Bridge to Answer: Structure-aware Graph Interaction Network for Video Question Answering -Supplementary Materials-

Jungin Park Jiyoung Lee Kwanghoo Sohn*
Yonsei University, Seoul, South Korea
{ newrun, easy00, khsohn }@yonsei.ac.kr

In this document, we briefly summarize graph convolution networks [1], and provide more details of *Bridge to Answer* and more qualitatively results that demonstrate the advantage of our method.

1. Graph Convolution Networks

A graph $\mathcal{G}$ can be represented by a tuple $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges representing the connectivity between vertices, such that the vertices v_i and v_j are connected with the edge weight e_{ij}. An adjacency matrix $\mathbf{A}$ contains the connectivity between vertices and their edge weights in a fully-connected or a sparse matrix form. The graph convolution networks (GCNs) [1] have been proposed to learn richer representations of vertices by aggregating the representations from their neighborhoods.

In standard GCNs, the node representations $\mathbf{X}$ and the adjacency matrix $\mathbf{A}$ are taken as inputs of graph convolution operation. The output of the graph convolution operation is then obtained by following equation:

$$\mathbf{Z} = \sigma(\mathbf{D}^{-1/2}\hat{\mathbf{A}}\mathbf{D}^{-1/2}\mathbf{X}\mathbf{W}), \tag{1}$$

where $\mathbf{Z}$ is the output node representations, $\sigma(\cdot)$ is an activation function such as ReLU, and $\hat{\mathbf{A}}$ is the summation of $\mathbf{A}$ and the identity matrix $\mathbf{I}$, such that $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$. $\mathbf{D}$ is a diagonal matrix of $\hat{\mathbf{A}}$ and $\mathbf{W}$ is a trainable weight matrix of the graph convolution layer, respectively.

In this paper, we construct fully-connected appearance and motion graphs and a sparse question graph and apply consecutive graph convolution layers to perform question-to-visual and visual-to-visual interactions.

2. Channel Configurations

Our method is divided into three parts: *1) Graph construction* to generate appearance, motion, and question graphs, *2) question-to-visual interactions* to make question conditioned visual representations (*i.e.*, question-to-appearance and question-to-motion), and *3) visual-to-visual interactions* to propagate each visual information to relative visual graph (*i.e.*, appearance-to-motion and motion-to-appearance). To clarify the usage of each component in our method, we provide the overall notations and channel configurations as shown in Tab. 1.

In addition, we describe the dimensional configuration for the weights used in Q2V and V2V interactions, as shown in Tab. 2.

Notation	Channel
$\hat{\mathbf{V}}$, $\hat{\mathbf{M}}$	$\mathbb{R}^{8\times16\times512}$, $\mathbb{R}^{8\times512}$
$\mathbf{U}$	$\mathbb{R}^{w\times512}$
$\mathbf{W}^v$, $\mathbf{W}^m$, $\mathbf{W}^m$	$\mathbb{R}^{128\times128}$, $\mathbb{R}^{8\times8}$, $\mathbb{R}^{w\times w}$
$\mathbf{S}^v$, $\mathbf{S}^m$	$\mathbb{R}^{128\times w}$, $\mathbb{R}^{8\times w}$
$\hat{\mathbf{V}}$, $\hat{\mathbf{M}}$	$\mathbb{R}^{8\times16\times512}$, $\mathbb{R}^{8\times512}$
$\mathbf{U}_b^v$, $\mathbf{M}_b^m$	$\mathbb{R}^{w\times512}$, $\mathbb{R}^{w\times512}$
$\hat{\mathbf{U}}_b^v$, $\hat{\mathbf{M}}_b^m$	$\mathbb{R}^{w\times512}$, $\mathbb{R}^{w\times512}$
$\mathbf{S}_b^v$, $\mathbf{S}_b^m$	$\mathbb{R}^{8\times16\times w}$, $\mathbb{R}^{8\times w}$
$\mathbf{V}^f$, $\mathbf{M}^f$	$\mathbb{R}^{8\times16\times256}$, $\mathbb{R}^{8\times256}$
$\bar{\mathbf{v}}$, $\bar{\mathbf{m}}$	$\mathbb{R}^{256}$, $\mathbb{R}^{256}$

Table 1. Notations and channel configuration

Equ.	Notation	Dimensions
(5), (7)	$\mathbf{W}_f^v$, $\mathbf{W}_g^v$	$\mathbb{R}^{512\times512}$
(6), (7)	$\mathbf{W}_f^m$, $\mathbf{W}_g^m$	$\{\mathbb{R}^{512\times512}, \mathbb{R}^{512\times256}\}$
(9), (11)	$\mathbf{W}_{gb}^v$, $\mathbf{W}_{gb}^m$	$\{\mathbb{R}^{512\times512}, \mathbb{R}^{512\times256}\}$
(10), (11)	$\mathbf{W}_b^v$, $\mathbf{W}_b^m$	$\mathbb{R}^{512\times512}$
(13)	$\mathbf{W}_1$, $\mathbf{W}_2$	$\mathbb{R}^{512\times512}$, $\mathbb{R}^{1024\times512}$
(13)	$\mathbf{W}_y$, $\mathbf{W}_{y'}$	$\mathbb{R}^{512\times256}$, $\mathbb{R}^{256\times W}$
(14)	$\mathbf{W}_w$, $\mathbf{W}_a$	$\mathbb{R}^{512\times512}$
(14)	$\mathbf{W}_y$, $\mathbf{W}_{y'}$	$\mathbb{R}^{2048\times512}$, $\mathbb{R}^{512\times1}$

Table 2. Dimensional configuration for the weights

*Corresponding author.

