Lab CudaVision
Learning Vision Systems on Graphics Cards (MA-INF 4308)

Self-Supervised Learning

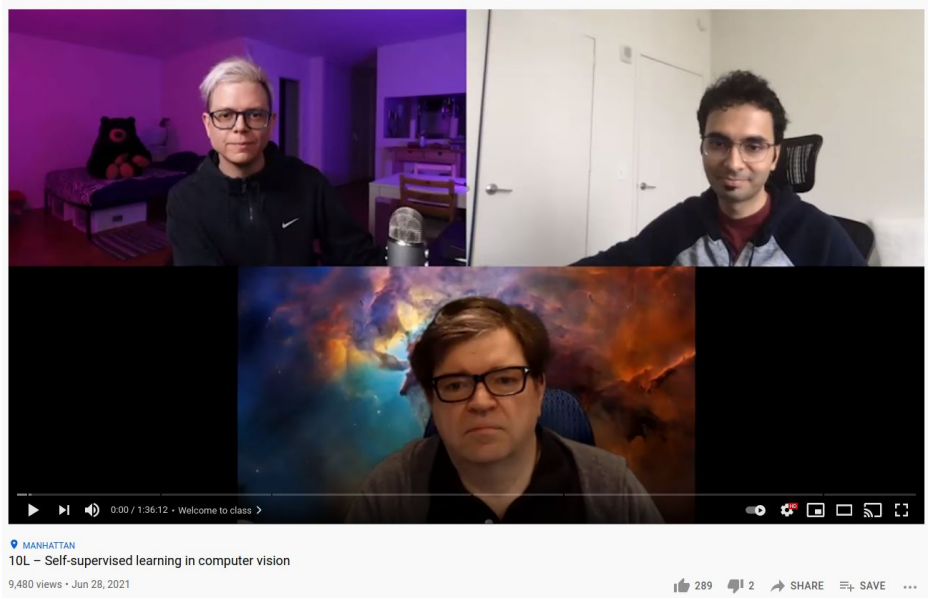
10.08.2021

PROF. SVEN BEHNKE, ANGEL VILLAR-CORRALES

Contact: villar@ais.uni-bonn.de

CVPR 2021 Tutorial on Self-Supervised Learning

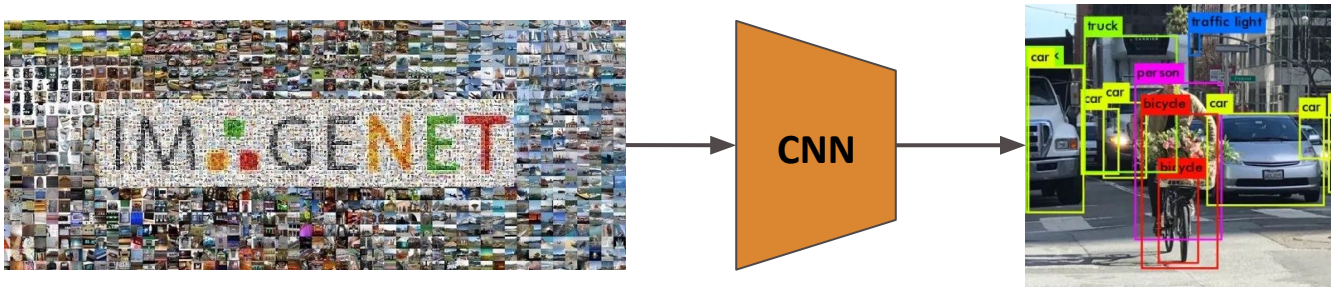
https://gidariss.github.io/self-supervised-learning-cvpr2021/



Talk by Ishan Misra (Facebook Research) at NYU

https://www.youtube.com/watch?v=8L10w1KoOU8

Deep Learning: How it Works?



