The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2021)

DisenQNet: Disentangled Representation Learning for Educational Questions

Zhenya Huang, Xin Lin, Hao Wang, Qi Liu, Enhong Chen*, Jianhui Ma, Yu Su, Wei Tong

Unversity of Science and Technology of China,

iFLYTEK Co., Ltd

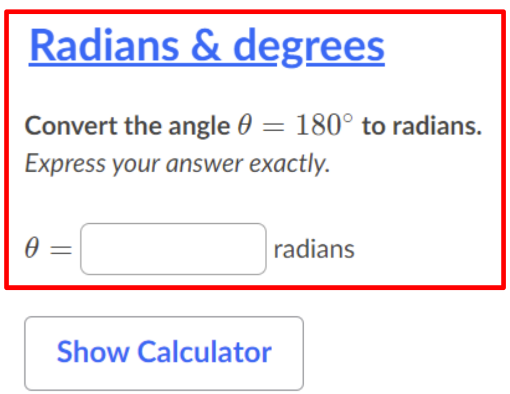
Outline

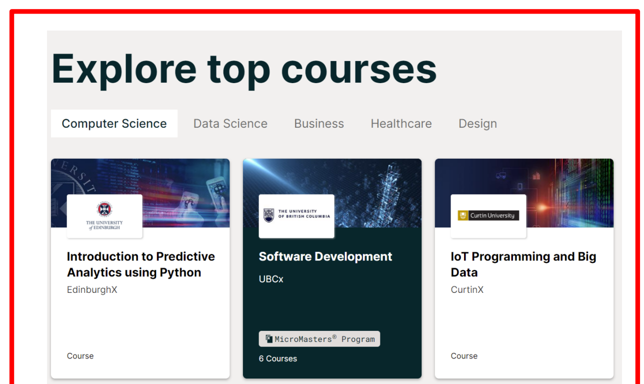
- **Introduction**
- Preliminary
- Our Method
- Experiment
- Conclusion

Introduction

- Online learning systems
  - Collect millions of learning materials
    - Course, question, test, etc
  - Provide intelligent services to improve learning experience
    - Students select suitable questions or courses to acquire knowledge
    - Systems provide personalized recommendations

Question





Course

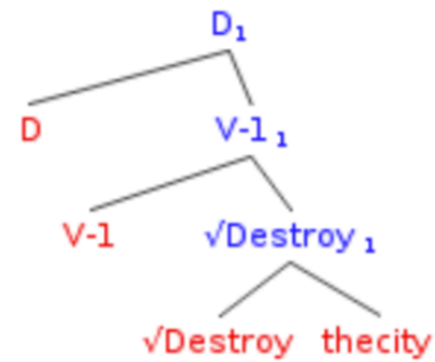
Introduction

- Real world challenges with millions of learning materials
  - How to organize, search and recommend questions?
  - How to promote question-based applications?
    - Search questions to find similar ones
    - recommend questions with property difficulty (for students)
    - ……

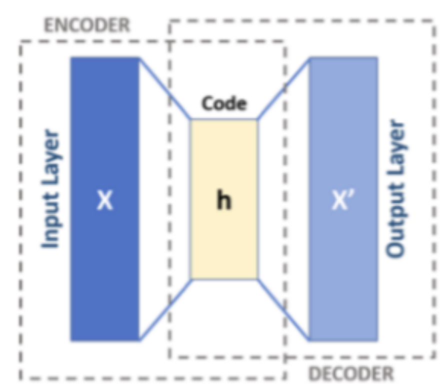
- Fundamental topic in AI education
  - **Question understanding** (automatic)
  - Goal: learning informative representations of question

Related work

- Traditional NLP work (earlier)
  - Lexical analysis or Semantic analysis
    - Design fine-grained rules or grammars
  - Representation: explicit trees or templates

