Efficient Vision-Language Pre-training by Cluster Masking

Supplementary Material

1. Visualization of attention-based masking

We extend the teaser image with examples from the attention-guided baseline (Figure 1). In contrast to our RGB model, the behavior of the attention-based method changes during training. In early iterations, it masks randomly, while later in training it produces fairly consistent clusters that do not vary much between iterations, since the attention maps change less over time, potentially limiting the diversity of training examples.

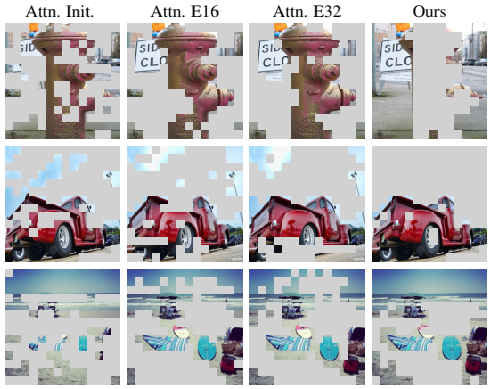

Figure 1. Visualization of attention-based masking. The images are the same from the teaser image.

2. Clustering Visualization

We provide more examples of our clustering masking visualization on COCO and CC3M datasets on Figure 2 and Figure 3 respectively. We mask out at least 50% patches in each image.

Figure 2. **Illustration of cluster based masks on COCO.**

Figure 3. **Illustration of cluster based masks on CC3M.**