An introduction to Self-supervised Learning in Computer Vision

Speaker: Xiangyu Peng
School of Computing, National University of Singapore
Advisor: Yang You

Supervised Learning

Supervised learning: data with annotations

- Classification, object detection, segmentation
- Labels are expensive
- Labels are limited

Self-supervised Learning

Self-supervised learning: data without annotations

- Visual data is everywhere: images, videos
- Much more unlabeled data than labeled data
- Can we make use of the enormous unlabeled data?

Self-supervised Learning

- How?
  - Design learning targets, or pretext tasks
- Goal?
  - Learn representations that carry good semantic or structural meanings
  - Trasferrable representations are preferred for downstream tasks.

Pretext Tasks

- Spatial relationship: rotation, context, jigsaw puzzle
- Contrastive learning: metric learning, clustering, instance discrimination
- Inpainting: colorization
- ……

Spatial Relationship

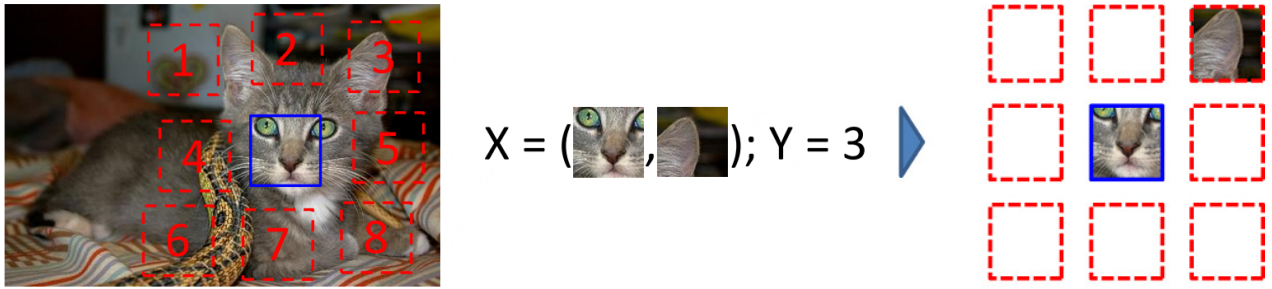
Predicting Rotation

$\theta = 0$ $\theta = \pi/2$ $\theta = \pi$ $\theta = 3\pi/2$

Unsupervised Representation Learning by Predicting Image Rotations. ICLR 2018

Spatial Relationship

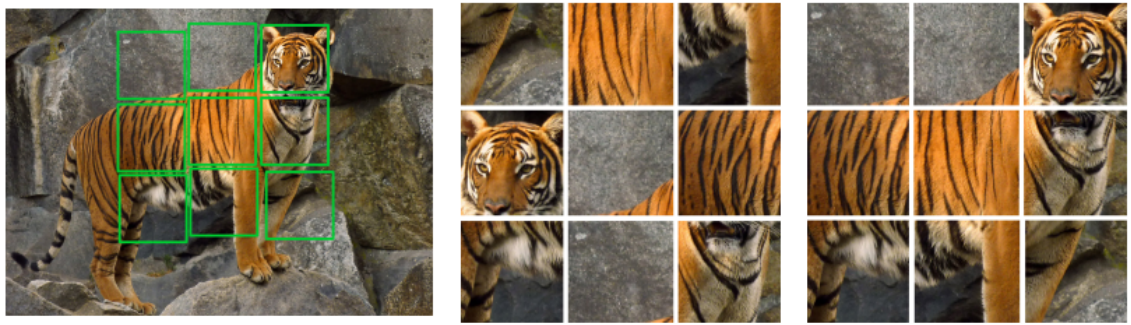
Predicting Relative Position



Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

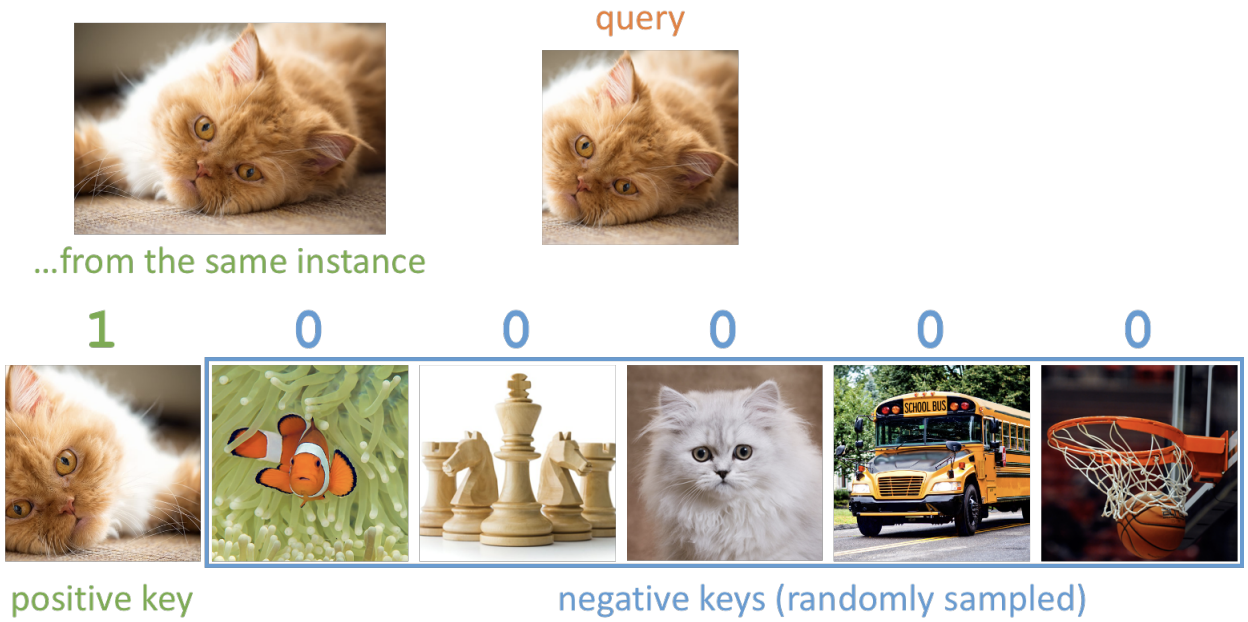
Spatial Relationship

Predicting Patch Organization



Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. ECCV 2016

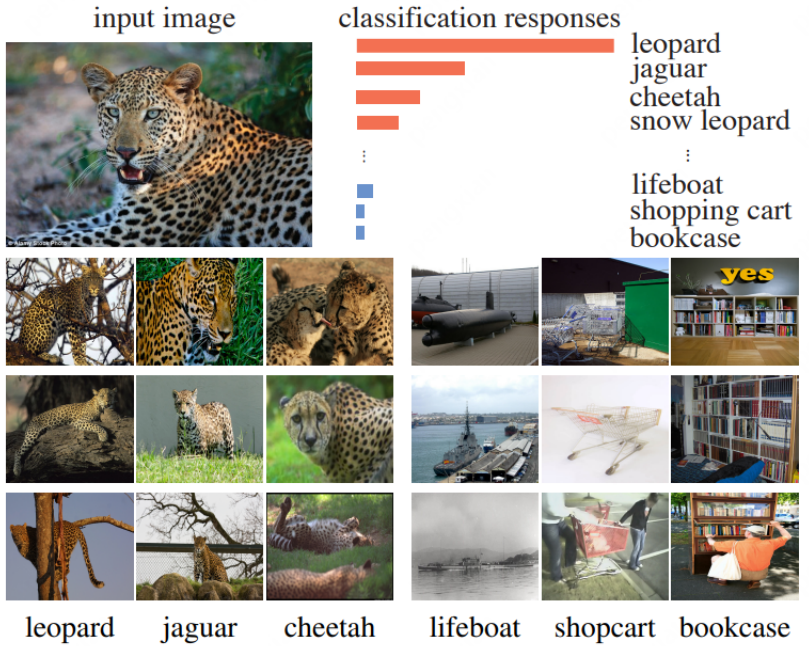
Contrastive Learning



Contrastive Learning

An early attempt: **InstDic**

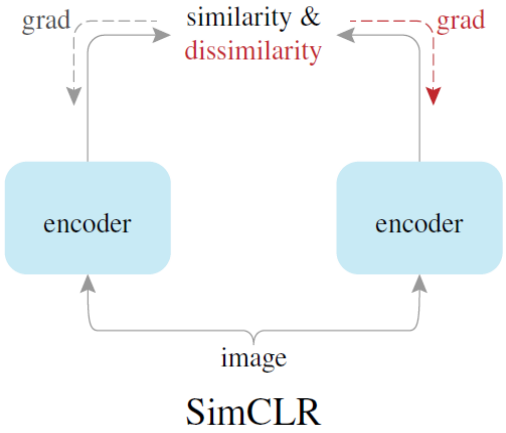
- Class-wise supervision
- Instance discrimination