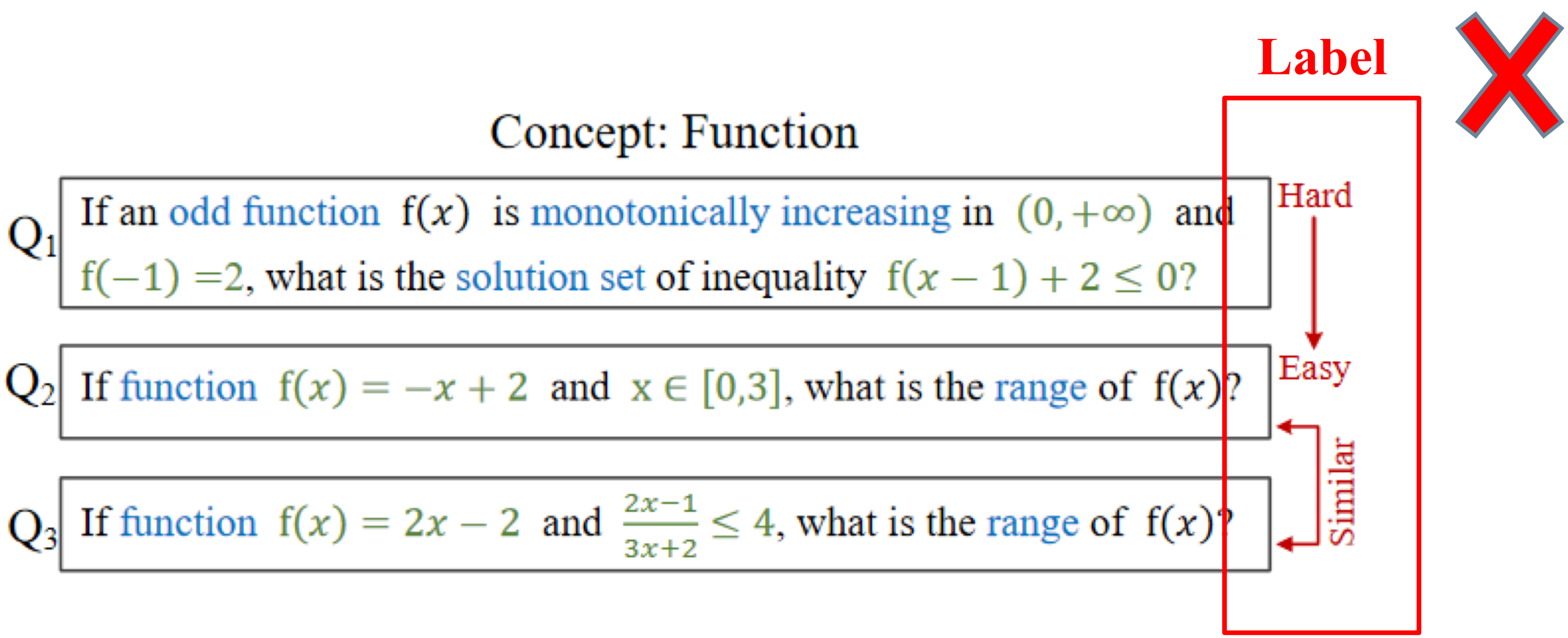


- NLP based work
  - End-to-end frameworks
    - Understand question content
    - Learn from application tasks, e.g., difficulty estimation, similarity search
  - Representation: latent semantic vector
- Recent Pre-training work
  - Pre-training with large question corpus
    - Enhance question semantics learning
  - Representation: latent semantic vector



