SemMAE: Semantic-Guided Masking for Learning Masked Autoencoders

基于语义指导的掩码用于学习掩码自编码器

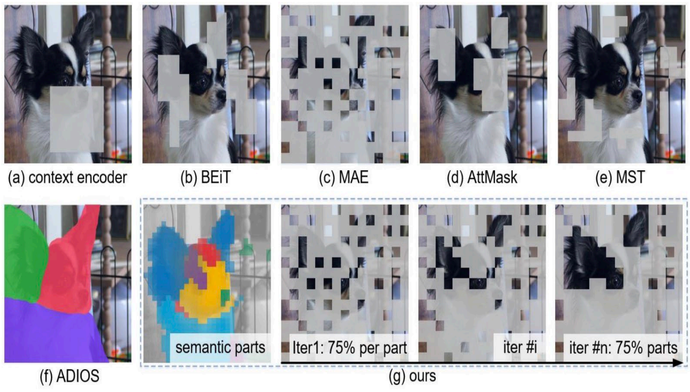
Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, Changwen Zheng

Neural Information Processing Systems (NeurIPS), 2022

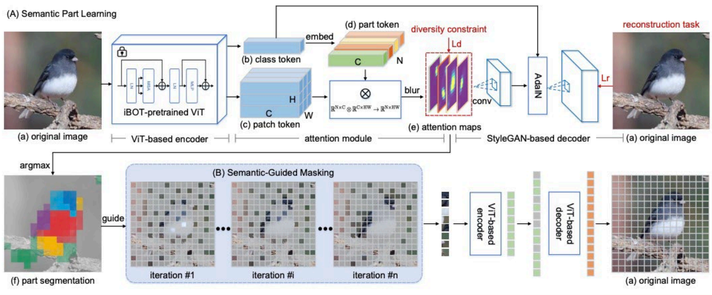
ucasligang@gmail.com, 13721371391

1.Research Motivation

The lack of semantic decomposition of images still makes masked autoencoding (MAE) different between vision and language. We explore a potential visual analogue of words, i.e., semantic parts, and we integrate semantic information into the training process of MAE by proposing a Semantic-Guided Masking strategy. Compared to widely adopted random masking, our masking strategy can gradually guide the network to learn various information, i.e., from intra-part patterns to inter-part relations.



2.Our Approach



SemMAE consists of two parts: :

（A）Semantic Part Learning:

We use the Mean squared error (MSE) loss function to optimize such econstruction task:

$$\mathcal{L}_{rec}(\mathbf{I},\hat{\mathbf{I}}) = \frac{1}{HW}\sum_{i,j}^{HW}(\mathbf{I}(i,j) - \hat{\mathbf{I}}(i,j))^2.$$

Moreover, to obtain diverse multiple attention maps, we add a diversity constraint over attention maps:

$$\mathcal{L}_{div}(\mathbf{M}) = \frac{1}{N^2}(\sum_{i\neq j}(0 - \frac{\mathbf{m}_i\mathbf{m}_j^T}{\|\mathbf{m}_i\|_2\|\mathbf{m}_j\|_2})^2 + \sum_{i=j}(1 - \frac{\mathbf{m}_i\mathbf{m}_j^T}{\|\mathbf{m}_i\|_2\|\mathbf{m}_j\|_2})^2),$$

The overall objective function can be denoted by:

$$\mathcal{L} = \mathcal{L}_{rec}(\mathbf{I},\hat{\mathbf{I}}) + \lambda\mathcal{L}_{div}(\mathbf{M})$$

(B) Semantic-Guided Masking:

We define two masking settings, i.e., 1) mask a portion of patches in each part and 2) random select some parts to mask (the whole part). The number of masked patches for each semantic part can be calculated for these two settings. A interpolation hyper-parameter α is introduced to adjust the weight of two settings. Finally, we random mask a certain number of patches based on the calculated masking number.

3.Summary

In this work, we study the visual analogue of words and propose a semantic-guided masked autoencoder model to reduce the gap between masked language modeling and masked image modeling. Our proposed self-supervised semantic part learning method can generate promising semantic parts on ImageNet and we show that the learned semantic parts can facilitate the learning of MAE. Unlike the main-stream random masking strategy, our semantic-guided mask strategy can effectively integrate semantic information in the pre-training process.

作者Google Scholar主页：
https://scholar.google.com/citations?user=StWrqHIAAAAJ&hl=en
欢迎大家引用和交流！