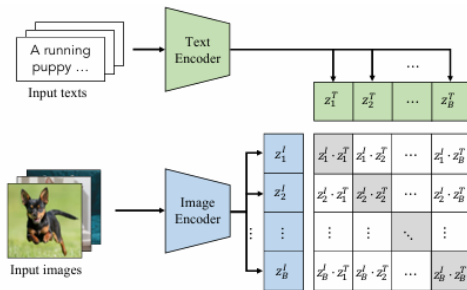


Image-text Contrastive Learnings



pulling the embeddings of paired images and texts close while pushing others

Loss Formulation

$$\mathcal{L}_{I \rightarrow T} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(z_i^I \cdot z_i^T / \tau)}{\sum_{j=1}^B \exp(z_i^I \cdot z_j^T / \tau)}$$

$$\mathcal{L}_{T \rightarrow I} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(z_i^T \cdot z_i^I / \tau)}{\sum_{j=1}^B \exp(z_i^T \cdot z_j^I / \tau)}$$