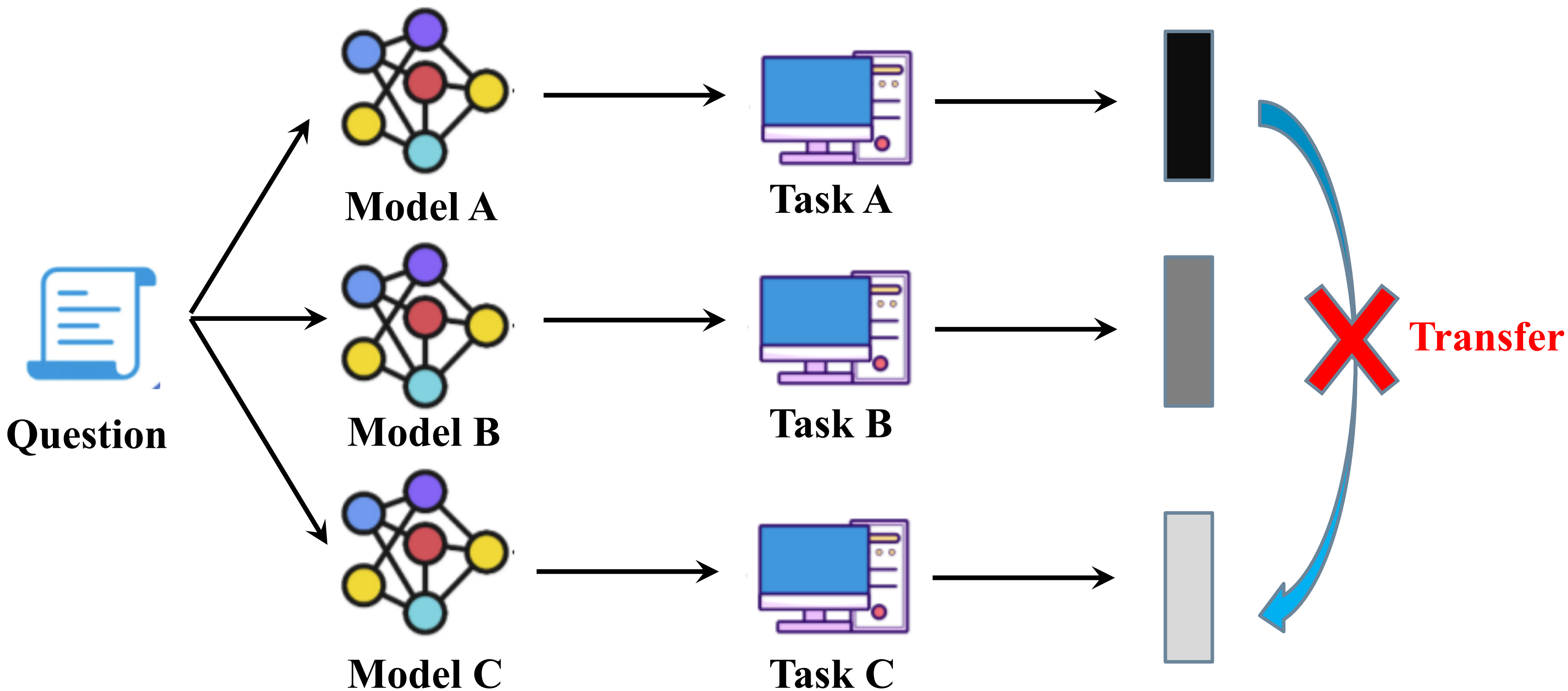
Related work—Limitations

- **Supervised** manner
  - Requiring sufficient labeled data
    - E.g., question difficulty, question pair similarity
  - Scarcity of labels with high quality
    - E.g., difficulty is being examined in standard tests (GRE)



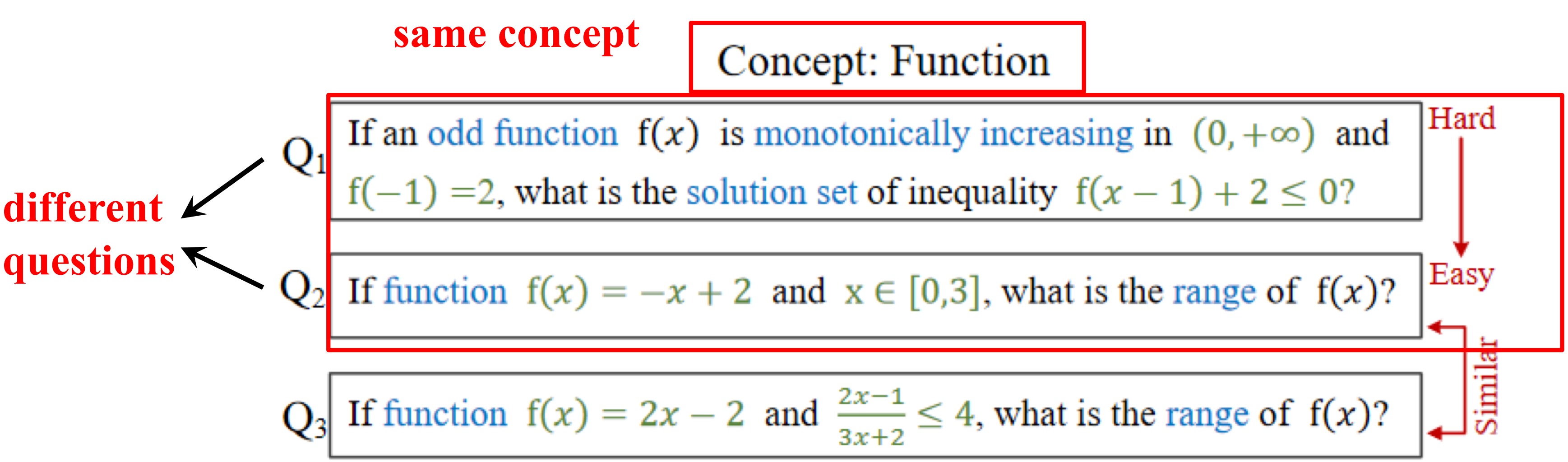
Related work—Limitations

- **Task-dependent** representation
  - Different models for same questions in different application tasks
  - Poor transferability across tasks



Related work—Limitations

- **One unified** vector representation
  - All the information are integrated together
  - Question with same concept are quite different
    - **Concept**
    - **Personal properties (difficulty, semantics)**



Introduction

- Ideal question representation model
  - Get rid of **labels** in specific tasks
    - Try to learn information of question **on their own**
  - **Distinguish** the different characteristics of questions
    - Reduce noise
    - Explicit way to get **good interpretability**
  - Question representations should be **flexible**
    - Can be applied in different downstream tasks
    - **Improve the applications** in online learning systems

Our work—main idea

- **Disentangled representation**
  - Disentangle question information into two representations
    - **Concept** representation
    - **Individual** representation
  - Concept representation
    - high dependency to concept information (knowledge)
  - Individual representation
    - high dependency to individual information (difficulty, semantics, et al.)
  - Two representations with high independency to each other
    - Contain no information from each other

Outline

- Introduction
- **Preliminary**
- Our Method
- Experiment
- Conclusion

Problem definition

- Unsupervised question representation Learning
  - **Given**: $Q = \{q_1, q_2, \dots, q_N\}$, $q = \{x_1, x_2, \dots, x_M\}$ with $k \in K$
  - **Goal**: disentangled question representation
    - Concept representation $v_K \in R^d$
    - Individual representation $v_I \in R^d$
- Question-based supervised tasks
  - **Given** $Q = Q^L \cup Q^U$ and $|Q^L| \ll |Q^U|$
    - Labeled $Q^L = \{q_1, q_2, \dots, q_L\}$ with $\{y_1, y_2, \dots, y_L\}$
    - Unlabeled $Q^U = \{q_1, q_2, \dots, q_U\}$
  - **Goal**: predict properties of unknown questions
    - e.g., difficulty of one question
    - e.g., similarity of question pair

Concept: Function

Q_1: If an odd function $f(x)$ is monotonically increasing in $(0, +\infty)$ and $f(-1) = 2$, what is the solution set of inequality $f(x-1) + 2 \leq 0$?

