(Supplementary Materials) Neighbor Does Matter: Global Positive-Negative Sampling for Vision-Language Pre-training

Anonymous Authors

Table 1: Results on fine-tune(FT) and zero-shot(ZS) retrieval tasks on Flickr30k dataset.

Method	Flickr30k-FT		Flickr30k-ZS	
	TR@1	IR@1	TR@1	IR@1
ALBEF	94.3	82.8	90.5	76.8
+ours	**94.9**	**83.2**	**92.8**	**78.7**
TCL	94.9	84.0	93.0	79.6
+ours	**95.8**	**84.7**	**93.1**	**79.9**

This Supplementary Material presents the results of fine-tune and zero-shot retrieval tasks on the Flickr30k[1] dataset in Table 1. Additionally, the pseudocode of GPN-S is provided in Algorithm 1.

Algorithm 1 Global Positive-Negative Sampling.

```
Input: 1) dataset with $D$ image-text pairs $\{(v_i, t_i)\}_{i=1}^{D}$.
    2) an off-the-shelf pre-trained model that consists of a visual encoder $g_V(\cdot)$ and a text encoder $g_T(\cdot)$.
    3) the model $f$.
% encoding image and text by the off-the-shelf encoders
$\mathbf{V} = \{\mathbf{v}_i^g\}_{i=1}^{D} = \{g_V(v_i)\}_{i=1}^{D}$
$\mathbf{T} = \{\mathbf{t}_i^g\}_{i=1}^{D} = \{g_T(t_i)\}_{i=1}^{D}$
% calculating neighbors and clusters
$V2V_k, V2T_k, T2V_k = Retrieve\_Topk\_Neighbor(\mathbf{V}, \mathbf{T})$
$Clusters = KMeans(\mathbf{V})$
% calculating global positive samples
for all $(v_i, t_i)$ do
    calculate $obj(v_i)$ by equation (7)
    $t_i^r =' \ '.join(obj(v_i))$
    calculate $SE_i$ by equation (9)
end for
% training.
for all Epoch do
    % organizing a N-pair batch with hard negatives
    $\mathcal{B} = \mathrm{GN\_S}(\ \{(v_i, t_i)\}_{i=1}^{D}\ ,\ Clusters)$
    for all $B_j \in \mathcal{B}$ do
        % GP-S
        for all $(v_i, t_i) \in B_j$ do
            $v_i^{\mathrm{pos}} = Random\_choice(SE_i)$
            % equation (8)
            if $(\mathbf{v}_i^g)^T \mathbf{t}_i^g < \alpha$ then
                $t_i = t_i^r$
            end if
        end for
        $v = \{v_i\}_{i=1}^{N}$
        $v^{\mathrm{pos}} = \{v_i^{\mathrm{pos}}\}_{i=1}^{N}$
        $t = \{t_i\}_{i=1}^{N}$
        % equation (10), (11)
        $\mathcal{L}(v, t, v^{\mathrm{pos}}) = \mathcal{L}_f(v, t) + \mathcal{L}_{\mathrm{uni}}(v, v^{\mathrm{pos}})$
        $\mathcal{L}.backward()$
    end for
end for
```

REFERENCES

[1] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.