Fine-Grained Similarity Measurement of Educational Videos and Exercises

Xin Wang[1], Wei Huang[1], Qi Liu[1,*], Yu Yin[1], Zhenya Huang[1], Le Wu[2], Jianhui Ma[1], Xue Wang[3]
[1]University of Science of Technology of China
[2]Hefei University of Technology
[3]Nankai University

Scan to view details!

Speaker: Xin Wang

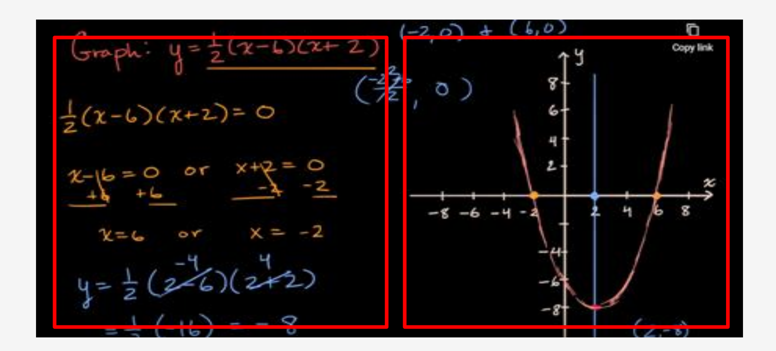
Email:
shenai@mail.ustc.edu.cn

Lab Homepage:
http://bigdata.ustc.edu.cn

Introduction

Motivation

- Measuring the similarity between educational videos and exercises is a fundamental task with **broad application prospects**.
- In most cases, an exercise is only similar to parts of the video. Therefore, it would be of great significance to application and user experience if we could further measure the similarity at segment-level, which we call **fine-grained similarity measurement**.
- The problem remains pretty much **open** due to several domain-specific challenges.
- Thus, fully considering the effects of multimodal information, we proposed the **VENet** to measure the similarity at both video-level and segment-level by just exploiting the coarse-grained labeled data on videos.

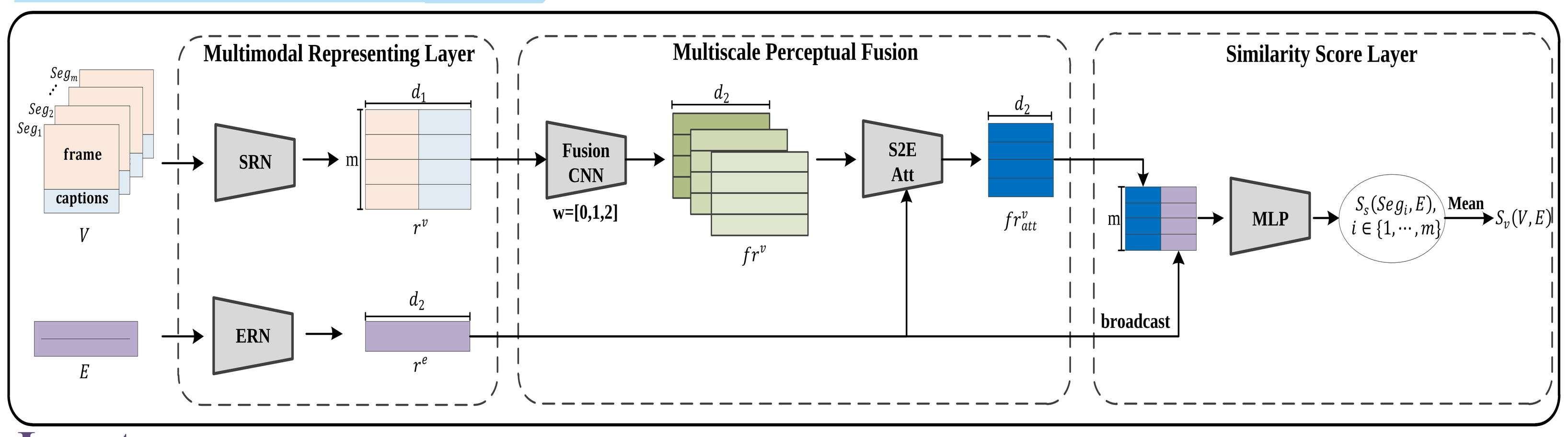


Challenges

- Educational videos contain not only graphics but also text and formulas, which have a fixed reading order. How to model the **spatial structure** and **temporal information?**
- How to perceive and incorporate the **semantic associations** among segments?
- How to learn the fine-grained similarity by just exploiting the coarse-grained labeled data on videos?

VENet

Framework



Input

Video: $V = \{Seg_1, \dots, Seg_m\}, Seg_i = \{f_i, c_i\}, c_i = \{w_{i1}, \dots, w_{it}\}$

Exercise: $E = \{w_1, \dots, w_n\}$

Output

$S_s(Seg_i, E)$: The similarity score between the i-th segment and the exercise.

$S_v(V, E)$: The similarity score between the whole video and the sxercise.