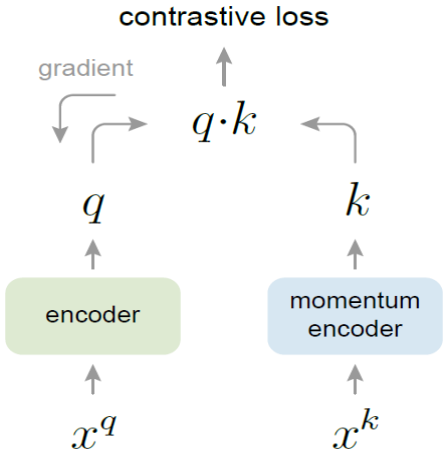
Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. CVPR 2018

Contrastive Learning

Two milestones: Surpassing supervised pre-training



SimCLR

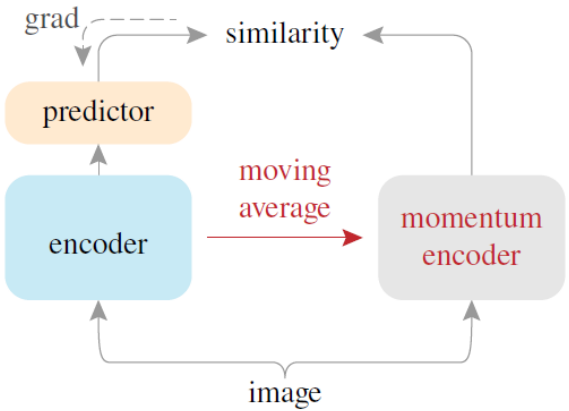


MoCo

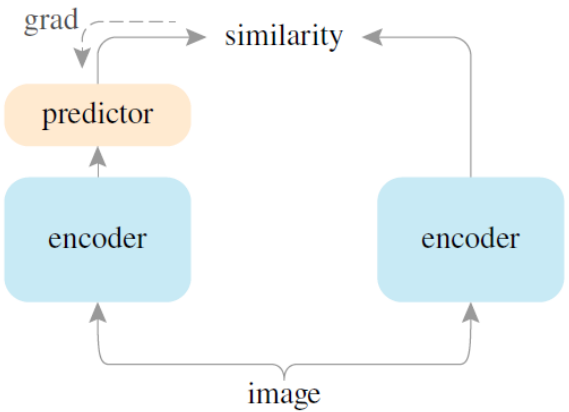
A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020
Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020 Oral

Contrastive Learning

Without negative samples: **BYOL, SimSiam**



BYOL

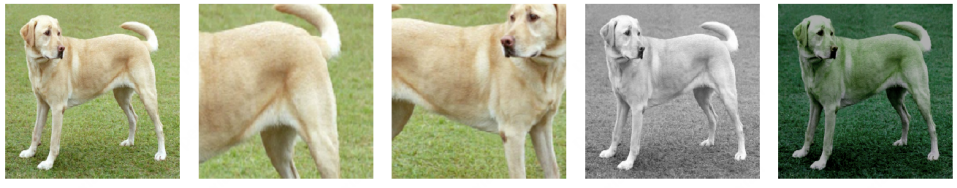


SimSiam

Bootstrap your own latent: A new approach to self-supervised Learning. NeurIPS 2020 Oral
Exploring Simple Siamese Representation Learning. CVPR 2021 Oral

