Negative-Aware Attention Framework for Image-Text Matching (Supplementary Material)

Kun Zhang[1], Zhendong Mao[1]*, Quan Wang[2], Yongdong Zhang[1]
[1]University of Science and Technology of China, Hefei, China; [2]MOE Key Laboratory of Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications, Beijing, China
kkzhang@mail.ustc.edu.cn, {zdmao, zhyd73}@ustc.edu.cn, wangquan@bupt.edu.cn

1. More Analysis of Visualization

We show more visualizations to verify the effectiveness and superiority of our proposed negative-aware attention framework (NAAF). Generally, existing methods mainly focus on discovering all matched word-region pairs, maximally improving the relevance of matched word-region fragments as meaningful scores while simply weakening or even erasing the mismatched word-region fragments, to measure image-text similarity. However, this will be inevitably prone to produce false-positive matching, as shown in Fig.1, where the false image-text pairs containing many matched word-region fragments can still obtain the high similarity and may rank quite the top as correct, which leads to inaccurate matching. From the visualization results of Q1-Q8, we can see that the negative role of these crucial image-text mismatching clues, *e.g.*, 'long hair' in Q1, 'camera' in Q2, and so on, in the existing methods is typically underestimated or neglected.

In contrast, with respect to these false-positive image-text pairs in our NAAF, as shown in Fig.1, we can observe that NAAF not only focuses on matched fragments but also discriminates subtle mismatched ones across modalities. For example, the 'girl' in Q5 is wrongly regarded to match the image region boy in the existing method, but it can be accurately located as the mismatched region in our NAAF. As a result, these crucial mismatched textual clues are able to explicitly reflect their negative effects to make more accurate image-text matching performance.

2. Derivation of Penalty Parameter α^*

The penalty parameter α is the weight of distinguishing errors of mismatched fragments, which determines the ability to mine mismatched fragments during the training process. As we argued in the paper that the learning boundary is expected to converge to the state that guarantees the maximum mining of mismatched fragments and avoids misjudgment of matched fragments causing performance degradation, we therefore want to find an appropriate penalty parameter α^* to achieve this. Thus, we described the detailed derivation of α^* in this section.

Assuming that the learning boundary t_k (Eq.4 in the paper) is a function of decision variable α, *i.e.*, $t_k(\alpha)$, the problem can be formulated as the following constrained probability optimization problem:

$$\begin{aligned} \alpha^* = \quad & \max_{\alpha} \int_{-\infty}^{t_k(\alpha)} f_k^-(s)ds, \\ & \text{s.t. } \int_{t_k(\alpha)}^{+\infty} f_k^+(s)ds \approx 1,\ \alpha > 0, \end{aligned} \tag{1}$$

where the objective function means to maximally mining the mismatched fragments in the similarity distribution $f_k^-(s)$, the constraints indicate that the judgment probability for matched fragments is as close to 1 as possible to ensure the mining accuracy.

The solution process of α^* can be divided into two steps. In brief, we first solve the feasible solution set, and then obtain the optimal solution through projection. 1) In order to meet the constraints, we determine the lower range of $t_k(\alpha)$ according to the probability limit theory: $[0, t^*]$, where t^* can be obtained based on the empirical lower bound, *i.e.*, $\mu - 3\sigma$, of matched similarity distribution with a probability of near 1. 2) Since the objective function is an increasing function about $t_k(\alpha)$, it is optimal as long as $t_k(\alpha)$ approximates the maximum value of feasible solutions, *i.e.*, $\lim_{\alpha \to \alpha^*} t_k(\alpha) = t^*$. According to the specific formula of the learning boundary, we can obtain the theoretically optimal penalty parameter as:

$$\alpha^* = \sigma_k^- [\sigma_k^+ exp^{\frac{\beta_k}{2(\sigma_k^{+2} - \sigma_k^{-2})}}]^{-1} \tag{2}$$

where $\beta_k = [\sigma_k^+(\mu_k^+ - \mu_k^-)/\sigma_k^- - 3(\sigma_k^{+2} - \sigma_k^{-})]^2 - (\mu_k^+ - \mu_k^-)^2$. It is worth noting that in the later stage of training, the range of matched and mismatched similarity distributions has stabilized, so only one adjustment of the penalty parameter is sufficient to meet the requirements. Empirically, it can

*Zhendong Mao is the corresponding author.

Figure 1. Visual comparison of negative effects of mismatched words (blue) and positive effects of matched words (red) in our NAAF and existing methods. The darker the color, the greater the effect. With respect to the query text, the given image is the false-positive one, which means the incorrect candidate is ranked top-1 to be considered as the matching one in existing method. Our NAAF can explicitly exploit both negative effects of mismatched clues and positive effects of matched clues to well eliminate these false-positive pairs.

be adjusted at three-quarters of the total number of training epochs. The premature adjustment will lead to insufficient mining of mismatched clues, because as we verified in the ablation study, small penalty parameters cannot be effectively mined and have relatively poor performance.