LREC-COLING 2024

Event Representation Learning with Multi-Grained Contrastive Learning and Triple-Mixture of Experts

Tianqi Hu, Lishuang Li, Xueyang Qin, Yubo Feng

大连理工大学
DALIAN UNIVERSITY OF TECHNOLOGY

Outline

I Task Definition & Motivation

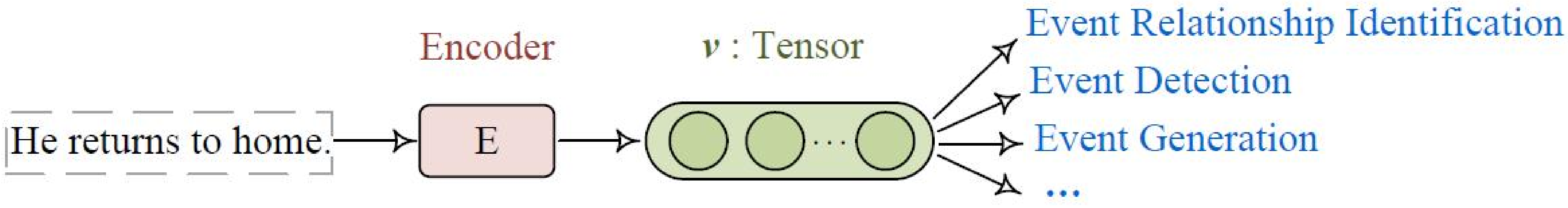
II Methods

III Experiments

IV Conclusion

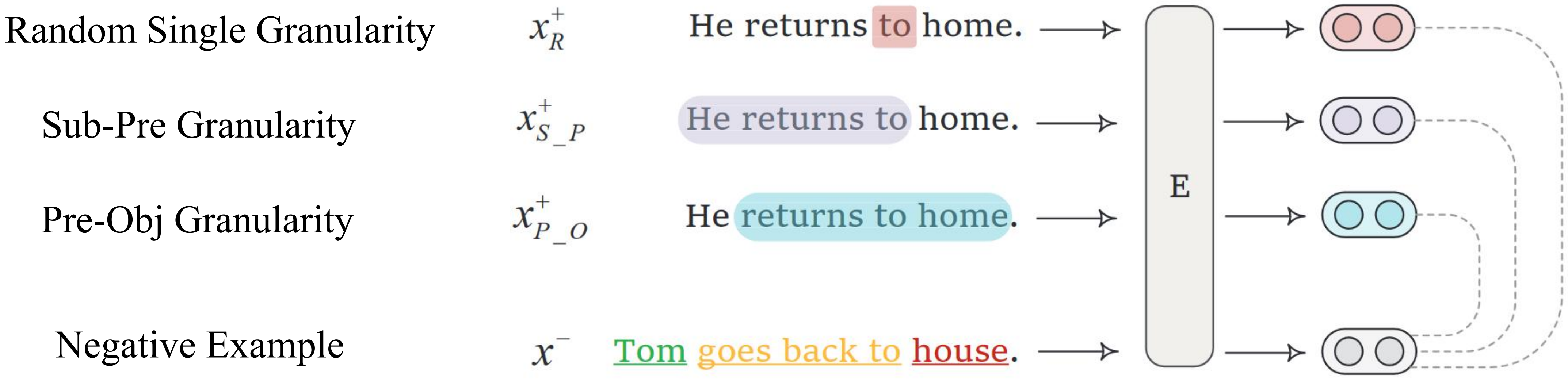
Task Definition

Event Representation Learning (ERL)