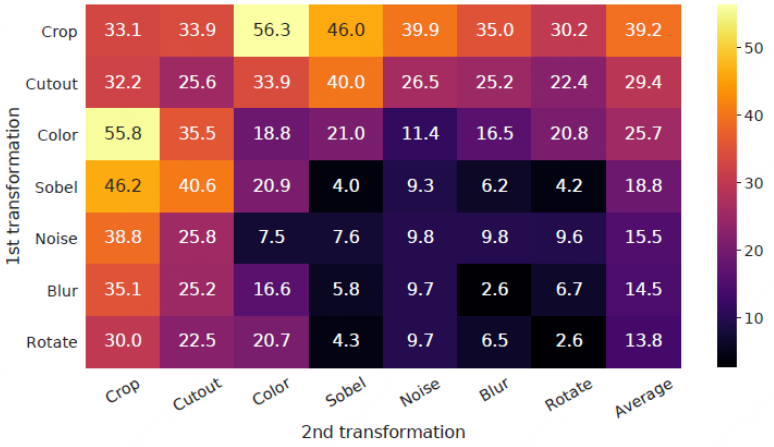
Contrastive Learning

Key: Augmentation



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

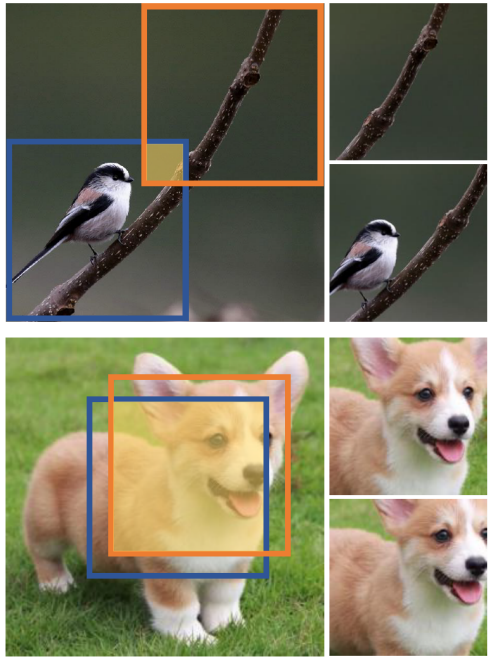
(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering



Contrastive Learning

Conventional approach: ***RandomCrop***

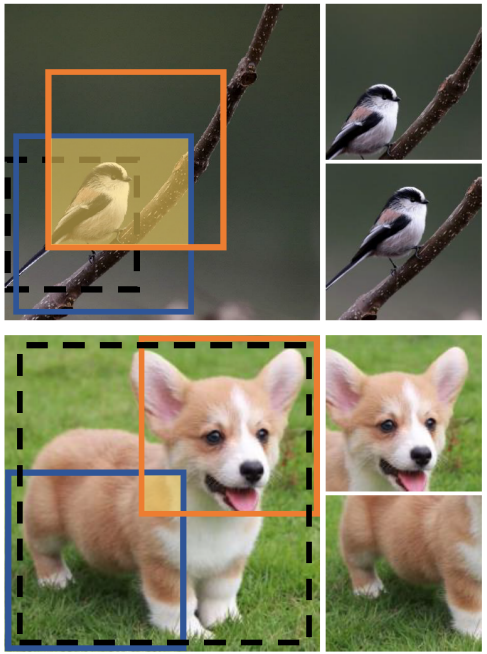
- False positive: object vs. background
- Trivial pair: too similar for optimization



Contrastive Learning

Our method: ***ContrastiveCrop***

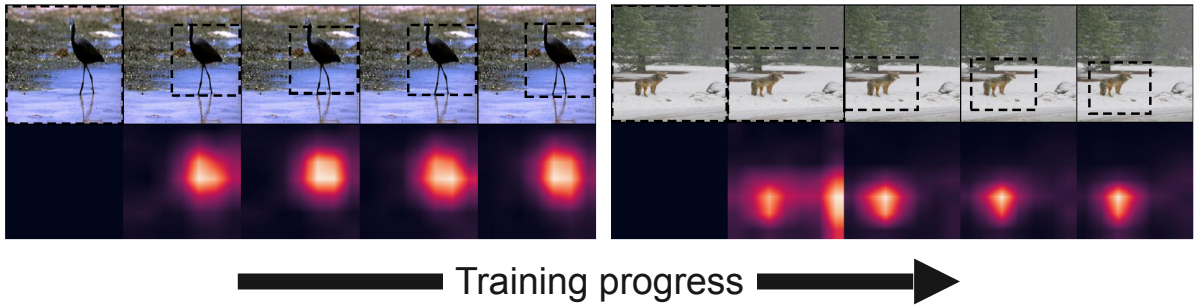
- Take semantic information into account
- Increase variance between positive views



ContrastiveCrop

Method: Semantic-aware Localization

- Generate a bounding box of the object from the heatmap
- Use the bounding box as a guidance to generate crops



Crafting Better Contrastive Views for Siamese Representation Learning. CVPR 2022 Oral

ContrastiveCrop

Center-suppressed distribution:

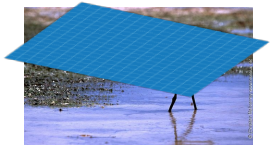
- Lower probability near the center, higher probability at other positions
- Larger sampling variance ⟶ Smaller overlap, less similarity

