Memory-Augmented Neural Networks for Knowledge Tracing from the Perspective of Learning and Forgetting

Heonseok Ha, Uiwon Hwang, Yongjun Hong, and Sungroh Yoon∗

Electrical and Computer Engineering
Seoul National University
Seoul 08826, Korea
{heonseok.ha, uixon.hwang, yjhong, sryoon}@snu.ac.kr

Abstract

Knowledge tracing (KT) refers to a machine learning technique to assess a student’s level of understanding (or knowledge state) based on the student’s past performance in exercise-solving. KT accepts a series of question-answer pairs as input and iteratively updates the knowledge state of the student, eventually returning the probability of the student solving a given question. To estimate the accurate knowledge state, a KT model should imitate the learning and forgetting mechanisms of the student. Deep learning-based KT models, proposed recently, show a higher predictive performance than traditional machine learning-based KT models due to the representative power of neural networks. The dynamic key value memory network (DKVMN), a kind of memory augmented neural network (MANN), is a state-of-the-art KT model, but it has some limitations. DKVMN does not utilize information from a current knowledge state and overestimates the amount of forgetting when updating the knowledge state. To improve the learning and forgetting mechanism of the DKVMN, we propose a knowledge tracing model that incorporates: (1) an adaptive knowledge growth depending on the current knowledge state, and (2) an additional loss term that can regularize the degree of forgetting. To measure the degree of forgetting of the KT model, we define a positive update ratio (PUR) that can complement the predictive performance metric (AUC). According to our experiments using four public benchmarks, the proposed approaches outperform the original DKVMN in terms of both AUC (predictive performance) and PUR (degree of forgetting).

Knowledge tracing (KT) is a machine learning-based task that identifies the current knowledge states of students based on their past performance (Corbett and Anderson 1994, Piech et al. 2015, Zhang et al. 2017a). KT plays crucial roles for intelligent tutoring systems (ITSs) (Brustolovsky, Schwarz, and Weber 1996, Goodkovsky 2004, Burns et al. 2014) to provide high-quality education service to students. Figure 1 shows the framework of an ITS. A tutoring system provides appropriate educational services (e.g., lectures and exercises) to a student. The student then gives feedback to the system through solving the exercise, and KT estimates the knowledge state of the student. Based on the estimated understanding of the student, the tutoring system recommends an educational item to the student.

Several studies on improving the predictive performance of KT models, such as item response theory (IRT) (Embreton and Reise 2013), Bayesian knowledge tracing (BKT) (Corbett and Anderson 1994), and performance factor analysis (PFA) (Pavlik Jr, Cen, and Koedinger 2009), have been proposed. These approaches have limitations including the fact that exercises should be labelled by experts and an exercise can be mapped to only one concept.

Deep learning-based models have a very high capacity compared to traditional machine learning-based models and can model components that traditional models cannot. Recently, deep learning-based KT models, such as deep knowledge tracing (DKT) (Piech et al. 2015) and dynamic key value memory networks (DKVMN) (Zhang et al. 2017a) which is state-of-the-art, have been proposed, showing a large performance improvement over previous hand-crafted models. Deep learning-based KT models do not need expert labors and can map one exercise to many concepts.

To be good at tracing students’ knowledge, a KT model requires modeling of both learning and forgetting mechanisms of the student. Learning is an increase in concept mastery level over studying (Rohrer 2009) and forgetting is a decrease in concept mastery level over time, reducing the probability of answering correctly (Ebbinghaus 2013).

In this paper, we analyze the DKVMN model from the learning and forgetting perspective. Firstly, we improve the learning process of the model by exploiting the insight of the cognitive science and increase the predictive performance. Secondly, we identify the cause of the forgetting that occurs in DKVMN and eliminate the forgetting caused by the limitations of the model, not by the data.

Several researches of cognitive science suggest that students have different degrees of knowledge growth even if they solve the same problem because of their different knowledge states (Wandersee, Mintzes, and Novak 1994). A probabilistic graphic model based KT (Reddy, Labutov, and Joachims 2016) assumes that the more knowledge a student have, the larger the growth is, and calculates the knowledge growth based on current knowledge state. In DKVMN, knowledge growth depends on question-answer pairs only and is independent of the current knowledge state. In this paper, we propose an adaptive knowledge growth that reflects
1. We improve the learning process of DKVMN by applying an adaptive knowledge growth derived from the insight of cognitive science.

2. We introduce the two types of forgetting (data-oriented and model-oriented) to explain the forgetting of a model, and propose the regularization term to reduce the effect of model-oriented forgetting.

