Invariant Grounding for Video Question Answering

Supplementary Material

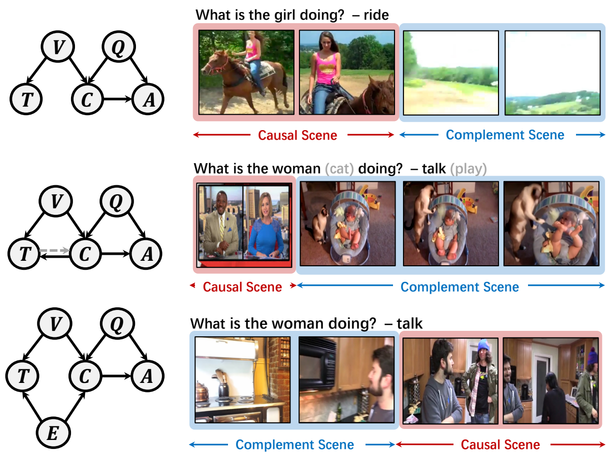

Figure 7. Illustration of complement type.

A. Example of context type

As shown in Figure 7, we classify the relation between causal scene and its complement (*e.g.* $T \dashleftarrow\dashrightarrow C$) into three types, where each row encompasses a causal graph (left) that depicts typical causal-complement relation demonstrated in the example (right):

- In the first row, C and T has no causal relation (*i.e.* $T \perp C$).
- The second row shows a scenario that C is the direct cause of T (*i.e.* $C \rightarrow T$), or vise versa if the question is modified (*e.g.*'What is the cat doing?')
- Similar to the example in Figure 1, the third row demonstrates how shortcut deviate the prediction from the gold answer (*e.g.* "talk") to false prediction (*e.g.* "cook") via common cause E (*e.g.* visual concept "kitchen") since LMI between visual concept "kitchen" and candidate answer "cook" is much higher than it is with "talk".

B. Our backbone

Most VideoQA architectures from the state of the art are compatible with our IGV learning strategy. To testify, we design a simple and effective architecture inspired by [15]. Specifically, $f_{\hat{A}}$ is presented as a combination of a visual-question mixer and an answer classifier. The mixer first encode $\hat{c}$:

$$\mathbf{v}_g^{\hat{c}}, \mathbf{v}_l^{\hat{c}} = \mathrm{LSTM}_5(\hat{c}) \tag{15}$$

where outputs $\mathbf{v}_g^{\hat{c}} \in \mathbb{R}^d$, $\mathbf{v}_l^{\hat{c}} \in \mathbb{R}^{N\times d}$ denote the global and local feature of $\hat{c}$ respectively. Then, based on the concatenation of local representation $\mathbf{q}_l$ (*cf*. Equation (6)) and $\mathbf{v}_l^{\hat{c}}$, we construct an undirected heterogeneous graph that propagates information over each video shot and each question token. Typically, the adjacency matrix $\mathcal{G}_{\hat{c}} \in \mathbb{R}^{(L+N)\times(L+N)}$ is computed as the node-wise correlation scores in form of dot-product similarity, where $N \leq K$ is the sequence length of casual scene. The output of the graph is assembled as holistic local factor $\mathbf{s}_l^{\hat{c}} \in \mathbb{R}^d$ via a attention pooling operator. More Formally, the process is as follows:

$$\mathbf{x}_{\hat{c}} = [\mathbf{v}_l^{\hat{c}}; \mathbf{q}_l], \quad \mathcal{G}_{\hat{c}} = \sigma(\mathrm{MLP}_5(\mathbf{x}_{\hat{c}})) \cdot \sigma(\mathrm{MLP}_6(\mathbf{x}_{\hat{c}}))^\top \tag{16}$$

$$\mathbf{z}_{\hat{c}} = \mathrm{GCN}(\mathbf{x}_{\hat{c}}, \mathcal{G}_{\hat{c}}) \tag{17}$$

$$\mathbf{s}_{\hat{c}}^l = \mathrm{Pooling}(\mathbf{z}_{\hat{c}}) \tag{18}$$

where $\mathbf{x}_{\hat{c}}$, $\mathbf{z}_{\hat{c}}$ $= \in \mathbb{R}^{(L+N)\times d}$ denote the input and output of graph reasoning, MLP_5 and MLP_6 denote is affine projection followed by ReLU activation $\sigma(\cdot)$. To capture the global information, our mixer integrates two global factors $\mathbf{v}_g^{\hat{c}}$ and $\mathbf{q}_g$ into holistic representation via BLOCK fusion [4]:

$$\mathbf{s}_{\hat{c}}^g = \mathrm{Block}(\mathbf{v}_g^{\hat{c}}, \mathbf{q}_g) \tag{19}$$

Similarly, we obtain the final representation by applying the BLOCK again to global and local factor, which is further decoded into answer space with classifier Ψ:

$$\mathbf{s}_{\hat{c}} = \mathrm{Block}(\mathbf{s}_{\hat{c}}^g, \mathbf{s}_{\hat{c}}^l) \tag{20}$$

$$\hat{y}_{\hat{c}} = \Psi(\mathbf{s}_{\hat{c}}) \tag{21}$$

Analogously, we can obtain the predictive answer for $\hat{t}$ and v^* via the shared backbone predictor.

C. Baselines

We compare our design against some existing work, which can be categorized into three categories: 1) **Memory-based** methods that perform multi-step reasoning via updating the recurrent unit, which refines the cross-modal representation iteratively. Specifically, AMU [37], Co-Mem [10] apply this module to encode the visual representation, and HME [9] managed better exploitation for both modalities; 2) **Graph-based** methods like HGA [15] and B2A [19] adopt graph reasoning on the clip-level, whose adjacent matrix is built on node-wise visual similarity. Comparatively, B2A additionally establishes a text graph through question parsing, and abridge two modalities via message passing; 3) **Hierarchical-based** methods HOSTR [8] and HCRN [17] have similar hierarchical conditional architectures. Their discrepancy lies in the feature granularity, where HCRN grounds the temporal relation between frames, while HOSTR roots in object trajectories.

D. Implementation details

All experiments are conducted on GPU NVIDIA Tesla V100 installed on Ubuntu 18.0.4. In terms of complexity, our algorithm matched equally with the corresponding baseline. As a comparison, the default backbone model is trained for 2 hours till convergence on MSRVTT-QA, whereas IGV takes 2.6 hours. For space complexity, since we use the same predictor for the causal, complement, and intervened prediction, IGV only takes 10% more parameters than the default backbone model.