SeCo: Exploring Sequence Supervision for Unsupervised Representation Learning

Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, Tao Mei
JD AI Research, Beijing, China
{tingyao.ustc, yihengzhang.chn, zhaofanqiu, panyw.ustc}@gmail.com, tmei@jd.com

Abstract

A steady momentum of innovations and breakthroughs has convincingly pushed the limits of unsupervised image representation learning. Compared to static 2D images, video has one more dimension (time). The inherent supervision existing in such sequential structure offers a fertile ground for building unsupervised learning models. In this paper, we compose a trilogy of exploring the basic and generic supervision in the sequence from spatial, spatiotemporal and sequential perspectives. We materialize the supervisory signals through determining whether a pair of samples is from one frame or from one video, and whether a triplet of samples is in the correct temporal order. We uniquely regard the signals as the foundation in contrastive learning and derive a particular form named Sequence Contrastive Learning (SeCo). SeCo shows superior results under the linear protocol on action recognition (Kinetics), untrimmed activity recognition (ActivityNet) and object tracking (OTB-100). More remarkably, SeCo demonstrates considerable improvements over recent unsupervised pre-training techniques, and leads the accuracy by 2.96% and 6.47% against fully-supervised ImageNet pre-training in action recognition task on UCF101 and HMDB51, respectively. Source code is available at https://github.com/YihengZhang-CV/SeCo-Sequence-Contrastive-Learning.

Introduction

Supervised learning has made significant progress and is still dominant in visual representation learning. Despite having high quantitative performances, the achievements rely heavily on the requirement to have large number of expert annotations for training deep neural networks, and the acquisition of annotations is an intellectually expensive and time-consuming process. Moreover, the representations especially learnt on very specific tasks in a supervised manner may suffer from generalization problem and transfer poorly to other objectives. In contrast, unsupervised representation learning alleviates the issues by completely exploiting the inherent structures and correlations from the data as the supervision. This is particularly applicable to video, which is an information-intensive media with spatiotemporal coherence and variation. Such facts motivate the explorations of building unsupervised learning models to yield powerful and generic representations.

The supervision in the video sequence generally originates from three types: spatial, spatiotemporal, and sequential. In between, spatial supervision is derived from the structures in static frame, spatiotemporal supervision reflects the correlation across different frames, and sequential supervision verifies the temporal coherence. In the literature, unsupervised learning methods for videos often involve different proxy tasks, e.g., predicting the pixel-level displacement across consecutive frames (Liu et al. 2017; Vondrick, Pirsiavash, and Torralba 2016; Wang et al. 2019), or reconstructing/predicting the input/future frame through decoder (Han, Xie et al. 2019; Luo et al. 2017; Srivastava, Mansimov, and Salakhutdinov 2015), and execute representation learning through optimizing such tasks with the supervision. Here, without loss of simplicity and generality, we present one simple proxy task on each type of supervision. From the spatial standpoint, we extend the instance discrimination task in (He et al. 2019; Wu et al. 2018; Cai et al. 2020) to an intra-frame instance discrimination task, which distinguishes whether two frame patches are from the same video frame, as shown in Figure 1(a). From the spatiotemporal perspective, we remould an inter-frame instance discrimination task, which determines whether two frame patches are derived from an identical video, as depicted in Figure 1(b). For sequential supervision, we develop a task of temporal order validation (Figure 1(c)) and verify whether a series of frame patches are in the correct temporal order.

To materialize the exploitation of supervision in the sequence through the three proxy tasks, we present a new Sequence Contrastive Learning (SeCo) approach for unsupervised representation learning. Considering that contrastive learning is at the core of recent advances (He et al. 2019; Wu et al. 2018) on unsupervised learning, we build SeCo on this recipe. The basic principle is to make positive/negative query-key pairs similar/dissimilar. Specifically, for each video, we randomly sample three frames and take either first frame or the last frame in time order as the "anchor" frame. In both intra-frame and inter-frame instance discrimination tasks, we perform data augmentation on the "anchor" frame to generate two image patches. One is taken as query and the other patch plus the augmentations of another two frames are used as keys. Moreover, inspired by

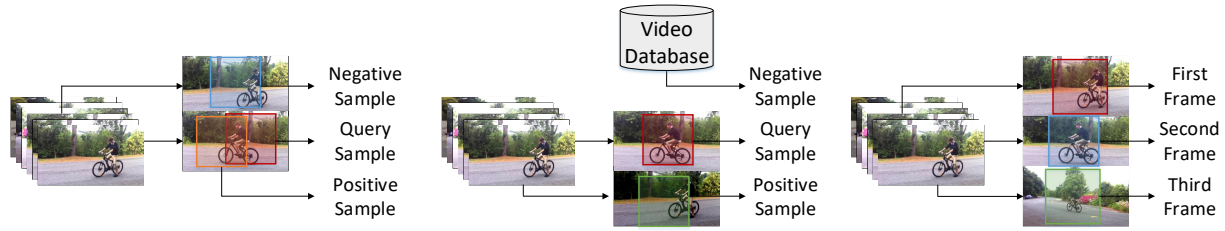

Figure 1: One proxy task on each type of supervision in video sequence for unsupervised learning.

(He et al. 2019), we additionally build a memory to track keys across mini-batches for inter-frame instance discrimination task. InfoNCE (Oord, Li, and Vinyals 2018), as one form of contrastive formulation, serves as the loss function in the two tasks. For the task of temporal order validation, we take the "anchor" frame as query and the rest as keys. We involve a linear classifier to predict if the query is in front of or behind keys (two-class classification). The classifier takes the concatenation of the features of query and keys as the input and is learnt via cross-entropy loss. Overall, SeCo is end-to-end trained by jointly optimizing the three proxy tasks.

