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 and Technology of China
[2]Hefei University of Technology
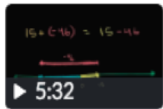[3]Nankai University

01 Introduction

Introduction

➤ Related Content Recommendation

Related content recommendation on Khan Academy

▪|▪ Introduction

➢ **Partial Similar**

An example from Khan Academy

Graphing Quadratics

S_1: find zeros | S_2: find vertice | S_3: Graphing

E_1	E_2	E_3
Find the zeros of the function: $f(x) = \frac{1}{2}(-x-2)(2x-3)$	Graph the function: $f(x) = \frac{2}{3}(x+1)(x-5)$	Graph a parabola whose x intercepts are at x=-3 and x=5 and whose minimum value is y=-4.

Q1: Are they similar?
Q2: Which segments are similar to this exercise? (fine-grained)

Introduction

> **Fine-Grained Similarity Measurement**

Segments	Similarity
S1	0.72
S2	0.43
S3	0.61
...	...

Input   **Model**   **Output**

02 Research Contents

📊 Challenge 1

➢ How to model the multimodal segment?
  • Captions
  • Keyframes

➢ Spatial and Temporal Information
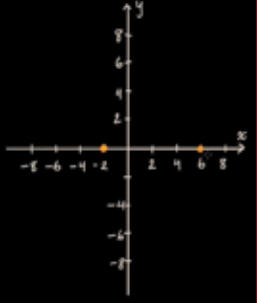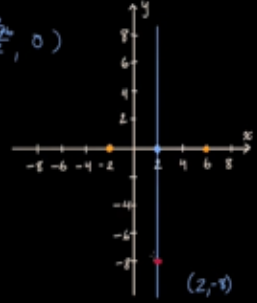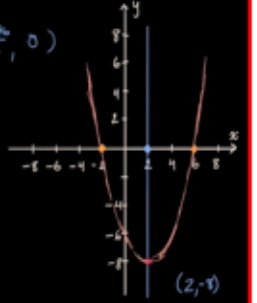
Segment Representation Network

