

SAINT+: Integrating Temporal Features for EdNet Correctness Prediction

DONGMIN SHIN, Riiid! AI Research, Republic of Korea

YUGEUN SHIM, Riiid! AI Research, Republic of Korea

HANGYEOL YU, Riiid! AI Research, Republic of Korea

SEEWOO LEE, Riiid! AI Research, Republic of Korea

BYUNGSOO KIM, Riiid! AI Research, Republic of Korea

YOUNGDUCK CHOI, Riiid! AI Research, Republic of Korea

We propose SAINT+, a successor of SAINT which is a **Transformer based** knowledge tracing model that separately processes exercise information and student response information. Following the architecture of SAINT, SAINT+ has an **encoder-decoder structure** where the encoder applies self-attention layers to a stream of exercise embeddings, and the decoder alternately applies self-attention layers and encoder-decoder attention layers to streams of response embeddings and encoder output. Moreover, SAINT+ incorporates two temporal feature embeddings into the response embeddings: **elapsed time, the time taken for a student to answer, and lag time, the time interval between adjacent learning activities**. We empirically evaluate the effectiveness of SAINT+ on EdNet, the largest publicly available benchmark dataset in the education domain. Experimental results show that SAINT+ achieves state-of-the-art performance in knowledge tracing with an improvement of 1.25% in area under receiver operating characteristic curve compared to SAINT, the current state-of-the-art model in EdNet dataset.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence; Neural networks**; • **Applied computing** → **Interactive learning environments**; • **Social and professional topics** → **Student assessment**.

Additional Key Words and Phrases: Education, Personalized Learning, Knowledge Tracing, Deep Learning, Transformer

ACM Reference Format:

Dongmin Shin, Yugeun Shim, Hangeol Yu, Seewoo Lee, Byungsoo Kim, and Youngduck Choi. 2018. SAINT+: Integrating Temporal Features for EdNet Correctness Prediction. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

The recent COVID-19 pandemic has raised needs for social distancing, leading many organizations to develop virtual and remote services to prevent widespread infection of the disease. Accordingly, educational systems have developed remote learning environments including Massive Open Online Courses and Intelligent Tutoring Systems (ITSs). Knowledge tracing, modeling a student's knowledge state based on the history of their learning activities records, is a fundamental task for creating ITS that aims to provide personalized learning experiences to each student. Traditionally,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Woodstock '18, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

Bayesian Knowledge Tracing [4, 5, 10, 15–17, 19, 20, 24, 26] and Collaborative Filtering based models [11, 22] were common approaches for knowledge tracing. However, with the widespread adoption of Deep Learning (DL) in many machine learning problems, DL based models [2, 6, 9, 12–14, 18, 21, 23, 25, 27] have become the de facto standard for knowledge tracing by capturing complex relations among interactions in students' learning activities records. The strength of the DL based models has become amplified with the advent of large scale public benchmark dataset in Artificial Intelligence in Education. EdNet [3] is such a publicly available dataset which is the largest in scale with more than 131M learning activities records from around 780K students.

In this paper, we propose SAINT+, a successor of SAINT [2] which enhances knowledge tracing with temporal feature embeddings, and empirically verify the effectiveness of the model on EdNet dataset. SAINT is a Transformer [25] based knowledge tracing model that separately processes information of exercise and student response. Specifically, a stream of exercises that a student consumes is fed to an encoder, and a decoder gets a corresponding response sequence and encoder output sequence, computing the final output sequence of the model. SAINT+ augments SAINT by integrating two temporal feature embeddings: elapsed time, the time taken for a student to answer, and lag time, the time interval between adjacent learning activities. Empirical evaluations conducted on EdNet dataset show that SAINT+ improves SAINT, the current state-of-the-art knowledge tracing model on EdNet dataset, by 1.25% in area under receiver operating characteristic curve (AUC). Also, the experimental results show that incorporating the temporal features into the decoder input achieves the best AUC compared to incorporating them into the encoder input, and both the encoder and decoder input, verifying the hypothesis that separately processing exercise information and student response information is appropriate for knowledge tracing.

