Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning

Shizhe Chen¹, Yida Zhao¹, Qin Jin¹, Qi Wu²
¹Renmin University of China, ²University of Adelaide
Video-Text Cross-modal Retrieval

Text: a man pours oil into a preheated stir fry pan and then carefully add some raw chicken from a small bowel.

• Dominant approach: learning joint embedding space
  • Global visual-semantic matching
    • One vector is hard to encode fine-grained details
  • Local visual-semantic matching
    • Relationships between local vectors are not well captured via sequential modeling
Hierarchical Graph Reasoning Model (HGR)

• Multi-level Video-Text Matching
  • Event
  • Actions
  • Entities

• Hierarchical Textual Encoding
  • Decompose sentence into semantic role graph
  • Capture relationships via graph reasoning

• Hierarchical Video Encoding
  • Guided by different levels of text to learn diverse video representations
Experiments

• In-domain Cross-modal Retrieval
  • Better performance across three datasets

• Cross-domain Generalization
  • Generalize better across datasets

• Fine-grained Binary Selection
  • Differentiate fine-grained difference between positive and negative sentences
Conclusion

• Decompose videos and texts into hierarchical semantic levels

• Utilize graph reasoning to generate hierarchical embeddings

• Evaluate on in-domain, cross-domain and fine-grained binary selection to demonstrate model’s effectiveness

Codes and datasets will be released at:
https://github.com/cshizhe/hgr_v2t