- Predefine set of visual concepts to be learned
- Collect and annotate a diverse & large number of images
- Initialize model parameters by pretraining on ImageNet
- Train a model for several GPU hours or days

Meanwhile, in the real world…

Annotating Datasets is Hard

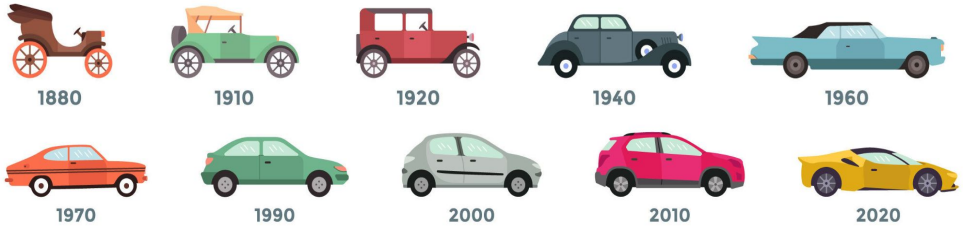
- Labeled data is scarce
- Labeling is time consuming and expensive
- Noisy annotations (human mistakes)

20 s/img

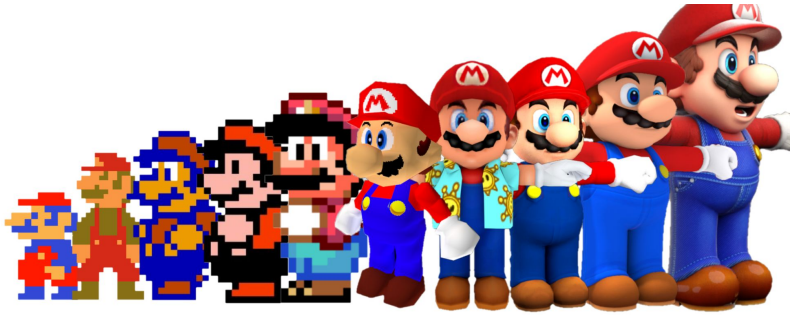
85 s/instance

1.5 hours/img

Ever Changing World



Evolution of cars

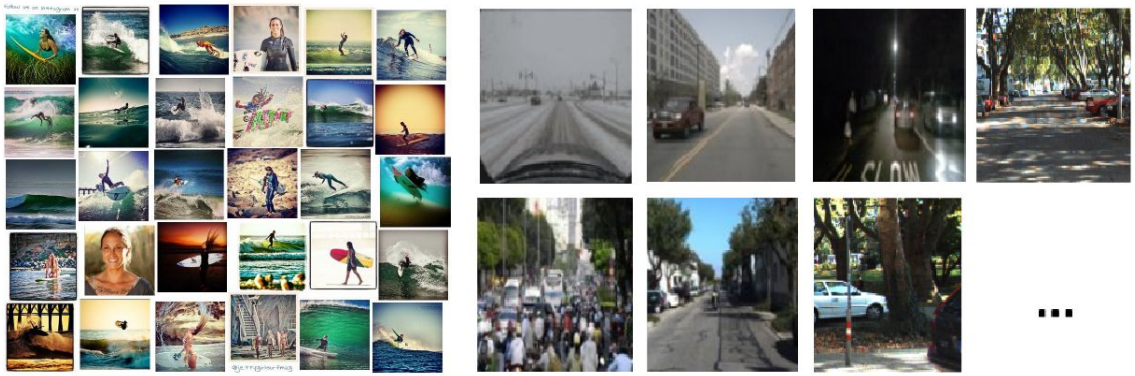


Super Mario from 1981 to 2017

- Data distributions shift all the time
- Cameras and sensors are frequently upgraded
- Infeasible to annotate new datasets each time!

Self-Supervised Learning (SSL)

Unlabelled Data



- Collecting unlabelled data is usually easy
- Supervised learning cannot learn from this
- Unsupervised methods do not scale
  - Autoencoders, GANs do not learn meaningful representations