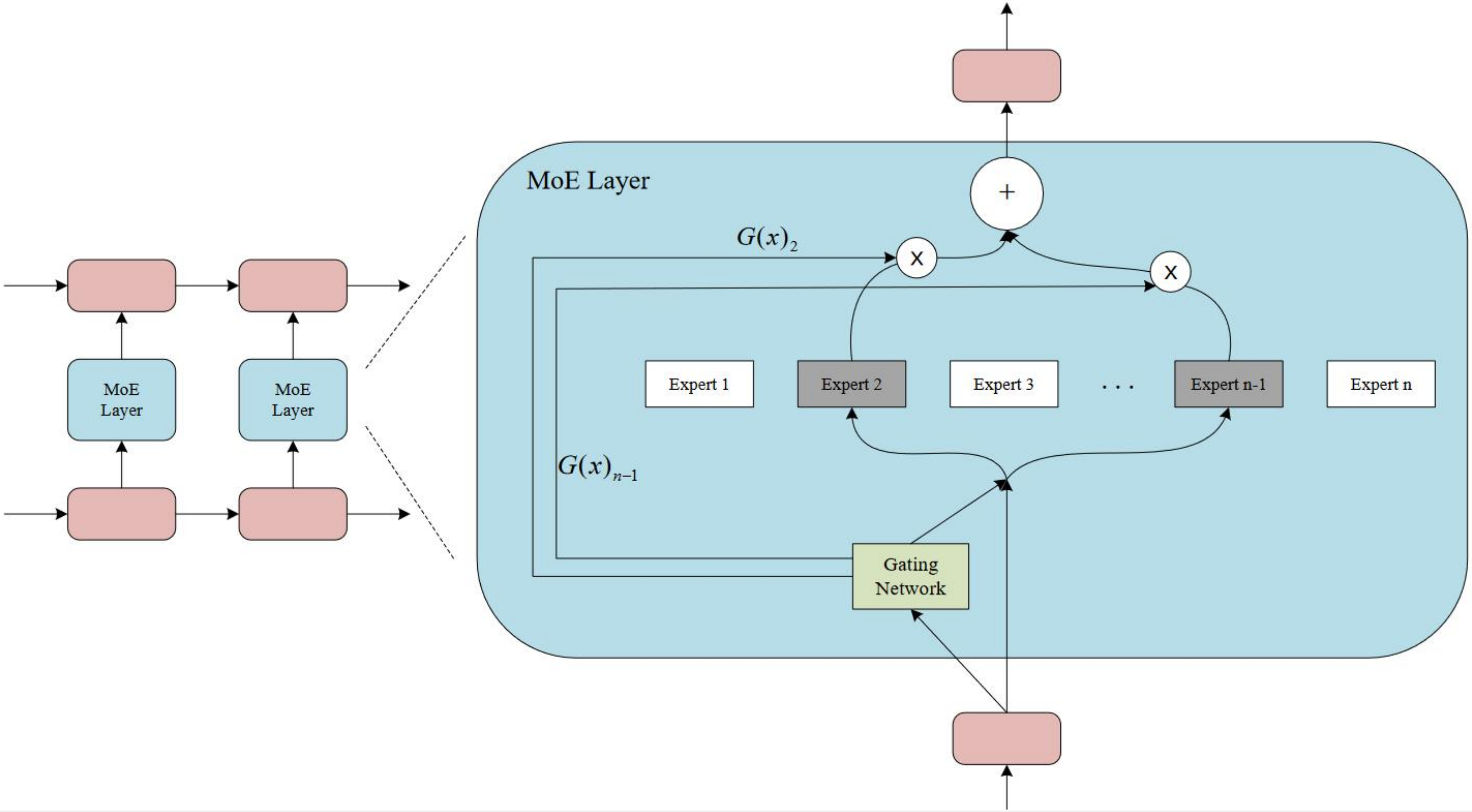
ERL aims to learn from the text how to converts events into a form that computers can <u>understand and process</u>.