Uniform

Center-suppressed

ContrastiveCrop

RandomCrop

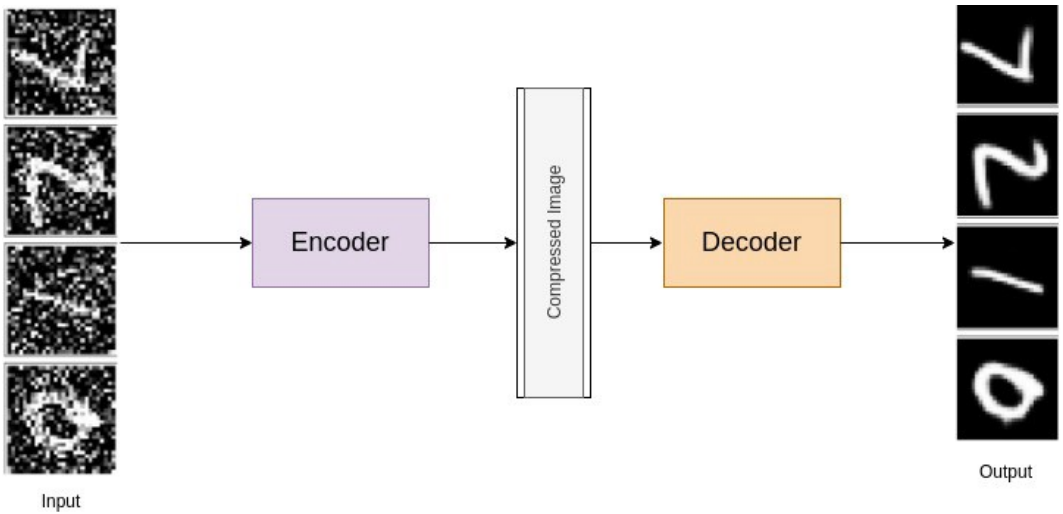


ContrastiveCrop

Crafting Better Contrastive Views for Siamese Representation Learning. CVPR 2022 Oral

