Cold Start Knowledge Tracing with Attentive Neural Turing Machine

Jinjin Zhao
Amazon.com
Seattle, USA
jinjzhao@amazon.com

Shreyansh Bhatt
Amazon.com
Seattle, USA
bhattshr@amazon.com

Candace Thille
Amazon.com
Seattle, USA
cthille@amazon.com

Neelesh Gattani
Amazon.com
Seattle, USA
neeleshg@amazon.com

Dawn Zimmaro
Amazon.com
Austin, USA
dzimmaro@amazon.com

ABSTRACT

Deep learning based knowledge tracing approaches achieve high accuracy in mastery prediction with pattern extraction on a large learning behavior data set. However, when there is little training data available, these approaches either fail to extract the key patterns or result in over fitting. Ideally, we aim to provide a similar learning experience to both the first group of learners, who interact with a new course or a new activity with little learning behavior data to provide personalized guidance, and the learners who interact with the course later. We propose a novel architecture, Attentive Neural Turing Machine (ANTM), to solve the cold start knowledge tracing problem. The proposed ANTM comprises an attentive controller module and differential reading and writing processes with extra memory bank. Accuracy (ACC) and Area Under Curve (AUC) measures are used for model performance comparison. Results show the proposed approach can learn fast and generalize well to unseen data. It achieves around 95% ACC trained with only 3 learners, while conventional deep learning based approaches achieve only 65% ACC with over prediction issues.

Author Keywords

Knowledge tracing; Neural turing machine; Attention mechanism.

CCS Concepts

•**Human-centered computing** → *User models; User studies;*
•**Applied computing** → **Interactive learning environments;**
•**Computer systems organization** → **Neural networks;**

INTRODUCTION AND RELATED WORK

Some online learning systems provide decision-making support to learners with mastery predictions and guidance on selecting the next set of learning activities. In these systems, mastery prediction information is critical, which can help learners understand their current knowledge state at any given time. The usefulness of the guidance on selecting the next set of learning activities heavily relies on the accuracy of knowledge state estimation. With the help of machine learning, the learner’s knowledge state can be estimated and predicted by modeling the relation between sequential learning activities and the correctness of the learning attempts. The estimated knowledge state is then used to infer knowledge mastery. Cold start knowledge tracing is a scenario where there is insufficient observable data from learners to train a model to predict knowledge state accurately. For example, when a new course is launched or a new activity is introduced, there is no observable data until learners start to interact with the system. Without a solution to accurately predict knowledge mastery in the cold start setting, the first group of learners who take the course won’t be able to get the guaranteed learning experiences. This work is focused on providing a novel solution to solve the cold start problem in knowledge tracing.

The state-of-the-art technique in achieving high mastery prediction accuracy is deep learning based approaches. DKT proposed by Piech et al. [7] and its variants proposed by Wang et al. [12] and Yeung et al. [14] have demonstrated the strong capability of neural network, especially LSTM, in modeling the sequential patterns and performing predictions. Memory augmented neural network proposed by Graves et al. [4] and Weston et al. [13] and its variants by Zhang et al. [15] and Abdelrahman et al. [1] have also demonstrated its capability in memorization and generalization in knowledge tracing. Learning factor models proposed by Cen et al. [3] and its variants by Pavlik et al. [6] have demonstrated the effectiveness in knowledge state prediction with predefined statistical models with sufficient data. However, there is little research work to solve the cold start problem, which is to predict knowledge mastery accurately and reliably for the first group of learners who interact with the system.

Meta learning is attracting more attention recently in Computer Vision and NLP applications. Metric based approach

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 the author(s) 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.
L@S ’20, August 12–14, 2020, Virtual Event, USA.
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7951-9/20/08 ...$15.00.
http://dx.doi.org/10.1145/3386527.3406741

from Koch et al. [5] and Vinyals et al. [11], optimization based approach proposed by Ravi et al. [8], and model based approaches Neural Turing Machine (NTM) from Grave et al. [4] and Memory Networks from Weston et al. [13] are three major research directions in meta learning. Metric based approaches are a set of algorithms similar to nearest neighbor algorithms, which are suitable for multiclass classification problems. Optimization based approaches try to optimize the gradient decent procedure to enable the optimization procedure to find the global optimum. Model based approaches try to use different surrogate models with different neural network architectures to approximate the internal mechanisms to achieve better performance. For knowledge tracing, which can be formulated as a sequence prediction problem, a model based approach is the most suitable solution out of the three sets of solutions.