2 RELATED WORKS

Knowledge tracing is a fundamental task for many computer-aided educational applications and has been studied extensively in the field of AIED. Traditional approaches addressed knowledge tracing based on Bayesian Knowledge Tracing (BKT) [4, 5, 10, 15–17, 19, 20, 24, 26] and Collaborative Filtering (CF) [11, 22]. Basically, BKT is the Hidden Markov Model where a latent variable represents evolving student knowledge. BKT assumes the latent student knowledge as a set of binary variables: either the student mastered the knowledge or not. Each latent variable is updated based on observations of the student correctly applying the knowledge which are also binary: either the student correctly or incorrectly answered a given exercise. On the other hand, CF based approaches model students and exercises as low-rank matrices. Each vector in the student matrix and exercise matrix represents latent traits of each student and latent knowledge required for each exercise, respectively. The probability of a student correctly answers to an exercise is calculated by applying the sigmoid function to the dot product between the corresponding student and exercise vectors.

The advances of Deep Learning (DL) have given rise to neural network based knowledge tracing models. DKT [18] is the first DL based knowledge tracing model. DKT models students' evolving knowledge state through Recurrent Neural Network (RNN) which compresses their past learning activities in a hidden layer. Like many RNN based models that commonly leverage attention mechanism, NPA [12] models student knowledge through Bidirectional Long-Short Term Memory (Bi-LSTM) network equipped with an attention layer that weighs more importance to relevant parts of their learning history for prediction. EKT [9] is also a Bi-LSTM knowledge tracing model with an attention layer. However, EKT addresses the cold start problem in knowledge tracing by exploiting not only students' learning activities records but also text descriptions of exercises. Not all DL based knowledge tracing models are based on RNN architecture. DKVMN [27] is a memory-augmented neural network knowledge tracing model where the key matrix stores knowledge concepts and the

value matrix stores students' mastery levels of corresponding concepts. CKT [21] is a knowledge tracing model that applies hierarchical convolutional operations to extract learning rate features from student's learning activities history. Applying self-attention mechanism in Transformer [25] architecture, which is de facto standard to many sequential prediction tasks, to knowledge tracing is also an actively studied area. SAKT [14] is the first knowledge tracing model with self-attention layers. In each self-attention layer of SAKT, each query is an exercise embedding vector, and key and value are interaction embedding vectors. SAINT [2] is the first Transformer based knowledge tracing model which leverages encoder-decoder architecture composed of stacked self-attention layers. Unlike SAKT, SAINT gets separated streams of exercises and responses as inputs where a sequence of exercises are fed to the encoder, and a sequence of encoder outputs and responses are fed to the decoder. AKT [6] also adopts self-attention layers for knowledge tracing. The attention weights in AKT are decayed exponentially based on the context-aware relative distance measure. Moreover, AKT uses the Rasch model based exercise and exercise-response embeddings to avoid overparameterization and overfitting. Recently, several works [13, 23] attempt to incorporate graph structure to the knowledge tracing model. [13] formulate knowledge tracing as a time series node-level classification task in graph structure and proposes GKT which extracts representation of each knowledge concept by aggregating representations of neighboring concepts. HGKT [23] applies graph neural network to get hierarchical exercise graph which better represent groups of similar exercises.

3 KNOWLEDGE TRACING

We formulate knowledge tracing as a task of predicting the probability of a student's answer being correct to a particular exercise given their previous interaction histories. Formally, the student's learning activity is recorded as an interaction sequence I_1, \dots, I_T . Each interaction $I_t = (E_t, R_t)$ is a tuple of *exercise information* E_t , the t -th exercise given to the student with related metadata, such as the type of the exercise, and *response information* R_t , the student's response to the exercise E_t along with related metadata including the correctness of the response, the duration of time the student took to respond and the time interval between the current and previous interactions. The student's response correctness $c_t \in \{0, 1\}$ is equal to 1 if the student answered the t -th exercise correctly and 0 if not. Thus, knowledge tracing aims to estimate the probability,

