log the student's practiced questions on the related concepts. For instance, if the student correctly

TABLE IV
PERFORMANCE COMPARISONS WITH DIFFERENT IMPLEMENTATIONS OF δ AND τ.

	$\delta = 0.5$	$\delta = 1$	$\delta = 10$	$\tau = 0.1$	$\tau = 1$	$\tau = 10$
	$\tau = 0.5$			$\delta = 1$		
ASSIST2009	0.8284	0.8383	0.8253	0.8381	**0.8404**	0.8336
ASSIST2015	0.7439	0.7451	0.7333	**0.7545**	0.7514	0.7334
ASSIST2017	0.7437	0.7585	0.7409	0.7567	**0.7650**	0.7470

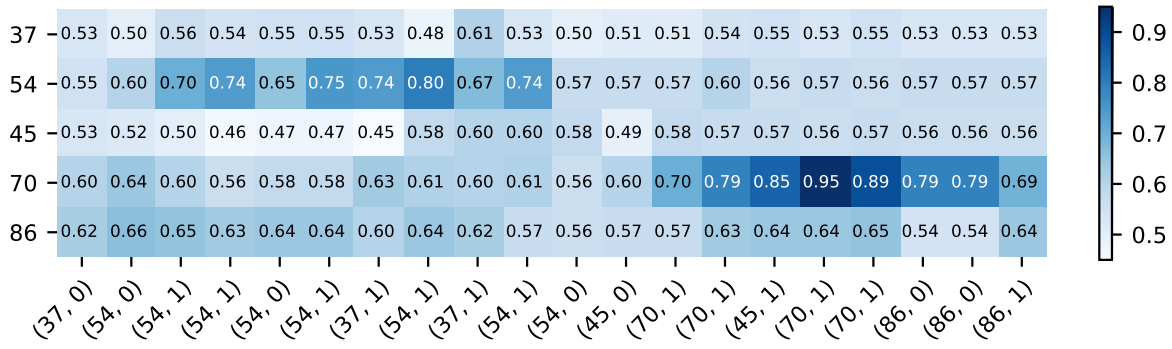

Fig. 3. The evolution of the knowledge state trend.

answers a question related to concept 37, we note it down as (37, 1), otherwise (37, 0).

As Fig. 3 shows, the student's mastery level of each concept is around 0.5 at the beginning, which is in accord with the student's guessing behavior. When the student answers a question correctly, the mastery level of the corresponding concept will be promoted slightly, and vice versa. Besides, the changes are generally subtle between the adjacent timestamps, which fits the property of consistency. If a concept has not been tested for a long time during the learning process, the corresponding mastery level will slowly drop around 0.5, which will still be higher than the initial state. For example, the student had answered two exercises that were related to concept 45, and the final mastery level of concept 45 was slightly higher than the initial state.

Secondly, we show the clustering result of selected students' knowledge states by using t-distributed stochastic neighbor embedding (t-SNE) [28], which is a statistical method for visualizing high-dimensional data by giving each embedding a location in two or three-dimensional map. To obtain more representative representations of knowledge states, we utilize contrastive learning to regularize them. We randomly choose eight students from ASSIST2009 and visualize their knowledge states. Fig. 4 shows the clustering result, we mark different students as different colors, which means that knowledge states belonging to the same student are marked as the same color. We can see that the knowledge states of the same student are scattered along a curve, which shows the property of consistency. And the knowledge states of different students are spread out, which shows the property of separability. With the learning process going on, the knowledge states h_t will change subtly once students practise new questions.

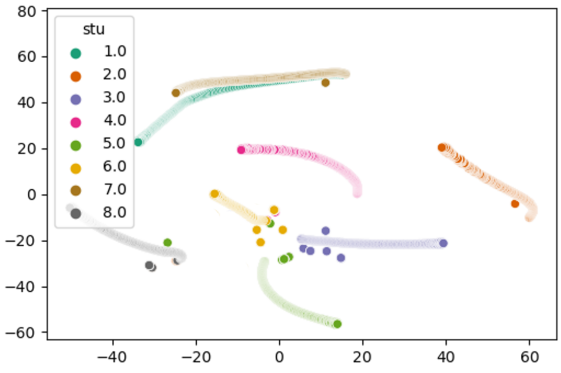

Fig. 4. The t-SNE result of selected students. We select eight students' knowledge states and mark them as different colors. Each point denotes one knowledge state at a certain timestamp.

