ConLearn: Contextual-knowledge-aware Concept Prerequisite Relation Learning with Graph Neural Network

Hao Sun* Yuntao Li† Yan Zhang‡

Abstract

Prerequisite relations among concepts are important for a wide range of educational applications, such as intelligent tutoring and curriculum planning. However, concept prerequisite relation learning is not trivial due to the sparsity of prerequisite relations. In this paper, we propose a contextual-knowledge-aware concept prerequisite relation learning approach called **ConLearn**. Four unique properties of the proposed approach are: (1) It transfers knowledge from large language model BERT to improve contextual representations of concepts; (2) It captures concept prerequisite transition patterns by applying graph neural network on concept prerequisite graph; (3) It is equipped with self-attention mechanism to fuse information from related concepts for target concept prerequisite relation classification; (4) No handcrafted features are used in our model, which makes our model easy to implement in downstream applications. Extensive experiments on three representative datasets demonstrate that our approach significantly outperforms the state-of-the-art methods.

1 Introduction

With the growth of available educational materials and the requirement of self-regulated learning, there is a rising need to organize knowledge in a reasonable order. Prerequisite relations among concepts are essentially considered as the dependency among knowledge concepts. It is crucial for people to learn, organize, apply, and generate knowledge [20]. Consider the running example shown in Figure 1, if someone wants to learn the knowledge about *Hidden Markov Model*, he should learn *Maximum Likelihood* first. Consequently, *Maximum Likelihood* is a prerequisite of *Hidden Markov Model*. Once this kind of prerequisite relation among concepts is learned, these relations can be used for a wide range of downstream educational tasks such as intelligent tutoring [3, 31], curriculum planning [17, 1] and automatic sequencing of learning materials [2].

Due to the importance of the investigated problem, lots of efforts have been made to extract prerequisite relations among concepts from textbooks [32, 15], MOOCs (Massive Open Online Courses) [23], courses [14, 17, 16, 9, 26] and scientific papers [6]. The earliest work focuses on utilizing handcrafted features to detect the prerequisite relations among concepts. For example, [23] proposes contextual, structural, semantic features, and [15] proposes graph-based features and text-based features. To avoid the tedious process of feature engineering, [16, 18] propose objective functions based on several observations and address the problem in an optimization way. Recently, more researchers turn to deep learning methods for concept prerequisite relation learning. For example, [9] applies variational graph autoencoders to learn concept prerequisite relations from courses. [26] develops a supervised learning approach called PREREQ while [7] proposes a weak supervision framework called CPRL.

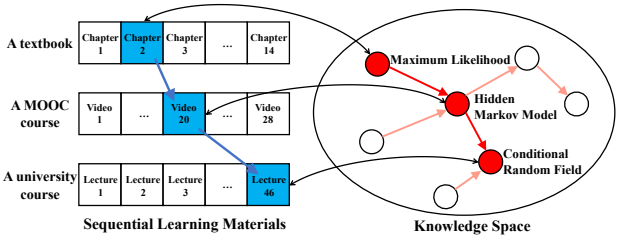

Figure 1: An example of concept prerequisite relations.

Despite the inspiring results of the deep learning methods mentioned above, concept prerequisite relation learning is still challenging for the following reasons. Firstly, previous work [16] reveals that concept prerequisite relations are extremely sparse, which means the number of prerequisite relations is much smaller than the total number of concept pairs. This can pose a challenge to existing deep learning methods that require a large amount of training data to achieve good representations of concepts. Secondly, prerequisite relations among concepts exhibit some common prerequisite transition patterns. For example, if *Maximum Likelihood* is a prerequisite of *Hidden Markov Model* and *Hidden Markov Model* is a prerequisite of *Conditional Random Field*, then *Maximum Likelihood* should be the prerequisite of *Conditional Random Field*. Existing methods

*Peking University, sunhao@stu.pku.edu.cn
†Peking University, li.yt@pku.edu.cn
‡Peking University, zhyzhy001@pku.edu.cn

cannot model this kind of multi-hop relation. Thirdly, existing methods are mainly designed for domains with well-structured learning materials such as textbooks, which limits the application scenario of these methods. Besides, the external features used in these methods bring about too much preprocessing work, which degrades the effectiveness of these methods.

To address the challenges mentioned above, we propose **ConLearn**, a contextual-knowledge-aware approach to learn concept prerequisite relations from sparse and unstructured educational data. Firstly, to deal with the prerequisite relation sparsity problem, we fine-tune the pre-trained language model BERT [4] using domain corpus, essentially achieving transfer learning from language to knowledge, to improve contextual representations of concepts. Based on our observation that related concepts have similar prerequisite relations, we utilize self-attention network to incorporate information from related concepts for target concept prerequisite relation classification. Secondly, we construct concept prerequisite graph using available concept prerequisite pairs. Based on the graph, gated graph neural network is able to capture common prerequisite transition patterns among concepts. Moreover, no handcrafted features are used in the model, which makes our model easy to implement in downstream applications.