The main contribution of this work is the proposal of exploring sequence supervision for unsupervised representation learning. Ours is among the first to systematically analyze the supervisory signals behind the rich structures in video sequence. This also leads to the elegant views of how to design simple proxy tasks which perform as a prism through which to leverage the supervision, and how to nicely capitalize on such proxy tasks for learning a generic representation, which are problems not yet fully understood. We demonstrate the effectiveness of SeCo on several downstream video applications and SeCo unsupervised pre-training also surpasses the ImageNet supervised pre-training on two video benchmarks for action recognition.

Related Work

Unsupervised Learning from Video aims to learn a generic representation without using any explicit semantic labels, which can be briefly grouped into three major categories. The first group learns feature representation by leveraging appearance variations in videos. For example, the most common constraint is to enforce the learnt representation to be temporally smooth (Mobahi, Collobert, and Weston 2009; Pan et al. 2016; Wang and Gupta 2015; Zou et al. 2012). Moving beyond only temporal smoothness, ego-motion constraints (Agrawal, Carreira, and Malik 2015; Jayaraman and Grauman 2015), object tracking (Wang and Gupta 2015) and temporal order verification (Misra et al. 2016) have been employed to further regularize the learning process. The recent works also attempt to learn the representation by predicting the pixel-level displacement across consecutive frames (Liu et al. 2017; Vondrick, Pirsiavash, and Torralba 2016; Wang et al. 2019). The second group focuses on temporal prediction and frame reconstruction tasks (Finn, Goodfellow, and Levine 2016; Han, Xie et al. 2019; Luo et al. 2017; Srivastava, Mansimov, and Salakhutdinov 2015). (Srivastava, Mansimov, and Salakhutdinov 2015) utilizes a LSTM-based encoder-decoder structure to reconstruct current frame or predict future frames. (Finn, Goodfellow, and Levine 2016) further upgrades (Srivastava, Mansimov, and Salakhutdinov 2015) by merging appearance information from previous frames with motion cues. Luo *et al.* (Luo et al. 2017) present to describe the motion between frames as a sequence of atomic 3D flows to predict long-term motion. More recently, (Han, Xie et al. 2019) learns a dense encoding of spatio-temporal blocks by recurrently predicting future representations. The third group attempts to predict the transformation parameters from the transformed video (Ahsan, Madhok, and Essa 2019; Jing et al. 2018). Jing *et al.* (Jing et al. 2018) introduce a pretext task which is defined as the prediction of the rotations applied to videos. Ahsan *et al.* divide multiple video frames into grids of patches and train a network to solve jigsaw puzzles on these patches from multiple frames in (Ahsan, Madhok, and Essa 2019).

Self-Supervised Learning is a form of unsupervised learning. It relies only on the data itself for some form of supervision without human-annotated labels. One mainstream of self-supervised learning focuses on the pretext tasks which are designed under various scenarios only for learning a good data representation. Some pretext tasks, e.g., relative patch prediction (Doersch, Gupta, and Efros 2015; Goyal et al. 2019; Noroozi and Favaro 2016), affine transformation prediction (Gidaris, Singh et al. 2018), and colorization (Deshpande, Rock, and Forsyth 2015; Zhang, Isola, and Efros 2016), are proven to be helpful for representation learning. Recently, contrastive learning is at the core of several works on self-supervised learning (Bachman, Hjelm, and Buchwalter 2019; Hjelm et al. 2019; Wu et al. 2018). The design principle is to maximize/minimize the similarity between the instances in positive/negative pairs and various pretext tasks can be represented in a contrastive manner. For instance, both contrastive multiview coding (CMC) (Tian, Krishnan, and Isola 2019) and colorization (Deshpande, Rock, and Forsyth 2015) attempt to make the representation be invariant to the color in images. For self-supervised contrastive video representation learning, Contrastive Predictive Coding (CPC) (Lorre et al. 2020) is proposed to learn long-term relations underlying the raw signal and predict the latent representation of future segments in the video. The most closely related work is Momentum Contrast (MoCo) (He et al. 2019), which builds dynamic dictionaries for contrastive learning and leverage the instance

discrimination task for unsupervised image feature learning. Our method is different in the way that we explore the generic supervision in the video sequence from spatial, spatiotemporal, and sequential perspectives, for unsupervised video representation learning.

Preliminary—Contrastive Learning for Unsupervised Feature Learning

We briefly review contrastive learning and its recent practical variant (MoCo (He et al. 2019)), which learn feature embedding in an unsupervised manner by making positive/negative query-key pairs similar/dissimilar. Formally, suppose we have an encoded *query* $\boldsymbol{q} \in \mathcal{R}^d$, and a group of encoded *key* vectors $\mathcal{K} = \{\boldsymbol{k}^+, \boldsymbol{k}_1^-, \boldsymbol{k}_2^-, ..., \boldsymbol{k}_K^-\}$ consisting of one positive key $\boldsymbol{k}^+ \in \mathcal{R}^d$ and K negative keys $\mathcal{K}^- = \{\boldsymbol{k}_i^-\}$, where d denotes the dimension of the embedding space. Note that the positive key $\boldsymbol{k}_i^+$ comes from the same distribution as the query $\boldsymbol{q}$, while the negative keys are derived from an alternative noise distribution. The objective of typical contrastive loss is to reflect the incompatibility of each query-key pair: returns low value when query $\boldsymbol{q}$ is similar to its positive key $\boldsymbol{k}^+$ and remains distinct to all negative keys $\{\boldsymbol{k}_i^-\}$. By measuring the query-key similarity via dot product, a prevailing form of contrastive loss (InfoNCE (Oord, Li, and Vinyals 2018)) is calculated in a softmax formulation:

$$\mathcal{L}_{NCE}(\boldsymbol{q}, \boldsymbol{k}^+, \mathcal{K}^-) = -\log \frac{\exp(\boldsymbol{q}^T \boldsymbol{k}^+/\tau)}{\exp(\boldsymbol{q}^T \boldsymbol{k}^+/\tau) + \sum_{i=1}^{K} \exp(\boldsymbol{q}^T \boldsymbol{k}_i^-/\tau)}, \tag{1}$$

where τ is the temperature hyper-parameter. The rationale behind such formulation is to train a classifier that could correctly classify query $\boldsymbol{q}$ as positive key $\boldsymbol{k}^+$.

