Team HUGE: Image-Text Matching via Hierarchical and Unified Graph Enhancing

Bo Li1,2,3, You Wu1,2, Zhixin Li*1,2

1Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China
2Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China
3School of Computer Science and Engineering, Guilin University of Aerospace Technology, Guilin 541004, China

* Zhixin Li is the corresponding author

Limitations of Image-Text Matching

- The local-level methods emphasize fine-grained matching, while the global-level methods emphasize the credibility of overall semantic scores. However, there are few works that analyze this task from two perspectives, even multi-perspectives.
- GNN can represent rich information, while its complex structure and easy overfitting limit performance, potentially leading to disappointing training that is difficult to ground, for example.

Our contributions:

- A hierarchical graph learning model is proposed to effectively determine the multi-perspective knowledge and preserve the discriminative semantics in cross-modal features, i.e. hierarchical graph learning.
- By analyzing the principles of different graph sub-models and leveraging their different functions, HUGE can improve model retrieval ability and reduce the probability of overfitting, i.e. unified graph enhancing.
- In addition, we designed a two-stage similarity representation learning strategy that combines the advantages of cosine similarity and vector similarity.

Hierarchical Graph Learning

- (1) Relationship graph analysis (RGA).
- (2) Weak-similarity graph pruning (WGP).
- (3) Sparse graph matching (SGM).
- (4) Global knowledge guiding (GKG).
- This design ensures sufficient learning from multiple angles, and the sub-modules can assume a single function to prevent overall model overfitting.

Unified Graph Enhancing

\[ L = a_1L_{RGA} + a_2L_{WGP} + a_3L_{SGM} + a_4L_{GKG} \]

HUGE trains four graph sub-modules separately. After integrated training, these sub-modules can guide each other to improve the performance of the overall model while preventing overfitting.

Two Stage Similarity Representation

- We implement cosine similarity representation learning in the feature sampling and adopt vector-based similarity representation learning in subsequent graph model training.
- The two representation learning strategies focus on different keynotes of samples and are capable of exploring the correlation of multi-media instances from multiple perspectives.

Experimental results