3. We define a positive update ratio (PUR) as a new metric for measuring the forgetting of the KT model.

4. The extensive experiments on the four published benchmarks show the performance improvement in terms of AUC and PUR.

Related Work

Learning and forgetting

In order to track students’ knowledge state well, it is important to know how students learn and forget. There have been numerous studies on human learning in many fields, such as cognitive science and neuroscience (Atkinson and Shiffrin 1968; Brod, Werkle-Bergner, and Shing 2013). Learning means that the level of concept mastery increases by studying while forgetting means that a student’s knowledge decreases over time. As for learning, there is a representative study that the knowledge growths differ depending on the order where exercises are solved (Rohrer 2009), and with respect to forgetting, there is a research that forgetting occurs along an exponential curve (Ebbinghaus 2013). Learning and forgetting can be estimated indirectly by observing the change in the probability of answering correctly.

Traditional knowledge tracing

There are many machine learning-based researches to estimate the performance of a student such as regression-based models (Embretson and Reise 2013; Cen, Koedinger, and Junker 2006; Pavlik Jr, Cen, and Koedinger 2009), matrix factorization-based models (Abdi, Khosravi, and Sadiq) and Bayesian network-based models (Corbett and Anderson 1994; Käser et al. 2017).

Bayesian knowledge tracing (BKT) (Corbett and Anderson 1994), which is one of the most prominent KT model, assumes the student’s knowledge state as a binary state and models the level of understanding using a hidden Markov model (HMM) (Sonnhammer et al. 1998) for each concept. Since the BKT tracks the level of understanding separately for each concept, it does not consider the entanglement between concepts and, hence, dealing with a mixture of exercises involving various concepts is difficult.
Traditional KT models basically require experts to label exercise tags directly. Such models have low complexity, and are not enough to express a level of understanding continuously (Corbett and Anderson 1994).

**Deep learning-based knowledge tracing**

The first studies that used deep learning in KT were DKT (Piech et al. 2015) based on RNN (LeCun, Bengio, and Hinton 2015) and LSTM (Hochreiter and Schmidhuber 1997). The KT treats a hidden state of RNN as the student’s knowledge state, assuming that a hidden state represents the level of understanding of whole concepts. After DKT was proposed, there were several studies comparing DKT and BKT (Khajah, Lindsey, and Mozer 2016; Wilson et al. 2016). DKT differs from BKT in that DKT can deal with several concepts simultaneously for various questions and express the student’s knowledge state in a continuous manner instead of a binary form. (Montero et al. 2018) showed that DKT’s excellence was due to its capacity to model what BKT could not model, and its ability to interpret an exercise as multiple skills. In the research of (Khajah, Lindsey, and Mozer 2016), the authors believed that the success of DKT is due to the degree of freedom of the model, not the high-level representation of deep learning. Therefore, they demonstrated that the variants of a shallow BKT model showed better predictive performance than or similar performance to DKT through experiments. In particular, the variants applying forgetting and skill discovery concepts showed a high performance.

The knowledge tracing model can utilize not only \((q_t, r_t)\), but also various side information, such as a description of exercises or the first response time of students. On the traditional machine learning basis, there was a difficulty for the expert to make hand-crafted features directly in order to analyze these data. Due to the development of deep learning, latent representation can be effectively extracted from side information. Auto-encoder (Hinton and Salakhutdinov 2006) converts high-dimensional side information into low-dimensional features (Zhang et al. 2017b), and a bidirectional LSTM-based exercise-enhanced recurrent neural network (EERNN) used text description as the side information (Su et al. 2018).

Memory augmented neural network (Santoro et al. 2016) has a larger capacity (degree of freedom) than RNN and LSTM as it uses external memory, which shows excellent performance in some tasks. A DKVMN (Zhang et al. 2017a), which is a memory augment neural network-based model, can analyze the level of understanding of each concept as BKT and utilize the correlation between concepts as DKT. The DKVMN adopts a key memory and a value memory. The key memory stores the representation of \(N\) concepts involved in exercises, and the value memory stores the student’s mastery level for each concept. With read and write operations to these two memories, the DKVMN updates the student’s knowledge state.