Our contributions can be summarized as follows:

- We propose ConLearn, a contextual-knowledge-aware concept prerequisite relation learning approach to learn prerequisite relations from sparse and unstructured educational data. ConLearn takes the advantage of both pre-trained language model BERT and GNN to enhance model's performance on the concept prerequisite relation classification task.
- We empirically prove that a proper knowledge transferring from pre-trained language model BERT is able to enrich contextual concept representations, which can further greatly improve the performance of concept prerequisite relation classification.
- We conduct extensive experiments on three real-world datasets from different domains. Our proposed ConLearn achieves new state-of-the-art performance on all datasets, with an absolute improvement of more than 0.1 on average in terms of F-score compared with previous state-of-the-art methods.

2 Preliminaries

In this section, we aim to give some necessary notations and then formulate the problem of concept prerequisite relation learning.

For convenience, we use the following notations:

- Educational data $D = \{o_1, o_2, ..., o_M\}$ is composed by M learning objects, where o_i is i-th learning object in D and is presented as a document. The document can be the text from a book chapter, a slide, or the speech script of a MOOC video.
- Concepts $C = \{c_1, c_2, ..., c_N\}$ is a set of concepts in D. Concept prerequisite pairs $P = \{<c_i, c_j> \mid i \rightarrow j\}$ is a set of prerequisite relations among concepts, where $i \rightarrow j$ represents concept c_i is a prerequisite of concept c_j.

The problem could be formally defined as: given an educational data D, its concept set C, and its corresponding concept prerequisite pairs P, the goal is to learn a function $f : C^2 \rightarrow \{0, 1\}$ that maps a concept pair $<c_i, c_j>$ to a binary class that indicates whether concept c_i is a prerequisite of concept c_j.

3 Related Work

3.1 Concept Prerequisite Relation Learning. Learning prerequisite relations among concepts has recently attracted much attention. Existing works can be classified into three categories: feature-based methods, recovery-based methods and deep learning methods. For feature-based methods, several hand-crafted features are proposed to explicitly measure the prerequisite relations among concepts. For example, [23] proposes contextual, structural and semantic features, and [15] proposes graph-based features and text-based features. In contrast, recovery-based methods do not need to extract features. Instead, they design objective functions based on some observations and recover the prerequisite relations in an optimization way. Recently, with the popularity of deep learning, more researchers turn to use deep learning methods to deal with this problem. For example, [9] applies variational graph autoencoders to learn concept prerequisite relations from courses. [26] develops a supervised learning approach called PREREQ while [7] proposes a weak supervision framework called CPRL. However, these methods are mainly designed for domains with well-structured learning materials such as textbooks, which limits the application scenarios of these methods. In this paper, we propose a contextual-knowledge-aware method ConLearn that has no requirement for the structure of learning materials, thus broadening the application scenarios of our method.

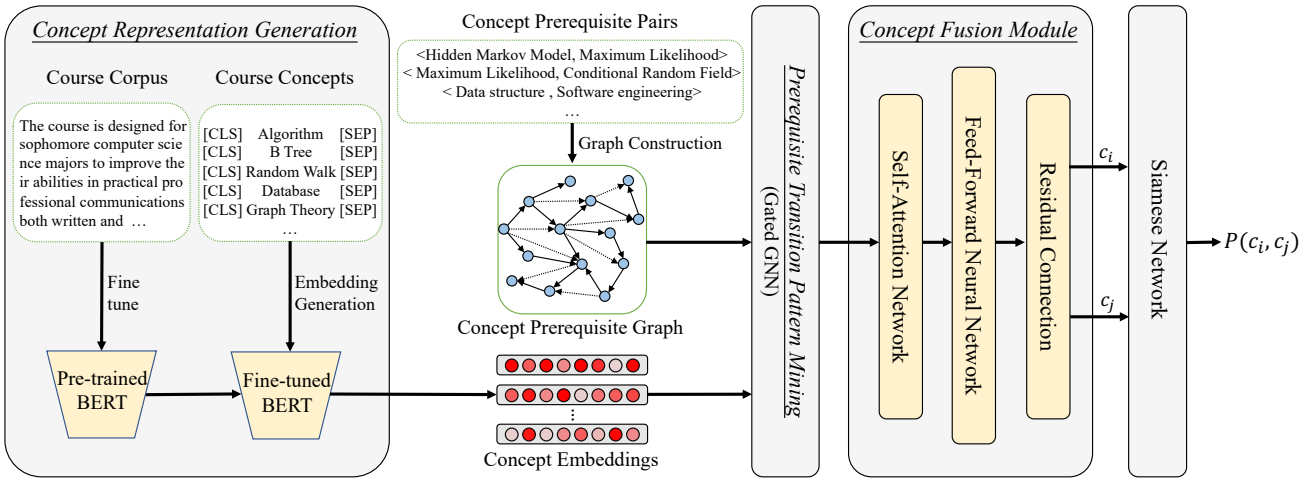

Figure 2: Model architecture of our ConLearn, where the dashed lines in the concept prerequisite graph refer to the synthetic edges. We first fine-tune the pre-trained BERT model and utilize it to precompute concept embeddings, then we construct concept prerequisite graph from concept prerequisite pairs and update concept representations using Gated GNN. We apply self-attention network to fuse information from related concepts. The final concept representations are fed into Siamese network for concept prerequisite relation classification.