Challenge 2

➤ Semantic Associations among Video Segments

Graphing Quadratics

S_1: find zeros	S_2: find vertice	S_3: Graphing

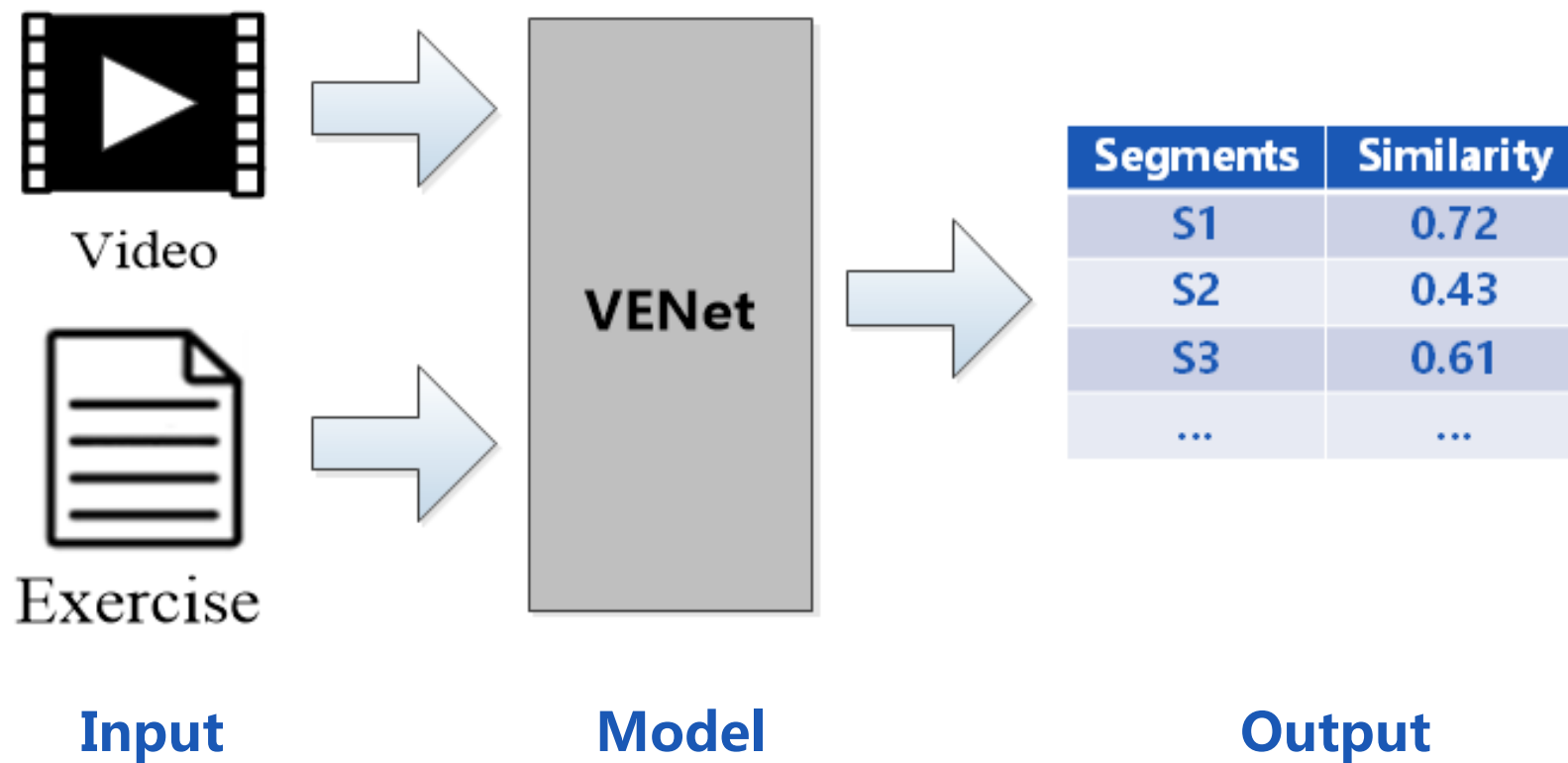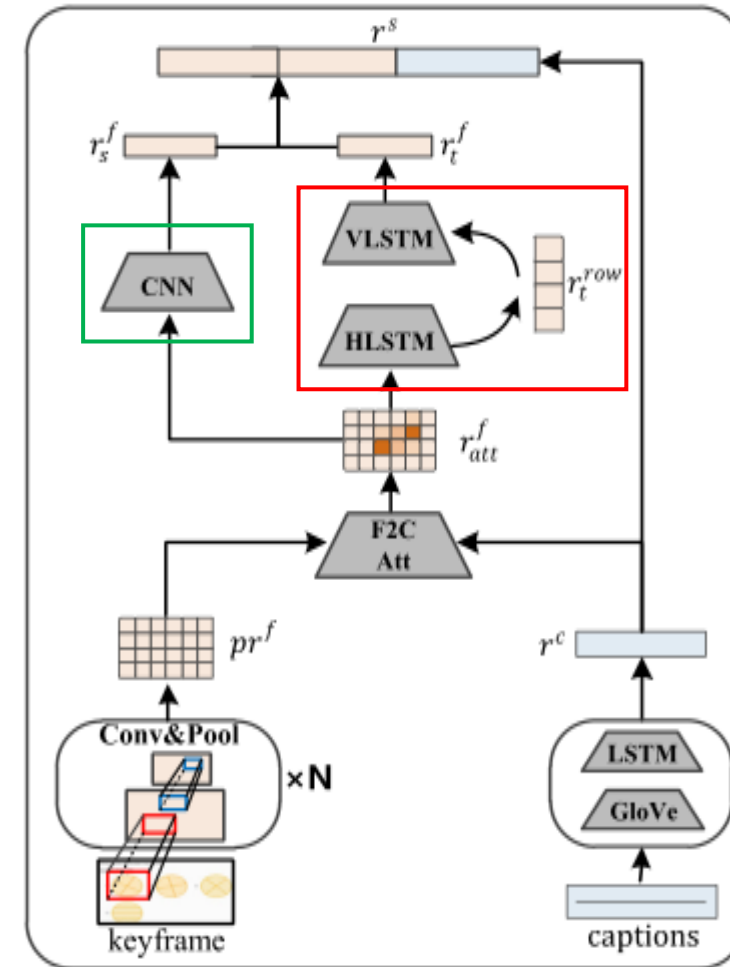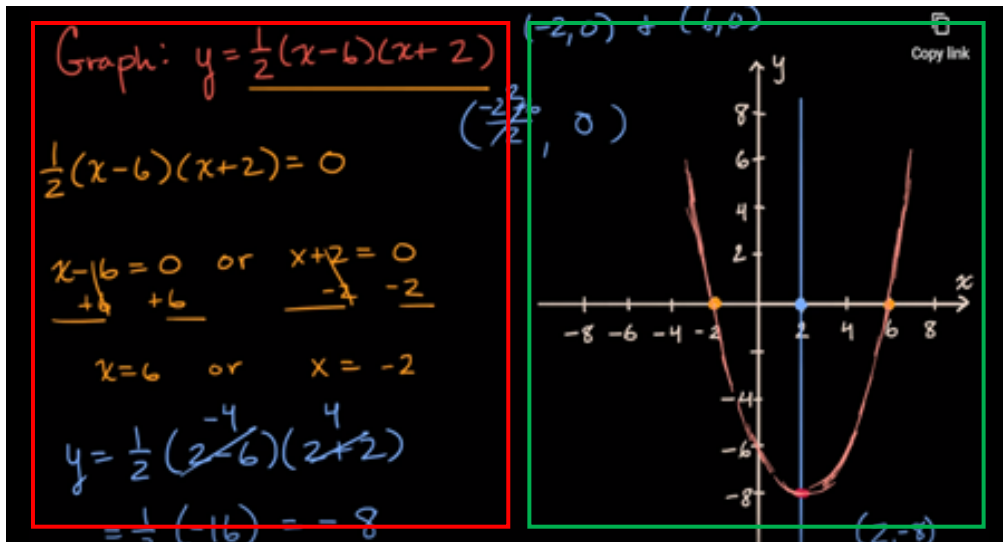| E_1	E_2	E_3
Find the zeros of the function: $f(x) = \frac{1}{2}(-x - 2)(2x - 3)$	Graph the function: $f(x) = \frac{2}{3}(x + 1)(x - 5)$	Graph a parabola whose x intercepts are at x=-3 and x=5 and whose minimum value is y=-4.