**Methodology**

A student interplays with an ITS, and the ITS can observe an interaction (so-called knowledge growth signal) \(x_t = (q_t, r_t)\) at time step \(t\), where \(q_t \in [1, Q]\) is an ex-
Exercise tag (ID), where $Q$ is the number of exercises, and $r_t \in \{0, 1\}$ is the correctness of the student response. KT is a supervised learning problem in that given past interactions $X = (x_1, x_2, \ldots, x_t)$ and a new exercise $q_{t+1}$, predicts the probability of answering correctly (i.e., $p(r_{t+1} = 1|q_{t+1}, X)$) (Corbett and Anderson 1994; Piech et al. 2015; Zhang et al. 2017a). In every time step, KT updates the knowledge state of the student given the knowledge growth signal $(q_t, r_t)$.

Figure 2 shows our proposed knowledge tracing model which improves the learning and forgetting mechanism of the DVKMN (Zhang et al. 2017a).

The DKVMN model has two types of memory: key memory and value memory. The key memory $M^K_t \in \mathbb{R}^{N \times d_k}$ stores the high dimensional ($d_k$) embeddings of each concept ($N$) in each slot. Each slot of the value memory $M^V_t \in \mathbb{R}^{N \times d_v}$ represents a student’s mastery level of a concept.

The DKVMN model performs four processes: attention, read, write and optimization. Given an exercise tag $q_t$ and the student’s knowledge state $S_t$ at a time step $t$, the attention process produces an attention vector $w_t \in \mathbb{R}^N$ between $q_t$ and each $N$ latent concepts. The read process receives the $w_t$ from the attention process and outputs $p(r_t = 1|S_t, q_t)$. The write process receives $(q_t, r_t)$, and updates the value memory ($M^K_t \rightarrow M^K_{t+1}$) by adding and erasing values. In the optimization process, the loss function are calculated, and then the parameters of DKVMN are updated.

We introduce adaptive knowledge growth that utilize the current knowledge state in the write process, and we also added negative influence loss to regularize the degree of forgetting in the optimization process.

**Attention process**

The input $q_t$ is embedded to a key vector $k_t \in \mathbb{R}^{d_k}$ by multiplying an embedding matrix $A \in \mathbb{R}^{d_k \times d_k}$. $w_t$ is computed by taking the softmax of the inner product of $k_t$ and $M^K_t(i)$ as follows (Zhang et al. 2017a):

$$w_t(i) = \text{softmax}(k_t^T M^K_t(i)), \quad (1)$$

where $M^K_t(i)$ and $w_t(i)$ are key memory slot and the attention weight of the $i^{th}$ concept.

**Read process**

The read process retrieves the attended knowledge state of the student from a value matrix $M^V_t$ using $w_t$, and predicts the probability of answering $q_t$ correctly.

The read content vector $r_t \in \mathbb{R}^{d_r}$ provides the overall understanding level of the student for each concept by the concept-wise weighted sum of $M^K_t(i)$ and $w_t(i)$ (Zhang et al. 2017a) as follows:

$$r_t = \sum_{i=1}^{N} w_t(i) M^K_t(i), \quad (2)$$

where $M^K_t(i)$ means the knowledge state of the $i^{th}$ concept.

To utilize the information of $q_t$, the DKVMN concatenates $r_t$ and $k_t$ to represent the summary vector $f_t$ as follows (Zhang et al. 2017a):

$$f_t = \text{sigmoid}(W_1^T [r_t, k_t] + b_1), \quad (3)$$

where $W_1$ and $b_1$ denote the weight and the bias of the fully connected layer respectively. The probability $p_t = P(r_t = 1|S_t)$ of $q_t$ is computed from $f_t$.

$$p_t = \text{sigmoid}(W_2^T f_t + b_2), \quad (4)$$

where $W_2$ and $b_2$ denote the trainable parameters of the last fully connected layer.

**Write process**

The write process updates $M^K_t$ to $M^K_{t+1}$. To update $M^K_t$, the given knowledge growth signal $(q_t, r_t)$ is embedded to a knowledge growth vector $v_t \in \mathbb{R}^{d_v}$ by multiplying with an embedding matrix $B \in \mathbb{R}^{d_Q \times d_v}$ (Zhang et al. 2017a).