Within model based meta learning branch, NTM and Memory Networks are two similar solutions, both of which aim to improve performance by coupling conventional deep learning approaches to external memories. NTM and its variant (MANN) proposed by Santoro et al. [9] has shown great performance improvement in solving image classification and regression problems in cold start setting. The existing research work, using MANN and its variants to perform knowledge tracing task with large dataset, only achieves slight (less than 2%) performance improvement compared to LSTM based approaches. Based on the existing MANN work in knowledge tracing, we propose a novel solution with attention based controller and NTM mechanism for solving the cold start problems in knowledge tracing. We conduct experiments on mastery prediction by predicting the knowledge state to demonstrate its effectiveness. The knowledge state is proposed in our preliminary research result [2]. In the referred paper, the proposed knowledge state representation is the probability of mastery with the ratio of the number of successful future attempts to the total number of attempts. We also argued in the referred paper that the proposed knowledge state representation is a better approximation of the probability of knowledge mastery.

Main contributions of the work are described as follows.

- A novel Attentive Neural Turing Machine framework (ANTM) for knowledge state estimation and mastery prediction.
- More than 15% gain in prediction accuracy over state-of-the-art deep learning based approaches in knowledge state estimation.

APPROACH

The overall framework consists of three major components as shown in Fig 1, external memory bank for coupling the conventional neural network, controller neural network feeding forward step, and memory retrieval and updating with differential operations. The external memory bank stores the sequential information extracted from learning activities, which can be selectively accessed with read and write operations. The read operation, with its content based attention mechanism, interacts with memory to selectively choose the useful information as the auxiliary information. The controller takes the input (as well as the information retrieved from the memory bank with read operation), feed forward its own multi-headed attention neural network, conduct loss computation, and back propagation. The writing operation, with its location based addressing attention mechanism, updates the memory with differential adding and erasing operations. The updated memory is then available for the next read operation in the next prediction step.

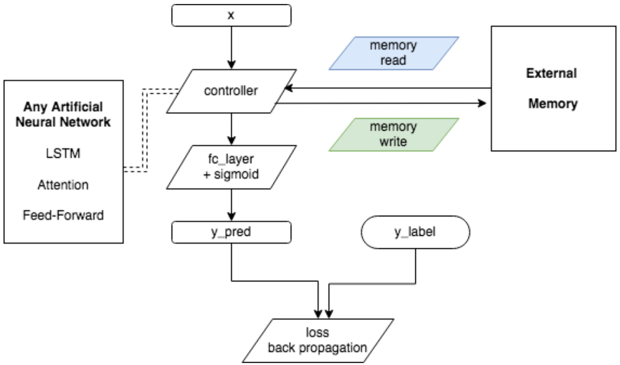


Figure 1: ANTM Framework Overview for cold start knowledge tracing

Memory retrieval with content based attention mechanism

In the proposed approach, memory is retrieved with a content based addressing mechanism. Content based addressing is essentially a similarity measurement step between the controller output vector C_t and the existing memory vectors M_t. Attention weight in reading is generated as in Eq.1,

$$w_t^r(i) = \frac{exp(\beta \cdot sim(C_t, M_t(i)))}{\sum_j (exp(\beta \cdot sim(C_t, M_t(j))))} \tag{1}$$

where β is the strength variable, which can amplify or attenuate the precision of focus. The similarity measurement is defined as the cosine similarity.

Multi-headed attention based controller

The controller can be designed with any artificial neural network. We propose to use multi-headed attention layer to model the relation between input and output sequences. The comparison of different neural network designs will be illustrated in result session.

We use the scaled dot-product attention [10] as described in equation 2.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{n}}\right) \tag{2}$$

where Q, K, and V are query, key, and value matrices and n is the total number of dimensions. Attention is applied on the input sequence I to compute a weighted sum for each i_t based on $i_1, i_2, \ldots, i_t$, where t is the current time stamp. As the output, the model learns a weight vector associated with each attempt and the learned weight vector is used to compute the output prediction for i_t. We use MSE loss function to train the network as in Eq.3,

$$L = \sum_b \sum_{k=1}^{n} \sum_{t=1}^{T} MSE(a_t^k, gt_t^k) \tag{3}$$

where a_t^k is the attempt correctness prediction, gt_t^k is the label information, b is the batch size, k is the sequence length, and t is the time stamp.

