DisenQNet: Disentangled Representation Learning for Educational Questions

Zhenya Huang[1], Xin Lin[1], Hao Wang[1], Qi Liu[1], Enhong Chen[1,*], Jianhui Ma[1], Yu Su[1,2], Wei Tong[1]
[1]Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China, [2]iFLYTEK Research, iFLYTEK, Co., Ltd
{huangzhy,qiliuql,cheneh,jianhui}@ustc.edu.cn; yusu@iflytek.com; {linx,wanghao3,tongustc}@mail.ustc.edu.cn

ABSTRACT

Learning informative representations for educational questions is a fundamental problem in online learning systems, which can promote many applications, e.g., difficulty estimation. Most solutions integrate all information of one question together following a supervised manner, where the representation results are unsatisfactory sometimes due to the following issues. First, they cannot ensure the presentation ability due to the scarcity of labeled data. Then, the label-dependent representation results have poor feasibility to be transferred. Moreover, aggregating all information into the unified may introduce some noises in applications since it cannot distinguish the diverse characteristics of questions. In this paper, we aim to learn the disentangled representations of questions. We propose a novel unsupervised model, namely DisenQNet, to divide one question into two parts, i.e., a concept representation that captures its explicit concept meaning and an individual representation that preserves its personal characteristics. We achieve this goal via mutual information estimation by proposing three self-supervised estimators in a large unlabeled question corpus. Then, we propose another enhanced model, DisenQNet+, that transfers the representation knowledge from unlabeled questions to labeled questions in specific applications by maximizing the mutual information between both. Extensive experiments on real-world datasets demonstrate that DisenQNet can generate effective and meaningful disentangled representations for questions, and furthermore, DisenQNet+ can improve the performance of different applications.

CCS CONCEPTS

• **Information systems** → **Information extraction**; • **Computing methodologies** → *Knowledge representation and reasoning*.

KEYWORDS

question learning, disentangled representation, mutual information

ACM Reference Format:
Zhenya Huang, Xin Lin, Hao Wang, Qi Liu, Enhong Chen, Jianhui Ma, Yu Su, Wei Tong. 2021. DisenQNet: Disentangled Representation Learning for Educational Questions. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14-18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548.3467347.

*Enhong Chen is the corresponding author.

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467347

1 INTRODUCTION

Online learning systems, such as coursera.org, edX.org and xuetangx.com, have attracted a large number of learners around the world [1]. In 2020, the Covid19 pandemic has impact promoted the proliferation of online learning. According to the statistics of Cousera (https://www.coursera.org/), over 77 million users are now learning and practicing on the platform.

Nowadays, online learning systems collect millions of learning materials (e.g., course and question), and facilitate many personalized applications to improve learning experiences of students [1, 18, 36]. On one hand, students can select suitable courses or questions to acquire knowledge according to their needs, e.g., selecting similar questions to review required concepts after answering one question incorrectly [18, 21, 27]. On the other hand, systems personalize necessary services based on students' feedbacks, e.g., recommending easier questions if noticing that students are struggling with current materials [14, 24]. To support such services, it is necessary to well organize the learning materials in advance [4], especially educational questions. This brings out a fundamental research topic of question understanding in AI education [30, 35], with the goal of learning informative representations of educational questions.

In the literature, focusing on different question-based applications, such as difficulty estimation [14, 24] and similarity analysis [18, 21], many efforts have been developed for understanding question content by taking advantage of natural language processing (NLP) techniques. In general, they design different models to learn question representations as syntactic patterns or semantic encodings, which are directly optimized in specific application tasks. For example, Qiu et al. [24] extracted semantic representations of multiple-choice questions to predict the difficulty. Though they have gained some achievements, there exist some limitations in practical systems as follows. First, existing models follow supervised manner, which requires sufficient labeled data (e.g., difficulty or similarity in Figure 1) for optimization. However, getting high-quality labels for questions is extremely hard in practice because experts to be competent should acquire enough professional knowledge (so we cannot take crowdsourcing in many traditional domains like e-commerce) [37]. As indicated in the literatures, for example, labeling similar questions should understand their psychological purpose [18] and calculating the difficulty scores even requires being examined in standard tests (e.g., GRE test) [5, 26]. Therefore,

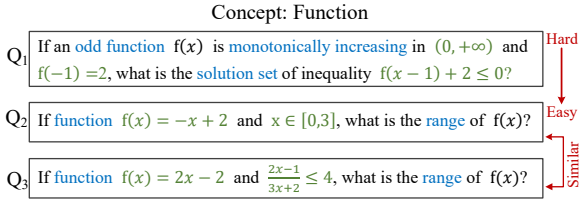


Figure 1: Three educational question examples. Blue contents show they are related to "Function" concept. Green parts present their personal information. Right red labels tell us that Q_1 is harder than Q_2 and Q_2 is similar to Q_3.

such models cannot ensure the representation ability sometimes due to the scarce labels. Second, their representation results are task-dependent, which has poor feasibility to be applied across different applications [6]. That is to say, we have to design different models to represent same questions for different applications. Third, although Yin et al. [36] pre-trained question representations to enhance the model ability with unlabeled data, all existing solutions equally represent a certain question as one unified vector, where all the information are integrated together. However, questions with same concepts (e.g., "Function") can be quite different from its content to show personal properties (e.g., difficulty) [14]. For example in Figure 1, question Q_1 is harder than Q_2 as it has more complicated expressions. Therefore, if we cannot distinguish such differences, it may introduce some noises and mislead the applications.

To this end, we argue that an ideal question representation model should satisfy three desirable abilities: 1) It can get rid of the labels in specific tasks, which can be optimized by learning questions on their own. 2) It should distinguish the diverse characteristics of questions inherent by an explicit way. 3) The learned representations should be flexible for being applied in different downstream tasks.

In this paper, we propose a novel unsupervised model, namely DisenQNet, for question representation learning. In DisenQNet, we aim to disentangle one question into two parts: a concept representation and an individual representation. Specifically, we develop three self-supervised estimators to optimize two disentanglements. First, we learn the question semantics by maximizing the mutual information between itself and our two disentangled representations. Second, we enforce the concept representation with explicit meaning by making prediction of its concepts. Third, we propose an adversarial process to ensure two disentanglements independent so that the individual one can preserve the question personality itself. Particularly, DisenQNet can be optimized by learning questions on their own without any additional annotations.

In addition, we propose another enhanced model, namely DisenQNet+, that applies our learned representations to several application tasks. Specifically, since our individual disentanglement especially leaves the personal information of one question without concept, it can be adapted to improve the representation ability of supervised models in different tasks even if owning very limited labeled data. We achieve this goal by maximizing the mutual information between the unsupervised representations and the supervised representations so that the knowledge can be transferred from DisenQNet to supervised models in different applications.