Hard

Easy

Q_2: If function $f(x) = -x + 2$ and $x \in [0,3]$, what is the range of $f(x)$?

Similar

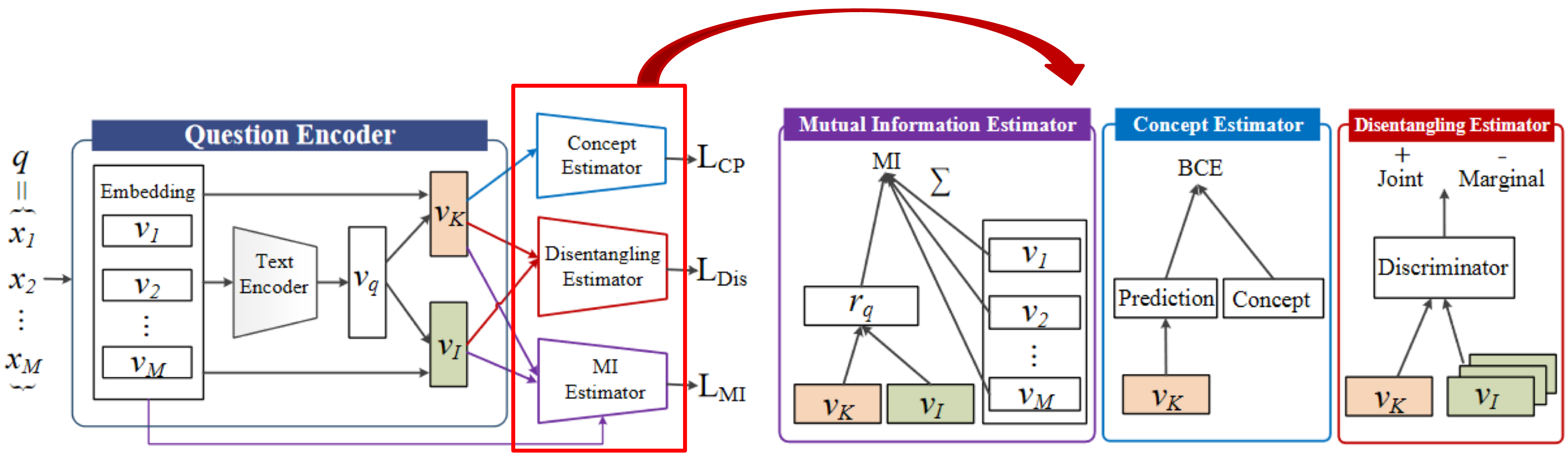
Q_3: If function $f(x) = 2x - 2$ and $\frac{2x-1}{3x+2} \leq 4$, what is the range of $f(x)$?

Outline

- Introduction
- Preliminary
- **Our Method**
- Experiment
- Conclusion

DisenQNet: glance

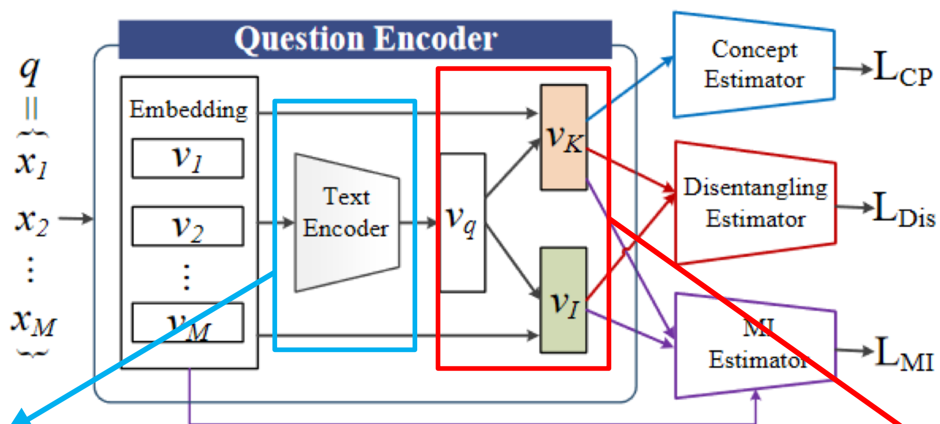
- Disentangled Question Network (DisenQNet)
  - Unsupervised model without labels
  - Question encoder
    - Learn to disentangle one question into two ideal representations
  - Self-supervised optimization
    - Three information estimators



Model architecture **Model optimization**

DisenQNet

- Question Encoder
  - Learn to disentangle one question into two ideal representations
    - Concept representation v_K
    - Individual representation v_I
  - Key: they focus on different content
    - e.g., concept: “function”, individual: “f(-1)=2”



Integrated semantics

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

e.g., **TextCNN**, CNN, LSTM.