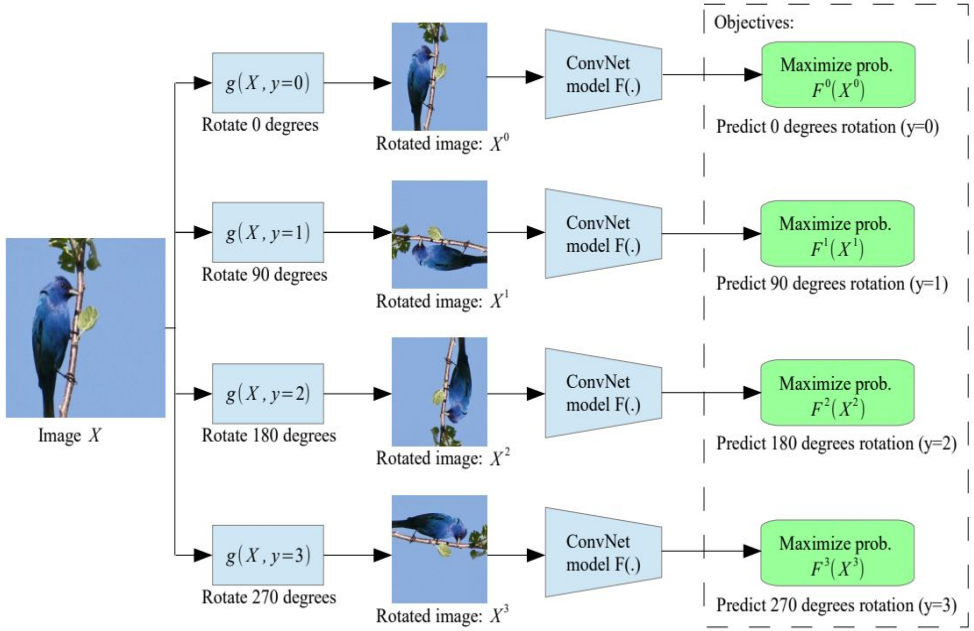
Self-Supervision

- A form of unsupervised learning where **the data provides the supervision**

- Define a **pretext task** to force a network to learn good representations:
  - Similarity learning
  - Inverse tasks: denoising, impainting,...
  - Predictions

- Features learned on the pretext task are subsequently used for a different **downstream task**, usually where some annotations are available.

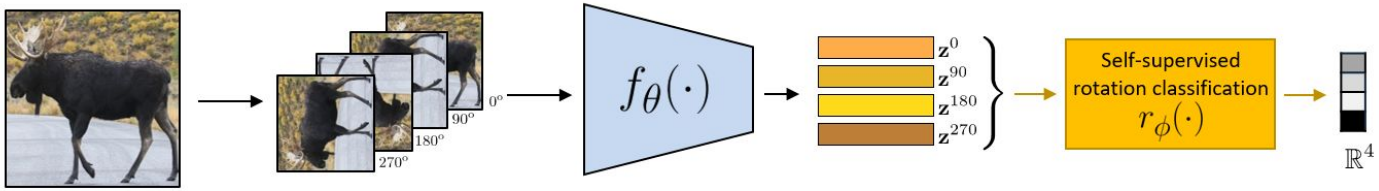
Example: Rotation Prediction



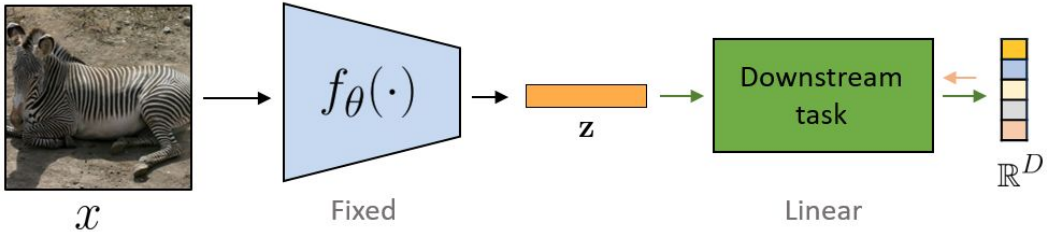
Gidaris et al. "Unsupervised representation learning by predicting image rotations." 2018

Self-Supervised Learning Pipeline

1. Train network on pretext task
   - large amount of unlabelled data



2. Train new model heah for downstream tasks
   - Small annotated dataset



The Karate Kid and SSL

The Karate Kid (1984)

Stage 1: Train on Pretext Tasks

Mr. Miyagi: Deep Learning Practitioner

Daniel LaRusso: Conv. Net

Daily chores: Pretext task

Learning karate: Downstream task

Stage 2: Rapidly Fine-tuning

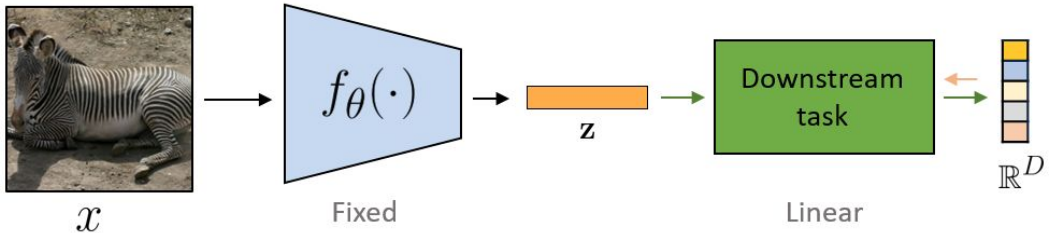
Evaluating SSL

Training & Evaluation

- Standard approach:
  - Perform pretext task on ImageNet
  - Fine-tune on other ImageNet or other dataset