However, $v_t$ is independent of the current knowledge state of the student since $v_t$ only depends on $(q_t, r_t)$. As shown in Figure 2B, we expand $v_t$ to adaptive knowledge growth $v_{t}^{\text{adap}}$ that contains the student’s current knowledge state $S_t$. There are some candidates of $S_t$, such as the read content $r_t$ and summary vector $f_t$. We choose $f_t$ as the student’s current knowledge since $f_t$ involves the concept mastery level of the student with $k_t$ and $c_t$, and $v_{t}^{\text{adap}}$ is defined as follows:

$$v_{t}^{\text{adap}} = [v_t, S_t] = [v_t, f_t]. \quad (5)$$

Figure 3: The diagram of the proposed methods. (A) Adaptive knowledge growth to utilize the PURrent knowledge state. (B) Negative influence loss term to reduce the model-oriented forgetting.
Motivated from the operations of LSTM (Hochreiter and Schmidhuber 1997), an erase vector \( e_t \in \mathbb{R}^{ne} \) and an add vector \( a_t \in \mathbb{R}^{ad} \) are exploited to erase unnecessary information and add new information (Zhang et al. 2017a) as follows:

\[
e_t = \text{sigmoid}(E_t^T v_t^{\text{adap}} + b_e), \tag{6}
\]

\[
a_t = \tanh(D_t^T v_t^{\text{adap}} + b_a)^T, \tag{7}
\]

where \( D, E, b_e, \) and \( b_a \) are the trainable parameters of each fully connected layer.

Then \( M^v \) is updated for each concept as follows (Zhang et al. 2017a):

\[
M_{t+1}^v(i) = M_t^v(i)[1 - w_t(i)e_t] + w_t(i)a_t, \tag{8}
\]

where \( i \in [1, N] \) is an index for a concept. \( M^v \) updates adaptively because \( e_t \) and \( a_t \) depends on \( v_t^{\text{adap}} \) so that we can improve the learning mechanism of the DKVMN.

**Optimization process**

To improve the predictive performance for a given exercise, the DKVMN is trained with the cross-entropy loss \( L^{ce} \) (Zhang et al. 2017a):

\[
L^{ce} = - \sum_t (r_t \log p_t + (1 - r_t) \log (1 - p_t)). \tag{9}
\]

In this paper, we define the forgetting of \( (q_t, r_t) \) as the decrease in the probability of answering correctly as follows:

\[
P(r_{t+1} = 1|q_t = j, S_{t+1}) - P(r_t = 1|q_t = j, S_t) < 0. \tag{10}
\]

where \( S_{t+1} \) is updated from \( S_t \) given \( (q_t = i, r_t = 1) \) and \( i, j \in [1, Q] \) are exercise IDs.

In DKVMN, \( L^{ce} \) is calculated based on \( p_t = P(r_t = 1|q_t = j, S_t) \) from the read process. \( S_t \) is updated in the write process, and \( p_{t+1} = P(r_{t+1} = 1|q_{t+1} = j, S_{t+1}) \) is not computed and does not affect the loss function. That is, since the correct calculation of \( p_{t+1} \) does not directly reduce the loss, the probability of answering exercises correctly at the time \( t + 1 \) in the DKVMN model training process can be reduced, which is defined as model-oriented forgetting. As shown in Figure 3B, we add a negative influence loss term \( L^{ne} \) to reduce model-oriented forgetting, by analyzing the total prediction probability. Given \( S_t \), the total prediction probability vector \( p_t \in \mathbb{R}^Q \) is defined as follows:

\[
p_t(i) = p(r_t = 1|S_t, q_t = i), \forall i \in [1, Q]. \tag{11}
\]

\( L^{ne} \) is the squared error for the probability difference only when the positive knowledge growth signal \( (q_t = i, r_t = 1) \) has forgetting as follows:

\[
L^{ne}_{t+1} = \sum_j (p_{t+1}(j) - p_t(j))^2. \tag{12}
\]

The objective function is then \( L^{ce} + \alpha L^{ne} \) where \( \alpha \) is a hyper-parameter. If \( \alpha \) is excessively large, data-oriented forgetting as well as model-oriented forgetting are removed, reducing the predictive performance of the model. Conversely, if \( \alpha \) is too small, model-oriented forgetting with negative influence can not be removed sufficiently. It is important to find \( \alpha \) which can remove model-oriented forgetting without reducing AUC.

Computation of \( p_t \) requires \( Q \) times attention and write processes because the DKVMN can calculate only one \( p_t(i) \) for the given \( q_t = i, r_t = 1 \). To reduce the required computation, we propose \( L_{app}^{ne} \), that is approximated version of \( L^{ne} \) and perform experiments with \( L_{app}^{ne} \). The techniques and asymptotic analysis of approximation are provided in implementation detail of Supplementary.