Disentangled semantics

Attention Network

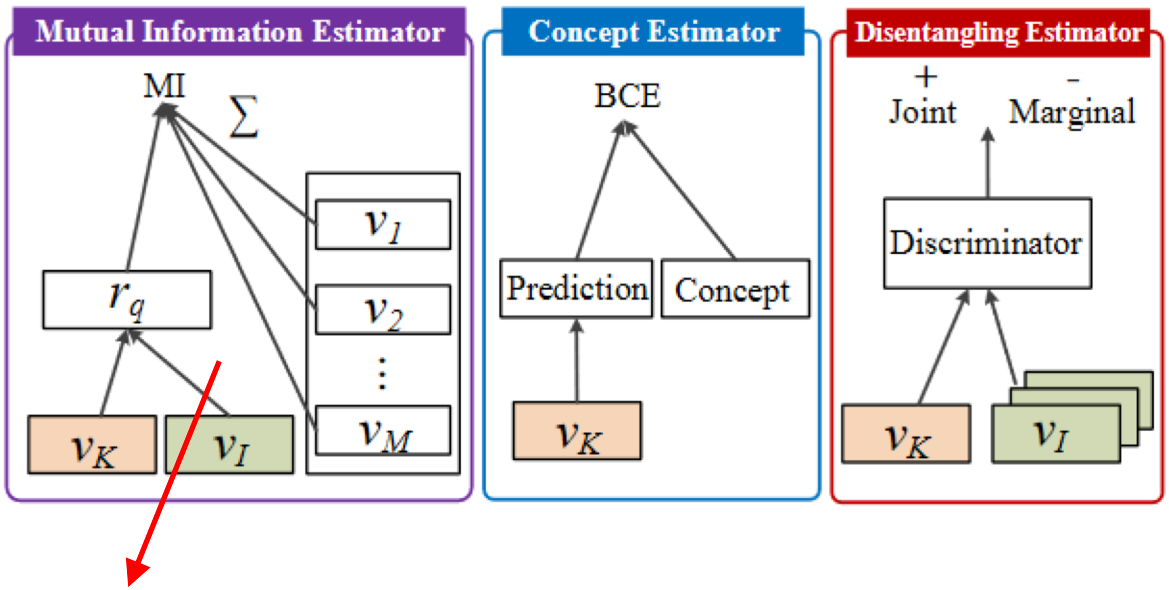
$r_q = (v_K, v_I)$

$v_K = \sum_{j=1}^{M} \alpha_j v_j$

$\alpha_j = \mathrm{Softmax}\left(\mathrm{MLP}(v_j, v_q)\right)$

DisenQNet

- How to optimize? — Self-supervised optimization
  - Three estimators to measure the information dependency



MI Estimator

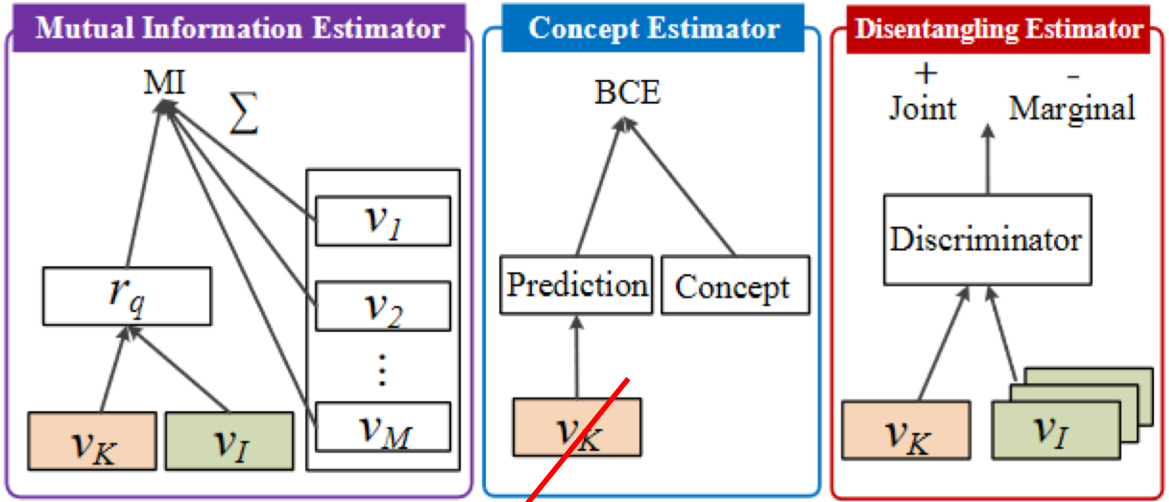
$r_q = (v_K, v_I)$ contains all information of one question