Because no human-annotated labels are available in unsupervised setting, one common practice is to produce two different augmentations (x_q, x_k^+) from the same instance (an image x), which correspond to the query $\boldsymbol{q}$ and positive key $\boldsymbol{k}^+$. The augmentations of other instances/images $\{x_k^-\}$ are taken as the negative keys $\{\boldsymbol{k}_i^-\}$. In this way, a simple instance discrimination task is designed for unsupervised visual representation learning: determining whether two image patches are derived from the same image. In the implementation, two encoders (query encoder f_q and key encoder f_k) are utilized to map query image x_q and each positive/negative key image x_k into the embedding space (i.e., $\boldsymbol{q} = f_q(x_q)$, $\boldsymbol{k} = f_k(x_k)$). Recently, MoCo (He et al. 2019) strengthens contrastive learning by involving an extreme large number of negative keys via maintaining a dynamic memory to track the keys across mini-batches. In addition, a momentum update strategy is leveraged to update the weights of the key encoder (in t-th iteration) conditioned on query encoder weights: $w_{f_k}^t = \alpha \cdot w_{f_k}^{t-1} + (1-\alpha) \cdot w_{f_q}^{t-1}$, where w_{f_k} and w_{f_q} are the weights of key encoder and query encoder. α is the momentum coefficient.

Sequence Contrastive Learning

In this work, we remould the contrastive learning under the sequence supervision from videos, namely Sequence Contrastive Learning (SeCo), for unsupervised representation learning. In SeCo, three kinds of basic and generic supervision in the video sequence (from spatial, spatiotemporal, and sequential perspectives) are exploited to learn powerful and generic visual representation. An overview of our sequence contrastive learning framework is illustrated in Figure 2.

Problem Formulation

In the scenario of unsupervised video feature learning, we are given a collection of video sequences $\mathcal{V} = \{v\}$ from a large-scale video benchmark. The goal is to pre-train a visual encoder over the video sequence data in an unsupervised manner to extract generic visual representations. The pre-trained visual encoder can be further utilized to support several video downstream tasks.

Inspired by recent success of contrastive learning in image domain (He et al. 2019; Wu et al. 2018), we formulate the unsupervised video feature learning in contrastive learning paradigm by exploiting the inherent supervision within sequential structure in videos. In particular, video is an information-intensive media with spatiotemporal coherence and variation across frames, which reflects three types of supervision from spatial, spatiotemporal and sequential perspectives. Accordingly, motivated by each type of supervision implicit in video sequence, we present one simple yet effective proxy task to guide the unsupervised feature learning with the corresponding supervision.

Formally, given an unlabeled video sequence v, we firstly sample three frames randomly (s^1, s^2, s^3) and take the first (or last) frame s^1 (or s^3) in time order as the anchor frame. The anchor frame is then transformed into two perturbed samples with different augmentations, one of which is taken as query $\boldsymbol{s}_q$ and the other is used as key $\boldsymbol{s}_k^1$. Meanwhile, we perform data augmentation over the other two frames, leading to two keys ($\boldsymbol{s}_k^2$, $\boldsymbol{s}_k^3$). In analogy to instance discrimination task in image domain that encourages a query matches a key if they are augmentations of an identical image, we consider **inter-frame instance discrimination task** that examines the compatibility of each query-key frame pair at video level, which is tailored for video understanding. That is, from the spatiotemporal perspective, the query $\boldsymbol{s}_q$ should be similar to all the keys ($\boldsymbol{s}_k^1$, $\boldsymbol{s}_k^2$, $\boldsymbol{s}_k^3$) in the same video, and dissimilar to the keys $\mathcal{K}^-$ sampled from other videos across mini-batches. Moreover, to characterize the temporal variation across frames in a video, a simple **intra-frame instance discrimination task** is particularly devised to determine whether two frame patches are derived from the same video frame, from the spatial standpoint. As such, the query $\boldsymbol{s}_q$ is enforced to match key $\boldsymbol{s}_k^1$ (augmented from the same frame s^1), and mismatch the keys ($\boldsymbol{s}_k^2$, $\boldsymbol{s}_k^3$) from other frames. Furthermore, from the sequential perspective, we involve the **temporal order validation task** to exploit the inherent sequential structure of videos by predicting the correct temporal order of a frame patch sequence. Specifically, given the input frame patch sequence consisting of the query $\boldsymbol{s}_q$ and two keys ($\boldsymbol{s}_k^2$, $\boldsymbol{s}_k^3$), a linear classifier is leveraged to judge whether the query $\boldsymbol{s}_q$ is in front of or behind keys ($\boldsymbol{s}_k^2$, $\boldsymbol{s}_k^3$).