3.2 Neural Network on Graphs. Nowadays, neural network has been applied on graph structured data such as social networks and knowledge bases. Extending the word2vec [22], an unsupervised algorithm DeepWalk [25] is proposed to learn representations of graph nodes based on random walk. The classical neural network CNN and RNN are also employed on graph structured data. For example, [5] designs a convolutional neural network that operates directly on graphs of arbitrary sizes and shapes. A scalable approach [8] chooses the convolutional architecture via a localized approximation of spectral graph convolutions, which is an efficient variant and can operate on graphs directly as well. However, these methods can only be implemented on undirected graphs. Previously, in form of recurrent neural networks, Recurrent Graph Neural Networks (RecGNNs) [27] are proposed to operate on directed graphs. Following RecGNNs, gated GNN [12] is proposed to operate on directed graphs, which uses gated recurrent units and employs back-propagation through time (BPTT) to compute gradients. Since then, gated GNN has established new state-of-the-art benchmarks in a wide range of applications including script event prediction [13], situation recognition [10], image classification [21] and recommender system [28]. Previous studies [33, 29] have verified the ability of gated GNN to capture complex transition patterns among nodes. In this paper, we leverage gated GNN to capture prerequisite transition patterns among concepts.

4 Framework and Methodology

The overview of our proposed ConLearn is shown in Figure 2. We first extract feature representations from fine-tuned BERT model to develop concept embeddings. Then we build concept prerequisite graph from existing prerequisite relations and apply gated graph neural network on it to update the concept representations. Motivated by our observation that related concepts should have similar prerequisite relations with other concepts, we design a concept fusion module to incorporate information from related concepts. Finally, concept representations are fed into the Siamese network for concept prerequisite relation classification.

It should be noted that no handcrafted features are used in our model, which makes our model easy to implement in downstream applications.

4.1 Concept Representation Generation. BERT is a recent language representation model that has surprisingly performed well in diverse language understanding benchmarks, which indicates the possibility that BERT captures semantic information about language. In this paper, we utilize BERT to develop contextual-knowledge-aware concept embeddings. Specifically, we follow the work of [19] and adapt BERT to each concept's naming style by fine-tuning BERT using Masked LM objective with the set of domain corpus extracted from education data D.

Then, for each concept, we construct the input of

[CLS] + e_i + [SEP], where e_i is the natural language phrase representation of the concept. We use the representation of the [CLS] token from the last layer of BERT model as initial concept embeddings in our model. Then, we transform concept embeddings to a low dimension by applying a linear transformation. We represent the transformed concept embedding matrix as $\mathbf{S} \in \mathbb{R}^{|C| \times d}$, where d is the dimensionality of concept embeddings.

4.2 Prerequisite Transition Pattern Mining. Previous studies [33, 29] show that gated GNN is capable of capturing complex transition patterns among nodes, which makes gated GNN suit our problem. In prerequisite transition pattern mining, we first construct a directed concept prerequisite graph using available prerequisite relations and then gated GNN is applied on the concept prerequisite graph to update the concept representations, which incorporates concept prerequisite transition patterns into the model.

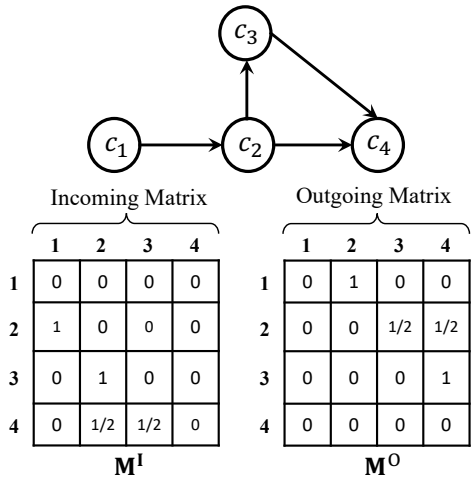

Figure 3: An example of a concept prerequisite graph structure and the connection matrices $\mathbf{M^I}$ and $\mathbf{M^O}$.

Graph Construction. Given concept prerequisite pairs $\{<c_i, c_j> \mid i \to j\}$ of the target domain, we treat each concept c_i as a node and each concept pair $<c_i, c_j>$ as a directed edge. Therefore, prerequisite relations in the target domain can be modeled as a directed graph. The graph structure is updated by promoting communications among different nodes.

Specifically, let $\mathbf{M^I}, \mathbf{M^O} \in \mathbb{R}^{|C| \times |C|}$ denote weighted connections of incoming and outgoing edges in the concept prerequisite graph, respectively. For example, consider concept prerequisite pairs $\{<c_1, c_2>, <c_2, c_3>, <c_3, c_4>, <c_2, c_4>\}$, the corresponding concept prerequisite graph and the connection matrices (i.e., $\mathbf{M^I}, \mathbf{M^O}$) are shown in Figure 3. Since concepts may appear in the concept prerequisite pairs repeatedly, we assign each edge with a normalized weight, which is calculated as the occurrence of the edge divided by the outdegree of that edge's start node.