VI. CONCLUSION

In this paper, we propose a model named CAKT that pursues high interpretable knowledge states and accurate predictive performance simultaneously. CAKT couples contrastive learning and attention networks for interpretable knowledge tracing. Specifically, we first utilized contrastive learning to learn more meaningful representations of learning interactions. Then, three types of learning sequences are fed into three attention-based encoders for capturing the dynamic factors of the IRT model. Moreover, we regularize the students' knowledge states by using contrastive learning again. In this way, we incorporate the semantic information of consistency and separability into the representations of learning interactions

and knowledge states. Extensive experiments have shown that CAKT yields better predictive performance than all baseline models on four widely-used datasets. And, CAKT can learn more explainable and meaningful knowledge states that are consistent with our cognitive process. In the future, we will try to explore more semantic information to enhance the representations of learning interactions and knowledge states.

VII. Acknowledgments

This research was partially supported by grants from the National Natural Science Foundation of China (No.61877051). We acknowledge all the developers and researchers for developing useful tools that enable our experiments.

References

[1] G. Abdelrahman, Q. Wang, and B. P. Nunes, "Knowledge tracing: A survey," *ACM Computing Surveys*, 2022.

[2] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized bayesian knowledge tracing models," in *International conference on artificial intelligence in education*, pp. 171–180, Springer, 2013.

[3] F. Drasgow and C. L. Hulin, "Item response theory.," 1990.

[4] P. I. Pavlik Jr, H. Cen, and K. R. Koedinger, "Performance factors analysis–a new alternative to knowledge tracing.," *Online Submission*, 2009.

[5] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," *Advances in neural information processing systems*, vol. 28, 2015.

[6] H. Nakagawa, Y. Iwasawa, and Y. Matsuo, "Graph-based knowledge tracing: modeling student proficiency using graph neural network," in *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 156–163, IEEE, 2019.

[7] S. Pandey and G. Karypis, "A self-attentive model for knowledge tracing," *arXiv preprint arXiv:1907.06837*, 2019.

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

[9] P. Cui, X. Wang, J. Pei, and W. Zhu, "A survey on network embedding," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 5, pp. 833–852, 2018.

[10] C.-K. Yeung, "Deep-irt: Make deep learning based knowledge tracing explainable using item response theory," *arXiv preprint arXiv:1904.11738*, 2019.

[11] F. Wang, Q. Liu, E. Chen, Z. Huang, Y. Chen, Y. Yin, Z. Huang, and S. Wang, "Neural cognitive diagnosis for intelligent education systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 6153–6161, 2020.

[12] S. Shen, Q. Liu, E. Chen, Z. Huang, W. Huang, Y. Yin, Y. Su, and S. Wang, "Learning process-consistent knowledge tracing," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1452–1460, 2021.

[13] T. Long, Y. Liu, J. Shen, W. Zhang, and Y. Yu, "Tracing knowledge state with individual cognition and acquisition estimation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 173–182, 2021.

[14] R. S. d Baker, A. T. Corbett, and V. Aleven, "More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing," in *International conference on intelligent tutoring systems*, pp. 406–415, Springer, 2008.

[15] G. Steuer and M. Dresel, "A constructive error climate as an element of effective learning environments," 2015.

[16] S. Lee, Y. Choi, J. Park, B. Kim, and J. Shin, "Consistency and monotonicity regularization for neural knowledge tracing," *arXiv preprint arXiv:2105.00607*, 2021.

[17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[18] Y. Choi, Y. Lee, J. Cho, J. Baek, B. Kim, Y. Cha, D. Shin, C. Bae, and J. Heo, "Towards an appropriate query, key, and value computation for knowledge tracing," in *Proceedings of the Seventh ACM Conference on Learning@ Scale*, pp. 341–344, 2020.

[19] R. Janssen, F. Tuerlinckx, M. Meulders, and P. De Boeck, "A hierarchical irt model for criterion-referenced measurement," *Journal of Educational and Behavioral Statistics*, vol. 25, no. 3, pp. 285–306, 2000.

[20] J. González-Brenes, Y. Huang, and P. Brusilovsky, "General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge," in *The 7th international conference on educational data mining*, pp. 84–91, University of Pittsburgh, 2014.

[21] X. Sun, X. Zhao, B. Li, Y. Ma, R. Sutcliffe, and J. Feng, "Dynamic key-value memory networks with rich features for knowledge tracing," *IEEE Transactions on Cybernetics*, 2021.

[22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18661–18673, 2020.

[23] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18661–18673, 2020.

[24] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[25] H. Wang, Y. Li, Z. Huang, Y. Dou, L. Kong, and J. Shao, "Sncse: Contrastive learning for unsupervised sentence embedding with soft negative samples," *arXiv preprint arXiv:2201.05979*, 2022.

[26] K. Nagatani, Q. Zhang, M. Sato, Y.-Y. Chen, F. Chen, and T. Ohkuma, "Augmenting knowledge tracing by considering forgetting behavior," in *The world wide web conference*, pp. 3101–3107, 2019.

[27] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[28] G. C. Linderman and S. Steinerberger, "Clustering with t-sne, provably," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 2, pp. 313–332, 2019.