Memory updating with attention mechanism

In the proposed approach, memory is updated with location based addressing mechanism. The location based addressing mechanism is designed with multiple steps to facilitate iteration across memory locations as well as random access jumps. The first step is to conduct an interpolation between the previous writing weight vector w_{t-1} and the content weighting vector produced by the content addressing mechanism w_t^c at the current time stamp using interpolation gate w_t^g.

$$w_t^g = g_t * w_t^c + (1 - g_t) * w_{t-1} \tag{4}$$

After interpolation, a shifting weighting s_t is applied to the gated weighting w_t^g with a circular convolution to adjust the memory, yielding to w_t^{ro}.

$$w_t^{ro} = \sum_{j=0}^{N-1} w_t^g(j)\, s_t(i-j) \tag{5}$$

where N is the memory length. At the last step, a sharpen operation is conducted to get the final weighting w_t,

$$w_t(i) = \frac{w_t^{ro}(i)}{\sum_j w_t^{ro}(j)} \tag{6}$$

where the sharpen is a normalization operation on w_t^{ro}.

EXPERIMENTS AND RESULTS

In this section, we describe datasets, experiment setup, evaluation measures, and results analysis.

Datasets

course 1: This dataset is from our online open navigation learning system. We use the first group of learners' interaction data to conduct the experiments. It has 13 students and 265 interactions for 6 skills.

course 2: This is another dataset from our online open navigation learning system. We also use the first group of learners' interaction data from two specific Knowledge Components (KC) to conduct the experiments. The first KC has interaction data with 15 students and 324 interactions and the second KC has 15 students and 632 interactions.

Evaluation Measures and Experiment Set up

We evaluate the proposed knowledge tracing framework based on its ability to predict the knowledge state of a learner. We use accuracy (ACC) and Area Under Curve (AUC) measures for the evaluation.

We conduct experiments on three scenarios where there are a) few learners, b) short learning activity sequences, and c) few learners with short learning sequences to demonstrate the effectiveness of ANTM. We also conduct experiments on single KC with different question difficulty design and observed learning behaviors. Knowledge state representation and estimation is defined as in Eq.7.

$$KS_t^k = \sum_{i=t+1}^{n} success_i^k / \sum_{i=t+1}^{n} total_i^k \tag{7}$$

where the averaged attempt correctness from similar learners who have achieved mastery is used to approximate the knowledge state for the current learner. We conduct experiments to compare with LSTM based conventional neural network on prediction accuracy.

Results

ANTM with few learners and short learning sequences

In the cold start setting, we want to achieve high prediction accuracy using as few learners' interactions as possible to train the model and yet generalize to new unseen samples. In Scenario1 in Table 1, we conducted experiments with different learner numbers ranging from 2 to 13 to compare the prediction performance between LSTM and ANTM. As results show, after 500 epochs of training for both, ANTM achieves 95% AUC and 93% ACC with only two learners. However, LSTM achieves 95% AUC and 67% ACC, where the over predicting (predicting most of the data into one class) happens. It illustrates that ANTM is performing better in memorizing and extracting the learning pattern even with few observations and unbalanced dataset. As learner number increases gradually to 13, ANTM continues performing at 95% AUC and 92% ACC score, while LSTM is performing at around 82% AUC and 73% ACC where the over predicting issue still exists even with more learners' data for pattern extraction. We also wanted to achieve high prediction accuracy with short learning sequences and yet generalize to longer sequences. In Scenario2 in Table 1, the result shows that with only 10% of the learning sequences (5 to 8 interactions), ANTM can achieve 87% AUC and 80% ACC while LSTM can only achieve 67% AUC and 63% ACC. It illustrates that with few learners' short and sparse interaction sequences, ANTM is outperforming in extracting the latent patterns and also be able to generalize to future unseen patterns. As Learning sequence length increases to 100% (around 80 interactions), ANTM continues outperforming in both AUC and ACC by 10% and 14%, respectively. In Scenario3 in Table 1, with 2 (10%) learners and 5 to 8 (10%) learning sequences, ANTM outperforms by 30% and 17% in AUC and ACC, respectively. With more learners and longer sequences, ANTM outperforms in both AUC and ACC by 10% and 14%, respectively. In conclusion, with limited number of learners interacting with limited number of practices, ANTM is able to extract latent learning patterns at high precision and generalize to unseen interactions with high prediction accuracy.