$$\mathbb{P}[c_t = 1 | I_1, I_2, \dots, I_{t-1}, E_t].$$

4 SAINT+

4.1 SAINT: Separated Self-Attentive Neural Knowledge Tracing

In this subsection, we give a brief review of SAINT, a Separated Self-Attentive Neural Knowledge Tracing. We refer the paper [2] for those who want to learn detailed aspects of SAINT. SAINT is a knowledge tracing model based on Transformer [25] architecture which consists of an encoder and a decoder. It separates a stream of student interactions into two sequences: an exercise sequence and a response sequence. Then, the encoder takes a sequence of exercise embeddings $E^e = [E_1^e, E_2^e, \dots, E_T^e]$ as input and pass an output sequence $O^{enc} = [O_1^{enc}, O_2^{enc}, \dots, O_T^{enc}]$ to the decoder. The decoder additionally takes a shifted response embedding sequence $R^e = [S^e, R_1^e, R_2^e, \dots, R_{T-1}^e]$ as input, whose first element is a start token embedding, to produce the final output sequence $\hat{c} = [\hat{c}_1, \hat{c}_2, \dots, \hat{c}_T]$. Each \hat{c}_t is an estimated probability of the student's answer to the t -th exercise being correct given the current exercise information E_t and the past interactions I_1, I_2, \dots, I_{t-1} .

The most fundamental part of SAINT is a multi-head attention layer. Let h be the number of heads. When input query matrix Q , key matrix K , and value matrix V are given, we compute

$Q_i = QW_i^Q$, $K_i = KW_i^K$, and $V_i = VW_i^V$ for each $1 \leq i \leq h$, where W_i^Q , W_i^K , and W_i^V are weight matrices of query, key and value, respectively. Then, the scaled dot-product attention computes each head matrix, and the final output is a linear transformation of concatenated head matrices,

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O$$

$$\text{where head}_i = \text{Softmax}\left(\text{Mask}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right)\right)V_i,$$

where d is a dimension of the query and key vectors, and W^O is a weight matrix. Note that the masking mechanism overwrites the region above the diagonal of the matrix $Q_i K_i^T$ with $-\infty$ so that the corresponding region of the softmax output becomes zero. This prevents the current position from attending to subsequent positions. In other words, SAINT uses no future information from the sequence while training.

The encoder block consists of sequentially aligned N copies of encoder layers. A single encoder layer is a multi-headed self-attention layer with an upper triangular mask followed by a feed forward network (FFN) which is defined by

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2,$$

where W_1, W_2 and b_1, b_2 are weights and biases, respectively. Suppose X is given as input to an encoder layer. Then, the output sequence O is computed as follows:

$$M = X + \text{Multihead}(\text{LayerNorm}(X, X, X))$$

$$O = M + \text{FFN}(\text{LayerNorm}(M)).$$

Here, the input sequence X for the foremost encoder layer is an exercise embedding sequence E^e while each subsequent layer takes the feed forward output of the previous layer. Note that we apply layer normalization [1] and skip connection [7] to every sub-layer.

The decoder is a sequence of N identical decoder layers, which consists of two multi-head attention layers with upper triangular masks followed by a feed forward network as well. Suppose X is an input sequence to a decoder layer. If the layer is the foremost, X is the response embedding sequence R^e . Otherwise, it is the output sequence from the previous decoder layer. The first layer is a multi-headed self-attention layer which only takes X . Then, its output M_1 serves as queries for the second attention whose keys and values are the encoder output O^{enc} . The computation can be summarized as follows:

$$M_1 = X + \text{Multihead}(\text{LayerNorm}(X, X, X))$$

$$M_2 = M_1 + \text{Multihead}(\text{LayerNorm}(M_1, O^{enc}, O^{enc}))$$

$$O = M_2 + \text{FFN}(\text{LayerNorm}(M_2)).$$

The output of the last decoder layer is passed to a fully connected layer to produce the final output of the model.

4.2 Enhancing Knowledge Tracing with Temporal Feature Embeddings

SAINT takes sequences of exercise embeddings and response embeddings, where each entry in the exercise embedding sequence is the sum of vectors of an exercise ID, an exercise category and the position, and each entry in the response embedding sequence is the sum of vectors of a response correctness and the position. SAINT+ augments response embeddings with two temporal feature embeddings: elapsed time, the duration of time a student took to respond, and lag time, the time interval between the current and previous interactions (Figure 1). The embedding vectors