Motivation



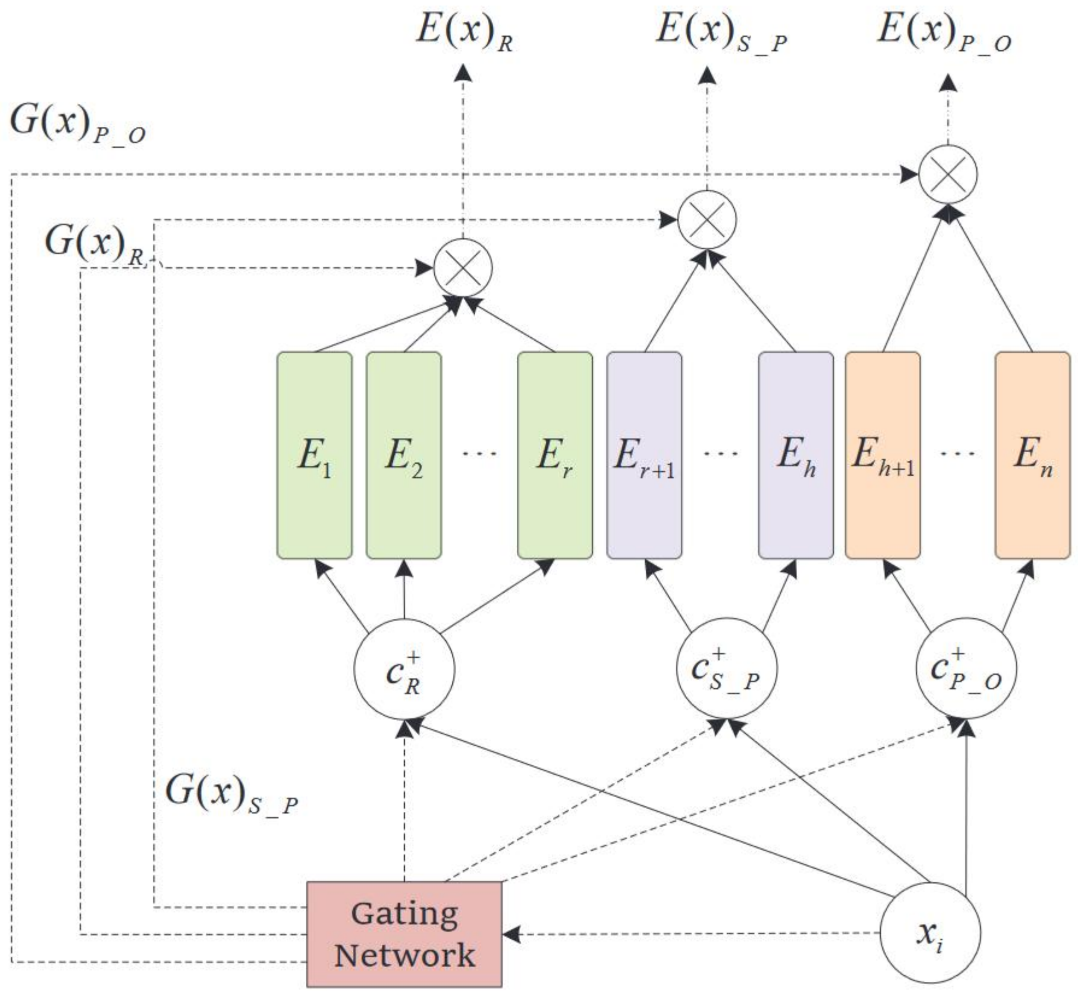
Random Single Granularity
Sub-Pre Granularity
Pre-Obj Granularity
Negative Example
x_R^+
$x_{S\_P}^+$
$x_{P\_O}^+$
x^-
He returns to home.
He returns to home.
He returns to home.
Tom goes back to house.
E

Motivation


MoE Layer
MoE Layer
MoE Layer
$G(x)_2$
$G(x)_{n-1}$
Expert 1
Expert 2
Expert 3
. . .
Expert n-1
Expert n
Gating Network

Methods

The overall architecture of **MCTM** :



Methods

Loss Funtion :

$$\mathcal{L}_{Final} = \varphi(\mathcal{L}_{info} + \mathcal{L}_{mlm}) + \lambda\mathcal{L}_E$$

InfoNCE Constractive Loss :

$$\mathcal{L}_{info} = \sum_{j}^{3} -\log \frac{\varepsilon_j g(c_i, c_{i_j}^{+})}{g(c_i, c_{i_j}^{+}) + \sum_{k\in\mathcal{M}(i)} g(c_i, c_k)}$$

Experts Loss :

$$\mathcal{L}_I = \omega_I \cdot CV(\sum_{x\in X} G(x))^2$$

$$\mathcal{L}_L = \omega_L \cdot CV(\sum_{x\in X} P(x,i))^2$$

$$\mathcal{L}_E = \mathcal{L}_I + \mathcal{L}_L,$$

Experiments

Results on Similarity Task :

Model	Original hard Sim.(%)	Extend hard Sim.(%)	Transitive sentence Sim.(ρ)
Predicate Tensor[32]	41.0	25.6	0.63
Role-factor Tensor[32]	43.5	20.7	0.64
SAM-Net[58]	51.3	45.2	0.59
KGEB[76]	52.6	49.8	0.61
FEEL[30]	58.7	50.7	0.67
NTN-IntSent[33]	77.4	62.8	0.74
UniFA-S[77]	78.3	64.1	0.75
SWCC[40]	80.9	72.1	0.82
MCTM (ours)	**81.7**	**75.2**	**0.85**

Experiments

Results on MCNC (downstream tasks) :

Model	Accuracy (%)
Random	20.00
PPMI	30.52
BiGram	29.67
Word2Vec	37.39
SWCC	44.50
MCTM	**46.15**

The purpose of the downstream task is to verify the **generalization** of the event representation learned by the model.