Graph Densification. Previous work [16] reveals that the number of prerequisite relations is much smaller than the total number of concept pairs, which leads to the sparsity of concept prerequisite graph. The sparsity makes it challenging for gated GNN to perform information propagation over a node's neighborhoods. To tackle this issue and prevent the graph structure from being too complex, we add a synthetic edge between two concepts when there exists a two-hop prerequisite relation between them. For example, if concept A is a prerequisite of concept B, and concept B is a prerequisite of concept C, then we add a directed edge pointing from concept A to concept C.

Concept Representation Updating. Next, we present how to update representations of concepts via gated graph neural network. We first embed each concept $c \in C$ into an unified low-dimensional latent space and the vector $\mathbf{s}_i \in \mathbb{R}^d$ denotes a d-dimensional real-valued latent vector of concept c_i.

For each concept c_i at step t in the concept prerequisite graph, given by the connection matrices $\mathbf{M}^I$ and $\mathbf{M}^O$, the information propagation can be formalized as:

$$\begin{aligned} \mathbf{a}_i^t = \mathbf{Concat}(&\mathbf{M}_i^I[\mathbf{h}_1^{t-1}, ..., \mathbf{h}_{|C|}^{t-1}]\mathbf{W}_a^I + \mathbf{b}^I, \\ &\mathbf{M}_i^O[\mathbf{h}_1^{t-1}, ..., \mathbf{h}_{|C|}^{t-1}]\mathbf{W}_a^O + \mathbf{b}^O), \end{aligned} \tag{4.1}$$

where $\mathbf{h}_i^0 \in \mathbb{R}^{1 \times d}$ corresponds to the i-th row of concept embedding matrix $\mathbf{S}$. $\mathbf{W}_a^I, \mathbf{W}_a^O \in \mathbb{R}^{d \times d}$ are learnable parameters. $\mathbf{b}^I, \mathbf{b}^O \in \mathbb{R}^d$ are the bias vectors. $\mathbf{M}_i^I, \mathbf{M}_i^O \in \mathbb{R}^{1 \times |C|}$ are the i-th row of incoming matrix and outgoing matrix corresponding to concept c_i. $\mathbf{a}_i^t$ extracts the contextual information of neighborhoods for concept c_i.

Then, two gates, i.e., update gate $\mathbf{z}_i^t$ and reset gate $\mathbf{r}_i^t$, decide what information to be preserved and discarded respectively. After that, we construct the candidate state $\tilde{\mathbf{h}}_i^t$ by the previous state $\mathbf{h}_i^{t-1}$, the current state $\mathbf{a}_i^t$, and the reset gate $\mathbf{r}_i^t$. The final state $\mathbf{h}_i^t$ is then the combination of the previous hidden state $\mathbf{h}_i^{t-1}$ and the candidate state $\tilde{\mathbf{h}}_i^t$, under the control of the update gate $\mathbf{z}_i^t$. After updating all nodes in the concept prerequisite graph until convergence, we can obtain the prerequisite-transition-aware concept

representations. The overall procedure can be expressed as follows:

$$\begin{aligned}\mathbf{z}_i^t &= \sigma\left(\mathbf{W}_z\mathbf{a}_i^t + \mathbf{U}_z\mathbf{h}_i^{t-1}\right),\\ \mathbf{r}_i^t &= \sigma\left(\mathbf{W}_r\mathbf{a}_i^t + \mathbf{U}_r\mathbf{h}_i^{t-1}\right),\\ \tilde{\mathbf{h}}_i^t &= \tanh\left(\mathbf{W}_h\mathbf{a}_i^t + \mathbf{U}_o\left(\mathbf{r}_i^t \odot \mathbf{h}_i^{t-1}\right)\right),\\ \mathbf{h}_i^t &= \left(1-\mathbf{z}_i^t\right)\odot \mathbf{h}_i^{t-1} + \mathbf{z}_i^t \odot \tilde{\mathbf{h}}_i^t,\end{aligned} \tag{4.2}$$

where $\mathbf{W}_z, \mathbf{W}_r, \mathbf{W}_h \in \mathbb{R}^{d\times 2d}, \mathbf{U}_z, \mathbf{U}_r, \mathbf{U}_o \in \mathbb{R}^{d\times d}$ are learnable parameters. $\sigma(\cdot)$ represents the sigmoid function and $\odot$ denotes element-wise multiplication.

4.3 Concept Fusion Module. Related concepts should have similar prerequisite relations with other concepts. For example, *Binary Tree* is a prerequisite of *DAG graph* and *AVL tree* is highly related to *Binary Tree*, then it is likely that *AVL tree* is also a prerequisite of *DAG graph*. To leverage this observation for concept prerequisite relation learning, one good solution is to select related concepts for the target concept and fuse the information from these related concepts for target concept prerequisite relation classification.