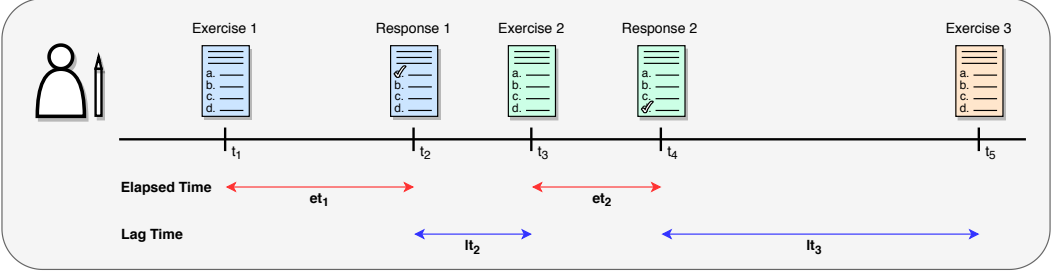


Fig. 1. Descriptions of elapsed time and lag time.

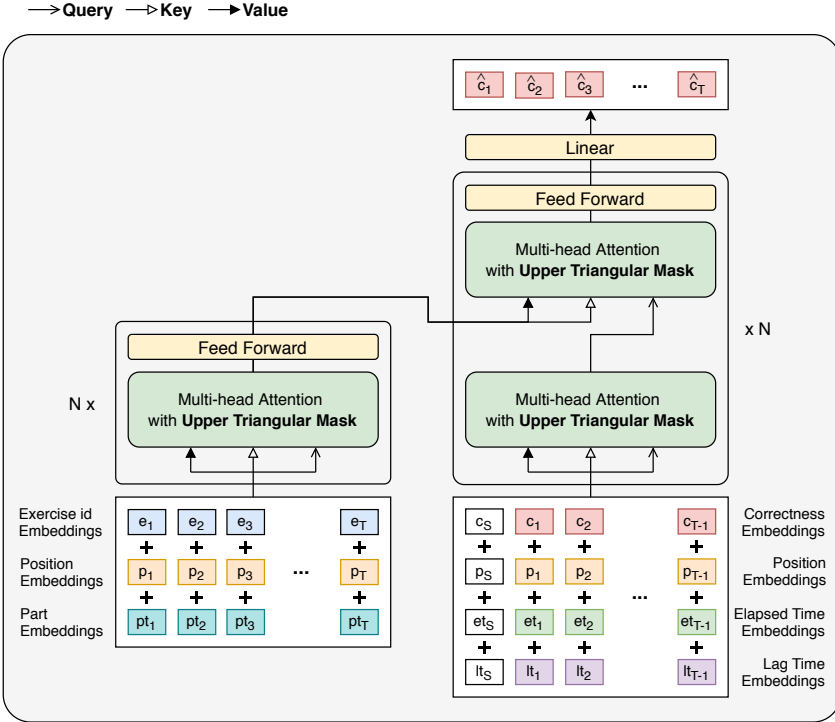


Fig. 2. Model architecture of SAINT+.

for elapsed time and lag time are added to the response embeddings (Figure 2). In the following subsections, we provide detailed explanations of the two temporal feature embeddings.

4.2.1 Elapsed Time. Elapsed time is an amount of time that a student spent on solving a given exercise. For example, in Figure 1, $et_1 = t_2 - t_1$ (resp. $et_2 = t_4 - t_3$) is the elapsed time for the exercise 1 (resp. exercise 2). If the student does not have enough knowledge and skills for the exercise, it would be hard to respond correctly within the recommended time limit. Hence, elapsed time provides strong evidence for a student's proficiency in knowledge and skills, and student's understanding of concepts associated with the exercise.

We propose two different approaches to embed elapsed times as latent vectors: **continuous embedding** and **categorical embedding**. In continuous embedding, a latent embedding vector for an elapsed time et is computed as $\mathbf{v}_{et} = et \cdot \mathbf{w}_{elapsed_time}$, where $\mathbf{w}_{elapsed_time}$ is a single learnable vector. For categorical embedding, **unique latent vectors are assigned to each integer seconds**. We set the maximum elapsed time as **300 seconds** and any time more than that is capped off to 300 seconds.