(a) **Question**: What does the fish do 4 times?

HCRN: hit an object
Ours: slap tail
Groundtruth: slap tail

(b) **Question**: What do the woman do 4 times?

HCRN: throw treat
Ours: turn off lights
Groundtruth: turn off lights

(c) **Question**: What does the dog do 2 times?

HCRN: lay head
Ours: turn around
Groundtruth: turn around

(d) **Question**: What does the man do 6 times?

HCRN: turn head
Ours: poke woman
Groundtruth: poke woman

(e) **Question**: What does the woman who wear a stripe shirt do before blow out air?

HCRN: headbutt
Ours: touch cheek
Groundtruth: touch cheek

(f) **Question**: What does the man who is sitting on the back seat of the from car do after look out window at tailgater?

HCRN: dance around
Ours: pull out machine
Groundtruth: pull out machine

Figure 1. Qualitative comparisons with the state-of-the-art method [2]. The results show the advantages of our method in various challenging cases: (a) Requiring more accurate answer. (b) Inferring action by appearance, not motion. (c) Capturing the rapid transition of the object. (d) Capturing the slow transition of the object. (e) Finding a target in multiple objects. (f) Processing a long and complex question.

3. More Results

In this supplementary, we provide the more quantitative comparisons with the state-of-the-art method [2]. As shown in Fig. 1, our model shows the advantages in various challenging cases. When the answer candidates are ambiguous, our model finds a more plausible answer. For example, the candidates "hit an object" and "slap tail" in Fig. 1-(a) both can be correct answers. Our model infer a more plausible answer "slap tail". Fig. 1-(b) shows an example of *inferring action by appearance*. The action of "turn off lights" is more related to appearance information, not motion. Since we learn the question conditioned appearance representation attributed to motion, our model adequately captures the action with the appearance changes. In addition, our model captures both rapid and slow transitions of an object as shown in Fig. 1-(c) and (d). The examples of the last row in Fig. 1 show the advantages of using the question graph. By associating question and visual graphs and propagating information, our model capture the object referred to in the question when multiple objects are in the video as shown in Fig. 1-(e). Also, more accurate processing for a long and complex question is possible by constructing the question graph that considers the compositional semantics of the question.

We depict A2M interaction that represents the connections of appearance-question and question-motion as shown in Fig. 2. Note that we depict the clips as frames sampled at each clip for visibility. The frames and words are placed according to temporal and word orders, and corresponding clips of each word are placed regardless of temporal order. Although cross-modal interactions are fully-connected, we display the connection with the largest value in each interaction matrix, such that two connected nodes are associated with the maximum interaction value. For example, the first frame is associated with the word "man" by the interaction value of 0.15, and the word "man" is related to the forth clip by the interaction value of 0.33. When all the

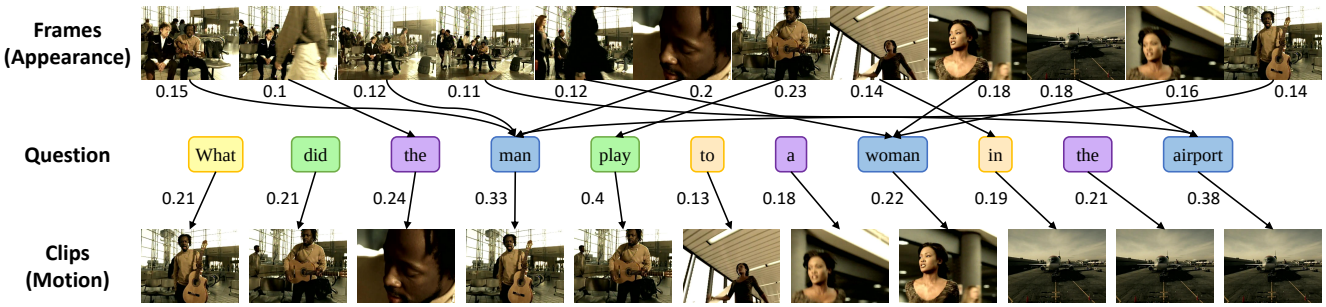

Figure 2. Visualization of A2M interaction. Although any two graphs are fully connected by interaction value, we only indicate the connection with the largest value in each interaction matrix for visibility. Note that the clips are represented by frames sampled from each clip.

nodes are connected with the equal weights, the values for appearance-question and question-motion interactions are 0.09 and 0.125, respectively.

References

[1] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *Int. Conf. Learn. Represent.*, 2017. 1

[2] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9969–9978, 2020. 2