**Proposed metric: positive update ratio (PUR)**

We propose a positive update ratio (PUR) \( \in [0, 1] \) as the metric for measuring data-oriented and model-oriented forgetting together. First, we define a positive influence \( P_l(i) \in [0, Q] \) as a metric that counts how many of the \( Q \) exercises increase \( p_{t+1} \) when \( q_t = i, r_t = 1 \):

\[
P_l(i) = \sum_j 1(p_{t+1}(j) - p_t(j)) > 0 \mid q_t = i, r_t = 1 \tag{13}
\]

where \( 1 \in [0, 1] \) denotes an indicator function. Based on this, \( PUR_t \) is defined as follows:

\[
PUR_t = \frac{1}{Q} \sum_j P_l(i)/Q. \tag{14}
\]

\( PUR_t \) depends on the time step \( t \) because the positive influence of the exercise can vary on current knowledge state. To compare PURs of different models in the same knowledge state, we define the PUR as \( PUR_0 \).

The closer the PUR is to 0, the more forgetting occurs, and vice-versa. When \( PUR = 1 \), there is no data-oriented forgetting and model-oriented forgetting. This can be interpreted

| Datasets     | # of students | # of questions | # of records | neg2pos (%) | pos2neg (%) |
|--------------|---------------|----------------|--------------|-------------|-------------|
| Synthetic-5  | 4,000         | 50             | 200,000      | 0           | 0           |
| ASSISTments2009 | 4,151       | 110            | 325,637      | 6.17±6.26   | 2.13±4.82   |
| ASSISTments2015 | 19,840      | 100            | 683,801      | 7.54±6.06   | 1.42±4.87   |
| Statics2011  | 333           | 1,223          | 189,297      | 0.17±0.39   | 0.17±0.41   |

Table 1: Data statistics. Neg2pos means the ratio of answering wrongly to the corrected exercise in the past, and pos2neg means the opposite case. The standard deviation of each case is given.
BKT model, extended to consider forgetting, showed an im-
prior study (Khajah, Lindsey, and Mozer 2016), the
thetic dataset has no forgetting pattern. According to a
neg2pos=0). Based on these facts, we can see that the syn-
does not consider forgetting. In addition, exercises do
Synthetic-5 does not have a forgetting pattern because IRT
is generated data based on IRT (Embretson and Reise 2013).

There is only model-oriented forgetting in the Synthetic-5
dataset, attributed to the characteristics of this dataset.

ASSISTment2009, ASSISTment2015, and Statics2011 are
real datasets, where a student can solve duplicate exer-
cises and pos2neg and neg2pos are calculated accordingly.
Particularly, pos2neg shows low values of 2.13%, 1.42%,
and 0.41%, respectively.

Experimental Results

Dataset

We used four widely used public benchmarks: Synthetic-5,
ASSISTments2009, ASSISTment2015, and Statics2011, the
statistics of which are reported in Table 1. If a student con-
tinues to correct the exercise and then answers correctly, it
is considered as a learning pattern and the ratio of this
pattern (positive to negative: pos2neg) can be calculated
from the dataset. Similarly, if a student continues to
miss the exercise and then answers correctly, it can be seen
as a learning pattern and the ratio of this pattern (negative
to positive: neg2pos) can be calculated.

As shown in Table 1, the Synthetic-5 dataset contains
4,000 students (the number of sequence of $q_i, r_j$), 50 ex-
ercise tags, and 200,000 records (the total number of $q_i, r_j$).
The ASSISTments2009 dataset has 4,151 students, 110 ex-
ercise tags, and 325,637 records. The ASSISTments2015
dataset contains 19,840 sequences, 100 exercise tags, and
683,801 $q_i, r_j$ pairs. The Statics2011 dataset has 333 stu-
dents, 1,223 questions, and 189,297 records.

Synthetic-5 is a dataset proposed in DKT cited, which
is generated data based on IRT (Embretson and Reise 2013).
Synthetic-5 does not have a forgetting pattern because IRT
itself does not consider forgetting. In addition, exercises do
not appear in duplicate in Synthetic-5 because it assumes
that students solve exercises 1 to 50 in turn (pos2neg=0,
neg2pos=0). Based on these facts, we can see that the syn-
thetic dataset has no forgetting pattern. According to a
previous study (Khajah, Lindsey, and Mozer 2016), the
BKT model, extended to consider forgetting, showed an
improvement in performance for all datasets apart from the