4.2.2 Lag Time. Lag time is the time gap between interactions, an important factor that affects complex phenomena occurring in students' learning process. For instance, students tend to forget what they have learned as time passes. If a lot of time passed after a student answers an exercise about certain concepts, it would be hard to respond to similar exercises correctly, even if they provided the correct answer to the exercise before. On the other hand, students need time to refresh. By taking a rest, their brains organize and arrange what they have learned, and prepare for the next learning session. We define lag time as the time interval between the moment a student encountered the current exercise and the moment the student consumed the previous exercise. For instance, in Figure 1, lag time for the exercise 2 (resp. exercise 3) is $lt_2 = t_3 - t_2$ (resp. $lt_3 = t_5 - t_4$).

Similar to the elapsed time embedding, we use continuous embedding and categorical embedding for lag time. In continuous embedding, a latent embedding vector for a lag time lt is computed as $\mathbf{v}_{lt} = lt \cdot \mathbf{w}_{lag_time}$, where \mathbf{w}_{lag_time} is a trainable vector. For categorical embedding, lag times are discretized as integer minutes 0, 1, 2, 3, 4, 5, 10, 20, 30, ..., 1440. As a result, there are a total of 150 trainable unique latent vectors assigned to each integer minute.

5 EXPERIMENTS

5.1 Dataset

EdNet [3] is the largest publicly available benchmark dataset in education domain consisting of interaction logs collected by Santa¹. We conduct experiments on an updated version of EdNet-KT1, which contains problem-solving logs from January 1st, 2019 to June 1st, 2020. Details of the dataset statistics are provided in Table 1. Also, the distributions of elapsed time and lag time of the dataset are shown in Figure 3. We use interaction logs of the most recent 100K students as test set and 80% (resp. 20%) of the remaining dataset are used as training (resp. validation) set.

Table 1. Statistics of an updated version of EdNet-KT1 dataset.

# of interactions	60294498
# of students	678128
# of exercises	14418

5.2 Baseline Models

We compare SAINT+ against benchmark knowledge tracing models, DKT [18], DKVMN [27], SAKT [14] and SAINT [2]:

- DKT is a simple RNN-based model that uses concept and response correctness as input features, and models student's knowledge status as RNN's hidden state vectors. We choose an LSTM [8] architecture. Also, we consider each unique exercise id as a concept associated with the exercise.

¹<https://aitutorsanta.com>

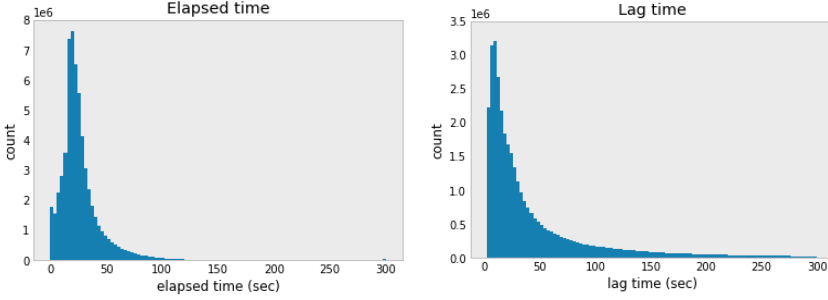


Fig. 3. Distributions of elapsed time and lag time in an updated version of EdNet-KT1.

- DKVMN is a memory augmented neural network based model where the key matrix stores knowledge concepts and the value matrix stores students’ mastery levels of the corresponding concepts. We use each exercise id as a concept of the corresponding exercise.
- SAKT is the first knowledge tracing model that utilizes Transformer’s self-attention architecture. It is a single encoder-based model where exercise ids are used as queries and interaction ids are used as keys and values.
- SAINT is the first Transformer-based knowledge tracing model which leverages an encoder-decoder structure to process information of question and student response separately. The encoder takes a sequence of question embeddings and the decoder gets a sequence of encoder outputs and response embeddings to compute the final output.

5.3 Model Training and Evaluation

We use the accuracy (ACC) and the area under the receiver operating characteristic curve (AUC) as the performance metric. We pick the weight with the best validation AUC and evaluate them on the test set. For SAINT+ and SAINT, we use Adam optimizer with $lr = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e-8$, and set warmup steps to 4000. The window size, number of layers, dimension of the model, dropout rate and batch size is set to 100, 4, 512, 0.0 and 64, respectively. For other benchmark models, both embedding and hidden dimensions of DKT, DKVMN and SAKT are searched over [50, 100, 150, 200, 256, 512], and the best results are reported on the test dataset. Also, we set the number of latent concepts as 64 for DKVMN.

