Quantizing Visual Features

Even though we were able to reduce the impact of the image-specific information in the teacher's [I_CLS] token by introducing the contrastive target loss, the problem remains the same. A repeated glance on the comparison between the training loss for the image and text component of the $\mathcal{L}_{\text{KD}}$ loss, which is now based on a contrastive loss with memory bank, shows that the loss for the image component $\mathcal{L}_{\text{KD}}^{\text{i2i}}$ is still significantly lower than that for the text component $\mathcal{L}_{\text{KD}}^{\text{t2i}}$. While our approach is able to achieve impressive results even with this imbalance, we push the boundaries by introducing an additional component that aims to further reduce the impact of the image-specific information.