| Knowledge growth | Synthetic-5 | ASSISTments2009 | ASSISTments2015 | Statics2011 |
|------------------|-------------|-----------------|----------------|-------------|
| $\alpha$         | AUC         | PUR             | AUC            | PUR         |
| 0.0              | 0.8308±0.0025 | 0.7483          | 0.8222±0.0005 | 0.8273       | 0.7273±0.0005 | 0.8386 | 0.8199±0.0003 | 0.5978 |
| 0.001            | 0.8323±0.0004 | 0.9840          | 0.8224±0.0007 | 0.9840       | 0.7278±0.0002 | 0.9349 | 0.8234±0.0003 | 0.9214 |
| 0.01             | 0.8321±0.0015 | 0.9986          | 0.8219±0.0008 | 0.9857       | 0.7279±0.0005 | 0.9856 | 0.8231±0.0005 | 0.9918 |
| 0.1              | 0.8370±0.0017 | 0.9990          | 0.8211±0.0003 | 0.9914       | 0.7272±0.0003 | 0.9927 | 0.8222±0.0007 | 0.9934 |
| 1                | 0.5467±0.0304 | 0.9130          | 0.5437±0.0271 | 0.9470       | 0.5563±0.0308 | 0.9557 | 0.5028±0.0064 | 0.9994 |
| 10               | 0.5501±0.0130 | 0.9541          | 0.5529±0.0097 | 0.8092       | 0.5342±0.0282 | 0.9497 | 0.5053±0.0111 | 0.9825 |

| $\nu$            | AUC         | PUR             | AUC            | PUR         |
|------------------|-------------|-----------------|----------------|-------------|
| 0.0              | 0.8230±0.0011 | 0.6317          | 0.8245±0.0007 | 0.9026       | 0.7284±0.0003 | 0.9334 | 0.8304±0.0002 | 0.6153 |
| 0.001            | 0.7645±0.1259 | 0.9798          | 0.8245±0.0007 | 0.9733       | 0.7289±0.0003 | 0.9109 | 0.8318±0.0002 | 0.9197 |
| 0.01             | 0.7699±0.1223 | 0.9874          | 0.8241±0.0005 | 0.9733       | 0.7289±0.0005 | 0.9473 | 0.8320±0.0002 | 0.9848 |
| 0.1              | 0.5998±0.1238 | 0.8708          | 0.7677±0.1169 | 0.9733       | 0.7292±0.0004 | 0.9465 | 0.8319±0.0003 | 0.9939 |
| 1                | 0.5451±0.0463 | 0.9512          | 0.5787±0.0977 | 0.9567       | 0.7059±0.0724 | 0.9291 | 0.872±0.1302  | 0.9568 |
| 10               | 0.5429±0.0155 | 0.9594          | 0.5422±0.0255 | 0.9709       | 0.5415±0.0285 | 0.9676 | 0.5049±0.0095 | 0.9795 |

Adaptive knowledge growth

According to Table 2, the predictive performances (AUC) of
all real datasets, except Synthetic-5, improve when $\nu_{adap}$
is used instead of $\nu$ for knowledge growth. In contrast, when
$\nu_{adap}$ is used instead of $\nu$ for the Synthetic-5 dataset, the
performance decreases from 0.8308±0.0025 to 0.8230±
0.0011. As previously mentioned, the Synthetic-5 dataset
is based on the IRT theory and assumes that the knowledge
state of the student is fixed. Therefore, it can be considered
that $\nu_{adap}$ has an adverse effect on prediction performance.
For real datasets, the AUC values are increased by taking
the current knowledge state into account when calculating
knowledge growth.

The adaptive knowledge growth improves the PUR on all
real datasets except Synthetic-5, although it is not aimed to
reduce forgetting. Therefore, it can be seen that updating
according to the current state also prevents unnecessary forget-
ting.

Negative influence loss

There is only model-oriented forgetting in the Synthetic-5
dataset, because the Synthetic-5 dataset has no data-oriented
forgetting. As shown in Table 2, when the knowledge growth
is $\nu$ for the Synthetic-5, the AUC increases from 0.8303 to
0.8320, and the PUR increases from 0.7483 to 0.9990 for
$\alpha = 0, 0.1$, respectively. The PUR increases as $L^{ne}_f$
reduces the overestimated forgetting (model-oriented forget-
ting), and then predictive performance also increases. When
$\alpha$ is larger than 1, we observe that both the AUC and the
PUR decrease. The predictive performance decreases because
the optimization process focuses on minimizing $L^{ne}_f$.
rather than $L^{nc}$. The PUR decreases due to the large step size from $\alpha$ that prevents a stable convergence.