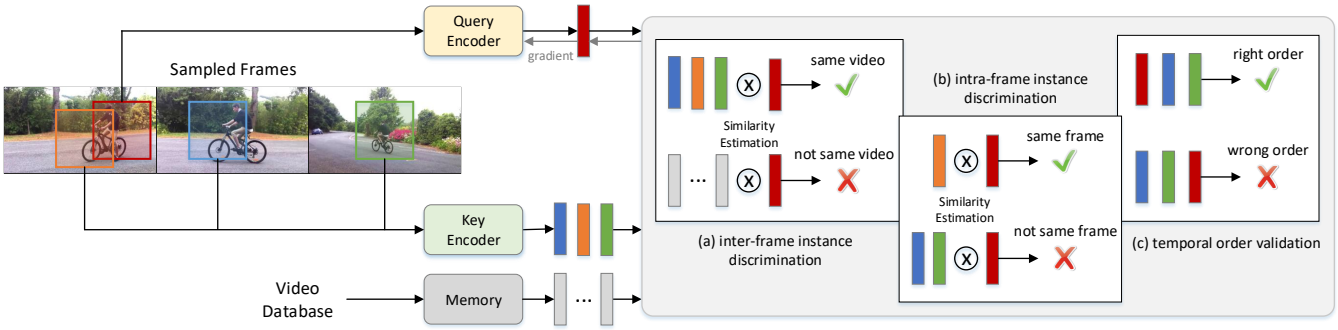

Figure 2: An overview of Sequence Contrastive Learning (SeCo) approach for unsupervised representation learning, which is composed of three proxy tasks: inter-frame instance discrimination task, intra-frame instance discrimination task, and temporal order validation task.

Inter-frame Instance Discrimination Task

Unlike (He et al. 2019) that exploits image-level query-key compatibility, we facilitate contrastive learning in video domain via the inter-frame instance discrimination task, which aims to exploit the video-level query-key compatibility. In this proxy task, the pre-trained visual encoder is learnt to not only differentiate the two augmented frame patches of the same frame in a video from the negative/mismatched frame patches in other videos, but also recognize the patches of other frames in the same video as positive/matched samples. Such design goes beyond the traditional supervision in a static image with data augmentation, and fetches more positive frame patches within the same video as supervision for contrastive learning, which sheds new light on objects with temporal evolution (e.g., new views/poses of objects). The way elegantly takes the advantage of spatiotemporal structure within videos and thus strengthens the unsupervised visual feature learning for video understanding.

Technically, suppose we have the encoded query $\boldsymbol{s}_q$ and key $\boldsymbol{s}_k^1$ belonging to the same frame, and two keys ($\boldsymbol{s}_k^2$, $\boldsymbol{s}_k^3$) from other frames in the same video. In the inter-frame instance discrimination task, our target is to determine whether two frame patches are from the same video. Therefore, we define all the keys ($\boldsymbol{s}_k^1$, $\boldsymbol{s}_k^2$, $\boldsymbol{s}_k^3$) within the same video as positive ones, and the frame patches sampled from other videos in neighboring mini-batches $\mathcal{K}^-$ are taken as the negative keys. Considering that the conventional formulation of contrastive learning (e.g., InfoNCE in Eq.(1)) only penalizes the incompatibility of each positive query-key pair at a time, we derive a particular form of contrastive learning that simultaneously match query $\boldsymbol{s}_q$ to multiple positive keys ($\boldsymbol{s}_k^1$, $\boldsymbol{s}_k^2$, $\boldsymbol{s}_k^3$) in our case. In particular, the new objective in this task is defined as the averaged sum of all the contrastive losses with regard to each positive query-key pair ($\boldsymbol{s}_q$,$\boldsymbol{s}_k^i$):

$$\mathcal{L}_{Inter-frame} = \frac{1}{3}\sum_{i=1}^{3}\mathcal{L}_{NCE}(\boldsymbol{s}_q, \boldsymbol{s}_k^i, \mathcal{K}^-). \qquad (2)$$

By minimizing the objective, the visual encoder is enforced to distinguish all the positive keys ($\boldsymbol{s}_k^1$, $\boldsymbol{s}_k^2$, $\boldsymbol{s}_k^3$) and query $\boldsymbol{s}_q$ within the same video from all the negative keys of other videos $\mathcal{K}^-$ at a time.

Intra-frame Instance Discrimination Task

In the inter-frame instance discrimination task, all sampled frame patches are holistically grouped as one generic class at video-level, while leaving the inherently spatial variation across frames within one video unexploited. To alleviate the issue, we additionally involve the intra-frame instance discrimination task to distinguish the frame patches of the same frame from the ones of the other frames in a video, which explicitly characters the variation from the spatial perspective. As such, by further steering unsupervised feature learning with the spatial supervision, the learnt visual representations are expected to be discriminative across frames in a video.

In particular, among all the four frame patches sampled from one video (query $\boldsymbol{s}_q$ and key $\boldsymbol{s}_k^1$ from an identical frame, and two keys $\boldsymbol{s}_k^2$ & $\boldsymbol{s}_k^3$ from another two frames), we take $\boldsymbol{s}_k^1$ as positive key and $\boldsymbol{s}_k^2$ & $\boldsymbol{s}_k^3$ as negative keys with regard to query $\boldsymbol{s}_q$. Note that since the previous proxy task has already exploited the incompatibility of negative query-key pairs derived from other videos, we exclude these negative keys for contrastive learning in this task for simplicity. Accordingly, we measure the objective of this task in the conventional form of contrastive loss:

$$\mathcal{L}_{Intra-frame} = \mathcal{L}_{NCE}(\boldsymbol{s}_q, \boldsymbol{s}_k^1, \{\boldsymbol{s}_k^2, \boldsymbol{s}_k^3\}). \qquad (3)$$

Such objective ensures that query $\boldsymbol{s}_q$ is similar to the positive key $\boldsymbol{s}_k^1$ augmented from the same frame and remains distinct to the negative keys $\{\boldsymbol{s}_k^2, \boldsymbol{s}_k^3\}$ of other frames, pursuing the temporally discriminative visual representation.