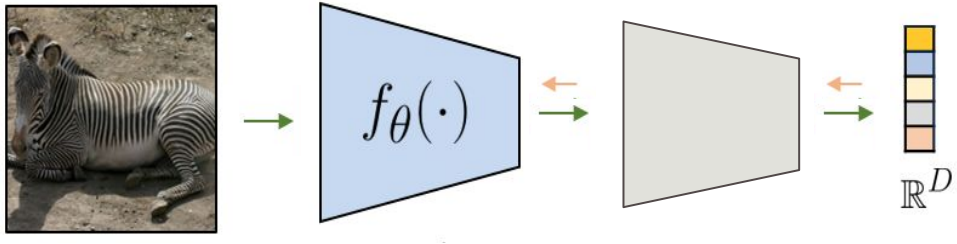
- Evaluations:
  - Linear probe
  - Learning with few labels
  - Transfer learning

Linear Probe



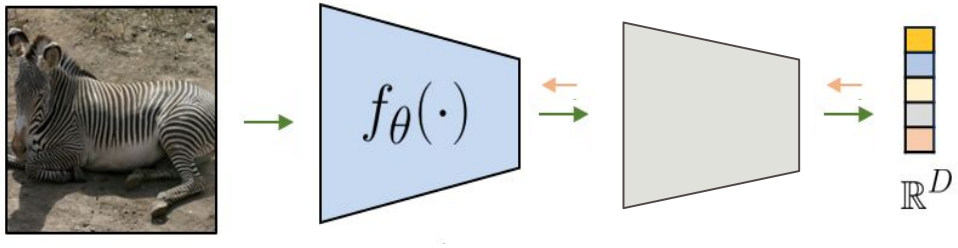
- Simplest evaluation approach
- Train a linear classifier (e.g., SVM) on top of the learned representations
  - Same as we did on our *Transfer Learning* class!
- Datasets: ImageNet, VOC07 (classification), COCO (classification)

Efficient Classification



- Fine-tune pretrained model with 1%-10% of the labels
- Datasets: ImageNet

Transfer Learning

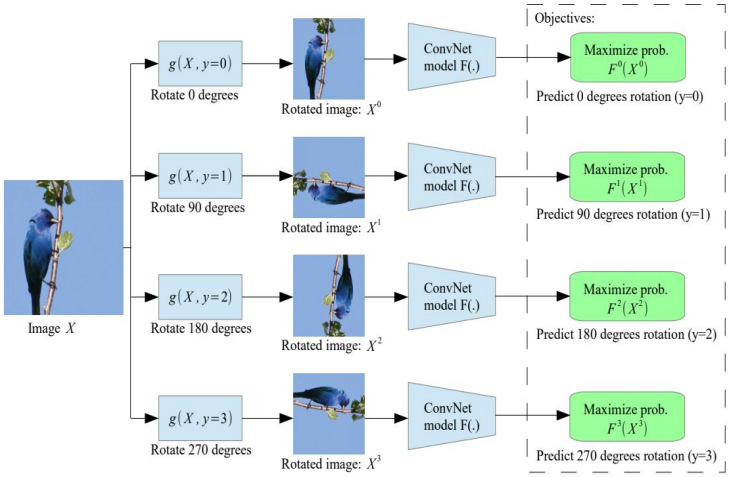


- Train pretrained backbone with a task-specific head (decoder, RPN, ...) on some other computer vision tasks
- Tasks and datasets:
  - Object detection: VOC07, COCO
  - Semantic segmentation: Cityscapes, ADE20K
  - ...