We define optimal PUR, measured when DKVMN has lowest degree of forgetting, while preserving the predictive performance. According to the extensive experiments, we find the optimal PUR (bold in Table 2) for each real dataset. Hence, $L^{nc}$ with a proper $\alpha$ can reduce model-oriented forgetting while modeling data-oriented forgetting. The optimal PURs are 0.98 0.99, which indicate the data-oriented forgetting is small. We can deduce that optimal PURs are high from the small pos2neg ratio of all three real datasets.

Average prediction probability

To observe how DKVMN models estimate the knowledge state of the student, we calculate average prediction probability $\sum p_t / Q$. At time step $t$, the student is given $q_t$ sequentially, and always answer correctly ($r_t = 1$) for the $q_t$. Finally at time step $Q$ (the number of total exercise), the student answers correctly all exercises without duplication.

As shown in Figure 4, all proposed models reach higher average prediction probability than original DKVMN. All proposed models outperform the original DKVMN in terms of AUC, and therefore the proposed models estimate student knowledge state more accurately. In addition, models with $L^{nc}$ show smaller degree of fluctuation than models without $L^{nc}$ that means $L^{nc}$ regularizes forgetting of the KT model effectively.

Discussion

The decrease in the probability of answering correctly due to model-oriented forgetting reduces the reliability that the DKVMN works like a real student. The DKVMN with model-oriented forgetting then cannot be used in ITS. In particular, it would be difficult to apply reinforcement learning (RL) [Sutton and Barto 1998; Zhao et al. 2017; Choi et al. 2018] to recommend appropriate contents to the current knowledge state when model-oriented forgetting prevails. In RL, it is crucial to define suitable rewards; however, a reward defined in any way based on the KT model with model-oriented forgetting may not function properly. Since the benefits from a proper exercise recommendation can be enormous, reducing model-oriented forgetting is important.

The optimal PUR is the PUR measured when the model-oriented forgetting is lowest while preserving the data-oriented forgetting (predictive performance). We find the optimal PUR by experiments, but it would be possible to calculate the optimal PUR theoretically. Furthermore, we believe that there is a correlation between the optimal PUR and the pos2neg ratio. Due to the complexity of the neural network, this theoretical approach remains as challenging future work.

Generally, the forgetting is decrease in concept mastery level. We define the forgetting as decrease in probability, not concept mastery level. Negative influence loss term $L^{nc}$ focuses on the change in probability of answering correctly. This can be considered as indirect regularization of the change in mastery level of concept through probability shifts. DKVMN assumes that the value memory represents the knowledge state of concepts. Regularizing the change in concept mastery level on the memory layer would be possible and might provide the insight of the internal operation.

We focus on the situation that the students answer correctly and do not consider wrong response. When the response of the student is wrong, the decrease in probability of answering correctly can be interpreted as tracing the knowledge state or forgetting. Factoring the result of answering wrongly to tracing and forgetting can be challenging research topic.

Conclusion

Our proposed knowledge tracing model introduces adaptive knowledge growth and a negative influence loss term to improve the learning and forgetting process of the original DKVMN, respectively. We have also proposed a new metric PUR that can be used to evaluate the forgetting of the KT model. We believe that the proposed approaches can closely resemble a real student’s learning and forgetting process and is more reliable than existing models.
References

[Abdi, Khoarsavi, and Sadiq] Abdi, S.; Khoarsavi, H.; and Sadiq, S. Predicting student performance: The case of combining knowledge tracing and collaborative filtering.

[Atkinson and Shiffrin 1968] Atkinson, R. C., and Shiffrin, R. M. 1968. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2. Elsevier. 89–195.

[Brod, Werkle-Bergner, and Shing 2013] Brod, G.; Werkle-Bergner, M.; and Shing, Y. L. 2013. The influence of prior knowledge on memory: a developmental cognitive neuroscience perspective. *Frontiers in behavioral neuroscience* 7:139.

[Brusilovsky, Schwarz, and Weber 1996] Brusilovsky, P.; Schwarz, E.; and Weber, G. 1996. Elm-art: An intelligent tutoring system on world wide web. In *International conference on intelligent tutoring systems*, 261–269. Springer.

[Burns et al. 2014] Burns, H.; Luckhardt, C. A.; Parlett, J. W.; and Redfield, C. L. 2014. Intelligent tutoring systems: Evolutions in design. Psychology Press.

[Cen, Koedinger, and Junker 2006] Cen, H.; Koedinger, K.; and Junker, B. 2006. Learning factors analysis—a general method for cognitive model evaluation and improvement. In *International conference on Intelligent Tutoring Systems*, 164–175. Springer.

