Masked Autoencoders Are Scalable Vision Leaners

Kaiming He et al.

Facebook AI Research

Nov 2021

Justin Kim

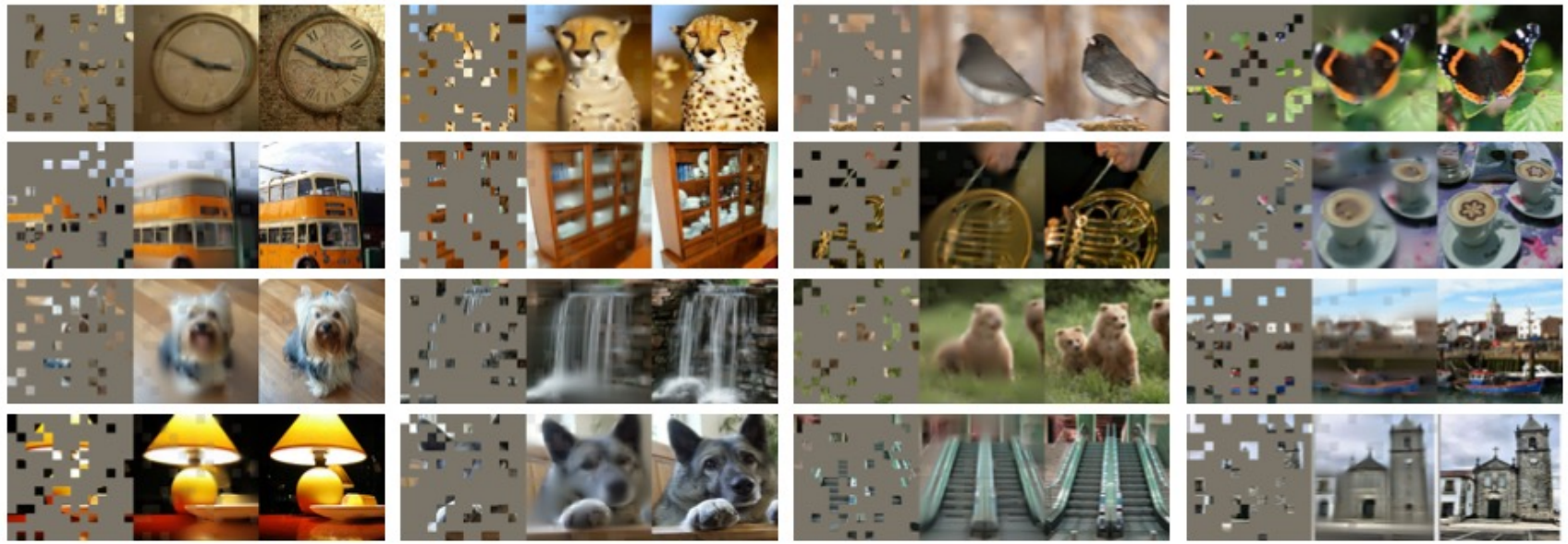


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction[†] (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix.
[†]*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

Figure 3. Example results on COCO validation images, using an MAE trained on ImageNet (the same model weights as in Figure 2). Observe the reconstructions on the two right-most examples, which, although different from the ground truth, are semantically plausible.

Motivation behind MAE

- architectures of continuously growing capability and capacity

-> issue = easy to overfit

-> solution?