Self-attention is a special case of attention mechanism and has been successfully applied in a lot of research fields such as NLP [30] and QA [11]. The self-attention mechanism can draw global dependencies between two sequences and capture concept-concept dependencies without regard to their distance, which makes self-attention mechanism suit our problem. After updating the latent vectors of all nodes involved in the concept prerequisite graph, i.e., $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{|C|}]$, we feed them into the concept fusion module to incorporate information from related concepts.

Here we implement m independent attention heads to stabilize the learning process and incorporate more information into the network.

$$\begin{aligned}\boldsymbol{\alpha} &= \frac{\left(\mathbf{H}\mathbf{W}_Q^{(h)}\right)\left(\mathbf{H}\mathbf{W}_K^{(h)}\right)^T}{\sqrt{d}},\\ \tilde{\mathbf{F}}^{(h)} &= \mathbf{Softmax}(\boldsymbol{\alpha})\left(\mathbf{H}\mathbf{W}_V^{(h)}\right),\\ \mathbf{F} &= \mathbf{Concat}\left(\tilde{\mathbf{F}}^{(1)}, \tilde{\mathbf{F}}^{(2)}, ..., \tilde{\mathbf{F}}^{(m)}\right)\mathbf{W}_O,\end{aligned} \tag{4.3}$$

where $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)} \in \mathbb{R}^{d\times d}$ are projection matrices under head h.

After that, we apply two linear transformations with a ReLU activation function to endow the model with nonlinearity and consider interactions between different latent dimensions. However, transmission loss may occur in self-attention operations. Thus we add a residual connection after the feed-forward network.

$$\tilde{\mathbf{E}} = \mathbf{ReLU}\left(\mathbf{F}\mathbf{W}_1 + \mathbf{b}_1\right)\mathbf{W}_2 + \mathbf{b}_2 + \mathbf{F}, \tag{4.4}$$

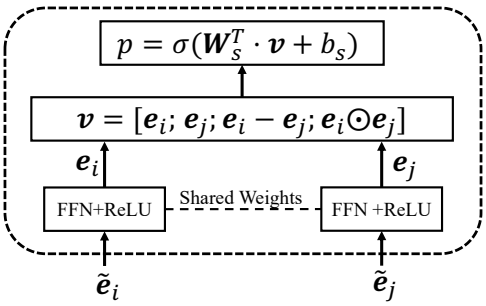

Figure 4: The architecture of Siamese Network.

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d\times d}$ are transformation matrices, $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^d$ are bias vectors.

4.4 Siamese Network. After obtaining final concept representations, Siamese network is used to predict whether concept c_i is a prerequisite of concept c_j.

The architecture of Siamese network is shown in Figure 4. As we can see, concept representations are fed into two feed-forward networks with shared weights. Then the outputs are concatenated for classification. The overall procedure can be expressed as follows:

$$\begin{aligned}\mathbf{e}_i &= \mathbf{ReLU}\left(\mathbf{W}_3\tilde{\mathbf{e}}_i + \mathbf{b}_3\right),\\ p(c_i, c_j) &= \sigma\left(\mathbf{W}_S^T\left[\mathbf{e}_i; \mathbf{e}_j; \mathbf{e}_i - \mathbf{e}_j; \mathbf{e}_i \odot \mathbf{e}_j\right] + b_S\right),\end{aligned} \tag{4.5}$$

where σ is the sigmoid operator, $\odot$ and $-$ are the element-wise multiplication and subtraction operator respectively, and $[\cdot;\cdot]$ denotes the concatenation of vectors.

Finally, we have the cross-entropy loss function:

$$\begin{aligned}\mathcal{L}_c = \frac{1}{|T|}\sum_{(c_i,c_j)\in T} &-[y_{ij}\log(p(c_i,c_j)) +\\ &(1-y_{ij})\log(1-p(c_i,c_j))],\end{aligned} \tag{4.6}$$

where T is the training dataset, $y_{ij} \in \{0,1\}$ indicates whether concept c_i is a prerequisite of concept c_j.

5 Experiments

5.1 Datasets. In order to validate the efficiency of our model, we conduct experiments on three representative datasets from different domains.

- **MOOC**[1]: This dataset [23] contains two MOOC courses: Data structure and Algorithm (DSA) and Machine Learning (ML).
- **LectureBank**[2]: This dataset [9] contains 1,352

[1] http://keg.cs.tsinghua.edu.cn/jietang/software/acl17-prerequisite-relation.rar

[2] https://github.com/Yale-LILY/LectureBank

English lecture files collected from university courses and the annotations of prerequisite relations on 208 concepts.

- **University Courses**[3]: This dataset [16] contains 654 courses from various universities in USA and the annotations of prerequisite relations on 407 concepts.

The set of concepts and prerequisite relations among them were annotated by experts and released with the datasets. The statistics of the datasets are listed in the Table 1.

Table 1: Dataset statistics.