Temporal Order Validation Task

Most video applications (e.g., action recognition and object tracking) capitalize on the understanding of inherent sequential structure of videos, which can not be directly captured via the aforementioned two tasks that only exploit the spatiotemporal/spatial supervision based on individual frame patches. Therefore, we devise the temporal order validation task from a sequential perspective, aiming to verify whether a series of frame patches is in the correct temporal order. The rationale behind is to encourage the pre-trained visual encoder to reason about the temporal ordering of frame patches and thus exploit the sequential structure of videos for unsupervised feature learning.

Specifically, recall that we randomly sample three frames from an unlabeled video sequence and take the first or last frame in time order as the anchor frame, there are two kinds of temporal orders between query (augmented from anchor frame) and two keys (derived from the other two frames): *in front of* or *behind*. Hence, given the input frame patch sequence consisting of query $\boldsymbol{s}_q$ and two keys ($\boldsymbol{s}_k^2$, $\boldsymbol{s}_k^3$), we concatenate the query and two keys as the holistic sequence representation and feed it into a binary classifier $g(\cdot)$, which predicts if the query is in front of or behind keys. The whole model is thus optimized with cross-entropy loss:

$$\mathcal{L}_{Temporal} = -y \log g(\boldsymbol{s}_q, \boldsymbol{s}_k^2, \boldsymbol{s}_k^3) - (1-y) \log(1 - g(\boldsymbol{s}_q, \boldsymbol{s}_k^2, \boldsymbol{s}_k^3)), \tag{4}$$

where $y \in \{0, 1\}$ represents the ground-truth label that indicates whether the query $\boldsymbol{s}_q$ is in front of or behind the two keys ($\boldsymbol{s}_k^2$, , $\boldsymbol{s}_k^3$).

Optimization

Training Objective. The overall training objective of our sequence contrastive learning integrates all the objectives of three proxy tasks (i.e., Eq.(2) for inter-frame instance discrimination task, Eq.(3) for intra-frame instance discrimination task, and Eq.(4) for temporal order validation task):

$$\mathcal{L} = \mathcal{L}_{Inter-frame} + \mathcal{L}_{Intra-frame} + \mathcal{L}_{Temporal}. \tag{5}$$

Weights Update. In our SeCo, the query encoder f_q is directly optimized with standard SGD algorithm by minimizing $\mathcal{L}$. The weights of key encoder f_k is accordingly updated conditioned on query encoder weights via a momentum update strategy:

$$w_{f_k}^t = \alpha \cdot w_{f_k}^{t-1} + (1 - \alpha) \cdot w_{f_q}^{t-1}, \tag{6}$$

where α denotes the momentum coefficient that controls the updating of key encoder weights.

Experiments

We empirically verify the merit of SeCo for unsupervised representation learning in three downstream tasks: action recognition, untrimmed activity recognition and object tracking. The first experiment is conducted respectively on action recognition (Kinetics), untrimmed activity recognition (ActivityNet) and object tracking (OTB-100) under "pre-trained representation + linear model" protocol. The second experiment transfers the network unsupervised pre-trained by SeCo as the initialization for fine-tuning in action recognition task (UCF101 and HMDB51). That is "pre-training + fine-tuning" protocol.

Datasets

Kinetics400 dataset (Kay et al. 2017) is one of the large-scale action recognition benchmarks which contains around 300K videos from 400 action categories. Each video clip in this dataset is cropped from the raw YouTube video and the duration is 10 seconds. All the videos are grouped into three subsets for training (240K), validation (20K), and testing (40K), respectively. Because the labels of testing set are not publicly available, the performances on the Kinetics400 dataset are reported on the validation set. **UCF101** (Soomro, Zamir, and Shah 2012) is one of the most popular action recognition benchmarks. This dataset consists of 13,320 videos from 101 action classes, which are split into about 9.5K and 3.7K videos in training and testing set, respectively. **HMDB51** (Kuehne et al. 2011) is another widely used action recognition dataset and includes 7K videos from 51 action categories. The dataset is split into training (3.5K) and testing (1.5K) sets.

ActivityNet dataset (Heilbron et al. 2015) is a large-scale human activity understanding benchmark. The latest released version (v1.3) consists of 19,994 videos from 200 activity categories and is utilized here for evaluation. All the videos in the dataset are divided into 10,024, 4,926, and 5,044 for training, validation, and testing sets, respectively. The labels of testing set are not publicly available and thus the performances on ActivityNet dataset are all reported on validation set.

The task of object tracking actually involves two widely adopted datasets in our case, including Generic Object Tracking Benchmark (GOT-10K (Huang, Zhao, and Huang 2019)) and Object Tracking Benchmark 2015 (OTB-100 (Wu, Lim, and Yang 2015)). **GOT-10K** dataset contains more than 10K real-world videos with moving objects and over 1.5M manually labeled bounding boxes. The dataset covers more than 560 categories of moving objects and 80+ categories of motion patterns. We exploit the training set of 9,335 videos to learn a linear feature transformer (1×1 convolution), whose outputs serve for the template matching in feature space to track the example object in the subsequent frames. **OTB-100** dataset includes 100 video sequences, which are utilized as the test set for the evaluation of object tracking.