> ➤ How to model the semantic associations between adjacent video segments?

Multiscale Perceptual Fusion
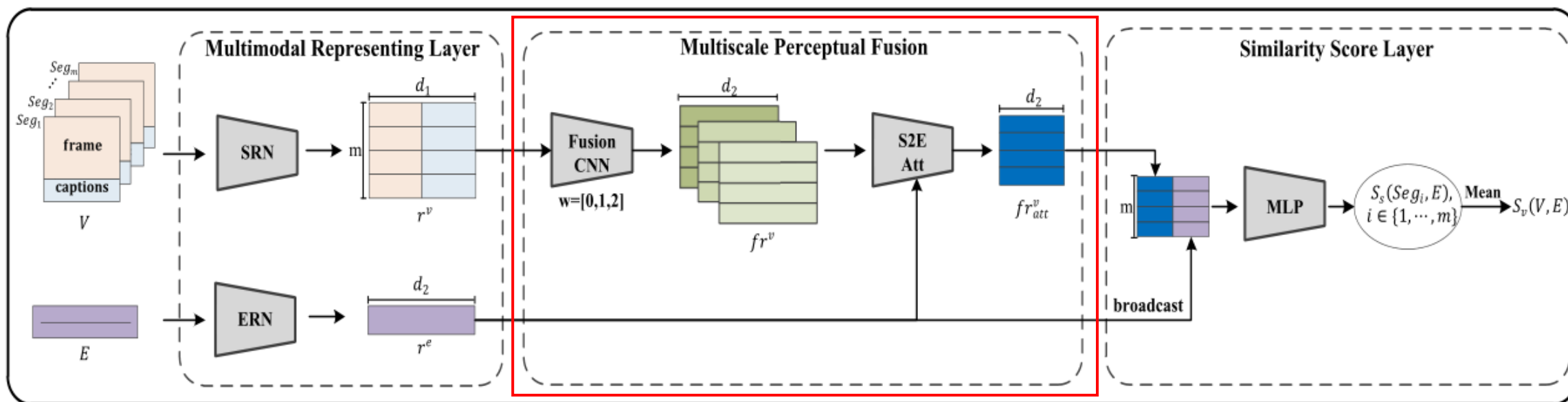
Figure 3: The VENet model architecture consists of three main parts: 1) Multimodal Representing Layer (MRL), 2) Multiscale Perceptual Fusion (MPF), and 3) Similarity Score Layer (SSL). VENet takes a video-exercise pair (V, E) as input and outputs the similarity score $S(Seg_i, E), i \in \{1, \cdots, m\}$ and $S(V, E)$.

