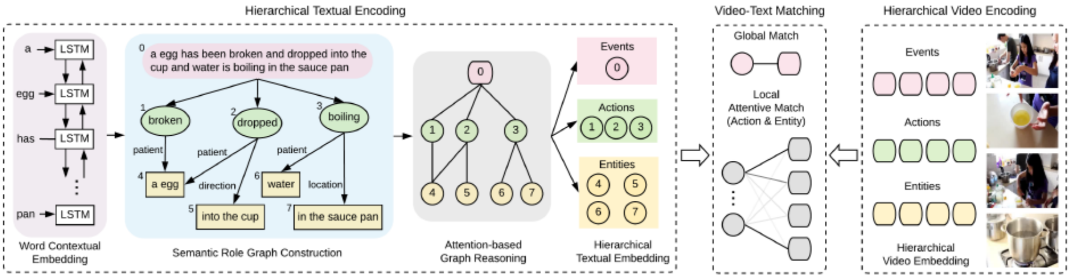
Tech Review: State of Art Video-Text Retrieval methods

Jianzhe Sun
jianzhe2

Nowadays, text retrieval is not only used for document retrieval, but also other cross-model retrievals like image-text retrieval and video-text retrieval. In this tech review I will mainly introduce two state-of-art papers of video-text retrieval.

The first paper is **Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning**. This paper proposed a Hierarchical Graph Reasoning (HGR) model. The model disentangles texts into hierarchical semantic graph including three levels of events, actions, entities and relationships across levels to form a global to local structure. The level of events is the whole sentence, which describes the structure of the sentence. The level of actions is a set of the verbs in the sentence, which describes the action of the events. Finally, the level of entities is a set of the phrases of nouns, including subject and object, which keeps the semantic information and the relation of the sentence.

The model can be divided by three parts: hierarchical textual encoding module, hierarchical video encoding module and video-text matching module.


Hierarchical Textual Encoding
a
egg
has
pan
LSTM
LSTM
LSTM
LSTM
Word Contextual Embedding
0 a egg has been broken and dropped into the cup and water is boiling in the sauce pan
1 broken
2 dropped
3 boiling
patient
patient
patient
4 a egg
direction
6 water
location
5 into the cup
7 in the sauce pan
Semantic Role Graph Construction
0
1
2
3
4
5
6
7
Attention-based Graph Reasoning
Events
0
Actions
1
2
3
Entities
4
5
6
7
Hierarchical Textual Embedding
Video-Text Matching
Global Match
Local Attentive Match (Action & Entity)
Hierarchical Video Encoding
Events
Actions
Entities
Hierarchical Video Embedding

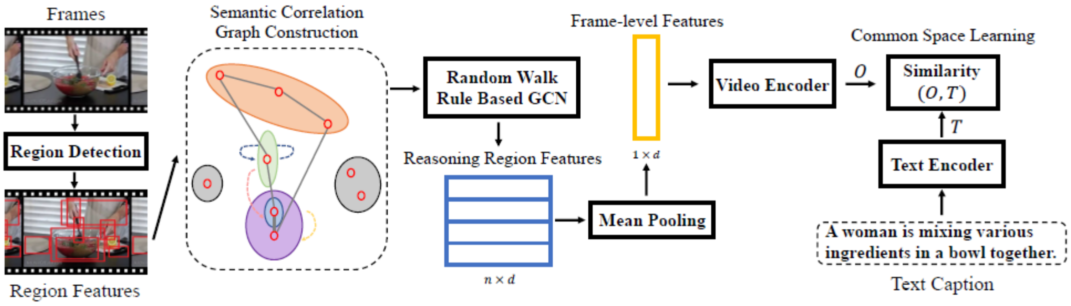
First, the hierarchical textual encoding module consists of two parts, semantic role graph construction and word contextual embedding. The input of the model is a sentence, in this case word embedding, and the model takes the sentence as an event node of a hierarchical graph. Then a semantic role parsing toolkit will be applied to extract the verbs and the nouns in the sentence. The verbs are action nodes which are sub-nodes of the event node, and the nouns are entity nodes which are sub-nodes of the action nodes. The paper embeds semantic meaning of each node into a dense vector. The word embeddings are generated by a bidirectional LSTM, which also utilizes the attention mechanism to give different weightings for different words. Finally, the embedding of each semantic node is determined after normalization.

After obtaining the Attention-based Graph, they do the Attention-based Graph Reasoning. They use GCN to obtain different weights of the edges. For example, the word “egg” may have different appearances, but the word “broken” constrains the state of “egg”, which means the action constrains the entity’s semantic. They use two matrices to define weights, one is a transformation matrix and the other is a role embedding matrix. Then they use them as the input of GCN (Graph Convolutional Networks), and after iterations the weights of the vectors will be updated, and GCN will output the nodes respectively.

Second, the Hierarchical Video Encoding is used to encode the video. They define three different weight matrices to represent event weight, action weight and object weight. Using the frame-wise features of the video, for global event level this module will output one global vector with attention mechanism and for action and object level, this module will output two sequences, and then send them to the matching module.

The third module is Video-Text Matching. They first use global match to match the event level, and use the cos similarity to compute the score. Then they do the Local Attentive matching. In the local Attentive matching, they first compute the local similarity between each pair of cross-modal local components. Then, they normalize the result and utilize it to dynamically align to video frames. Then, the score can be computed by the similarity above and the attention weights over video frames for each local textual node.

The second paper is **Exploiting Visual Semantic Reasoning for Video-Text Retrieval**. The paper states that previous works represent videos by directly encoding from frame-level features but fail to pay attention to the semantic relations. The main structure is as follows:



The paper utilizes the bottom-up attention model to generate frame regions and extract features from frames (using Faster RCNN). For regions, they construct a fully-connected semantic correlation graph with n region features. In the adjacency matrix, they use attention mechanism to generate the weight of an edge.

Subsequently, they do semantic reasoning between these regions by leveraging random walk rule based Graph Convolutional Networks (GCN) to generate region features with relation information. Random walks on a graph is a rule to visit a sequence of vertices together with a sequence of edges and widely applied in network representation learning. Since networks and graphs have similar topological structures, they introduce random walk statistics into GCN.

After applying random walk in GCN, GCN will output the set of the relation enhanced representations for vertices. After mean pooling, they use the Common Space Learning, which is a multi-level encoder to map text and video ultimate embedding representations in the same dimensional common space.

With the methods described above, the two papers reached state-of-art performance. In common, they both use semantic information to improve the accuracy of video-text retrieval, and we know that semantic information is hard to obtain by computers. With the development of AI, I believe that the accuracy of video-text retrieval will be higher and higher.

References:
[1] Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning
[2] Exploiting Visual Semantic Reasoning for Video-Text Retrieval