Simple Masked Image Modeling

Major componends

↳ Masking strategy - what to, how to mask

↳ Encoder - feature extraction, masked area prediction

↳ Prediction head - generative head

↳ Prediction target - target + loss

Masking Strategy

↳ replacing the masked patches w/ mask token vector

- Patch-aligned random masking

  ↳ The one described above, patch-level masking as it is convinient

- Other strategies have also been tried

  ↳ central region masking

  ↳ block-wise masking

Prediction head

↳ a single linear layer is sufficient

Prediction Targets

↳ Raw pixel regression

↳ Tested 1x, x32, x16, x8, x4, x2 downsampled targets

↳ $\mathcal{L}_1$-loss ($\mathcal{L}_2$ and $\mathcal{L}_1$ smooth also tried)

Encoder

↳ Vanilla ViT and SWin Transformer

(Some) Results

default
patch:32
ratio:0,6

↳ Simple masking strategy seemed to be the best, with large patchsize (32) and masking % of 10 - 70
↓
On image size of 192^2

↳ larger masking ratio seem to help on smaller patchsizes as well

↳ smaller heads (1-2 linear layers) result in better fine-tuned results as the head can't learn much