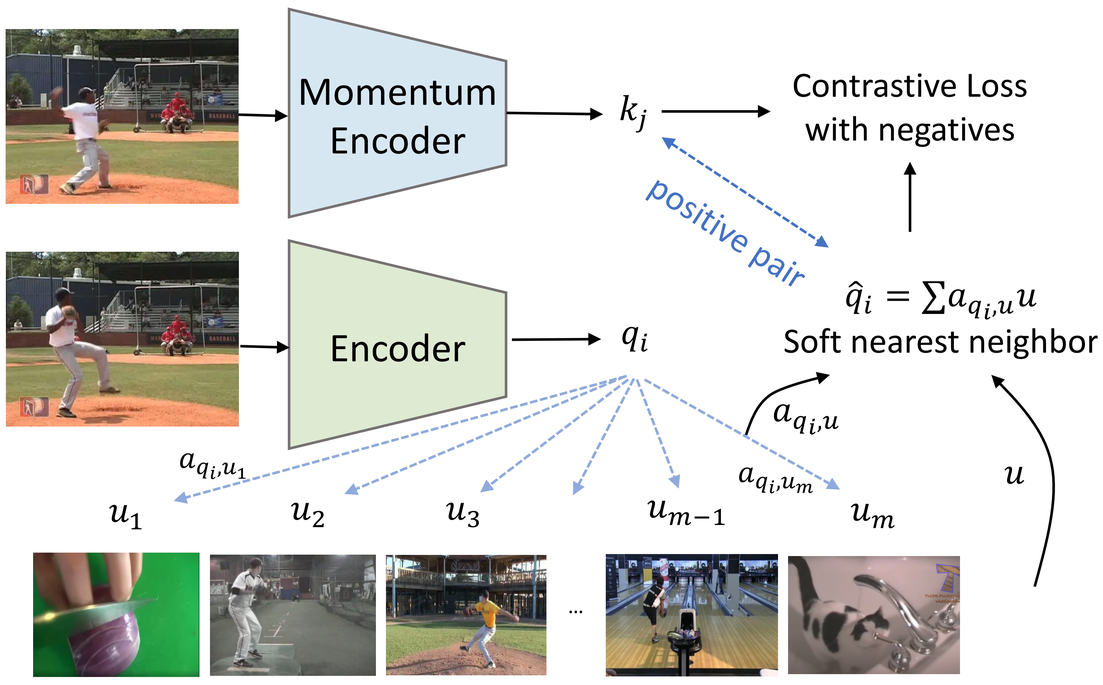
Contrastive Learning of Image Representations with Cross Video Cycle-Consistency

Haiping Wu Xiaolong Wang
McGill University, Mila UC San Diego

Introduction

- In this paper, we propose a novel contrastive learning method which explores the cross-video relation by using cycle-consistency for general image representation learning.
- We validate our method by transferring our image representation to multiple downstream tasks including visual object tracking, image classification, and action recognition.

Cross-video Cycle Consistent Learning



Given the encoded representation q_i for an image,

- Step 1. Cycle Forward. Find the soft nearest neighbor $\hat{q}_i$ in the candidate neighbor set U

$$\hat{q}_i = \sum_{u \in U} \text{sim}_\tau(q_i, u) u, \tag{1}$$

where $\text{sim}(\cdot)$ calculates the normalized cosine similarity with temperature τ.

- Step 2. Cycle Backward. $\hat{q}_i$ should be close to q_i or invariant set of q_i. Thus we use k_j, which is the representation of image within the same video to form the positive pair $(\hat{q}_i, k_j)$. The loss is

$$\mathcal{L}_{\text{cycle}} = -\log \frac{\exp\left(\text{sim}\left(\hat{q}_i, k_j\right)/\tau\right)}{\sum_{u \in \{U, k_j\}} \exp\left(\text{sim}\left(\hat{q}_i, u\right)/\tau\right)}, \tag{2}$$

Experiments & Results

Visual object tracking.

Method	Backbone	Dataset	OTB Precision	OTB Success
Supervised	ResNet-18	ImageNet	61.4	43.0
SimSiam	ResNet-18	ImageNet	58.8	42.9
MoCo	ResNet-18	ImageNet	62.0	47.0
VINCE	ResNet-18	R2V2	62.9	46.5
Ours	ResNet-18	R2V2-S	**65.6**	**48.6**
upervised	ResNet-50	ImageNet	65.8	45.5
SimSiam	ResNet-50	ImageNet	61.0	43.2
MoCo	ResNet-50	ImageNet	63.7	46.5
SeCo	ResNet-50	Kinetics	71.9	51.8
VINCE	ResNet-50	R2V2	40.2	30.0
Ours	ResNet-50	R2V2	69.3	49.2

Image Classification

Methods	Backbone	Dataset	ImageNet Top-1 (%)
Supervised	ResNet-50	ImageNet	76.2
MoCo	ResNet-50	ImageNet	67.7
MoCo	ResNet-50	R2V2	53.6
VINCE	ResNet-50	R2V2	54.4
Ours	ResNet-50	R2V2	**55.6**

Video Action Recognition

Method	Backbone (#Param)	Dataset	UCF101
3D-RotNet	3D-R18-full (33.6M)	K-400	62.9
SpeedNet	I3D(12.1M)	K-400	66.7
DPC	3D-R18(14.2M)	K-400	68.2
DPC	3D-R34	K-400	75.7
Video-Pace	R(2+1)D (33.3M)	K-400	77.1
CBT	S3D	K-400	79.5
MemDPC	R-2D3D (32.4M)	K-400	78.1
Temporal-ssl	R(2+1)D (33.3M)	K-400	81.6
VTHCL	3D-R50(31.7M)	K-400	82.1
Ours	R-18 (**11.69** M)	R2V2	76.8
Ours	R-50(**25.56** M)	R2V2	**82.1**