We perform extensive experiments on three datasets. We empirically show that DisenQNet can generate meaningful disentangled representations of educational questions, and DisenQNet+ can improve the performance on domain-specific applications including difficulty estimation and similarity analysis. To the best of our knowledge, this is the first few attempts to learn disentangled representations for educational questions by distinguishing the concept and individual effects via mutual information estimation.

2 RELATED WORK

In this section, we summarize the related work as follows.

Question Understanding. Online learning systems can collect abundant educational materials, e.g., courses and questions, so that provide several intelligent services, such as searching target questions and personalized recommendation [1, 19, 21], which attracts many participations from the public. It is necessary to help systems organize such materials, and thus triggers a fundamental issue in AI education of question understanding. In the earlier time, scholars try to design fine-grained rules or grammars to understand questions, where they can be organized by specific structures like semantic trees or templates [9]. However, such structures require manually design with strong expertise and cost much time, but could only match very limited questions with weak abilities. Obviously, they are not suitable for online systems nowadays that contain millions of question resources. Therefore, recent advances try to automatically understand question textual content with semantic representations via natural language processing techniques, and support several applications, such as difficulty estimation [14, 24], similarity search [18, 21, 27], question solving [15] and student performance prediction [19]. For example, Qiu et al. [24] extracted semantic representations of multiple-choice questions to predict their difficulty properties. Liu et al. [18] proposed an attention model to evaluate the similarities of question pairs integrated with the heterogeneous information. Though achieving some success, most of them generally follow a supervised manner, which suffer from two main problems. First, the models require enough labeled data to ensure the performance. However, labeling educational questions is hard in practice because experts to be competent should acquire enough professional knowledge. For example, estimating the difficulties of GRE questions require organizing the exams with volunteers [5]. Such effort is even harder than traditional domains, such as news, e-commerce [37], where crowdsourcing fails in practice. Second, the learned question representations are task-specific, which are incapable of being applied cross tasks. In other words, we should design many complicated models for different tasks.

Question Pre-training. To deal with the problems of supervised solutions above, pre-training, as a kind of typical unsupervised techniques, can make use of large-scale unlabeled data to enhance the representation ability for different data structures, such as text [8] and image [10]. Since we focus on how to learn question representation from its textual content, we generally review some NLP efforts. Generally, representative methods can be divided into two categories: feature-based methods, where text is represented by some sorts of feature extractors as fixed vectors [22, 23], and pre-training based methods, where parameters of models are pre-trained on corpus and then fine-tuned in specific tasks [7, 8, 13]. Specially, BERT [7, 8] and GPT [25] are two of the most successful methods, which have already performed impressively in many classic NLP tasks including question-answering machine translation,

etc, and continuously upgraded in recent years. To the best of our knowledge, Yin et al. [36] made the first attempt to pre-train the representations of educational questions from the heterogeneous data. In summary, the main goal of pre-training methods is to make full use of large corpus, which try to integrate all the information together to learn one unified representation for each input. However, in practical learning systems, students try to distinguish the differences of questions, e.g., questions with same concepts but have inconsistent difficulty levels. Therefore, these pre-training methods may be unsatisfactory.

Mutual Information. Mutual information is a fundamental quantity for measuring the dependency between random variables [3, 17]. Since the mutual information is historically difficult to compute for high-dimensional variables, recent works employ several non-linear estimators based on deep neural networks to maximize mutual information for representation learning in many domains, such as image [3, 12], graph [31, 32] and recommendation [29]. For example, Hjelm et al. [12] proposed Deep InfoMax to learn image representations via Jensen-Shannon divergence. Velickovic et al. [32] and Sun et al. [31] proposed DGI and InfoGraph for graph learning in terms of generating node level embeddings and graph level embeddings, respectively. On the basis, Sanchez et al. [28] employed mutual information estimation to learn image disentangled representations with the shared ones and exclusive ones via pairwise instance learning.

Our work targets at representation learning for educational questions. We try to propose an unsupervised model to disentangle one question by distinguishing its concept meaning and individual information via mutual information estimation on its own. We also propose a principle way to transfer our learned representations in several downstream tasks, and therefore, provide a solid backbone in use of questions in online learning systems.

3 PRELIMINARIES

In this section, we present the formal problem definition and introduce some basic knowledge of mutual information.

3.1 Problem Definition

We focus on two problems. First, we define unsupervised representation learning problem for educational questions. Then, we present question-based application tasks in online learning systems.

3.1.1 Unsupervised Question Representation Learning. Let Q denote a set of N questions $Q = \{q_1, q_2, \cdots, q_N\}$. Each question $q \in Q$ is given as its text content with a sequence of M word tokens $q = \{x_1, x_2, \cdots, x_M\}$, along with the concepts $k \in K$, where $|K|$ is the number of all concepts in the system. Please note that each exercise may contain multiple concepts as shown in Table 1. Traditionally, the representation learning for a certain question is to output one unified d-dimensional vector ($d \ll N$), which should capture as much information as possible. However, as we mentioned in Section 1, questions with same concepts can be quite different, so that it is worthwhile to distinguish such differences in an explicit way. Therefore, our goal of disentangled question representation learning transforms to output two d-dimensional vectors including one $v_k \in \mathbb{R}^d$ referring to its concept information and the other $v_i \in \mathbb{R}^d$ capturing the individual characteristics.

3.1.2 Question-based Supervised Tasks. The original question set Q can be divided into two subsets Q^L and Q^U. Specifically, $Q^L = \{q_1, q_2, \cdots, q_L\}$ represents the labeled questions whose properties have been obtained by expertise or organizations, i.e., $\{y_1, y_2, \cdots, y_L\}$. Comparatively, $Q^U = \{q_1, q_2, \cdots, q_U\}$ is the unlabeled questions whose properties remain unknown. Specifically, $Q = Q^L \cup Q^U$, and $|Q^L| \ll |Q^U|$ in most cases. In real-world scenarios, the properties can be specified with different labels, e.g., the difficulty level of one specific question or the similarity score between a pair[1]. Given both Q^L and Q^U, our general goal of supervised task can be formulated as learning a model to predict the properties of unknown questions. In this paper, we focus on two representative supervised application tasks including difficulty estimation and similarity analysis, which will be discussed in detail in Section 5.3.