➢ maximize **MI** between $\boldsymbol{r_q}$ and each word $\boldsymbol{v_i} \in \boldsymbol{V}$

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

DisenQNet

- How to optimize? — Self-supervised optimization
  - Three estimators to measure the information dependency



Concept Estimator

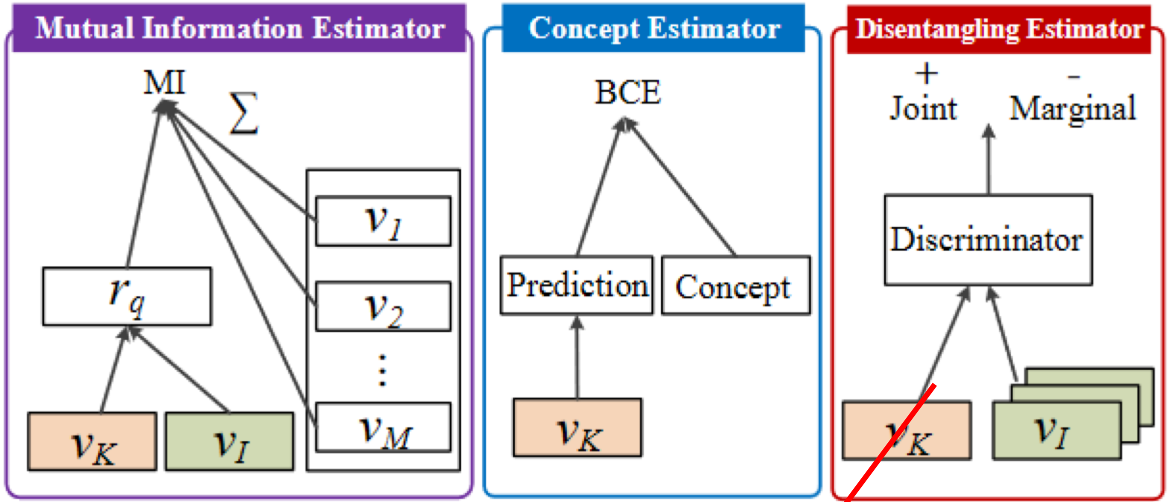
v_K contains the given concept meaning explicitly

- Multi-label concept classification task: predict concepts

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

DisenQNet

- How to optimize? — Self-supervised optimization
  - Three estimators to measure the information dependency



Disentangling Estimator

- Keep v_I and v_K **independent**: v_I must not contain the information by v_K
- Minimize the mutual information between v_I and v_K (cannot directly learn)
- Method: WGAN-like **adversarial** training

  Minimize Wasserstein distance between $P(v_I, v_K)$ and $P(v_I) \otimes P(v_K)$

  $=> P(v_I) \otimes P(v_K) \approx P(v_I, v_K) =>$ **Independent** v_I and v_K

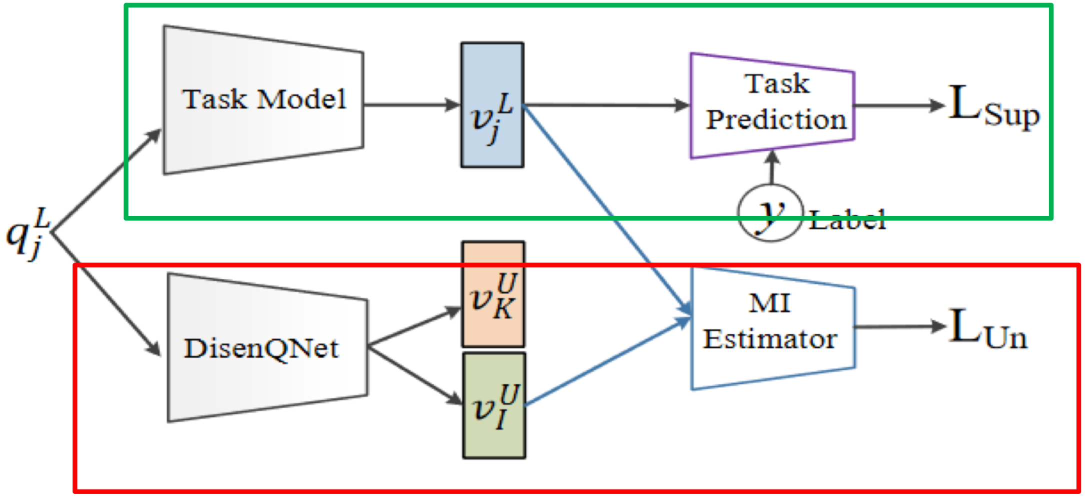
$$\mathcal{L}_{Dis} = \mathbb{E}_{\mathbb{P}(v_k, v_i)}\left[D_\phi(v_k, v_i)\right] - \mathbb{E}_{\mathbb{P}(v_k)\mathbb{P}(v_i)}\left[D_\phi(v_k, v_i)\right].$$