Ablation Study

Ablation experiment to verify the impact of MoE sparsity on model performance.

Model	OHS (%)	EHS (%)	TSS(ρ)
MCTM	**81.7**	**75.2**	**0.85**
Single Fine-grained	80.9 (↓0.8)	72.1 (↓3.1)	0.81 (↓0.04)
Single Coarse-grained	79.1 (↓2.6)	70.6 (↓4.6)	0.82 (↓0.03)
w/o MoE Layer	79.9 (↓1.8)	73.9 (↓1.3)	0.84 (↓0.01)

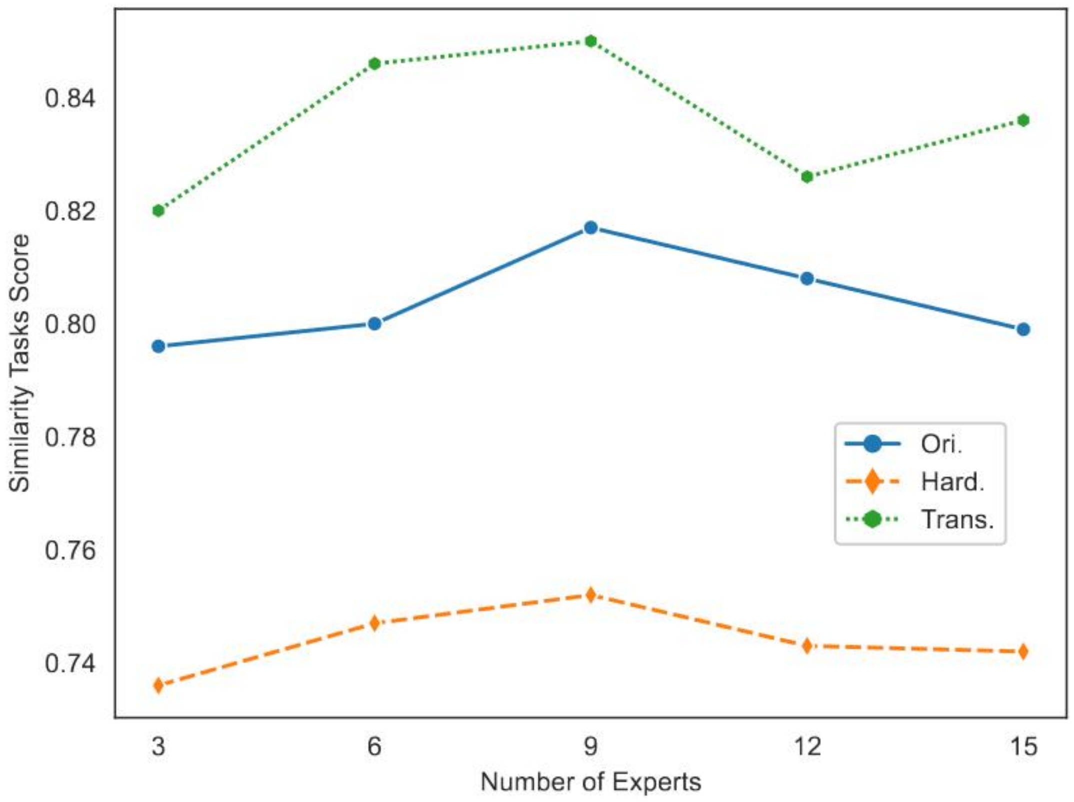
Ablation experiments of different methods on the event similarity task.