3.2 Mutual Information

In this subsection, we briefly introduce some related basic knowledge of mutual information. In information theory, mutual information $I(X, Y)$ measures the dependence between two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, which can be expressed as the decrease of the uncertainty in one given the other:

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{1}$$

where, $H(X)$ and $H(X|Y)$ are the Shannon entropy and the conditional entropy of X given Y, respectively. Generally, Eq. (1) is equivalent to the Kullback-Leibler (KL) divergence between the joint distribution $\mathbb{P}(X, Y)$ and the product of two marginals $\mathbb{P}(X) \otimes \mathbb{P}(Y)$ as: $I(X; Y) = D_{KL}(\mathbb{P}(X, Y)||\mathbb{P}(X) \otimes \mathbb{P}(Y))$. Intuitively, $I(X; Y)$=0 means variables X and Y are independent, and the larger the value is, the stronger the dependence they have.

However, directly computing the mutual information $I(X; Y)$, if variables X and Y are continuous and high-dimensional, is extremely difficult [3]. Therefore, Belghazi et al. [3] developed a nonlinear neural network estimator MINE to measure a tight lower bound of it, which is based on the Donsker-Varadhan representation, a dual representation of KL-divergence as:

$$\begin{aligned} \hat{I}_\theta^{(DV)} &:= D_{KL}(\mathbb{P}(X, Y)||\mathbb{P}(X) \otimes \mathbb{P}(Y)) \\ &:= \mathbb{E}_{\mathbb{P}(X,Y)}\left[T_\theta(x, y)\right] - \log \mathbb{E}_{\mathbb{P}(X)\mathbb{P}(Y)}\left[e^{T_\theta(x,y)}\right], \end{aligned} \tag{2}$$

where $T_\theta(x, y) : X \times Y \to \mathbb{R}$ is the neural network approximator with parameters θ. In practice, using estimator in Eq. (2) is not stable because it is sensitive to negative samples. To overcome this problem, noticing that the representation learning work does not focus on the precise value of mutual information, but on maximizing it, Hjelm et al. [12] introduced a approximate Jensen-Shannon (JS) divergence based estimator as:

$$\begin{aligned} \hat{I}_\theta^{(JS)} &:= D_{JS}(\mathbb{P}(X, Y)||\mathbb{P}(X) \otimes \mathbb{P}(Y)) \\ &:= \mathbb{E}_{\mathbb{P}(X,Y)}\left[-\mathrm{sp}(-T_\theta(x, y))\right] - \mathbb{E}_{\mathbb{P}(X)\mathbb{P}(Y)}\left[\mathrm{sp}(T_\theta(x, y))\right], \end{aligned} \tag{3}$$

where $\mathrm{sp}(x) = \log(1 + e^x)$ is the softplus function.

In this paper, we employ this method in a principle way as we focus on learning representations for educational questions. Readers who are interested in mutual information can refer to the literature [3, 12, 17] for more details.

[1] Please refer to [26] for more useful question properties like discrimination etc.

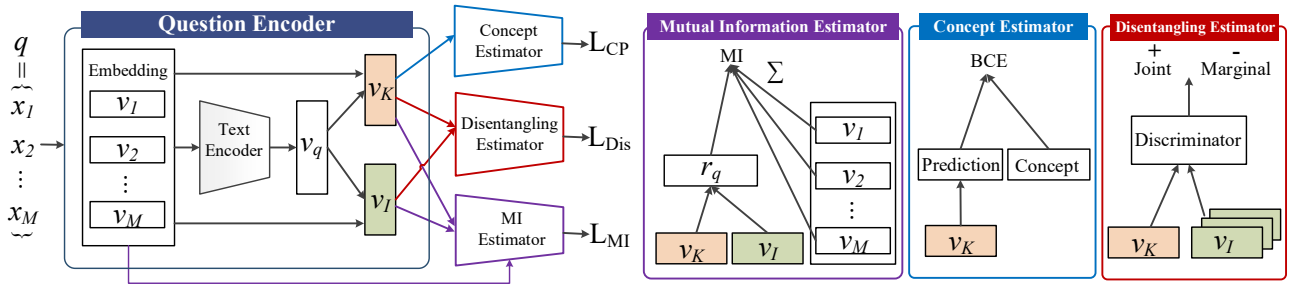


Figure 2: DisenQNet. Left part shows the model architecture. Right part illustrates three estimator details for optimization.

4 METHODOLOGY

4.1 DisenQNet

We propose a novel unsupervised model, namely DisenQNet, for our first problem of question representation learning. Different from most existing models which integrate all the information together [18, 36], our DisenQNet tries to distinguish both the concept information and individual characteristics of one question in an explicit way. Figure 2 illustrates its general model architecture, which consists of the question encoder and self-supervised optimization.

4.1.1 Question Encoder. We disentangle the final question representation into two parts, i.e., the concept representation v_K that captures its concept information and the individual representation v_I that preserves the personal characteristics, i.e., $r_q = (v_K, v_I)$.

Given a certain question $q = \{x_1, x_2, \cdots, x_M\}$, we initialize each of them $\{x_n\}$ by a d_0-dimensional word embedding, i.e., $V = \{v_1, v_2, \cdots, v_M, v_m \in \mathbb{R}^{d_0}\}$. Then, we perform a text encoder, via a neural network, i.e., $f_\theta : \mathbb{R}^{d_0 \times M} \to \mathbb{R}^{d_q}$, to generate its integrated semantic representation v_q from its content input V:

$$v_q = f_\theta(\{v_m : v_m \in V\}). \tag{4}$$

Note that the text encoder f_θ can be specified with several models like CNN, LSTM etc. We do not emphasize their differences and implement it by TextCNN [16], as it is computationally efficient.

After that, we try to split this integrate semantics v_q into two disentangled representations, i.e., the concept representation v_K and the individual representation v_I. As shown in Figure 1, such two disentanglements may focus on different content in the question. For example, some informative words like "odd function" tell what concepts the question Q_1 has, and therefore, contribute more to the concept part. Comparatively, mathematical expressions like "f(-1)=2" bring the detailed information of itself. Therefore, we perform two attention networks to capture such differences in the modeling. Specifically, the two networks are established with same architecture but different parameters, so as to quantify the dominant contents on the target disentanglements, guiding the modeling. As the representative, for obtaining the concept representation v_K, the corresponding attention network can be formulated as:

$$v_K = \sum_{j=1}^{M} \alpha_j v_j, \quad \alpha_j = \text{Softmax}(\text{MLP}(v_j, v_q)), \tag{5}$$

where α_j represents the weight score of word x_j to the concept vector v_K, which is normalized by Softmax function. MLP is the multi-layer perception network that captures the distance between two vectors (v_j, v_q). The individual presentation v_I is generated with another attention network similarly.

4.1.2 Self-supervised optimization. Now, we discuss how to optimize DisenQNet to learn two desirable disentanglements. The challenge here is that we can only extract the unique question characteristics itself without any additional explicit labels. Thus, as shown in Figure 2, we propose three estimators with self-supervised objectives to guide model learning including mutual information (MI) estimator, concept estimator and disentangling estimator.