Submodule

SRN

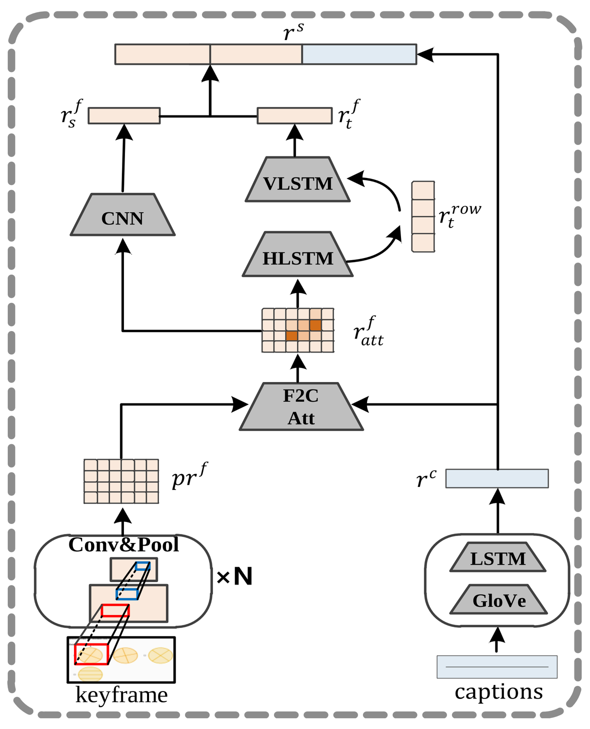
- Encode the **multimodal data** (keyframe and captions) of the segment into the segmantic vector.
- Align the keyframe and captions by **F2C Attention**.
- Model the **spatial structure** and **temporal information** embedded in the keyframe.

ERN

- Initialize the word embedding with **GloVe**.
- Model the word sequence of the exercise by **LSTM**.

MPF

- Fuse adjacent segments on multiple scales.
- Weight the fusional vectors according to the exercise by **S2E Attention**.



SRN

Fusion

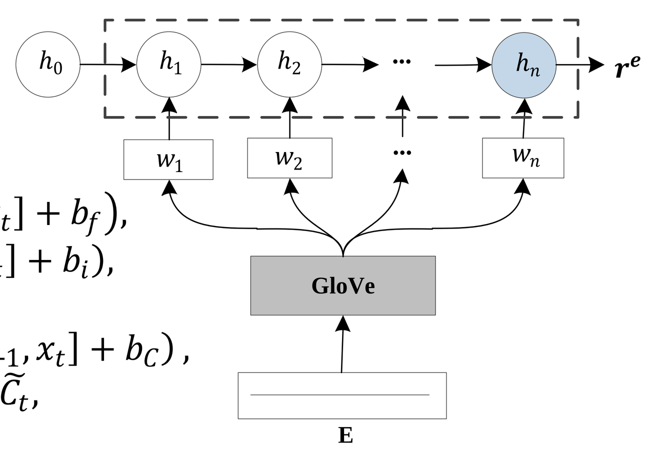
$$\widetilde{C}_i = [r^v_{i-w}, \cdots, r^v_i, \cdots, r^v_{i+w}],$$
$$fr^v_i = ReLU(W_{fuse}\,\widetilde{C}_i + b_{fuse}).$$

Attention

$$A = \sum_i \alpha_i \cdot v_i\,,$$
$$\alpha_i = \frac{\phi(v_i, q)}{\sum_i \phi(v_i, q)},$$
$$\phi(v_i, q) = \exp(V_* \cdot \tanh(W_* \cdot [v_i, q]))\,.$$

LSTM

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f),$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i),$$
$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)\,,$$
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t,$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o),$$
$$h_t = o_t \cdot \tanh(C_t)\,.$$



ERN

Experiments

Datasets

- There is no public data for this task. So we collect the real-word data from Khan Academy and show how to create a dataset using publically available educational services.
- All of our data is crawled from the math domain, which contains 17,116 math exercises and 1,053 educational videos with closed captions, covering 836 topics.
- We crawl 10,679 similar video-exercise pairs and build an equal number of dissimilar pairs by negative sampling.

Main Results

Model	Input: Text	Input: Frame	Task: Video-Level	Task: Segment-Level
MaLSTM	✓	×	✓	×
DeepLSTM	✓	×	✓	×
ABCNN	✓	×	✓	×
TextCNN	✓	×	✓	×
DeepLSTM (Seg)	✓	×	✓	✓
TextCNN (Seg)	✓	×	✓	✓
TextualVENet	✓	×	✓	✓
3DCNN	✓	✓	✓	×
JSFusion	✓	✓	✓	×
EarlyFusion	✓	✓	✓	✓
VENet	✓	✓	✓	✓

Model	Video-Level Auc	Video-Level NDCG	Segment-Level Auc	Segment-Level NDCG
MaLSTM	0.591	0.635	-	-
DeepLSTM	0.778	0.7503	-	-
ABCNN	0.764	0.7448	-	-
TextCNN	0.792	0.771	-	-
DeepLSTM (Seg)	0.844	0.7728	0.754	0.7437
TextCNN (Seg)	0.806	0.7658	0.7418	0.7415
TextualVENet	0.876	0.832	0.768	0.781
3DCNN	0.654	0.742	-	-
JSFusion	0.826	0.788	-	-
EarlyFusion	0.854	0.7806	0.7863	0.7494
VENet	**0.942**	**0.879**	**0.871**	**0.823**