Dataset		# Concepts	# Prerequisites
MOOC	DSA	175	354
	ML	216	1,446
LectureBank		208	913
University Courses		407	1,007

5.2 Baselines. We use the following state-of-the-art approaches as baselines.

- **Binary classifiers**: We compare our model with the binary classifiers used in [23], including Naive Bayes classifier (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest classifier (RF).
- **RefD**: RefD [14] is a simple link-based metric for measuring the prerequisite relations among concepts.
- **GAE**: GAE denotes graph autoencoder. [9] transforms concept prerequisite relation learning problem to link prediction problem and predicts links via using GAE to reconstruct adjacency matrix.
- **VGAE**: VGAE is an extension of GAE which is also used in [9] for concept prerequisite relation learning.
- **PREREQ**: PREREQ [26] obtains latent representations of concepts through the Pairwise Latent Dirichlet Allocation model and identifies prerequisite relations using Siamese network.
- **CPRL**: CPRL [7] is the state-of-the-art approach that combines concept representations learned from heterogeneous graph and concept pairwise features. Besides, CPRL is extended under weakly supervised settings to avoid costly training data labeling.

Consistent with many methods, we employ the widely used metrics Precision (P), Recall (R) and F-score (F_1) to evaluate the performance of ConLearn compared with all the baselines.

5.3 Experimental Settings. To evaluate the performance of ConLearn, we split the concept prerequisite pairs into training, validation and testing sets. In order to fairly compare with the previous researches, 80% samples of LectureBank are used for training, 10% are used for validation and the rest 10% are used for testing, while the proportion changes to 60%, 10% and 30% for other datasets. We randomly extract half of the concept prerequisite pairs to construct the concept prerequisite graph while the rest concept prerequisite pairs serve as the training data. We generate negative samples by sampling random unrelated pairs of phrases from the vocabulary in addition to the reverse pairs of original positive samples. For the task of concept prerequisite relation learning, positive samples are usually rare, leading to imbalanced datasets. We oversampled 1.5 times the number of positive samples in the training, validation and testing sets for all datasets to address the imbalance problem.

All the experiments are conducted on a server machine equipped with a Titan RTX GPU. We use AdamW as the optimizer and the learning rate is set to be 0.005 for University Course dataset and 0.001 for other datasets. We set batch size to be 512 for all datasets. We train our model for 100 epochs until its performance does not improve on the validation set for 15 consecutive epochs. We fine-tune the hidden size in {64, 128, 256, 512}, the head number and the step size in range 1 to 9 for different datasets. We adopt the learning rate of 3e-5 to fine-tune the uncased BERT-base model with the Masked LM objective. For baseline methods, we use default parameter settings as in their original implementations.

5.4 Overall Performance. In Table 2, we summarize the comparing results of different methods on different datasets. We find that our method ConLearn outperforms baseline methods on all datasets consistently. For example, the F-scores of our method ConLearn on DSA dataset and ML dataset outperform the best baseline CPRL by 30.0% and 22.5% respectively.

From the result, several observations can be made: (1) Four binary classifiers achieve satisfactory results on all datasets because it utilizes many hand-crafted features, which requires too much preprocessing work.

[3] https://github.com/suderoy/PREREQ-IAAI-19/tree/master/datasets/University%20Course%20Dataset

Table 2: Overall performance comparison in terms of Precision, Recall and F-scores.

Dataset		Metric	NB	SVM	LR	RF	RefD	GAE	VGAE	PREREQ	CPRL	ConLearn
MOOC	DSA	P	0.613	0.705	0.808	0.344	**0.92**	0.294	0.269	0.492	0.641	0.823
		R	0.696	0.624	0.168	0.715	0.252	0.715	0.657	0.462	0.619	**0.816**
		F_1	0.652	0.662	0.278	0.464	0.396	0.417	0.382	0.476	0.630	**0.819**
	ML	P	0.577	0.668	0.748	0.375	0.784	0.293	0.266	0.448	0.800	**0.895**
		R	0.623	0.577	0.270	0.669	0.188	0.733	0.647	0.592	0.642	**0.850**
		F_1	0.599	0.619	0.397	0.481	0.303	0.419	0.377	0.510	0.712	**0.872**
LectureBank		P	0.670	0.857	0.744	0.855	0.666	0.462	0.417	0.590	**0.861**	0.831
		R	0.640	0.692	0.744	0.681	0.228	0.811	0.575	0.502	0.858	**0.960**
		F_1	0.655	0.766	0.744	0.758	0.339	0.589	0.484	0.543	0.860	**0.891**
University Courses		P	0.478	0.796	0.595	0.739	**0.919**	0.450	0.470	0.468	0.689	0.611
		R	0.649	0.635	0.546	0.480	0.415	0.886	0.694	0.916	0.760	**0.966**
		F_1	0.550	0.707	0.569	0.582	0.572	0.597	0.560	0.597	0.723	**0.749**

(2) RefD achieves the highest precision on both DSA dataset and University Course dataset because the link-based metric it proposes can indeed measure the prerequisite relations between concepts. However, its performance on recall is low, leading to low F-score. (3) GAE and VGAE utilize GCN for adjacency matrix reconstruction but their performance is not satisfactory because they only utilize prerequisite relations among concepts and ignore contextual information from related concepts. Similarly, PREREQ's performance is not good because it only utilizes prerequisite relations among educational resources. (4) The state-of-the-art method CPRL beats other baselines but performs worse than our method ConLearn. The reason is that ConLearn transfers knowledge from large language model BERT to improve contextual representations of concepts and fuse information from related concepts to facilitate the prerequisite relation classification of the target concept.