MI Estimator. First, our learned disentangled representations, i.e., $r_q = (v_K, v_I)$, should capture the given question information in all. We perform the MI estimator on the given question, which maximizes the estimated mutual information between the concatenation r_q and the question content V. As illustrated in Figure 2, we perform this MI maximization on each word $v_j \in V$ in the question, in which every local word information can be estimated with the global concatenation r_q smoothly. This objective can be expressed as maximizing the JS-divergence based estimator (Eq. (3)) as:

$$\mathcal{L}_{MI} = \hat{I}^{(JS)}_{\theta_1}(r_q, V) = \frac{1}{M}\sum_{j=1}^{M} \{\mathbb{E}_{\mathbb{P}(r_q, v_j)}\left[-\text{sp}(-T_{\theta_1}(r_q, v_j))\right] - \mathbb{E}_{\mathbb{P}(r_q)\mathbb{P}(v_j)}\left[\text{sp}(T_{\theta_1}(r_q, v_j))\right]\}, \tag{6}$$

where the approximator T_{θ_1} is designed with a multi-layer fully connected network. In practice, it is not easy to directly get $\mathbb{P}(r_q)\mathbb{P}(v_j)$, but we take the similar technique in [12, 31] to achieve that by shuffling either of them in a batch sampling from $\mathbb{P}(r_q, v_j)$.

Concept Estimator. Next, we ensure our concept representation v_K with the explicit concept meaning of the given question. As shown in Figure 2, we perform the concept estimator to predict its given one-hot concept encoding $k \in \{0, 1\}^{|K|}$ by $h_{\phi_1}(v_K)$, where the network $h_{\phi_1} : \mathbb{R}^d \to \mathbb{R}^{|K|}$ projects the concept disentanglement v_K into the prediction. Therefore, this estimator is defined as minimizing the binary cross-entropy (BCE) objective:

$$\mathcal{L}_{CP} = \frac{1}{|K|}\sum_{j=1}^{|K|}(k_j \log(h_{\phi_1}(v_K)_j) + (1 - k_j)\log(1 - h_{\phi_1}(v_K)_j)). \tag{7}$$

Disentangling Estimator. Last, the disentangling estimator preserves the personal characteristics of the given question in its individual representation v_I. Recall the example in Figure 1, questions with the same concepts ("Function") are different reflected by their properties (e.g., difficulty). Along this line, our ideal individual representation v_I must not contain the information captured by the concept one v_K, and should be independent with v_K. Therefore, our intuitive goal here is to minimize the mutual information between v_K and v_I. However, as Sanchez et al. [28] suggested, we cannot directly achieve this goal by minimizing Eq. (3) since the estimator fails to converge when minimizing. As an alternative, in this work, we propose an adversarial process that minimizes the

distance between the joint distribution $\mathbb{P}(v_K, v_I)$ and the marginals $\mathbb{P}(v_K)\mathbb{P}(v_I)$. Specifically, as shown in Figure 2, our adversarial process contains two components. First, we train a discriminator D_Φ to classify the sampled representations drawn from the joint $\mathbb{P}(v_K, v_I)$ as the real and samples drawn from the marginals $\mathbb{P}(v_K)\mathbb{P}(v_I)$ as the fake. Then, we train our DisenQNet that can fool the discriminator D_Φ by shuffling the individual representations of samples in a batch from $\mathbb{P}(v_K, v_I)$. During this process, our DisenQNet tries to generate the combined disentanglements (v_K, v_I) that look like drawn from $\mathbb{P}(v_K)\mathbb{P}(v_I)$, and therefore, ensures the independence as we desire. More formally, we express such adversarial objective which is similar to WGAN with spectral normalization [2] since it is more stable in the learning process:

$$\mathcal{L}_{Dis} = \mathbb{E}_{\mathbb{P}(v_k, v_i)}\left[D_\phi(v_k, v_i)\right] - \mathbb{E}_{\mathbb{P}(v_k)\mathbb{P}(v_i)}\left[D_\phi(v_k, v_i)\right]. \quad (8)$$

By combining Eq. (6), Eq. (7) and Eq. (8), our final objective of DisenQNet is defined with the hyper-parameters λ_1, λ_2, and $\lambda_3$as:

$$\mathcal{L}_{\text{DisenQNet}} = -\lambda_1\mathcal{L}_{MI} + \lambda_2\mathcal{L}_{CP} + \lambda_3\mathcal{L}_{Dis}. \quad (9)$$

The overall objective can be minimized using SGD with the Adam optimizer. We will specify the details in the experiments.

In summary, our DisenQNet has the following advantages. First, it learns question representations only with the characteristics themselves without requiring additional labels, where large corpus of unlabeled questions could be well leveraged to enhance the ability. Second, it splits questions into two independent disentanglements via mutual information estimation so that it can distinguish the different effects of both concept meaning and personal information.

4.2 DisenQNet+

In this subsection, we deal with the second goal of question-based tasks based on our representation results. Most of existing solutions [14, 18] devote many efforts to several application tasks (e.g., difficulty estimation) in a supervised manner. In practice, they may suffer from the following two problems. First, they are label-hungry, as the performances rely on the sufficient annotations (e.g., difficulty). However, getting high-quality labels is costly with high expertise, which would be easily unsatisfactory. Second, their learned representations are label-dependent, which have poor feasibility to be applied across different tasks. To overcome such issues, we propose an enhanced model to apply our DisenQNet to the downstream tasks for improving the performance. We call it DisenQNet+.

Note that most applications focus on distinguishing the differences among questions. For example in Figure 1, even if the three are related to the same "Function" concept, Q_1 is harder than Q_2 since it has more complicated mathematical expressions, and Q_2 is similar to Q_3 due to possible same purpose ("What is the range..."). Therefore, an ideal model should devote more energy to capture the unique information of one question itself rather than the same part. Motivated by this intuition, in our DisenQNet+, we directly transfer our individual representations (from DisenQNet) to downstream models, since this disentanglement especially removes the concept information but leaves its personal characteristics of one question. (We show it experimentally in Section 5.3). The architecture of DisenQNet+ is illustrated in Figure 3.