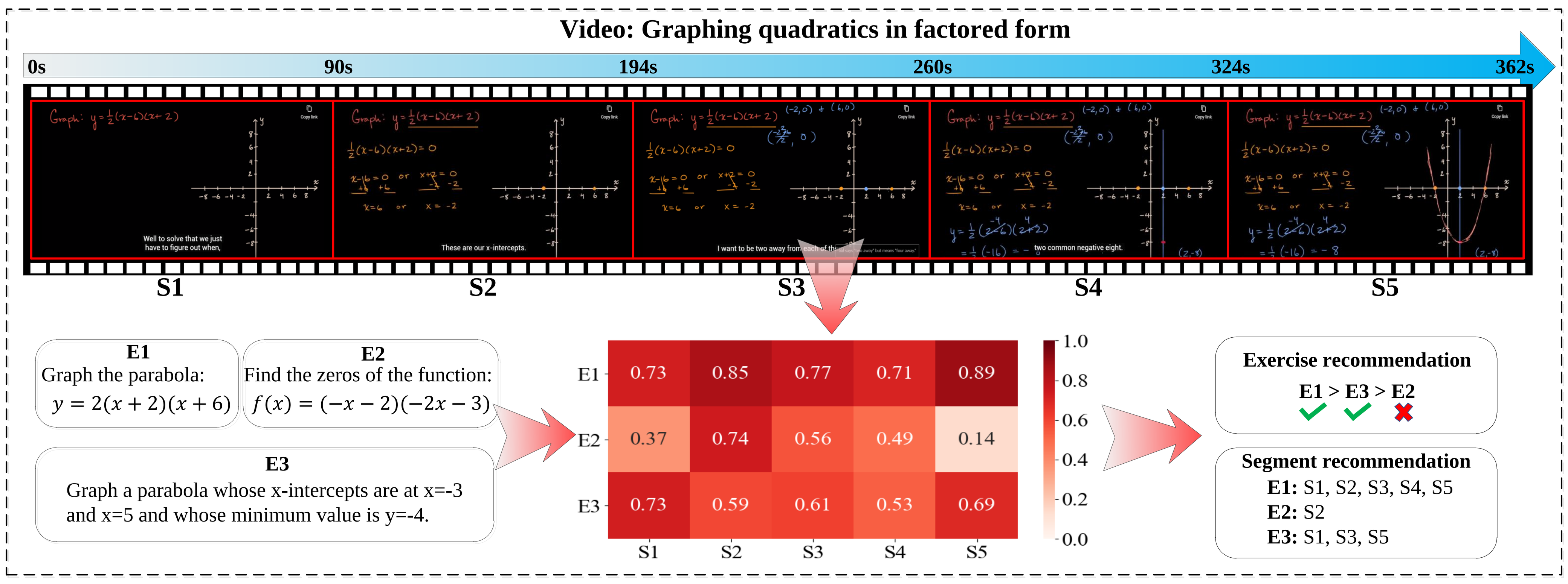
- Our proposed VENet achieves the best performance at both video-level and segment-level, with a significant improvement on all metrics compared to other methods.
- In all seven textual models, TextualVENet also performs best.
- Comparing DeepLSTM and DeepLSTM (Seg), we can find that dividing video into segments can improve the performance at video-level significantly.
- The performance of TextualVENet is worse than that of VENet, which shows that the visual data is helpful to accurately understand the video.

Ablation Experiments

Model	Video-Level Auc	Video-Level NDCG	Segment-Level Auc	Segment-Level NDCG
TextualVENet	0.876	0.832	0.768	0.781
VisualVENet	0.624	0.7328	0.6324	0.6931
VENet	**0.942**	**0.879**	**0.871**	**0.823**
VENet-F2C	0.9	0.855	0.8284	0.8198
VENet-S2E	0.91	0.851	0.846	0.8137
VENet-HVLSTM	0.89	0.802	0.803	0.795
VENet-MPF	0.866	0.815	0.789	0.7616

- The performance of VisualVENet is much worse than TextualVENet, which indicates the textual material is more important than the visual data.
- All the key modules (i.e., F2C, S2E, HVLSTM and MPF) have a significant impact on the final result, which shows the effectiveness of them.
- MPF has the biggest impact on the final results.

Case Study



Conclusion

- We explore the promising yet challenging problem of measuring the **fine-grained similarity** between educational videos and exercises by just exploiting the coarse-grained labeled data.
- We propose a novel **VENet** which make full use of visual and textual data to measure the fine-grained similarity accurately.
- We conduct extensive experiments on real-world datasets. The experimental results demonstrate that our proposed VENet outperforms other state-of-the-art methods.

Future

- Collect data and conduct experiments to test the performance on other subjects, such as Physics.
- Consider exploiting other meta information to enhance video understanding, such as topics and titles.