Handcrafted Pretext Tasks

Predicting Image Rotations

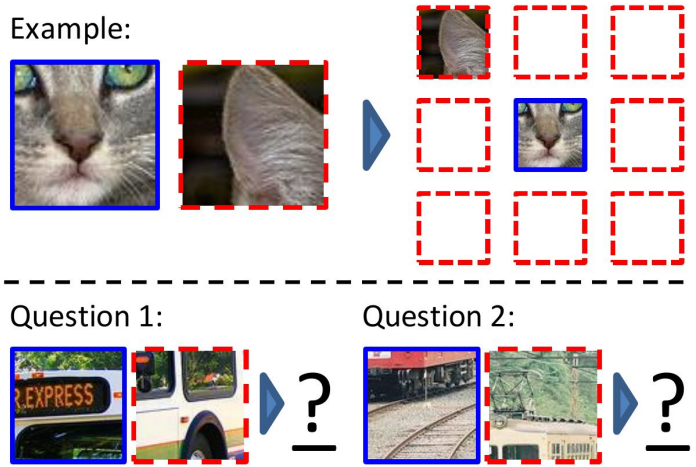
- Probably the simplest pretext task
- Pretext task:
  - Rotate images by multiple of 90°
  - Predict which rotation has been applied



Gidaris et al. "Unsupervised representation learning by predicting image rotations." 2018

Localize Image Patches

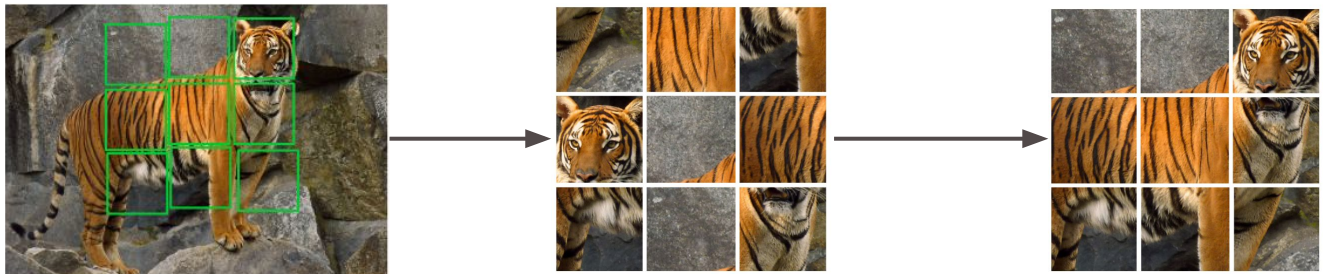
- One the first pretext-based SSL methods
- Pretext task:
  - Sample two random image patches
  - Predict relative position
  - 8-class classification
- Forces the model to recognize objects and their parts



Doersch et al. "Unsupervised visual representation learning by context prediction." 2015

Solving Jigsaw Puzzles

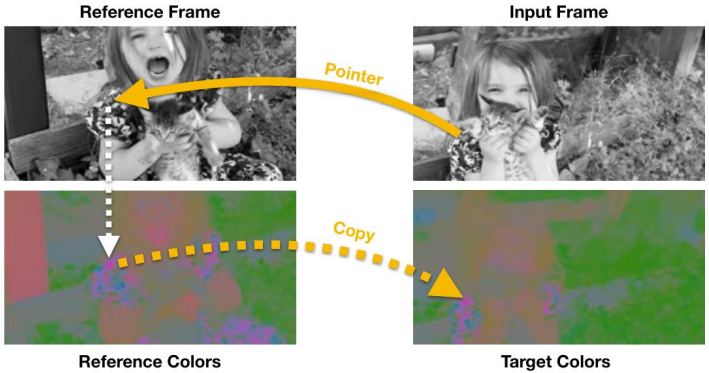
- Learning representations by solving a 9-piece Jigsaw puzzle
- Pretext task:
  - Extracting nine image patches
  - Shuffle using one out of 64 permutations
  - Siamese backbone processes each patch
  - MLP-heads predict permutation

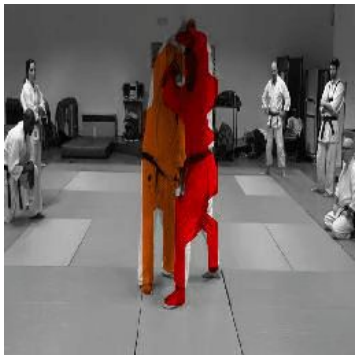


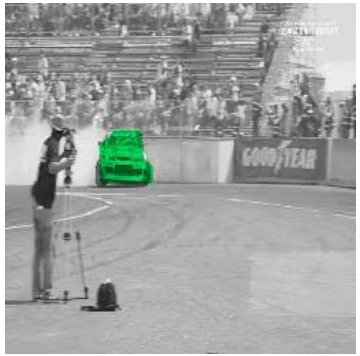
Noroozi et al. "Unsupervised learning of visual representations by solving jigsaw puzzles." 2016