NLP

remove portion of data and learn to predict the removed portion

GPT - autoregressive language modeling

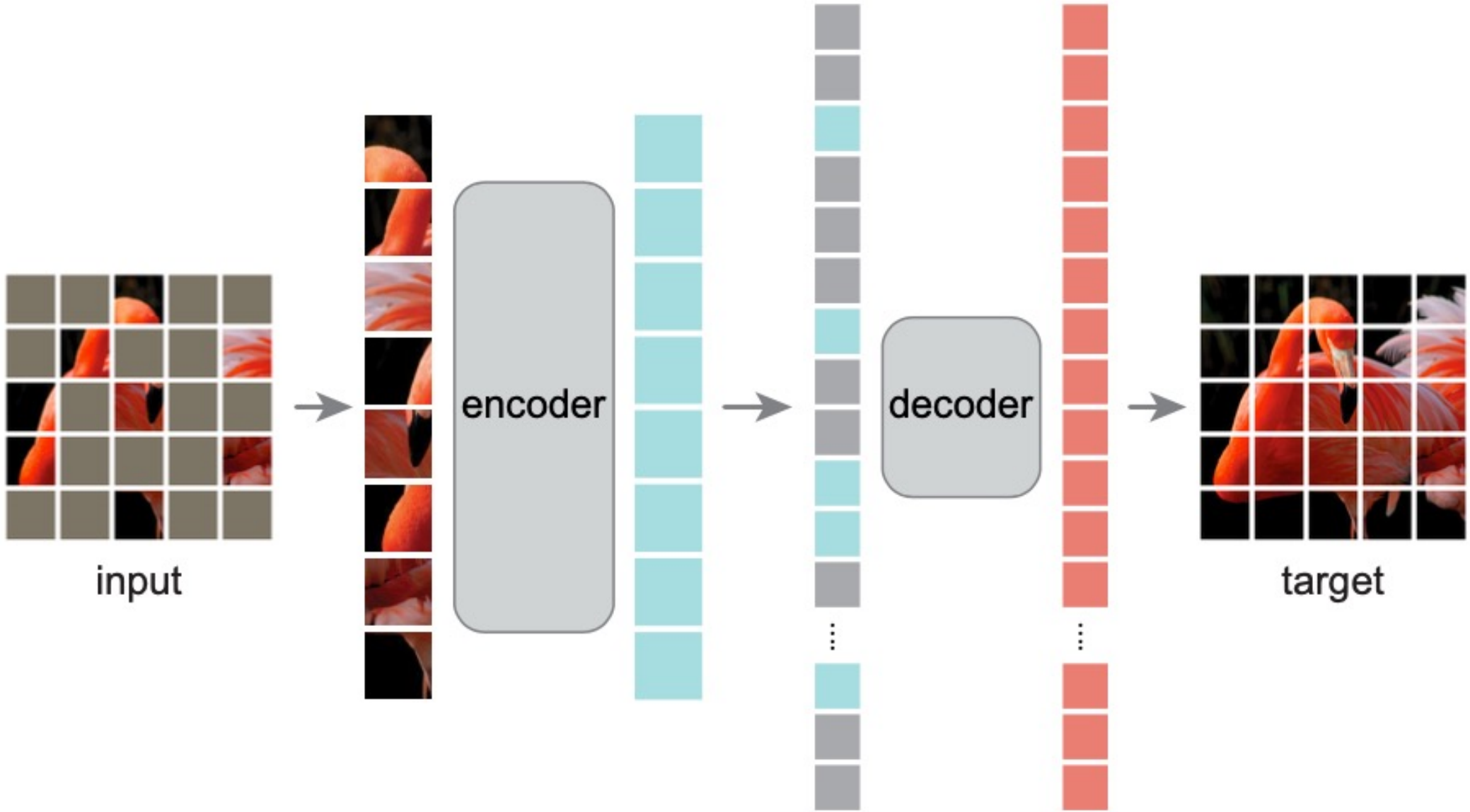
BERT – masked autoencoding

CV

???

Apply similar strategies to CV?

- Issues or why not so simple?
  - Architectural differences
    - CV = CNN
    - NLP = Transformer
  - Information differences
    - Image = heavy spatial redundancy
    - Language = highly semantic
  - Decoder differences
    - CV = pixel = low semantic
    - NLP = word = high semantic


input
encoder
decoder
target

Results

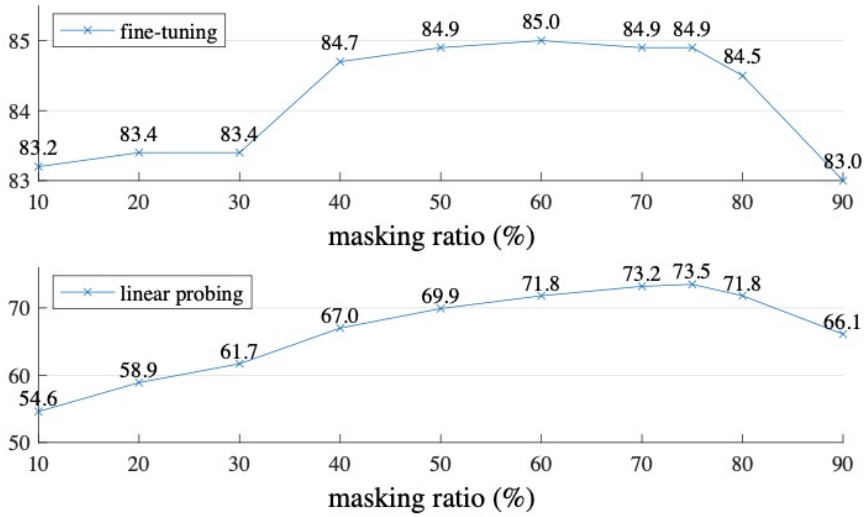


Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

Baseline: ViT-Large. We use ViT-Large (ViT-L/16) [16] as the backbone in our ablation study. ViT-L is very big (an order of magnitude bigger than ResNet-50 [25]) and tends to overfit. The following is a comparison between ViT-L trained from scratch *vs*. fine-tuned from our baseline MAE:

scratch, original [16]	scratch, our impl.	baseline MAE
76.5	82.5	84.9