Finally, our method ConLearn almost achieves the best performance on all datasets across all evaluation metrics except for precision metric. This coincides with our desired ambition because we would rather give the students more concepts which they need to know rather than fewer in which case they may miss important fundamental knowledge.

5.5 Ablation Study. We analyze the effects of each model component. We create ablations by removing them one by one. Specifically, we use 300-dimensional GloVe [24] as the pre-trained concept embeddings to replace BERT embedding initialization, directly remove graph neural network, concept fusion module and replace Siamese network with the concatenation of concept representations followed by MLP network, respectively.

Table 3: Impact of each component on University dataset, where Δ denotes the extent of performance decline.

Ablation	Precision (Δ)	Recall (Δ)	F-score (Δ)
-BERT Embedding	0.596(-2.5%)	**0.612(-36.6%)**	**0.604(-19.4%)**
-Graph Neural Network	0.602(-1.5%)	0.749(-22.5%)	0.667(-10.9%)
-Concept Fusion Module	0.604(-1.1%)	0.788(-18.4%)	0.683(-8.8%)
-Siamese Network	**0.573(-6.2%)**	0.752(-22.2%)	0.650(-13.2%)

As shown in Table 3, ConLearn outperforms all the ablations, which indicates the importance of each model component in improving model performance. Specifically, the performance on recall and F-score drops the most significantly when we replace BERT embedding initialization with GloVe embedding initialization. This is because by transferring knowledge from large language model BERT, we incorporate rich semantic information into the model, which greatly helps improve concept prerequisite relation classification accuracy. However, when we remove graph neural network and concept fusion module, no significant drop on precision is observed. This is because graph neural network enables each concept to interact with prerequisite concepts while self-attention network enables each concept to interact with all concepts. By setting multiple steps of gated graph neural network, each concept can interact with other concepts multiple hops away. So graph neural network and concept fusion module complement each other to some extent.

5.6 Effect of Training Data Size. Previous works reveal that the number of prerequisite relations is much smaller than the total number of concept pairs. It is partially because of the high cost for human experts to label prerequisite relations in educational data. Thus it is significant to pay attention to the model's robustness towards scenarios where the size of training data is limited. In this section, we conduct experiments to evaluate the effect of training data size. Specifically, we randomly reduce 20% to 80% of the original dataset as our new dataset and train our model on it. We report the performance on DSA and LectureBank dataset.

Table 4: Performance w.r.t training data size on DSA dataset and LectureBank dataset.

Training Data Size		80%	60%	40%	20%
DSA	P	0.655	0.644	0.634	0.604
	R	0.716	0.710	0.660	0.628
	F_1	0.684	0.675	0.646	0.616
LectureBank	P	0.819	0.775	0.763	0.744
	R	0.895	0.823	0.780	0.753
	F_1	0.855	0.798	0.772	0.748

As we can see, as the size of training data decreases, ConLearn's performance on both datasets drops. However, the performance only reduces marginally as the training data size reduces from 80% to 20%. It is worth mentioning that the performance of ConLearn trained with 40% of training data is better than the performance of most baselines trained with complete training data. This is because our model transfers knowledge from large language model BERT to improve contextual representations of concepts. Besides, the use of self-attention network also enables the model to fully utilize information from related concepts to facilitate prerequisite relation classification of the target concept.

5.7 Case Study. We further investigate our method by studying examples of concept prerequisite pairs recovered. Table 5 lists examples of both true positive pairs (TPP) and false negative pairs (FNP). TPP denotes the concept pairs our model correctly identifies as prerequisite pairs and FNP denotes the concept pairs our model mistakenly identifies as non-prerequisite pairs.

Looking closely at the TPP, we find that our model successfully predicts the prerequisite relations among *Numerical analysis*, *Discrete mathematics* and *Mathematics*, which verifies our model's ability to capture prerequisite transition patterns among concepts. Besides, by looking at *<Regular language, Automata theory>*, *<C programming language, Automata theory>* and *<Assembly language, Automata theory>*, we find that our model is also capable of capturing the intuition that related concepts have similar prerequisite relations. This is because that we give higher weights to related concepts, resulting in similar representations for related concepts and consequently similar prerequisite relation classification results.

Table 5: Examples of concept pairs we recovered, which are categorized as true positive pairs (TPP) and false negative pairs (FNP).