[Choi et al. 2018] Choi, S.; Ha, H.; Hwang, U.; Kim, C.; Ha, J.-W.; and Yoon, S. 2018. Reinforcement learning based recommender system using biclustering technique. *arXiv preprint arXiv:1801.05532*.

[Corbett and Anderson 1994] Corbett, A. T., and Anderson, J. R. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4(4):253–278.

[Ebbinghaus 2013] Ebbinghaus, H. 2013. Memory: A contribution to experimental psychology. *Annals of neurosciences* 20(4):155.

[Ebbinghaus, Schwarz, and Weber 1996] Ebbinghaus, H.; Schwarz, E.; and Weber, G. 1996. Elm-art: An intelligent tutoring system on world wide web. In *International conference on intelligent tutoring systems*, 261–269. Springer.

[Hinton and Salakhutdinov 2006] Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science* 313(5786):504–507.

[Hochreiter and Schmidhuber 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

[Käser et al. 2017] Käser, T.; Klingler, S.; Schwing, A. G.; and Gross, M. 2017. Dynamic bayesian networks for student modeling. *IEEE Transactions on Learning Technologies* 10(4):450–462.

[Khajah, Lindsey, and Mozer 2016] Khajah, M.; Lindsey, R. V.; and Mozer, M. C. 2016. How deep is knowledge tracing? *arXiv preprint arXiv:1604.02416*.

[LeCun, Bengio, and Hinton 2015] LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature* 521(7553):436.

[Montero et al. 2018] Montero, S.; Arora, A.; Kelly, S.; Milne, B.; and Mozer, M. 2018. Does deep knowledge tracing model interactions among skills? In *Proceedings of the Eleventh International Conference on Educational Data Mining*. Educational Data Mining Society Press.

[Pavlík Jr, Cen, and Koedinger 2009] Pavlík Jr., P. I.; Cen, H.; and Koedinger, K. R. 2009. Performance factors analysis—a new alternative to knowledge tracing. *Online Submission*.

[Piech et al. 2015] Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L. J.; and Sohl-Dickstein, J. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, 505–513.

[Reddy, Labutov, and Joachims 2016] Reddy, S.; Labutov, I.; and Joachims, T. 2016. Latent skill embedding for personalized lesson sequence recommendation. *arXiv preprint arXiv:1602.07029*.

[Rohrer 2009] Rohrer, D. 2009. The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education* 4–17.

[Santoro et al. 2016] Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*.

[Sonnhammer et al. 1998] Sonnhammer, E. L.; Von Heijne, G.; Krogh, A.; et al. 1998. A hidden markov model for predicting transmembrane helices in protein sequences. In *Ismb*, volume 6, 175–182.

[Su et al. 2018] Su, Y.; Liu, Q.; Liu, Q.; Huang, Z.; Yin, Y.; Chen, E.; Ding, C. H.; Wei, S.; and Hu, G. 2018. Exercise-enhanced sequential modeling for student performance prediction. In *AAAI*.

[Sutton and Barto 1998] Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

[Wandersee, Mintzes, and Novak 1994] Wandersee, J. H.; Mintzes, J. J.; and Novak, J. D. 1994. Research on alternative conceptions in science. *Handbook of research on science teaching and learning* 177:210.

[Wilson et al. 2016] Wilson, K. H.; Karklin, Y.; Han, B.; and Ekanadham, C. 2016. Back to the basics: Bayesian extensions of irt outperform neural networks for proficiency estimation. *arXiv preprint arXiv:1604.02336*.

[Zhang et al. 2017a] Zhang, J.; Shi, X.; King, I.; and Yeung, D.-Y. 2017a. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web*, 765–774. International World Wide Web Conferences Steering Committee.

[Zhang et al. 2017b] Zhang, L.; Xiong, X.; Zhao, S.; Botelho, A.; and Heffernan, N. T. 2017b. Incorporating rich features into deep knowledge tracing. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, 169–172. ACM.

[Zhao et al. 2017] Zhao, X.; Zhang, L.; Ding, Z.; Yin, D.; Zhao, Y.; and Tang, J. 2017. Deep reinforcement learning with feature mixing. *arXiv preprint arXiv:1703.02703*.

[Zhao et al. 2018] Zhao, X.; Zhang, L.; and Mozer, M. 2018. Does deep knowledge tracing model interactions among skills? In *Proceedings of the Eighth International Conference on Educational Data Mining*. Educational Data Mining Society Press.
learning for list-wise recommendations. *arXiv preprint arXiv:1801.00209.*