DisenQNet+

- DisenQNet+ — Question-based supervised tasks
  - Transfer v_I from DisenQNet to improve different applications
    - e.g., difficulty estimation, similarity search
  - Key: individual v_I focus more on **unique** information

Traditionally: end-to-end method

➢ may suffer from overfitting due to **insufficient data**



Our improve: force v_I from DisenQNet to task model via mutual information maximization

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

Outline

- Introduction
- Preliminary
- Our Method
- **Experiment**
- Conclusion

Experiment

- Dataset
  - System1: high school level questions
  - System2: middle school level questions
    - Concepts: “Function”, “Triangle”, “Set”, etc
  - Math23K: elementary school level questions
    - Concepts (five operations): $+, -, \times, \div, \wedge$

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%

Concept: Function

Q_1 If an odd function $f(x)$ is monotonically increasing in $(0,+\infty)$ and $f(-1)=2$, what is the solution set of inequality $f(x-1)+2 \leq 0$?

Hard

Easy

Q_2 If function $f(x)=-x+2$ and $x \in [0,3]$, what is the range of $f(x)$?

Similar

Q_3 If function $f(x)=2x-2$ and $\frac{2x-1}{3x+2} \leq 4$, what is the range of $f(x)$?

Problem: Robin was making baggies of cookies with 6 cookies in each bag. If she had 23 chocolate cookies and 25 oatmeal cookies, how many baggies could she make?
Expression: $x = (23+25) \div 6$
Answer: 8

Experiment

DisenQNet Evaluation (v_K and v_I)

- Task: Concept Prediction Performance
- Baseline: Text model, NLP pre-trained models, question pre-trained model

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

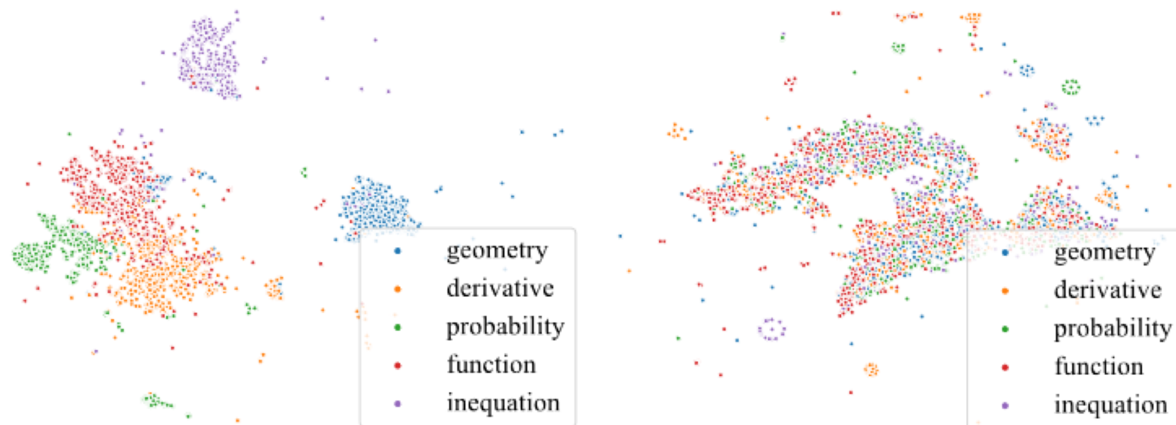
Disentangled representation learning is necessary

- **DisenQNet-v_K** is well predicted: v_K capture the concept information of questions
- **DisenQNet-v_I** fails to predict concepts: v_I removes the concept information

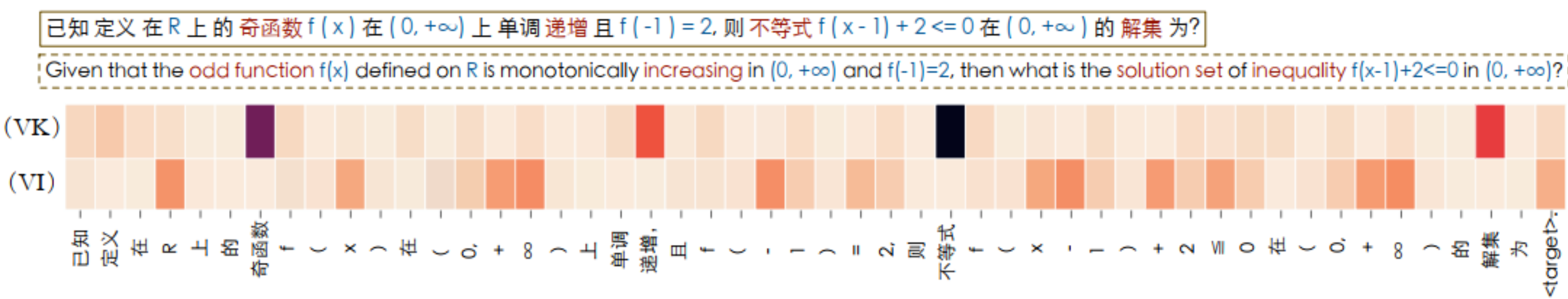
Experiment

□ DisenQNet Visualization (v_K and v_I)

