Supplementary Material for "Weakly-Supervised Spatio-Temporal Video Grounding with Variational Cross-Modal Alignment"

Yang Jin[1] and Yadong Mu[1*]

[1]Peking University
jiny@stu.pku.edu.cn, myd@pku.edu.cn

In this supplemental material, we firstly elaborate on the detailed proofs for Equation 8 of the main text in Section 1. We also provide more ablations to explore the different modules in the proposed framework in Section 2,. Finally, to give a more intuitive demonstration, more visualizations of the grounding results are provided in Section 3.

1 Proofs of Equation 8

To make the notation more concise, we use $\{z_i\}_{i=1}^{L}$ to substitute the latent variable $Z = (z^S, z^T)$, where $z_i \in \{z_t^S, z_t^T\}$ indicates the component of the latent variable Z. In the E-step, the optimal distribution q_ϕ satisfies the following term when adopting the mean-field approximation:

$$q_\phi(z_i|E,V) \propto \exp\left(\mathbb{E}_{-z_i}[\log p_\theta(z_i|R,E,V,-z_i)]\right), \tag{1}$$

where $-z_i$ indicates all the $L-1$ latent components except z_i. Recall that in this iteration step, the optimization objective is to minimize:

$$\mathcal{F}(\phi) = \sum_Z q_\phi(Z|E,V) \log \frac{q_\phi(Z|E,V)}{p_\theta(Z|R,E,V)}. \tag{2}$$

When θ is frozen, the log-likelihood function $\log p_\theta(R|E,V)$ is the constant. Therefore, minimizing $\mathcal{F}(\phi)$ is equivalent to maximizing the evidence lower bound:

$$\mathcal{O}(q_\phi) = \mathbb{E}_{q_\phi(Z|E,V)}\left[\log \frac{p_\theta(R,Z|E,V)}{q_\phi(Z|E,V)}\right]. \tag{3}$$

When $q_\phi(Z|E,V)$ satisfies the mean-field assumption, it has the following format:

$$q_\phi(Z|E,V) = \prod_{i=1}^{L} q_\phi(z_i|E,V). \tag{4}$$

* Corresponding Author.

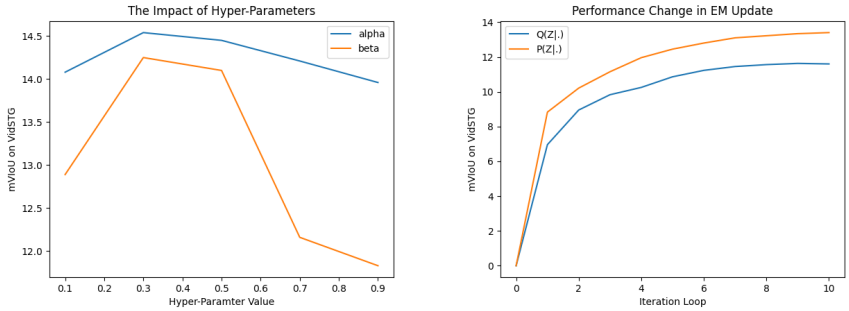

Fig. 1: Impact of hyper-parameters and EM update on VidSTG.

We reformulate p_θ based on the chain rule of conditional probability distribution and substitute Eq. 4 into the ELBO:

$$\begin{aligned}\mathcal{O}(q_\phi) = &\log p_\theta(R|E,V) \\ &+ \sum_{i=1}^{L}(\mathbb{E}_{q_\phi(Z|E,V)}\left[\log p_\theta(z_i|R,E,V,z_{1:(i-1)})\right] \\ &- \mathbb{E}_{q_\phi(z_i|E,V)}\left[\log q_\phi(z_i)\right]).\end{aligned} \tag{5}$$

We take the partial derivative of $q_\phi(z_i)$ with respect to $\mathcal{O}(q_\phi)$ and set it to zero:

$$\begin{aligned}\frac{d\mathcal{O}(q_\phi)}{dq_\phi(z_i)} =&\mathbb{E}_{-i}\left[\log p_\theta(z_i|z_{-i},R,E,V)\right] \\ &-\log q_\phi(z_i|E,V) - 1 = 0\end{aligned} \tag{6}$$

The final update rule for $q_\phi(z_i|E,V)$ is thus be obtained:

$$\begin{aligned}&q_\phi(z_i|E,V) \propto \exp\left(\mathbb{E}_{-z_i}[\log p_\theta(R,E,V,z_i,-z_i)]\right) \\ &\propto \exp\left(\mathbb{E}_{-z_i}\left[\log p_\theta(z_i|R,E,V,-z_i) + \log p_\theta(R,E,V,-z_i)\right]\right) \\ &\propto \exp\left(\mathbb{E}_{-z_i}[\log p_\theta(z_i|R,E,V,-z_i)]\right).\end{aligned} \tag{7}$$

where the third line is based on the independence property in the mean-field approximation.

2 More Experimental Results

Impact of Hyper-parameters: We have conducted the ablations to explore the impact of hyper-parameters α and β. As shown in the left part of Figure 1, the performance change is insignificant when the value of α changes. Hence, we take $\alpha = 0.5$ as the default configuration to take both cosine similarity and IoU distance into consideration when linking region proposals. As for β, the grounding performance will initially improve and then diminish when it

Fig. 2: Some illustration examples of the spatio-temporal video grounding predictions produced by our model on the VidSTG benchmark. The right part is the entities and relations in the query sentence.

increases, which proves the importance of modeling fine-grained relationships ($p_\theta(Z|E,R,V)$) among instances for settling the STVG task.

The Convergence of EM Update: We also visualized the changes of metric m_vIoU score of $p_\theta(Z|E,R,V)$ and $q_\phi(Z|E,V)$ in the validation set of VidSTG [1] benchmark during training, which reflects the correctness of the learned latent variables. The results are presented in the right part of Figure 1. As observed, these two update steps work synergistically during the optimization process, leading to continuously enhancing the performance of grounding. Therefore, the iteration stopped when we observed that the performance on the validation set started to fluctuate and no longer improve.

3 More Visualization Results

To provide a qualitative validation of our proposed approach, several additional illustrative examples of the results of spatio-temporal grounding on VidSTG [1] are presented in Figure 2. The visualizations clearly demonstrate that our model

generates highly precise predictions for the given sentence queries. Specifically, for the third visualization example, the desired object tube should contain be pulled by the adult woman. To ground the target, the model must learn to reason the sentence semantics and subtle clues "pull dog" in the video. Our framework can achieve this objective by reconstructing the detailed interactions between entities in the video, thereby localizing the target successfully. The same phenomenon can also be observed in the 4-th example, where the relationship "hold a dish" contributes to distinguishing the target woman from several candidates.

References

1. Zhang, Z., Zhao, Z., Zhao, Y., Wang, Q., Liu, H., Gao, L.: Where does it exist: Spatio-temporal video grounding for multi-form sentences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10668–10677 (2020)