5.4 Main Results

Table 2 shows the performance comparison of SAINT+ with the benchmark knowledge tracing models. Compared to the benchmark models, SAINT+ achieves increases of ACC and AUC maximally up to 2.72% and 3.61%, respectively. Also, SAINT+ improves SAINT with 1.03% and 1.25% gain in ACC and AUC, respectively, demonstrating the effectiveness of integrating the temporal features for knowledge tracing.

5.5 Ablation Test

5.5.1 Temporal Feature Embedding: Continuous vs. Categorical. We compare different approaches for embedding temporal features: continuous and categorical. Modeling elapsed time in continuous (resp. categorical) fashion assumes that the relationship between a student’s knowledge status and the difficulty of a question for the student is a smooth (resp. stepwise) function of time. Similarly, for lag time, continuous (resp. categorical) modeling addresses various aspects in a students’ learning process including forgetting, re-organizing concepts and improvement change smoothly (resp.

Table 2. Comparison of ACC and AUC between benchmark knowledge tracing models and SAINT+.

Model	ACC	AUC
DKT	0.7060	0.7638
DKVMN	0.7079	0.7668
SAKT	0.7073	0.7663
SAINT	0.7178	0.7816
SAINT+	0.7252	0.7914

discretely) over time. Table 3 show that the best result is obtained when using continuous embedding for elapsed time and categorical embedding for lag time.

Table 3. Comparison of different approaches for temporal feature embeddings: continuous and categorical for elapsed time and lag time.

Elapsed time	Lag time	ACC	AUC
Continuous	Continuous	0.7239	0.7898
Continuous	Categorical	0.7252	0.7914
Categorical	Continuous	0.7245	0.7911
Categorical	Categorical	0.7235	0.7900

5.5.2 Elapsed Time vs. Lag Time. We verify the contribution of each temporal feature by comparing performance improvements of two variants of SAINT: 1) SAINT with elapsed time only, and 2) SAINT with lag time only. The results are described in Table 4. When used alone, lag time increases performance more than elapsed time. Moreover, regardless of the temporal features used, utilizing only one feature is not good as using both features together (SAINT+), while certain improvements are obtained compared to SAINT.

Table 4. Contribution of each temporal feature. ET (resp. LT) stands for elapsed time (resp. lag time) and SAINT+ET (resp. SAINT+LT) only uses elapsed time (resp. lag time).

	ACC	AUC
SAINT	0.7178	0.7816
SAINT + ET	0.7213	0.7858
SAINT + LT	0.7239	0.7898
SAINT + ET + LT	0.7252	0.7914

5.5.3 Integrating Temporal Features: Encoder vs. Decoder. The temporal information from the interaction logs, elapsed time and lag time, are provided as decoder features in SAINT+ as described in Figure 2. Since both elapsed time and lag time arise as a result of student response, this approach is naturally aligned with the core idea of SAINT that providing exercise information to the encoder and response information to the decoder is appropriate for knowledge tracing. We compare SAINT+ (decoder only) with other two variants: feeding the temporal features to 1) the encoder only, and 2) both the encoder and the decoder. Table 5 summarizes the results. As expected, SAINT+ shows

the best performance among the variants. Also, SAINT+ and the variants show better results than SAINT, demonstrating that the temporal features provide useful information for estimating knowledge status.

Table 5. Comparison of methods for integrating temporal features: encoder only (Enc), decoder only (Dec) and both encoder and decoder (Enc+Dec).