ANTM with different KCs

In order to understand the complexity of extracting learning patterns for both ANTM and LSTM, we conduct another experiment to decouple the complexity and only train the models with one set of learning patterns instead of a mixed group of patterns. We used the course 2 dataset for this experiment, where we chose two different KCs and observed training and test performance for each of the KCs. KC1 is a relatively easier skill where learners make fewer attempts and achieve mastery

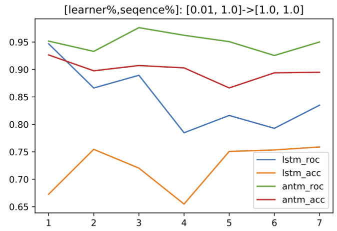
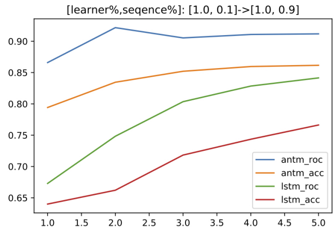
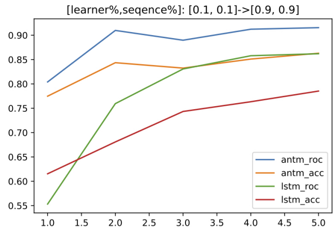
Scenario1	Scenario2	Scenario3
		

Table 1: Knowledge state prediction result comparison between ANTM and LSTM

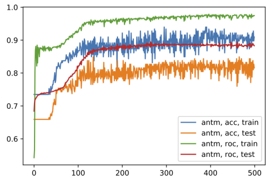
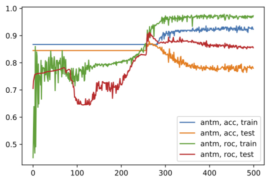
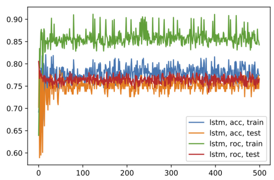
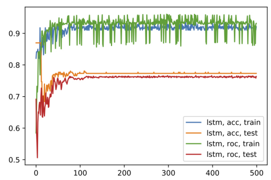
antm-KC1	antm-KC2	lstm-KC1	lstm-KC2
			

Table 2: Knowledge state prediction across different KCs with ANTM and LSTM

with similar patterns. For KC2, learners have more diverse behaviors. As the result shows in Table 2, for KC1, ANTM achieves 95% AUC and 85% ACC as an averaged measurement from training and test sets, where LSTM achieves 80% AUC and 76% ACC. Both models are able to achieve high accuracy in first 100 epochs. For KC2, ANTM achieves 90% AUC and 80% ACC for test set and LSTM achieves around 75% AUC and ACC as training with 500 epochs. Results show ANTM is able to jump out of the local minimum after around 270 epochs to continue the optimization towards its global optimum. However, LSTM ends up being stuck in its local optimum after around 80 epochs, which limits its accuracy.

CONCLUSIONS

In this paper, we proposed a novel architecture, Attentive Neural Turing Machine, to predict knowledge mastery in the cold start setting. The experimental result shows significant accuracy improvement compared to deep learning based approaches.

REFERENCES

[1] Ghodai Abdelrahman and Qing Wang. 2019. Knowledge Tracing with Sequential Key-Value Memory Networks. (2019).

[2] Shreyansh Bhatt, Jinjin Zhao, Candace Thille, Dawn Zimmaro, and Neelesh Gattani. 2020. A novel approach for knowledge state representation and prediction. In *Proceedings of the Seventh Annual ACM Conference on Learning at Scale*. ACM.

[3] Hao Cen, Kenneth Koedinger, and Brian Junker. 2006. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems*. Springer, 164–175.

[4] Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401* (2014).

[5] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, Vol. 2. Lille.

[6] Philip I Pavlik Jr, Hao Cen, and Kenneth R Koedinger. 2009. Performance Factors Analysis–A New Alternative to Knowledge Tracing. *Online Submission* (2009).

[7] 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.

[8] Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. (2016).

[9] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*. 1842–1850.

[10] 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.

[11] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, and others. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*. 3630–3638.

[12] Zhiwei Wang, Xiaoqin Feng, Jiliang Tang, Gale Yan Huang, and Zitao Liu. 2019. Deep Knowledge Tracing with Side Information. In *International Conference on Artificial Intelligence in Education*. Springer, 303–308.

[13] Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916* (2014).

[14] Chun-Kit Yeung and Dit-Yan Yeung. 2018. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. ACM, 5.

[15] 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.