Image Colorization

- Pretext task:
  - Remove image color channel
  - Model predicts image color in *CIELAB* color space
- Extension to tracking by colorization







Zhang et al. "Colorful image colorization." 2016 & Vondrick et al. “Tracking emerges by colorizing videos”. 2018

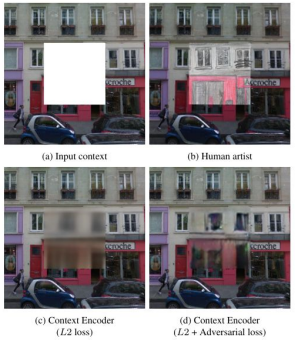
Many more...

Context Encoders: Feature Learning by Inpainting

Deepak Pathak Philipp Krähenbühl Jeff Donahue Trevor Darrell Alexei A. Efros
University of California, Berkeley
{pathak,philkr,jdonahue,trevor,efros}@cs.berkeley.edu

Abstract

We present an unsupervised visual feature learning algorithm driven by context-based pixel prediction. By analogy with auto-encoders, we propose Context Encoders – a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings. In order to succeed at this task, context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s). When training context encoders, we have experimented with both a standard pixel-wise reconstruction loss, as well as a reconstruction plus an adversarial loss. The latter produces much sharper results because it can better handle multiple modes in the output. We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of our learned features for CNN pre-training on classification, detection, and segmentation tasks. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.

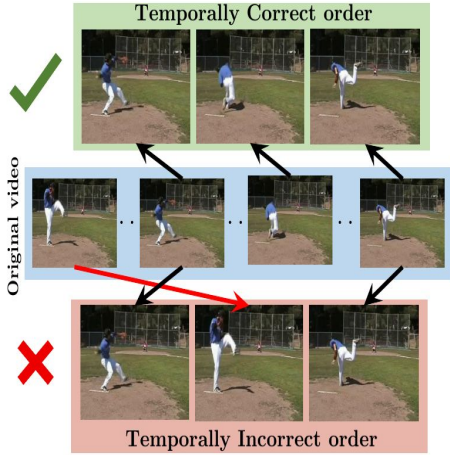


Shuffle and Learn: Unsupervised Learning using Temporal Order Verification

Ishan Misra[1] C. Lawrence Zitnick[2] Martial Hebert[1]

[1] The Robotics Institute, Carnegie Mellon University
[2] Facebook AI Research
{imisra, hebert}@cs.cmu.edu, zitnick@fb.com

Abstract. In this paper, we present an approach for learning a visual representation from the raw spatiotemporal signals in videos. Our representation is learned without supervision from semantic labels. We formulate our method as an unsupervised sequential verification task, i.e., we determine whether a sequence of frames from a video is in the correct temporal order. With this simple task and no semantic labels, we



Many more...

Objects that Sound

Relja Arandjelović[1] and Andrew Zisserman[1,2]

[1] DeepMind
[2] VGG, Department of Engineering Science, University of Oxford

Abstract. In this paper our objectives are, first, networks that can embed audio and visual inputs into a common space that is suitable for cross-modal retrieval; and second, a network that can localize the object that sounds in an image, given the audio signal. We achieve both these objectives by training from unlabelled video using only *audio-visual correspondence* (AVC) as the objective function. This is a form of cross-modal self-supervision from video.

Self-Supervised Feature Learning by Learning to Spot Artifacts

Simon Jenni Paolo Favaro
University of Bern, Switzerland
{jenni,favaro}@inf.unibe.ch