Inpainting

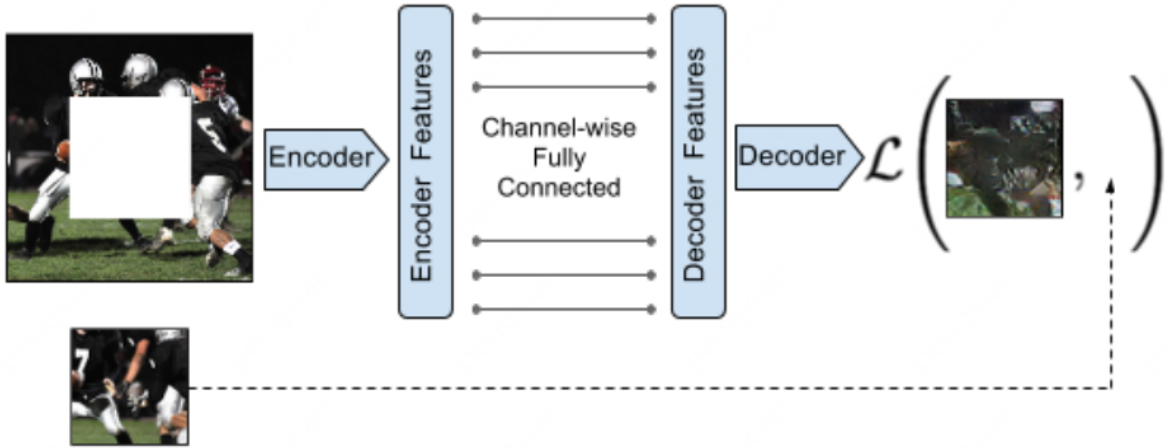
Denoising autoencoder



Extracting and Composing Robust Features with Denoising Autoencoders. ICML 2008

Inpainting

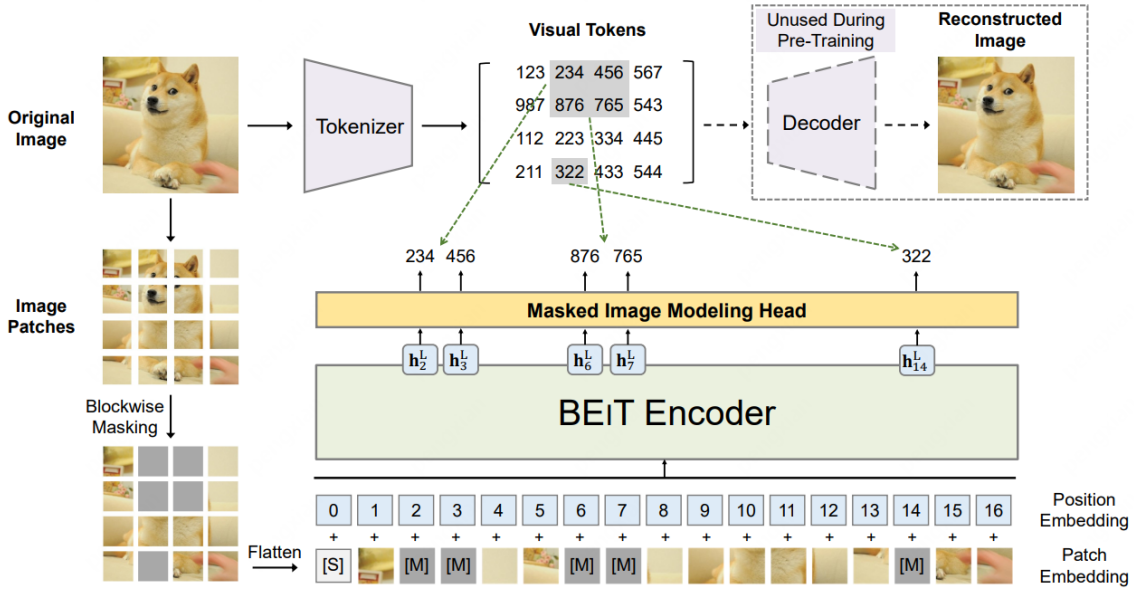
Context Encoder



Context encoders: Feature learning by inpainting. CVPR 2016

Inpainting

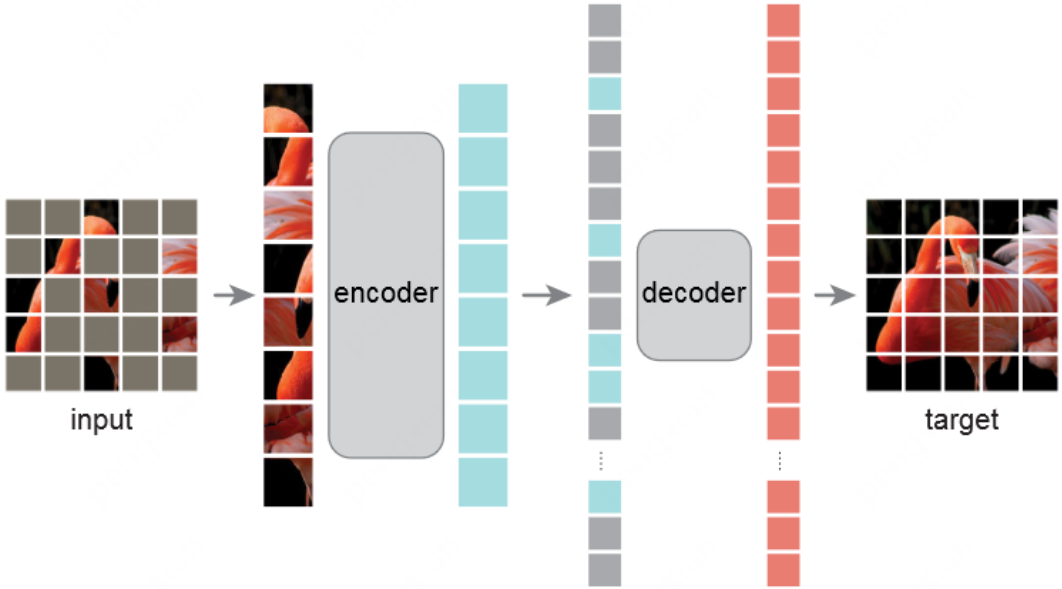
BEiT



BEiT: BERT Pre-Training of Image Transformers. ICLR 2022

Inpainting

Masked AutoEncoder (MAE)



Masked Autoencoders Are Scalable Vision Learners. CVPR 2022 Oral

Problems

What do we really want to learn from visual data?

- Not well defined
- Evaluation metrics:
  - ➢ Linear probing
  - ➢ Fine-tuning on downstream tasks

Problems

Can we scale to large-scale unlabeled data?

- Currently pure vision lags far behind NLP
  - ➢ NLP: GPT3, Wu Dao, …
  - ➢ CV: None
- Vision & Language is promising
  - ➢ DaLL·E 2, Imagen, ...

Conclusion

Y. LeCun

How Much Information is the Machine Given during Learning?

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

THANK YOU

ContrastiveCrop: https://github.com/xyupeng/ContrastiveCrop