Experimental Settings

SeCo Training. We perform the training of our SeCo on the training set of Kinetics400 dataset and utilize the backbone of ResNet50 plus an MLP head. Note that the MLP head only influences SeCo training and is not involved in downstream tasks. The image patches input to the backbone are with the size of 224×224, and the head takes the global pooling feature as the input and embeds the feature into $128d$ with two fully-connected layers (2048×2048 and 2048×128). The output vector of the MLP head is normalized by its L2-norm and then exploited as the encoded representation of query or keys. In our implementations, the size of the mini-batch is set to 512 and the size of memory is 131,072. The momentum coefficient α for momentum update of the encoder is set to 0.999 and the temperature τ in infoNCE loss is 0.1. Following (He et al. 2019), shuffling BN is utilized for multi-GPU training. To optimize the parameters in the encoder, we use the momentum SGD with initial learning rate 0.2 which is annealed down to zero following a cosine decay. The network is trained for 400 epoch base on the network initialized with MoCo (He et al. 2019) on ImageNet. For data augmentation, we employ random cropping with random scales, color-jitter, random grayscale, blur, and mirror.

Action Recognition and Untrimmed Activity Recognition under "Pre-trained Representation + Linear

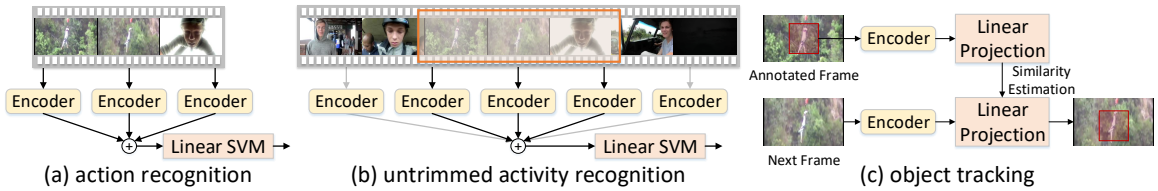

Figure 3: The detailed procedures of three downstream tasks, i.e., (a) action recognition, (b) untrimmed activity recognition, and (c) object tracking, under "pre-trained representation + linear model" protocol.

	Action Recognition	Untrimmed Activity Recognition	Object Tracking	
Dataset	Kinetics 400	ActivityNet	OTB-100	
Learnable Module	Linear SVM	Linear SVM	1x1 Convolution	
Metric	Top-1	Top-1	Precision	Success
MoCo-ImageNet	51.30	66.17	59.91	43.06
Supervised ImageNet Pre-training	52.34	67.19	69.54	48.01
VINCE (Gordon et al. 2020)	36.20	-	62.90	46.50
SeCo-Inter	58.97	66.69	67.92	48.03
SeCo-Inter+Intra	60.74	68.31	70.29	50.48
SeCo-Inter+Intra+Order	**61.91**	**68.55**	**71.86**	**51.78**

Table 1: Performance comparison of the representations pre-trained by different mechanisms in three downstream tasks under "Pre-trained Representation + Linear Model" protocol.

Model" Protocol. We directly exploit the backbone of unsupervised learnt network by SeCo on Kinetics400 as the feature extractor, and verify the frozen representation via linear classification on both downstream tasks of action recognition and untrimmed activity recognition. For each video in Kinetics400 and ActivityNet, we uniformly sample 30 and 50 frames, respectively, resize each frame with short edge of 256, and crop the resized version to 224×224 by using center crop. As shown in Figure 3 (a) and (b), we extract the frame-level feature by feature extractor and average all the frame-level features to obtain the video-level representation. A linear SVM is finally trained on the training videos of Kinetics400 or ActivityNet and evaluated on each validation set. We adopt the top-1 accuracy as the performance metric of the two tasks.

Object Tracking under "Pre-trained Representation + Linear Model" Protocol. Given the initial bounding box of an object in the first frame of a video, the task of object tracking is to locate the object in the subsequent frames. We exploit SiamFC (Gordon et al. 2020) as our tracking algorithm and execute object tracking on the representation pre-learnt by SeCo, as illustrated in Figure 3 (c). Following the setting in SiamFC that the spatial resolution of the output feature map is 1/8 of the input image, we modify the configuration of ResNet50. Specifically, for the convolution with "stride 2" in the last two stages $\{res4, res5\}$, the "stride" is changed to 1, and for the 3×3 convolutions in $res4$ and $res5$, the dilation rate is modified from 1 to 2 and 4, respectively. Note that the weights of the layers in ResNet50 remain unchanged during such modification and thus the representations are still considered as frozen. Furthermore, an additional 1×1 convolution is placed on the top of the backbone to transform the frozen representation for SiamFC tracking algorithm. In this sense, only 1×1 convolution is learnable and we also regard such protocol as linear model. The 1×1 convolution is optimized with the training set of GOT-10K, and object tracking is evaluated on OTB-100 in terms of two performance metrics: Area Under the Curve (AUC) of precision and success.

Action Recognition with "Pre-training + Fine-tuning" Protocol. Another essential function of unsupervised learning is for the purpose of network pre-training, which serves as the network initialization for fine-tuning in downstream tasks. Here, we initialize ResNet50 with the backbone in the unsupervised training of SeCo and fine-tune the network with the standard supervised setting (Qiu, Yao, and Mei 2017; Qiu et al. 2019) on UCF101 and HMDB51 for action recognition.

Evaluations on Pre-trained Feature+Linear Model

We first validate our SeCo under the protocol of "Pre-trained Representation + Linear Model," which is to manifest the generalization capability of representations learnt by SeCo. We compare the following three training mechanisms: (1) MoCo-ImageNet train the network on ImageNet in an unsupervised manner by using MoCo (He et al. 2019) algorithm. (2) Supervised ImageNet Pre-training capitalizes on the supervision of human-annotated labels on the images and learns the network in a fully-supervised fashion. (3) VINCE (Gordon et al. 2020) forms multiple anchor-positive pairs from multiple frames in a video and also executes contrastive training for unsupervised representation learning.