	Concept Pairs
TPP	<Numerical analysis, Discrete mathematics>
	<Discrete mathematics, Mathematics>
	<Numerical analysis, Mathematics>
	<Regular language, Automata theory>
	<C programming language, Automata theory>
	<Assembly language, Automata theory>
FNP	<Query optimization, Database transaction>
	<Congestion control, Communications protocol>
	<Quantum cryptography, Computer security>

Looking closely at the FNP, we find that these concepts are relatively more complex technical terms that usually share sparse prerequisite relations with other concepts. We hypothesize the errors are due to the lack of professional knowledge in our model. Such concept prerequisite relations are more likely to be recovered by feature-based methods.

6 Conclusion

In this paper, we propose a contextual-knowledge-aware concept prerequisite relation learning approach called ConLearn, which transfers knowledge from large language model BERT to improve contextual representations of concepts, captures concept prerequisite transition patterns using graph neural network and utilizes self-attention network to fuse information from related concepts for target concept prerequisite relation classification. More importantly, no handcrafted features are used in our model, which makes our model easy to implement in downstream applications. Extensive experiments on three datasets show that the proposed approach achieves the state-of-the-art performance compared with existing methods.

Acknowledgement

This work was supported by National Key Research and Development Program of China under Grant No. 2018AAA0101902, and NSFC under Grant No. 61532001.

References

[1] Rakesh Agrawal, Behzad Golshan, and Evangelos Papalexakis. Datadriven synthesis of study plans. *Data Insights Laboratories*, 2015.

[2] Sahar Changuel, Nicolas Labroche, and Bernadette Bouchon-Meunier. Resources sequencing using automatic prerequisite–outcome annotation. *TIST*, 2015.

[3] Penghe Chen, Yu Lu, Vincent W Zheng, and Yang Pian. Prerequisite-driven deep knowledge tracing. In *ICDM*, 2018.

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

[5] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *arXiv preprint arXiv:1509.09292*, 2015.

[6] Jonathan Gordon, Linhong Zhu, Aram Galstyan, Prem Natarajan, and Gully Burns. Modeling concept dependencies in a scientific corpus. In *ACL*, 2016.

[7] Chenghao Jia, Yongliang Shen, Yechun Tang, Lu Sun, and Weiming Lu. Heterogeneous graph neural networks for concept prerequisite relation learning in educational data. In *NAACL*, 2021.

[8] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[9] Irene Li, Alexander R Fabbri, Robert R Tung, and Dragomir R Radev. What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. In *AAAI*, 2019.

[10] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. Situation recognition with graph neural networks. In *ICCV*, 2017.

[11] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI*, 2019.

[12] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

[13] Zhongyang Li, Xiao Ding, and Ting Liu. Constructing narrative event evolutionary graph for script event prediction. *arXiv preprint arXiv:1805.05081*, 2018.

[14] Chen Liang, Zhaohui Wu, Wenyi Huang, and C Lee Giles. Measuring prerequisite relations among concepts. In *EMNLP*, 2015.

[15] Chen Liang, Jianbo Ye, Shuting Wang, Bart Pursel, and C Lee Giles. Investigating active learning for concept prerequisite learning. In *AAAI*, 2018.

[16] Chen Liang, Jianbo Ye, Zhaohui Wu, Bart Pursel, and C Lee Giles. Recovering concept prerequisite relations from university course dependencies. In *AAAI*, 2017.

[17] Hanxiao Liu, Wanli Ma, Yiming Yang, and Jaime Carbonell. Learning concept graphs from online educational data. *Journal of Artificial Intelligence Research*, 2016.

[18] Weiming Lu, Yangfan Zhou, Jiale Yu, and Chenhao Jia. Concept extraction and prerequisite relation learning from educational data. In *AAAI*, 2019.

[19] Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. Commonsense knowledge base completion with structural and semantic context. In *AAAI*, 2020.

[20] Eric Margolis, Stephen Laurence, et al. *Concepts: core readings*. Mit Press, 1999.

[21] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*, 2016.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

[23] Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. Prerequisite relation learning for concepts in moocs. In *ACL*, 2017.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[25] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *KDD*, 2014.

[26] Sudeshna Roy, Meghana Madhyastha, Sheril Lawrence, and Vaibhav Rajan. Inferring concept prerequisite relations from online educational resources. In *AAAI*, 2019.

[27] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 2008.

[28] Hao Sun, Zijian Wu, Yue Cui, Liwei Deng, Yan Zhao, and Kai Zheng. Personalized dynamic knowledge-aware recommendation with hybrid explanations. In *DASFAA*, 2021.

[29] Hao Sun, Changjie Yang, Liwei Deng, Fan Zhou, Feiteng Huang, and Kai Zheng. Periodicmove: Shift-aware human mobility recovery with graph neural network. In *CIKM*, 2021.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[31] Shuting Wang and Lei Liu. Prerequisite concept maps extraction for automaticassessment. In *WWW*, 2016.

[32] Shuting Wang, Alexander Ororbia, Zhaohui Wu, Kyle Williams, Chen Liang, Bart Pursel, and C Lee Giles. Using prerequisites to extract concept maps fromtextbooks. In *CIKM*, 2016.

[33] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. In *AAAI*, 2019.