Without loss of generality, we can take the common process for one specific application in the following. Specifically, given

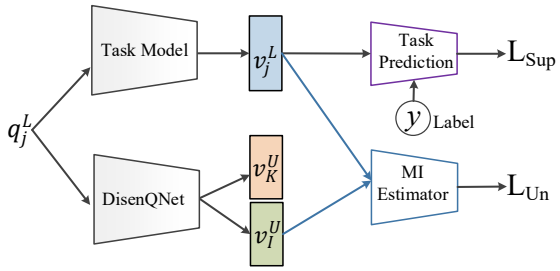


Figure 3: DisenQNet+ for downstream application tasks.

one labeled question $q_j^L \in Q^L$, we first design a task model $f_\delta : \mathbb{R}^{d_0 \times M} \to \mathbb{R}^{d_t}$ to produce its task-specific semantic representation $v_j^L \in R^{d_t}$, i.e., $v_j^L = f_\delta(q_j^L)$. Then, we apply one prediction network $h_{\phi_2} : \mathbb{R}^{d_t} \to \mathbb{R}$ that outputs the task prediction, i.e., $p_j = h_{\phi_2}(v_j^L)$. In summary, the task loss on label set Q^L can be formulated in a supervised manner as:

$$\mathcal{L}_{Sup} = \sum_{j=1}^{|Q^L|} \mathcal{L}_P(p_j, y_j), \quad p_j = h_{\phi_2}(v_j^L), \quad (10)$$

where $\mathcal{L}_P(p_j, y_j)$ represents the loss function between the prediction p_j and the true label y_j of the given labeled question q_j^L.

Obviously, due to the scarcity of labeled set, models only trained with Eq. (10) may be overfitting, as the semantic representation v_j^L may not be optimized very well. Therefore, we design another component to enhance this representation ability, where a pre-trained unsupervised DisenQNet on all question set (Q^L, Q^U) is incorporated to generate the individual disentanglement v_I^U of the given in parallel for being transferred. To achieve the goal, a straightforward way is to concatenate both v_I^U and v_j^L, and then make the prediction. However, this simple way may lead to negative transfer problem [31] since the supervised task model and unsupervised DisenQNet may favor different information in the latent space. Therefore, we propose an effective way to alleviate this adaptation problem. Specifically, we perform another estimator to maximize the mutual information between v_I^U and v_j^L, so that the knowledge from DisenQNet can be transferred to the task models in applications more smoothly. Formally, this process can be formulated with the JS-divergence estimator as:

$$\begin{aligned}\mathcal{L}_{Un} = \hat{I}^{(JS)}_{\theta_2}(v_I^U, v_j^L) = \mathbb{E}_{\mathbb{P}(v_I^U, v_j^L)}\left[-\text{sp}(-T_{\theta_2}(v_I^U, v_j^L))\right] \\ - \mathbb{E}_{\mathbb{P}(v_I^U)\mathbb{P}(v_j^L)}\left[\text{sp}(T_{\theta_2}(v_I^U, v_j^L))\right]. \quad (11)\end{aligned}$$

where the statistics approximator T_{θ_2} is also designed with a multi layer fully connected network.

Therefore, the overall loss in DisenQNet+ can be summarized with the hyper-parameter λ_4 and λ_5 as:

$$\mathcal{L}_{\text{DisenQNet+}} = \lambda_4\mathcal{L}_{Sup} - \lambda_5\mathcal{L}_{Un}. \quad (12)$$

Similarly, this objective can be minimized using SGD with the Adam optimizer, which will be specified in details in the experiments.

Particularly, our DisenQNet+ is flexible for different downstream tasks with their original supervised loss functions (e.g., ranking loss or classification loss). We will discuss it in detail in Section 5.3. Moreover, the task model f_δ can also be specified with any related ones, e.g., TACNN [14] in difficulty estimation task or MANN [18] in question search task. In this paper, we do not put much emphasis

Table 1: The statistics of the datasets.

Dataset	SYSTEM1	SYSTEM2	Math23K
#Questions	108,137	25,293	23,096
#Concepts	31	21	5
Avg. question length	48.15	129.96	28.06
Avg. concepts per question	1.91	1.16	1.9
#Questions with difficulty label	5,291	/	2000
Avg. difficulty labels per concept	307	/	772
#Questions with similarity label	/	2944	/
#Labeled similar pairs	/	1900	/
Avg. similarity labels per question	/	1.29	/
Label sparsity	4.9%	11.6%	8.7%

on comparing the performances of complicated task models, and therefore, we implement f_δ by the commonly used and useful TextCNN [16] in a principle way.

5 EXPERIMENTS

We run all experiments on a Linux server with four 2.0GHz Intel Xeon E5-2620 CPUs and a Tesla K80 GPU. Our code is available at https://github.com/bigdata-ustc/DisenQNet.

5.1 Datasets

We use three datasets in the experiments, namely SYSTEM1, SYSTEM2, and Math23K. The SYSTEM1 and SYSTEM2 datasets collect the mathematical questions from the online learning system iFLYTEK Zhixue.com that respectively accord with high-school level and middle-school level. Specifically, SYSTEM1 dataset contains 108,137 questions and 31 concepts in total, while SYSTEM2 consists of 25,293 questions with 21 concepts. The concepts are those commonly required being mastered for high-school and middle-school students, such as "Function", "Geometry" and "Set". The Math23K is a public dataset which is primarily used for math word problem task in NLP [34]. It contains 23,162 questions for elementary school students. Specifically, questions in Math23K do not have explicit concepts, and are only supplied with mathematical expressions consisting of five elementary operations including addition (+), subtraction (−), multiplication (×), division (÷) and power (∧). Please refer to Wang et al. [34] for more details. Therefore, we treat such operators (5 in total) as the concepts since they can generally reveal the corresponding calculation knowledge.

We focus on several question-based application tasks in online learning systems, and therefore, we get some specific property labels. Specifically, in SYSTEM1, we follow [5, 14, 26] and calculate the difficulty scores of questions, which refers to the correct rate of students. To ensure confidence, we just leave the questions that have more than one hundred students answered, and finally, we get 5,291 questions labeled. In SYSTEM2, we invite three experts (i.e., high school teachers) to label similar questions, where each similar pair would only be left when more than two of them agree with the results. As a result, we get 2,944 questions labeled with several similar ones. In Math23K, we do not have the manual labels while we make the following preprocessing to get the difficulty labels. First, without loss of generality, we think that questions would be more difficult if they have longer mathematical expressions (which means students need more calculation steps to get the answers), and thus we treat the expression length as the difficulty. Second, we select 2,000 questions with difficulty labels in the application task. The difficulty scores are normalized in the range [0, 1].

Table 1 presents the deep statistics of all datasets. There are some observations. First, questions in SYSTEM2 are more difficult for representation learning since they have longer length on average (129.96) than those in other two. Second, our labeled data are limited, as the label sparsities in three datasets are 4.9%, 11.6%, and 8.7%, respectively. Note that, although questions having labels take more than 10% in SYSTEM2, they only have 1.29 similar ones on average, which means that comparing with the total corpus, one question still has very limited (unbalanced) annotations in the tasks.