	ACC	AUC
SAINT	0.7178	0.7816
Enc	0.7216	0.7866
Dec	0.7235	0.7900
Enc + Dec	0.7221	0.7872

6 CONCLUSION

In this paper, we proposed SAINT+, a Transformer based knowledge tracing model that processes exercise information and student response information separately, and integrates two temporal feature embeddings into the response embeddings: elapsed time and lag time. Experiments conducted on EdNet dataset show that SAINT+ improves SAINT, the former state-of-the-art knowledge tracing model, in both ACC and AUC. Furthermore, the best result was obtained by incorporating the temporal features into the decoder input, verifying the hypothesis that separately processing exercise information and student response information is appropriate for knowledge tracing. **Avenues of future work include 1) modeling not only students' problem-solving records, but also various learning activities, such as watching lectures and studying explanations for each exercise, 2) exploring architectures for knowledge tracing models other than Transformer based encoder-decoder model that separately processes exercise information and student response information.**

REFERENCES

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [2] Youngduck Choi, Youngnam Lee, Junghyun Cho, Jineon Baek, Byungsoo Kim, Yeongmin Cha, Dongmin Shin, Chan Bae, and Jaewe Heo. 2020. Towards an Appropriate Query, Key, and Value Computation for Knowledge Tracing. *arXiv preprint arXiv:2002.07033* (2020).
- [3] Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. 2020. Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education*. Springer, 69–73.
- [4] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.
- [5] Ryan SJ d Baker, Albert T Corbett, and Vincent Aleven. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *International conference on intelligent tutoring systems*. Springer, 406–415.
- [6] Aritra Ghosh, Neil Heffernan, and Andrew S Lan. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2330–2339.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [9] Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, Guoping Hu, et al. 2019. EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction. *IEEE Transactions on Knowledge and Data Engineering* (2019).
- [10] Jussi Kasurinen and Uolevi Nikula. 2009. Estimating programming knowledge with Bayesian knowledge tracing. *ACM SIGCSE Bulletin* 41, 3 (2009), 313–317.

- [11] Kangwook Lee, Jichan Chung, Yeongmin Cha, and Changho Suh. 2016. Machine Learning Approaches for Learning Analytics: Collaborative Filtering Or Regression With Experts?. In *NIPS Workshop, Dec.* 1–11.
- [12] Youngnam Lee, Youngduck Choi, Junghyun Cho, Alexander R Fabbri, Hyunbin Loh, Chanyou Hwang, Yongku Lee, Sang-Wook Kim, and Dragomir Radev. 2019. Creating A Neural Pedagogical Agent by Jointly Learning to Review and Assess. *arXiv preprint arXiv:1906.10910* (2019).
- [13] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. 2019. Graph-based Knowledge Tracing: Modeling Student Proficiency Using Graph Neural Network. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 156–163.
- [14] Shalini Pandey and George Karypis. 2019. A Self-Attentive model for Knowledge Tracing. *arXiv preprint arXiv:1907.06837* (2019).
- [15] Zachary Pardos and Neil Heffernan. 2010. Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm. In *Educational Data Mining 2010*.
- [16] Zachary A Pardos, Yoav Bergner, Daniel T Seaton, and David E Pritchard. 2013. Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX. *EDM 13* (2013), 137–144.
- [17] Zachary A Pardos and Neil T Heffernan. 2010. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 255–266.
- [18] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in neural information processing systems*. 505–513.
- [19] Yumeng Qiu, Yingmei Qi, Hanyuan Lu, Zachary A Pardos, and Neil T Heffernan. 2011. Does Time Matter? Modeling the Effect of Time with Bayesian Knowledge Tracing.. In *EDM*. 139–148.
- [20] Michael Sao Pedro, Ryan Baker, and Janice Gobert. 2013. Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In *Educational Data Mining 2013*.
- [21] Shuanghong Shen, Qi Liu, Enhong Chen, Han Wu, Zhenya Huang, Weihao Zhao, Yu Su, Haiping Ma, and Shijin Wang. 2020. Convolutional Knowledge Tracing: Modeling Individualization in Student Learning Process. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1857–1860.
- [22] Nguyen Thai-Nghe, Lucas Drumond, Artus Krohn-Grimberghe, and Lars Schmidt-Thieme. 2010. Recommender system for predicting student performance. *Procedia Computer Science* 1, 2 (2010), 2811–2819.
- [23] Hanshuang Tong, Yun Zhou, and Zhen Wang. 2020. HGKT: Introducing Problem Schema with Hierarchical Exercise Graph for Knowledge Tracing. *arXiv preprint arXiv:2006.16915* (2020).
- [24] Brett van De Sande. 2013. Properties of the bayesian knowledge tracing model. *Journal of Educational Data Mining* 5, 2 (2013), 1.
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [26] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. 2013. Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education*. Springer, 171–180.
- [27] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 765–774.