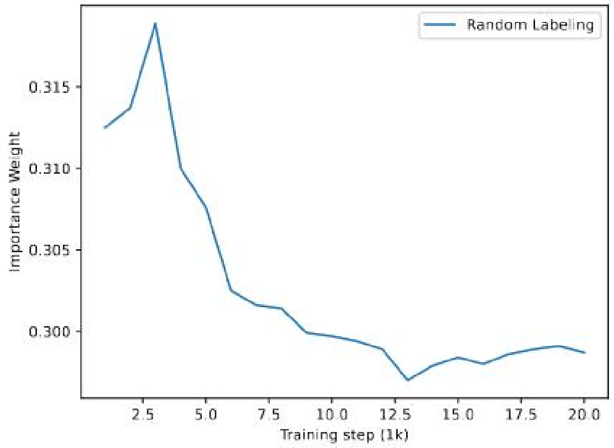
Model	OHS (%)	EHS (%)	TSS(ρ)
MCTM	**81.7**	**75.2**	**0.85**
Active all experts	80.2(↓1.5)	73.8(↓1.4)	0.84(↓0.01)
Multiple experts	79.9(↓1.8)	73.5(↓1.7)	0.84(↓0.01)
Random experts	80.1(↓1.6)	72.9(↓2.3)	0.83(↓0.02)

Ablation experiment to verify the impact of some loss function.

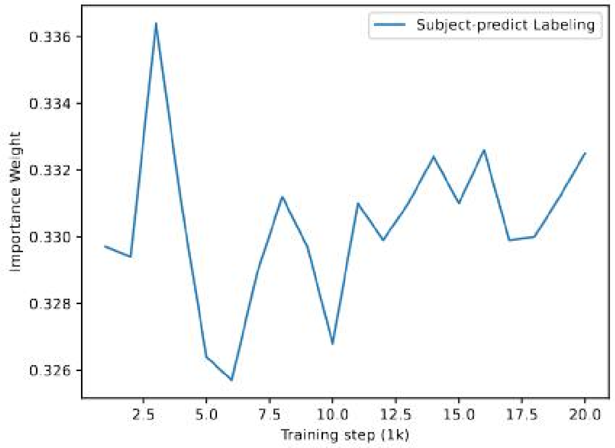
Model	OHS (%)	EHS (%)	TSS(ρ)
MCTM	**81.7**	**75.2**	**0.85**
w/o $\mathcal{L}_I$	76.7(↓5)	70.3(↓4.9)	0.79(↓0.06)
w/o $\mathcal{L}_L$	77.4(↓4.3)	72.2(↓3)	0.79(↓0.06)

Analysis

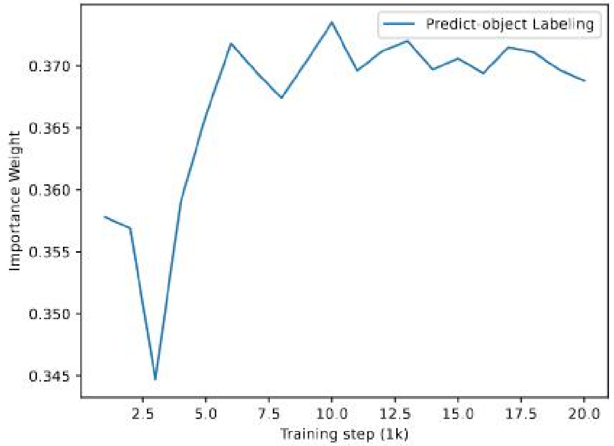




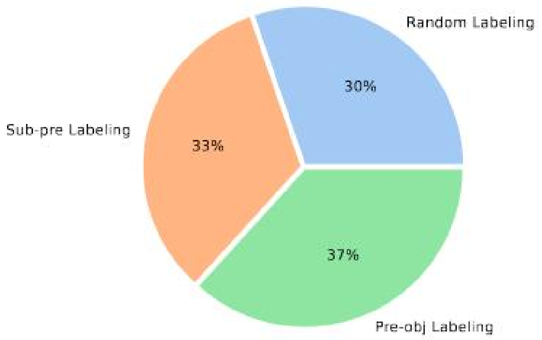
(a) Random Labeling



(b) Sub-Pre Labeling



(c) Pre-Obj Labeling



(d) Weight Graph

Conclusion

- We propose MCTM, which leverages multigrained labels in the positive examples of contrastive learning, enabling an understanding of deep event features from multiple perspectives.
- We adopt a triple-Mixture of Experts layer structure to optimize the model structure so that the model can independently learn the importance weights of each label granularity to achieve better results.
- MCTM achieves SOTA results on the event similarity task and the best results so far on the MCNC dataset, which validates the validity of MCTM.

Thanks for listening!