Method	Top-1 (%)
OPN† (Lee et al. 2017)	20.86
RotNet† (Gidaris, Singh et al. 2018)	23.33
3DRotNet† (Jing et al. 2018)	19.33
VIE-Single (Zhuang et al. 2019)	44.41
VIE-TRN (Zhuang et al. 2019)	44.91
VIE-3DResNet (Zhuang et al. 2019)	43.40
VIE-SlowFast (Zhuang et al. 2019)	47.37
VIE-Full (Zhuang et al. 2019)	48.53
SeCo (ResNet18)	**50.81**

Table 2: Performance comparisons of unsupervised representation learning on Kinetics400.

	Top-1 (%)	
	UCF101	HMDB51
Shuffle&Learn (Misra et al. 2016)	50.20	18.10
OPN (Lee et al. 2017)	59.60	23.80
ClipOrder (Xu et al. 2019)	72.40	30.90
3DRotNet (Jing et al. 2018)	66.00	37.10
DPC (Han, Xie et al. 2019)	75.70	35.70
CBT (Sun et al. 2019)	79.50	44.60
VIE-Full (Zhuang et al. 2019)	80.40	52.50
Supervised ImageNet Pre-training	85.30	49.08
SeCo	**88.26**	**55.55**

Table 3: Performance comparisons of pre-training + fine-tuning protocol on UCF101 and HMDB51.

Table 1 summarizes performance comparisons of different representation learning mechanisms in three downstream tasks. Overall, the performances across the three tasks consistently indicate that our SeCo leads to performance boost against other training mechanisms. Particularly, by doing classification on the representations prelearnt by SeCo achieves 61.91% and 68.55% on action recognition (Kinetics400) and untrimmed activity recognition (ActivityNet), respectively, making the absolute improvement over Supervised ImageNet Pre-training by 9.57% and 1.36% in terms of top-1 accuracy. Furthermore, SeCo benefits from three types of supervision, and models the spatiotemporal coherence and variation in videos better, therefore leading the precision by 2.32% in object tracking (OTB-100). The results clearly demonstrate the advantage of our SeCo unsupervised pre-training for learning representations that are more generic across various downstream tasks. As expected, SeCo-Inter remoulds MoCo-ImageNet in the context of video and exhibits better performance than MoCo-ImageNet on video tasks. SeCo-Inter+Intra constantly outperforms SeCo-Inter and SeCo learnt through the three proxy tasks performs the best. The results also verify the complementarity between three supervision in the sequence for representation learning.

Table 2 further details the comparisons with state-of-the-art unsupervised representation learning methods on Kinetics400. † denotes that each method is implemented and learnt on Kinetics400 as reported in (Zhuang et al. 2019). Please also note that here we exploit ResNet18 as the backbone in our SeCo training for fair comparisons. Specifically, VIE learns deep nonlinear embeddings to group similar videos and push different videos apart in the embedding space and such idea is similar to our SeCo-Inter in spirit. As indicated by the results, VIE-Single leads to a large performance gain over OPN and RotNet, and all the three runs select one frame from each video, which is input into a 2D network for classification. VIE-3DResNet further extends 2D ResNet18 to 3D and VIE-SlowFast employs the advanced SlowFast structure of two 3D networks. By combining VIE-Single and VIE-SlowFast, VIE-Full achieves 48.53% top-1 accuracy, which is still lower than 50.81% of SeCo learnt only on a 2D ResNet18. That again proves the impact of our SeCo for unsupervised representation learning.

Evaluations on Pre-training+Fine-tuning

Next, we evaluate SeCo from the aspect of network pre-training, which is taken as network initialization for fine-tuning on downstream tasks. Such protocol is to examine the transferability of the pre-trained structure. Table 3 shows the comparisons of pre-training the networks by different methods and then supervised fine-tuning on UCF101 and HMDB51 as the backbone in TSN (Wang et al. 2018) for action recognition. Compared to the best competitor VIE-Full, SeCo improves the top-1 accuracy from 80.40%/52.50% to 88.26%/55.55% on the two datasets. Notably, SeCo unsupervised pre-training leads the accuracy by 2.96% and 6.47% against fully-supervised ImageNet pre-training, which is really impressive.

Conclusions

We have presented Sequence Contrastive Learning (SeCo) method which explores the generic supervision in the video sequence for unsupervised representation learning. Particularly, we study the sequence supervision systematically from three aspects: spatial, spatiotemporal and sequential. To verify our claim, we devise one simple proxy task, i.e., intra-frame/inter-frame instance discrimination task or temporal order validation task, to present and leverage each supervision. In between, intra-frame/inter-frame instance discrimination task is to determine whether two frame patches are from one frame or an identical video, respectively, and temporal order validation examines whether a series of frame patches are in chronological order correctly. We materialize the three proxy tasks and build our SeCo on contrastive learning framework. Experiments conducted on both "pre-trained representation + linear model" and "pre-training + fine-tuning" protocols, validate our proposal and analysis. More remarkably, SeCo pre-training leads to an increase of accuracy by 2.96% and 6.47% over ImageNet supervised pre-training on UCF101 and HMDB51 datasets for action recognition task.

Ethics Statement

Video understanding (e.g., action recognition and object tracking) is one of the fundamental problems in numerous real-world applications, ranging from video surveillance, indexing and retrieval to human computer interaction. However, the achievements of these video applications rely heavily on the assumption that large quantities of human annotations are accessible for neural model learning. The assumption becomes impractical when cost-expensive and labor-intensive manual labeling is required. This significantly limits and discourages the motivations for relatively small research communities without adequate financial supports. We demonstrate in this paper that the challenge can be mitigated by pre-training a visual encoder via our Sequence Contrastive Learning (SeCo) in an unsupervised manner without any human-annotated labels. Such pre-trained visual encoder can be further utilized to facilitate a wide variety of video applications. Notice that our SeCo, an unsupervised learning approach, even surpasses the supervised ImageNet pre-training counterpart in action recognition task. Nevertheless, one potential risk lies in that if the use of unsupervised visual representation learning in videos means video understanding systems may now be easily developed by those with lower levels of domain or ML expertise, this could increase the risk of the video understanding model or its outputs being used incorrectly.