5.2 DisenQNet Evaluation

In this subsection, we first evaluate DisenQNet, where we aim to show the effectiveness of our learned two disentanglements, i.e., the concept representation v_K and the individual representation v_I. To achieve the goal, we perform the concept prediction, which could be treated as the classification task, with the goal to predict the concepts of questions by model representations.

Experimental Settings. We initialize the DisenQNet as follows. We set the attention network (Eq. (5)), the MI statistics network f_{θ_1} (Eq. (6)), the concept network h_{ϕ_1} (Eq. (7)) and the discriminator D_ϕ (Eq. (8)) all as the MLP with 2 layers. We also pre-train the word2vec [20] tool on all our question corpus to ensure better word embedding. Then, we set the dimension d_0 of word embedding vector, d of both disentangled representations all as 128.

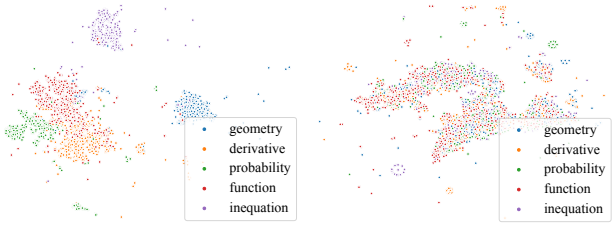
In the training process, we randomly partition all questions into training/test sets with 80%/20%. We follow [11] and set up the model parameters with He initialization. We set the hyper-parameters in Eq. (9) as: λ_1=1, λ_2=1.5, λ_3=2. The learning rates are 0.0002, 0.001, 0.001 in SYSTEM1, SYSTEM2, Math23K, respectively. We set mini batches as 128 for training and used dropout (with probability 20%).

Comparison Methods. Please note that we do not put much emphasis on providing the complicated networks for text classification, but perform the disentangled representation effectiveness for educational questions. Therefore, we introduce baseline models as: one commonly used text model TextCNN, two typical pre-training NLP methods ELMo and BERT, and one SOTA pre-training question representation model QuesNet. We also introduce our two disentanglements (i.e., v_K and v_I) from DisenQNet into evaluation:

- *TextCNN*: TextCNN [16] is a classical textual model for sentence level classification with convolutional neural network.
- *ELMo*: It is a LSTM based feature extraction method with bidirectional language model as pre-training strategy [23].
- *BERT*: It is a popular pre-trained method in NLP featuring Transformer structure and masked language model. As our question content are based on Chinese, we selected Chinese BERT in the experiments [7].
- *QuesNet*: QuesNet [36] is the SOTA pre-trained model for educational question representation learning, with considering heterogeneous data including text, image and side information. We simplify it as just modeling question text.
- *DisenQNet-v_K*: We use the concept representation v_K in our DisenQNet with a 2 layer MLP for prediction.
- *DisenQNet-v_I*: We use the individual representation v_I in our DisenQNet with a 2 layer MLP for prediction.

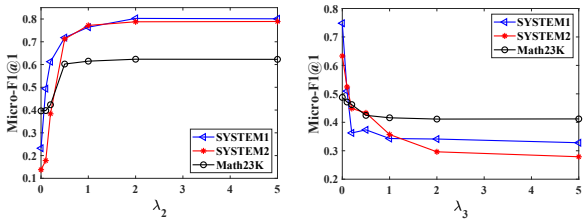
Table 2: Concept prediction performance of all methods on three datasets.

Datasets	SYSTEM1				SYSTEM2				Math23K			
Metrics	Micro-F1@k		Macro-F1@k		Micro-F1@k		Macro-F1@k		Micro-F1@k		Macro-F1@k	
	1	2	1	2	1	2	1	2	1	2	1	2
TextCNN	0.6772	0.5402	0.2287	0.2406	0.6311	0.5407	0.4263	0.4339	0.5001	0.6544	0.3589	0.4926
ELMo	0.6944	0.5622	0.2742	0.2657	0.7702	0.6313	0.6638	0.6329	0.5719	0.7242	0.4366	0.5727
BERT	0.6908	0.5407	0.3875	0.3539	0.7760	0.6352	0.6920	0.6318	0.5906	0.7510	0.5790	0.7210
QuesNet	0.7252	0.6081	0.3291	0.3338	0.7734	0.6321	0.6903	0.6485	0.6236	0.7867	0.4834	0.6818
DisenQNet-v_K	**0.8133**	**0.6498**	**0.3815**	**0.3544**	**0.7996**	**0.6499**	**0.7115**	**0.6655**	**0.6311**	**0.7989**	**0.5654**	**0.7536**
DisenQNet-v_I	0.3672	0.3933	0.1743	0.2228	0.2996	0.3153	0.1941	0.2395	0.4360	0.5916	0.2553	0.3864



(a) Concept representation v_K (b) Individual representation v_I

Figure 4: Disentanglement visualization with concepts.



(a) Concept Representation v_K (b) Individual Representation v_I

Figure 5: Results with different parameters λ_2 and λ_3.

5.2.1 Concept Prediction Performance. We treat the concept prediction task as multi-label classification, as questions in the datasets may related to not only one concept (Table 1). Therefore, we select the widely-used Micro-F1@k and Macro-F1@k as metrics. We repeat all experiments five times and report the average results in Table 2. There are some observations. First, DisenQNet performs the best on all datasets. This demonstrates distinguishing different information, rather than integrating them all, is reasonable for question representation learning. Second, our learned disentangled representations achieve the expected results. Specifically, the concept representations reach the best compared with all since they capture the concept information of questions. Comparatively, the individual ones, which preserves the personal characteristics, fail on this task. Third, we see that the pre-training models (ElMo, BERT, QuesNet) perform better than the traditional TextCNN, which means that they can extract more semantics with their more sophisticated architectures. Last, there is an interesting result that QuesNet, as the domain-specific model for question pre-training, does not perform consistently better than BERT, especially on SYSTEM2. This is possibly because we just use it as the text-only model, so that overlooking some important effects from the heterogeneous data like image. In summary, DisenQNet can distinguish the concept and individual effects for question representation learning.

5.2.2 Disentanglement Visualization. As we mentioned, DisenQNet can disentangle one question into two parts: a concept representation that captures its concept meaning and an individual representation that preserves its personality. Here, we intuitively demonstrate such representation ability. We choose top 5 frequent concepts, and randomly sample 2000 questions for each in SYSTEM1. Then we project their two disentanglements, i.e., v_K and v_I, into 2D space by t-SNE for visualization. We mark questions with their concepts using different colors. Figure 4 shows the results. Generally, questions with same concepts from their concept representations are easily to be grouped, meaning that they have well maintained the concept information. Comparatively, their individual representations are scattered because they are independent with the concept ones that preserve the other personal characteristics.