- v_K are easier to be grouped by concepts
- v_I are scattered



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



- v_K is more related to **concept words** (“Odd function”, “solution set”, “inequality”)
- v_I focuses more on **mathematical expressions** (“f (-1) = 2”)

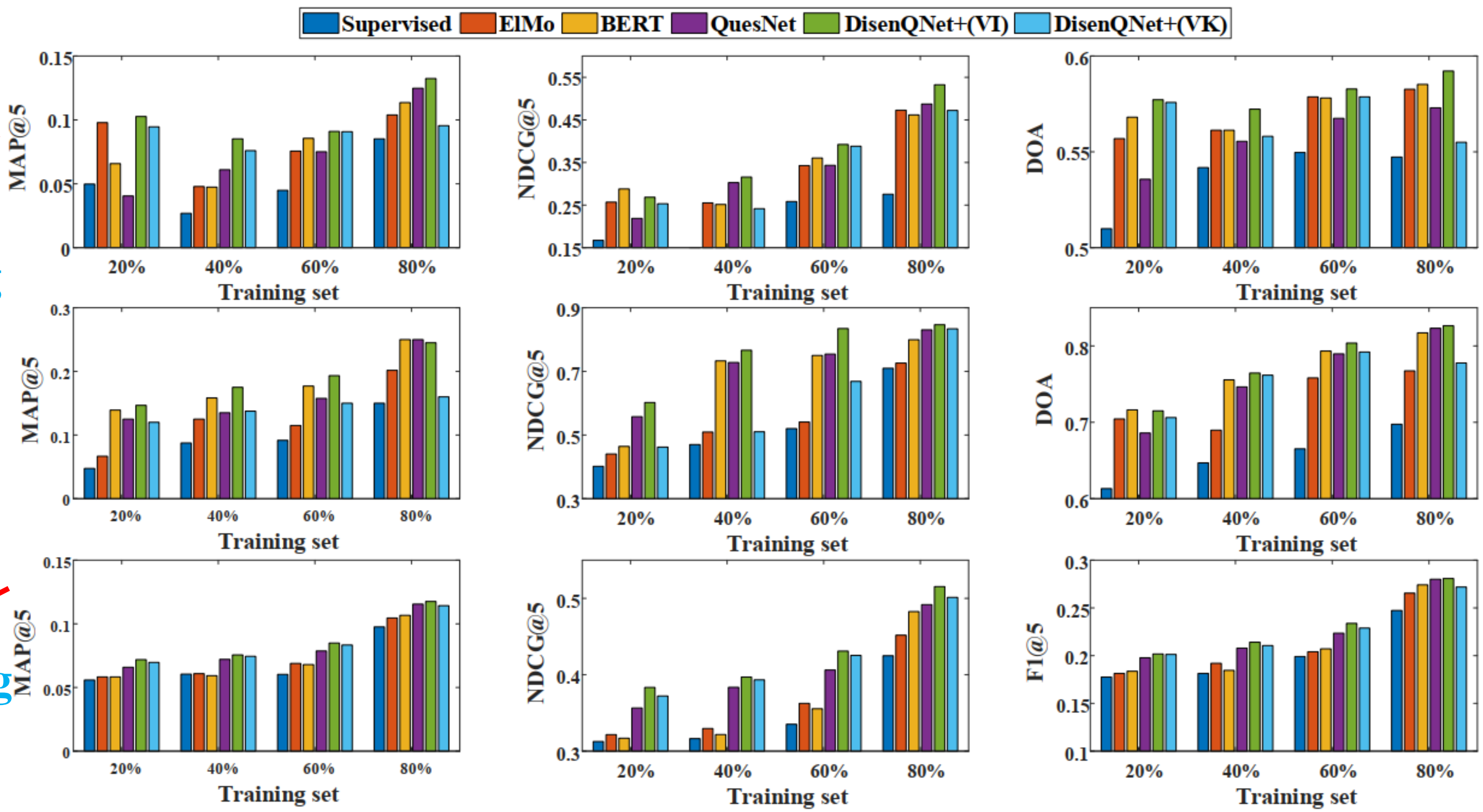
Experiment

DisenQNet+ evaluation

Two tasks

Difficulty ranking

Similarity ranking



- Disentangled learning is better than integrated learning
- v_I improves the application performance (best)
  - It can preserve personal information of questions
  - It has good ability to be transferred across different tasks

Outline

- Introduction
- Preliminary
- Our Method
- Experiment
- **Conclusion**

Conclusion

- **Summary**
  - Disentangled representation learning for educational questions
  - Unsupervised DisenQNet
    - Distinguish concept and individual information of questions
    - Good interpretability
  - Semi-supervised DisenQNet+
    - Improve the performance of different tasks
    - Good transferability

- **Future work**
  - More sophisticated models for disentanglement implementation
  - Heterogeneous questions, e.g., geometry
  - Deeper knowledge transferring

Thanks!

huangzhy@ustc.edu.cn
linx@mail.ustc.edu.cn