> ➤ How to learn the fine-grained similarity by just exploiting the video-level labeled data?
> - The segment-level labeled data is scarce and costly.
> - The video-level labeled data is much easier to obtain.

$$\mathcal{L}(V, E_s, E_{ds}; \Theta) = \sum_{Seg_i \in V} \sum_{Seg_j \in V} \max\left(0, \mu - \left(S_s\left(Seg_i, E_s\right)\right.\right.$$
$$\left.\left.- S_s\left(Seg_j, E_{ds}\right)\right)\right) + \lambda \|\Theta\|^2,$$

E_s	similar exercise
E_{ds}	dissimilar exercise
μ	margin distance
λ	regularization hyperparameter

03 Experiments

📊 Dataset

Table 1: The statistics of the dataset

Data	Statistics	Values
Exercise	Num of exercises	17,116
	Avg. words per exercise	34.95
	Avg. similar videos per exercise	1.67
Video & Captions	Num of videos	1,053
	Avg. length per video	383.79s
	Total length	404,130s
	Total size	22.6GB
	Avg. words per closed captions	831.78
	Avg. similar exercises per video	17.04
Label	Num of similar pairs	10,679
	Num of dissimilar pairs (negative sampling)	10,679

All the data were crawled from the Khan Academy's math domain (https://www.khanacademy.org/math**)**

Results

Table 2: Characteristics of the comparison methods

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

Table 3: Performance of comparison methods

Model	Video-Level		Segment-Level	
	Auc	NDCG	Auc	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**

Ablation Experiments

Table 4: Ablation Experiments

Model	Video-Level		Segment-Level	
	Auc	NDCG	Auc	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

➢ Visual Information is helpful
➢ Textual Information is more important
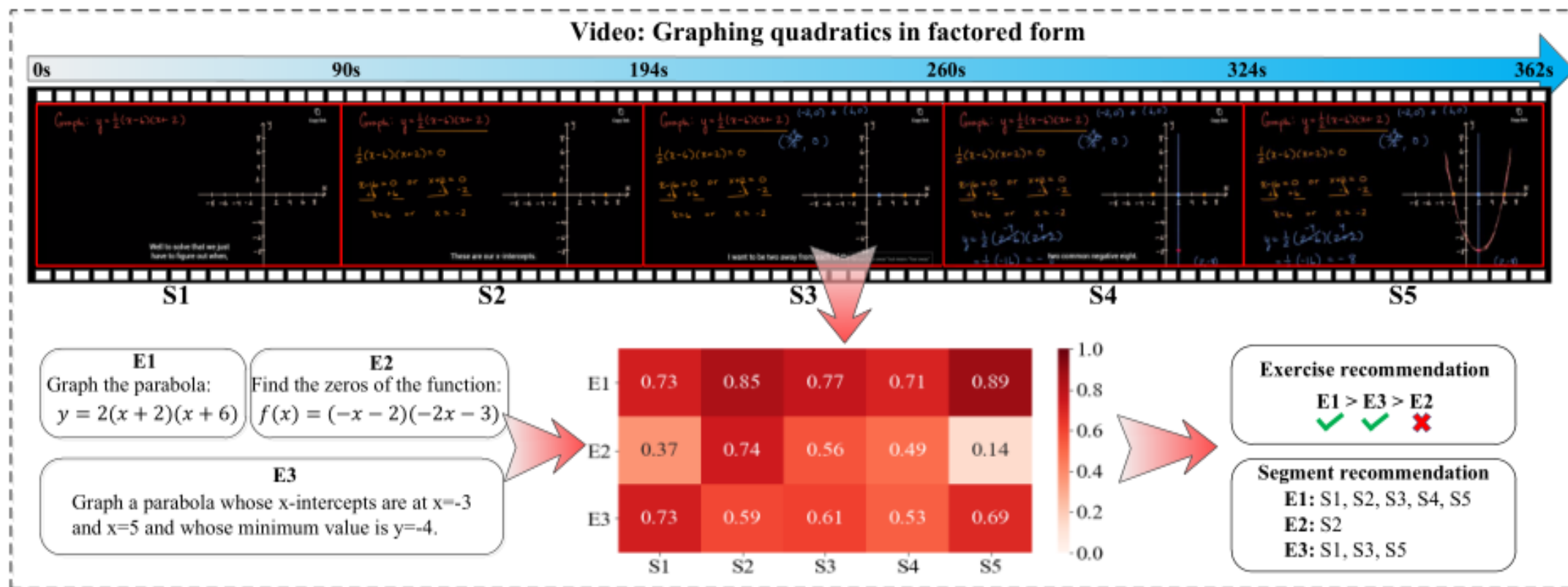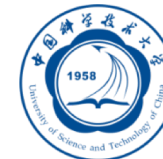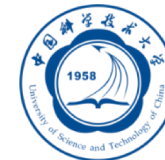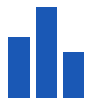➢ All the key modules are eddective

Figure 6: A case study of the similarity measurement for an educational video and three exercises.

Thanks for Listening!

ACM MM 2020
QA Session

Fine-Grained Similarity Measurement of Educational Videos and Exercises

Any Questions?
Just be free to let me know!