5.2.3 Parameter Sensitivity. In DisenQNet, both λ_2 and λ_3 in the objective function Eq. (9) play the important role for modeling learning. Specifically, λ_2 regularizes how much information the concept representations capture, while λ_3 controls the degree of personality in questions to be preserved by the individual representations. Figure 5 shows the model performance with both parameters selecting from $\{0, 0.1, 0.2, 0.5, 1, 2, 5\}$. As λ_2 increases, the concept representations perform better and better since they capture concept information of questions as much as possible. However, the individual presentations gradually get lost with the increase of λ_3. This is because they distinguish both concept and personal information of questions, and just leave the personalities if λ_3 is large. Therefore, they work ideally that help DisenQNet learn good disentangled representations for educational questions.

5.2.4 Case Study. DisenQNet can quantify the dominant content on the learned disentangled representations for educational questions via different attention scores in Eq. (5). Here, we visualize attention results of one question example in SYSTEM1 dataset in Figure 6. In the figure, we present the original question text and the English translation on the top. We also mark the words with higher attention scores in its both representations using different colors, i.e., red for the concept representation and blue for the individual one. We can clearly see they focus on different parts. Specifically, the concept representation v_K is more related to four words ("Odd function", "monotonically increasing", "inequality", "solution set") which show the concept meaning. Comparatively, the individual representation v_I concerns more on mathematical expressions (e.g., "$f(-1) = 2$"), which means that expression details can reveal personal characteristics of the question itself.

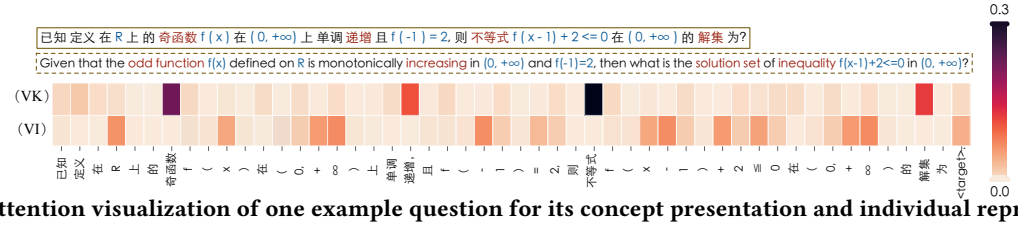


Figure 6: Attention visualization of one example question for its concept presentation and individual representation.

5.3 DisenQNet+ Evaluation

We now evaluate DisenQNet+, where we aim to show the effectiveness of our question representations in the downstream tasks. Based on our datasets, we perform the difficulty estimation task [14, 24] in SYSTEM1 and Math23K, where the goal is to sort the questions from harder ones to the easier under concepts. Then, we perform the similarity analysis task [18, 21, 27] in SYSTEM2, which targets at ranking questions that are similar to one specific.

Experimental Settings. We treat the both as ranking tasks. Therefore, the supervised objective in Eq. (10) can be rewritten as:

$$\mathcal{L}_{Sup} = \sum_{(q,q_+,q_-)\in Q} \max(0, \mu - h_{\phi_2}(v_q, v_{q_+}) + h_{\phi_2}(v_q, v_{q_-})), \quad (13)$$

where μ=1 is the margin hyperparameter. q, q_+, q_- mean the pivot, positive and negative questions. For difficulty ranking (SYSTEM1 and Math23K), we sample question pairs, where q_+ and q_- represent the harder and easier ones (we set q as NULL), so that Eq. (13) lets the estimated score of the positive q_+ be larger than the negative q_-. For similarity ranking (SYSTEM2), given one pivot q, we sample the positive q_+ as its similar questions, and q_- as the others, so that Eq. (13) makes the estimated distance between positive pairs (q, q_+) closer, and separates negative pairs (q, q_-) farther.

In DisenQNet+, we set the MI network f_{θ_2} (Eq. (12)) and the prediction network h_{ϕ_2} (Eq. (13)) as 2-layer MLP. We set d_t=128 for task-specific representation. In the training process, we set λ_4=1, $\lambda_5 \in [0, 0.1]$ in Eq. (10). Other settings are the same with DisenQNet.

We train our unsupervised DisenQNet model on all questions in the datasets. In both tasks, we partition the labeled questions into training/test sets with 20%/80%, 40%/60%, 60%/40%, 80%/20% to show the model robustness with different sparsity ratios.

Comparison Methods. In our work, we aim to show a rigorous comparative analysis of our disentangled question representation effectiveness in a common framework since the task models can be implemented by any ones, as mentioned in Section 4.2. Therefore, we introduce the following comparison models. We use the task model only with labeled data, namely "Supervised". Then, we pre-train ElMo, BERT and QuesNet on all corpus (similar to DisenQNet), and then apply their enhanced representations in the task model. Last, we apply our two disentanglements v_K and v_I in DisenQNet+.

Experimental Results. Figure 7 shows the overall results on all datasets. Specifically, we adopt the ranking metrics including MAP@5, NDCG@5, and F1@5 [19, 33] to evaluate similarity analysis task, but replace F1@5 by DOA metric in difficulty estimation task (We can rank all questions in this task, so we use DOA to measure the result of total lists). We calculate the metric scores on each concept and report the average results of all. Generally, DisenQNet+(v_I) performs the best to improve the results significantly on all datasets. Therefore, it gains the better question representations, where the knowledge from DisenQNet can be transferred more effectively to both tasks. Moreover, it outperforms DisenQNet+(v_K). This demonstrates that the individual disentanglements from DisenQNet, preserving the personality of questions, are more capable of being applied to both tasks because they can distinguish the differences among questions without concept meaning. Then, only using the supervised model does not generate satisfactory results since it cannot ensure the representation ability with limited labeled data. Last, although traditional pre-training models (ElMo, BERT, QuesNet) improve the results, they do not perform as well as ours because they may introduce noises by integrating all question information together in application tasks. Our DisenQNet+ has potentials to support several online educational services.

6 CONCLUSION AND FUTURE WORK

In this paper, we learned disentangled representations of educational questions. We proposed an unsupervised model DisenQNet that divided one question into two parts: a concept representation which captured its explicit concept meaning and an individual representation which preserved its personal characteristics. We also proposed DisenQNet+ to transfer the representation knowledge from DisenQNet in several application tasks including difficulty estimation and similarity analysis. Experimental results showed that DisenQNet could distinguish unique concept and personality effects for question representation learning, and DisenQNet+ improved task performances by incorporating our individual representations.