Abstract

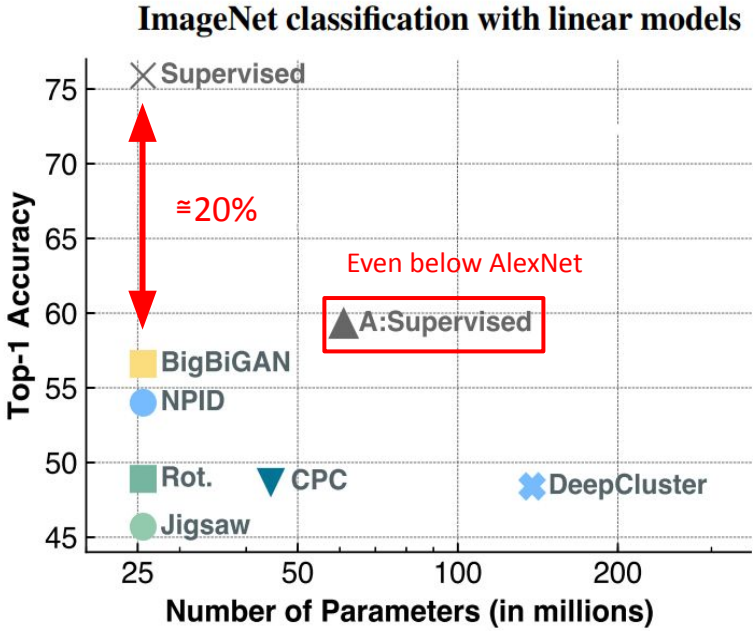
We introduce a novel self-supervised learning method based on adversarial training. Our objective is to train a discriminator network to distinguish real images from images with synthetic artifacts, and then to extract features from its intermediate layers that can be transferred to other data domains and tasks. To generate images with artifacts, we pre-train a high-capacity autoencoder and then we use a damage and repair *strategy: First, we freeze the autoencoder and* damage *the output of the encoder by randomly dropping its entries. Second, we augment the decoder with a* repair *network, and train it in an adversarial manner against the discriminator. The repair network helps generate more realistic images by inpainting the dropped feature entries. To make the discriminator focus on the artifacts, we also make it predict what entries in the feature were dropped. We demonstrate experimentally that features learned by creating and spotting artifacts achieve state of the art performance in several benchmarks.*

Figure 1. A mixture of real images (green border) and images with synthetic artifacts (red border). Is a good object representation necessary to tell them apart?

Towards this goal, we propose to learn features by classifying images as real or with artifacts (see Figure 1). We aim at creating image artifacts, such that a model capable of spotting them would require an accurate representation of objects and thus build features that could transfer well to

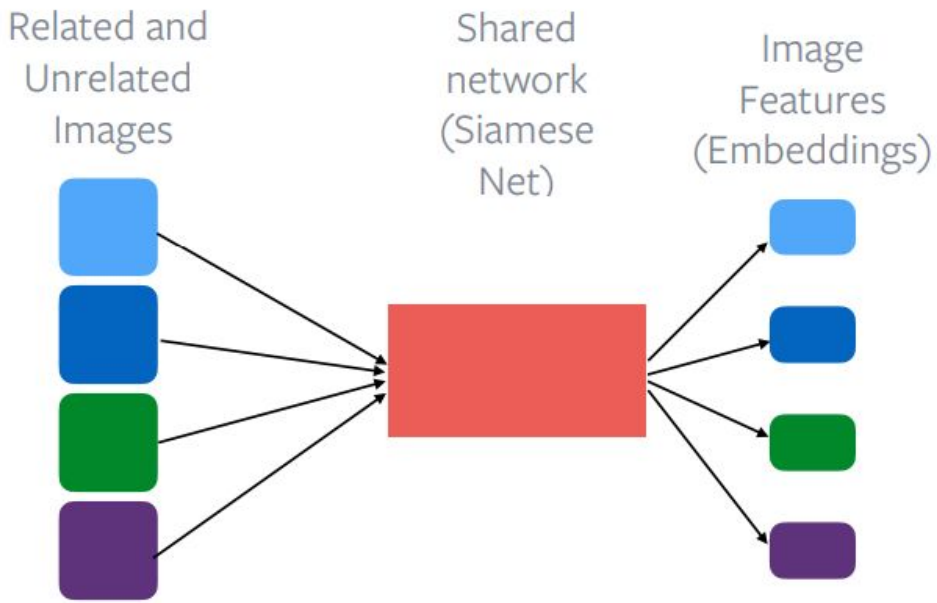
Problem with Handcrafted Tasks

- Features should represent how images relate
- Should be robust to “nuisance factors”:
  - Location of objects
  - Lighting and texture
  - ...