References

Agrawal, P.; Carreira, J.; and Malik, J. 2015. Learning to See by Moving. In *ICCV*.

Ahsan, U.; Madhok, R.; and Essa, I. 2019. Video Jigsaw: Unsupervised Learning of Spatiotemporal Context for Video Action Recognition. In *WACV*.

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019. Learning Representations by Maximizing Mutual Information Across Views. In *NeurIPS*.

Cai, Q.; Wang, Y.; Pan, Y.; Yao, T.; and Mei, T. 2020. Joint Contrastive Learning with Infinite Possibilities. In *NeurIPS*.

Deshpande, A.; Rock, J.; and Forsyth, D. 2015. Learning Large-Scale Automatic Image Colorization. In *ICCV*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised Visual Representation Learning by Context Prediction. In *ICCV*.

Finn, C.; Goodfellow, I. J.; and Levine, S. 2016. Unsupervised Learning for Physical Interaction through Video Prediction. In *NIPS*.

Gidaris, S.; Singh, P.; et al. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR*.

Gordon, D.; Ehsani, K.; Fox, D.; and Farhadi, A. 2020. Watching the World Go By: Representation Learning from Unlabeled Videos. *arXiv preprint arXiv:2003.07990* .

Goyal, P.; Mahajan, D.; Gupta, A.; and Misra, I. 2019. Scaling and Benchmarking Self-Supervised Visual Representation Learning. In *ICCV*.

Han, T.; Xie, W.; et al. 2019. Video Representation Learning by Dense Predictive Coding. In *ICCV Workshop*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2019. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv preprint arXiv:1911.05722* .

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*.

Huang, L.; Zhao, X.; and Huang, K. 2019. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Trans. on PAMI* .

Jayaraman, D.; and Grauman, K. 2015. Learning image representations tied to ego-motion. In *ICCV*.

Jing, L.; Yang, X.; Liu, J.; and Tian, Y. 2018. Self-Supervised Spatiotemporal Feature Learning via Video Rotation Prediction. *arXiv preprint arXiv:1811.11387* .

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; Suleyman, M.; and Zisserman, A. 2017. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950* .

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: A large video database for human motion recognition. In *ICCV*.

Lee, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2017. Unsupervised Representation Learning by Sorting Sequences. In *ICCV*.

Liu, Z.; Yeh, R. A.; Tang, X.; Liu, Y.; and Agarwala, A. 2017. Video Frame Synthesis Using Deep Voxel Flow. In *ICCV*.

Lorre, G.; Rabarisoa, J.; Orcesi, A.; Ainouz, S.; and Canu, S. 2020. Temporal Contrastive Pretraining for Video Action Recognition. In *WACV*.

Luo, Z.; Peng, B.; Huang, D.-A.; Alahi, A.; and Fei-Fei, L. 2017. Unsupervised Learning of Long-Term Motion Dynamics for Videos. In *CVPR*.

Misra, I.; et al. 2016. Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification. In *ECCV*.

Mobahi, H.; Collobert, R.; and Weston, J. 2009. Deep learning from temporal coherence in video. In *ICML*.

Noroozi, M.; and Favaro, P. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *ECCV*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* .

Pan, Y.; Li, Y.; Yao, T.; Mei, T.; Li, H.; and Rui, Y. 2016. Learning deep intrinsic video representation by exploring temporal coherence and graph structure. In *IJCAI*.

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In *ICCV*.

Qiu, Z.; Yao, T.; Ngo, C.-W.; Tian, X.; and Mei, T. 2019. Learning spatio-temporal representation with local and global diffusion. In *CVPR*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402* .

Srivastava, N.; Mansimov, E.; and Salakhutdinov, R. 2015. Unsupervised Learning of Video Representations using LSTMs. In *ICML*.

Sun, C.; Baradel, F.; Murphy, K.; and Schmid, C. 2019. Learning Video Representations using Contrastive Bidirectional Transformer. *arXiv preprint arXiv:1906.05743* .

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive Multiview Coding. *arXiv preprint arXiv:1906.05849* .

Vondrick, C.; Pirsiavash, H.; and Torralba, A. 2016. Generating Videos with Scene Dynamics. In *NIPS*.

Wang, J.; Jiao, J.; Bao, L.; He, S.; Liu, Y.; and Liu, W. 2019. Self-Supervised Spatio-Temporal Representation Learning for Videos by Predicting Motion and Appearance Statistics. In *CVPR*.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2018. Temporal segment networks for action recognition in videos. *IEEE Trans. on PAMI* .

Wang, X.; and Gupta, A. 2015. Unsupervised Learning of Visual Representations Using Videos. In *ICCV*.

Wu, Y.; Lim, J.; and Yang, M.-H. 2015. Object Tracking Benchmark. *IEEE Trans. on PAMI* .

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-parametric Instance Discrimination. In *CVPR*.

Xu, D.; Xiao, J.; Zhao, Z.; Shao, J.; Xie, D.; and Zhuang, Y. 2019. Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. In *CVPR*.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful Image Colorization. In *ECCV*.

Zhuang, C.; She, T.; Andonian, A.; and Yamins, D. 2019. Unsupervised Learning from Video with Deep Neural Embeddings. *arXiv preprint arXiv:1905.11954* .

Zou, W.; Zhu, S.; Yu, K.; and Ng, A. Y. 2012. Deep Learning of Invariant Features via Simulated Fixations in Video. In *NIPS*.