There are some directions for further studies. First, we will perform representation learning for educational questions with heterogeneous forms, which some geometry figures can be incorporated. Second, we will design more meaningful question-based online intelligent services. We hope this work could lead to more studies.

Acknowledgements. This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. 61922073 and U20A20229), the Fundamental Research Funds for the Central Universities (Grant No. WK2150110021), the Foundation of State Key Laboratory of Cognitive Intelligence (No. iED2020-M004), and the iFLYTEK joint research program.

REFERENCES

[1] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*. 687–698.
[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.
[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International Conference on Machine Learning*. PMLR, 531–540.
[4] Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. Introducing a framework to assess newly created questions with Natural Language Processing. In *International Conference on Artificial Intelligence in Education*. Springer, 43–54.
[5] Markus Broer. 2005. Ensuring the fairness of GRE writing prompts: Assessing differential difficulty. *ETS Research Report Series* 2005, 1 (2005), i–41.

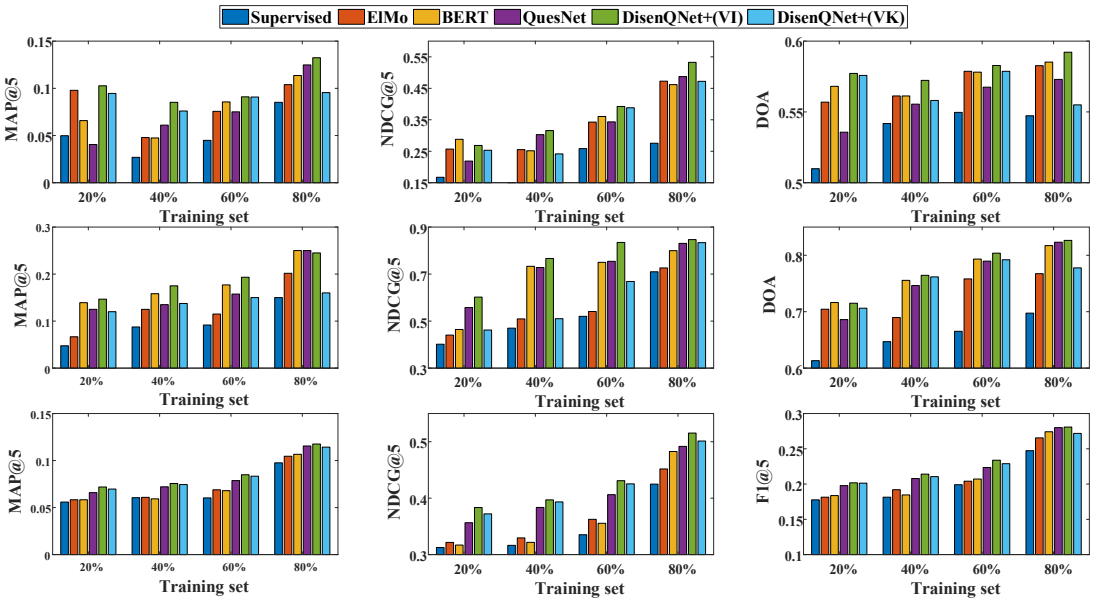


Figure 7: Task results: Difficulty estimation on SYSTEM1 (top), Math23K (medium). Similarity analysis on SYSTEM2 (bottom).

[6] Youngduck Choi, Youngnam Lee, Junghyun Cho, Jineon Baek, Dongmin Shin, Seewoo Lee, Youngmin Cha, Byungsoo Kim, and Jaewe Heo. 2020. Assessment Modeling: Fundamental Pre-training Tasks for Interactive Educational Systems. *arXiv preprint arXiv:2002.05505* (2020).

[7] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101* (2019).

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

[9] Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. 2008. Searching questions by identifying question topic and question focus. In *Proceedings of Acl-08: HLT*. 156–164.

[10] Kaiming He, Ross Girshick, and Piotr Dollár. 2019. Rethinking imagenet pre-training. In *IEEE/CVF ICCV*. 4918–4927.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IProceedings of the IEEE international conference on computer vision*. 1026–1034.

[12] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

[13] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).

[14] Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question Difficulty Prediction for READING Problems in Standard Tests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[15] Zhenya Huang, Qi Liu, Weibo Gao, Jinze Wu, Yu Yin, Hao Wang, and Enhong Chen. 2020. Neural mathematical solver with enhanced formula structure. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1729–1732.

[16] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*. 1746–1751.

[17] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E* 69, 6 (2004), 066138.

[18] Qi Liu, Zai Huang, Zhenya Huang, Chuanren Liu, Enhong Chen, Yu Su, and Guoping Hu. 2018. Finding similar exercises in online education systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1821–1830.

[19] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering* 33, 1 (2019), 100–115.

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

[21] Radek Pelánek. 2019. Measuring similarity of educational items: An overview. *IEEE Transactions on Learning Technologies* 13, 2 (2019), 354–366.

[22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[23] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[24] Zhaopeng Qiu, Xian Wu, and Wei Fan. 2019. Question difficulty prediction for multiple choice problems in medical exams. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 139–148.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[26] Mark D Reckase. 2009. Multidimensional item response theory models. In *Multidimensional item response theory*. Springer, 79–112.

[27] Jirí Rihák and Radek Pelánek. 2017. Measuring Similarity of Educational Items Using Data on Learners' Performance. *International Educational Data Mining Society* (2017).

[28] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. 2020. Learning disentangled representations via mutual information estimation. In *European Conference on Computer Vision*. Springer, 205–221.

[29] Aravind Sankar, Yanhong Wu, Yuhang Wu, Wei Zhang, Hao Yang, and Hari Sundaram. 2020. GroupIM: A Mutual Information Maximization Framework for Neural Group Recommendation. In *ACM SIGIR*. 1279–1288.

[30] Norbert Ed Schwarz and Seymour Ed Sudman. 1996. *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. Jossey-Bass/Wiley.

[31] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. 2019. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000* (2019).

[32] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep Graph Infomax.. In *ICLR (Poster)*.

[33] Hao Wang, Tong Xu, Qi Liu, Defu Lian, Enhong Chen, Dongfang Du, Han Wu, and Wen Su. 2019. MCNE: An end-to-end framework for learning multiple conditional network representations of social network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1064–1072.

[34] Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 845–854.

[35] Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 2020. Predicting item survival for multiple choice questions in a high-stakes medical exam. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 6812–6818.

[36] Yu Yin, Qi Liu, Zhenya Huang, Enhong Chen, Wei Tong, Shijin Wang, and Yu Su. 2019. Quesnet: A unified representation for heterogeneous test questions. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1328–1336.

[37] Jing Zhang and Xindong Wu. 2018. Multi-label inference for crowdsourcing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2738–2747.