Supplementary Material for "Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning"

Shizhe Chen[1], Yida Zhao[1], Qin Jin[1], Qi Wu[2]
[1]School of Information, Renmin University of China
[2]Australian Centre for Robotic Vision, University of Adelaide
{cszhe1, zyiday, qjin}@ruc.edu.cn, qi.wu01@adelaide.edu.au

1. Semantic Roles

We parse text into hierarchical semantic role graph consisting of global event node, action nodes and entity nodes. The action nodes are connected with global event node with edge type of action, and the entity nodes are linked with the corresponding action nodes with different edge types according to their semantic roles. Table 1 presents all semantic roles used in our graph and their descriptions based on linguistic experts [1].

Table 1: Semantic roles in parsed semantic role graph.

Semantic Role	Description
Event	global event description
Action	verb
ARG0	proto-agent
ARG1	proto-patient
ARG2	instrument, benefactive
ARG3	start point
ARG4	end point
ARGM-LOC	location (where)
ARGM-MNR	manner (how)
ARGM-TMP	time (when)
ARGM-DIR	direction (where to/from)
ARGM-ADV	miscellaneous
OTHERS	other argument types

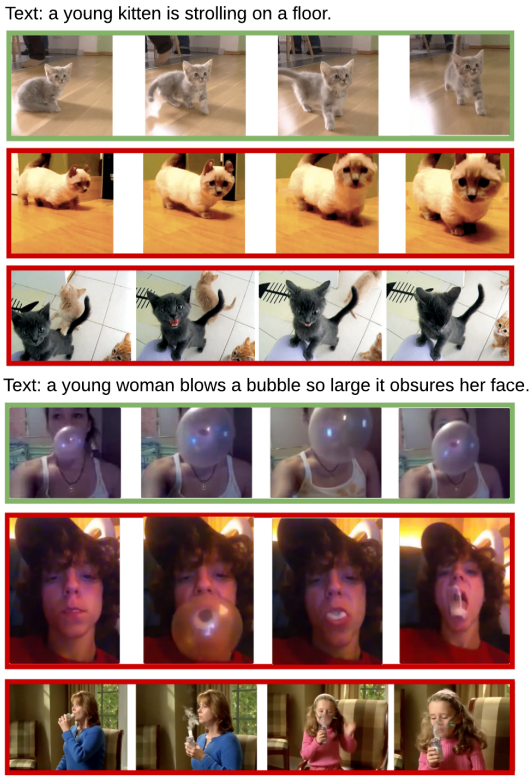

Figure 1: Cross-modal video-text retrieval results on TGIF and VATEX testing set.

2. Binary Selection Task

In order to evaluate fine-grained discrimination ability on texts of different video-text retrieval models, we propose a binary selection task which requires the model to select a sentence that better matches with a given video from two very similar but semantically different sentences. We utilize testing videos from the Youtube2Text dataset and randomly select one ground-truth video description for each video as the positive sentence. The negative sentence is generated by perturbing the ground-truth sentence as shown in Table 2.

3. Additional Qualitative Examples

We visualize some examples on cross-modal video-text retrieval on TGIF and VATEX datasets in Figure 1. Our model achieves robust and superior performance on different datasets for cross-modal retrieval.

Table 2: Examples of different perturbation types in binary selection task.

Task	Description	Example
switch roles	switching the agent and patient of an action in the sentence	positive: a woman is cutting an onion. negative: an onion is cutting a woman.
replace actions	replacing action in the sentence with a random one	positive: a person pours coconut water into a bowl. negative: a person drives coconut water into a bowl.
replace persons	replacing agent or patient in the sentence with a random one	positive: a man is keep the knife on the machine. negative: a man is keep a dog on the floor on the machine.
replace scenes	replacing scene in the sentence with a random one	positive: a man strums a violin on a stage. negative: a man strums a violin in the beach.
incomplete events	only keeping part of the description of an event	positive: men are dancing in towels. negative: men in towels.

References

[1] Daniel Jurafsky and James H. Martin. Speech and language processing, 2009. 1