MoCo

Momentum Contrast for Unsupervised Visual Representation Learning (2020) 5.6k

Outperform its supervised pre-training counterpart in 7 detection or segmentation tasks on PASCAL; close the gap between unsupervised and supervised representation learning

Contrastive learning:

To learn such an embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart

Example: instance discrimination

positive: x_i after transformation; negative: $x_j (j \neq i)$

$x_1 \rightarrow x_1^1(x^q)$ (anchor) $\rightarrow E_{11} \rightarrow f_{11}$ — query

$\searrow x_1^2(x^k)$ (positive) $\rightarrow E_{12} \rightarrow f_{12}$ $f_2, f_3, ..., f_N$ — key

$x_2, x_3, ..., x_N$ (negative) $\nearrow$ Embedding

Momentum: $y_t = my_{t-1} + (1-m)x_t$

Pretext tasks: the term “pretext” implies that the task being solved is not of genuine interest, but is solved only for the true purpose of learning a good data representation

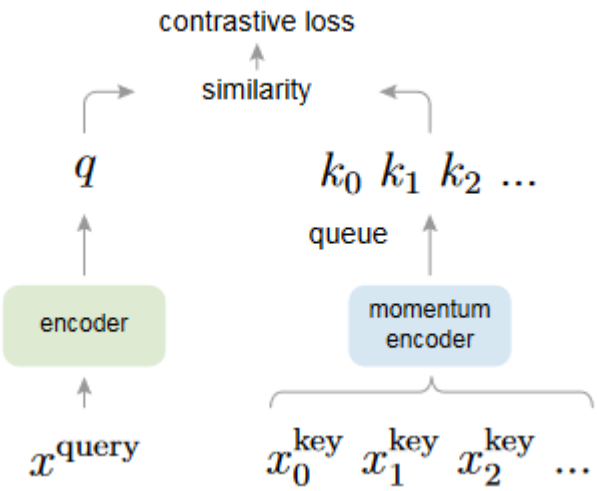
Contrastive losses: measure the similarities of sample pairs in a representation space

Adversarial losses: measure the difference between probability distributions

Contrastive learning can be viewed as a way of building a discrete dictionary on high-dimensional continuous inputs such as images. The dictionary is dynamic in the sense that the keys are randomly sampled, and that the key encoder evolves during training.

Dictionaries for contrastive learning:

- large
- consistent as they evolve during trainin → represented by the same or similar encoder



InfoNCE (noise contrastive estimation): $L_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}$ where τ is a temperature hyper-parameter

Dictionary as a queue: maintain the dictionary as a queue of data sample, so our dictionary size can be much larger than a typical mini-batch size

Momentum update:

encoder θ_q

decoder $\theta_k = m\theta_{k-1} + (1-m)\theta_q$

A relatively large momentum (e.g., m = 0.999) works much better than a smaller value, suggesting that a slowly evolving key encoder is a core to making use of a queue.

Experiments:

Linear Classification Protocol: unsupervised pre-training → freeze the features and train a supervised linear classifier (a fully-connected layer followed by softmax)

- Ablation: contrastive loss mechanisms
- Ablation: momentum
- Comparison with previous results

Transferring Features: A main goal of unsupervised learning is to learn features that are transferrable. Moco outperforms its ImageNet supervised pre-training counterpart in 7 detection or segmentation tasks.

Discussion: hope to adopt MoCo for pretext tasks like masked auto-encoder — MAE paper two years later!