Contrastive SSL Learning

Contrastive Learning



Loss Function

Embeddings from related images should be closer than embeddings from unrelated images

Contrastive Losses

Contrastive Loss

$$\mathcal{L}(A,B) = \begin{cases} d(A,B) & \text{; positive pair} \\ max(0, \alpha - d(A,B)) & \text{; otherwise} \end{cases}$$

Triplet Loss:

$$\mathcal{L}(A,P,N) = \max\{0, d(A,P) - d(A,N) + \alpha\}$$

InfoNCE Loss (or NT-Xent)

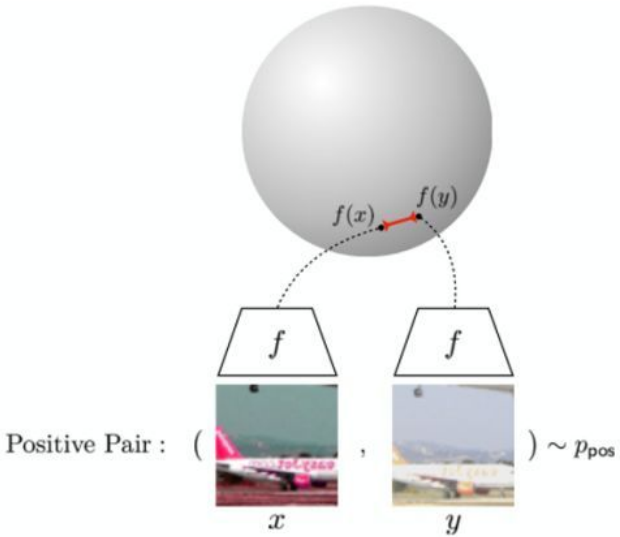
- Batch-wise contrastive loss: $\mathbb{B} = \{\mathbf{x}_1^1, \mathbf{x}_1^2, \mathbf{x}_2^1, \mathbf{x}_2^2, ..., \mathbf{x}_N^1, \mathbf{x}_N^2\}$
  - Positive pairs: $\mathbf{x}_i^1, \mathbf{x}_i^2$
  - Negative pairs $\mathbf{x}_i, \mathbf{x}_j \quad \forall j \neq i$
- Cosine similarity function: $\text{sim}(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{||\mathbf{x}|| \cdot ||\mathbf{y}||}$

Similarity for positive example

$$\mathcal{L}(\mathbf{x}_1^1) = -\log \frac{\exp(\text{sim}(\mathbf{x}_1^1, \mathbf{x}_1^2)/\tau)}{\sum_{\substack{\mathbf{y} \in \mathbb{B} \\ \mathbf{y} \neq \mathbf{x}_1^1}} \exp(\text{sim}(\mathbf{x}_1^1, \mathbf{y})/\tau)}$$

Similarity for all nontrivial examples

InfoNCE Loss (or NT-Xent)



Alignment: Similar samples have similar features.
(Figure inspired by Tian et al. (2019).)

Uniformity: Preserve maximal information.

Simple Contrastive SSL Learning

Self-Supervised Learning of Pretext-Invariant Representations

Ishan Misra Laurens van der Maaten
Facebook AI Research

Abstract

The goal of self-supervised learning from images is to construct image representations that are semantically meaningful via pretext tasks that do not require semantic annotations. Many pretext tasks lead to representations that are covariant *with image transformations. We argue that, instead, semantic representations ought to be* invariant *under such transformations. Specifically, we develop Pretext-Invariant Representation Learning (PIRL, pronounced as "pearl") that learns invariant representations based on pretext tasks. We use PIRL with a commonly used pretext task that involves solving jigsaw puzzles. We find that PIRL substantially improves the semantic quality of the learned image representations. Our approach sets a new state-of-the-art in self-supervised learning from images on several popular benchmarks for self-supervised learning. Despite being unsupervised,* ***PIRL outperforms supervised pre-training*** *in learning image representations for object detection. Altogether, our results demonstrate the potential of self-supervised representations with good invariance properties.*

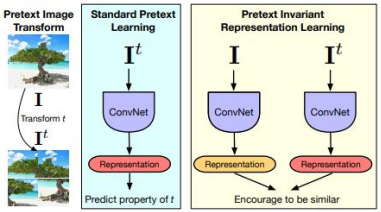


Figure 1: Pretext-Invariant Representation Learning (PIRL). Many pretext tasks for self-supervised learning [20, 54, 85] involve transforming an image **I**, computing a representation of the transformed image, and predicting properties of transformation t from that representation. As a result, the representation must *covary* with the transformation t and may not contain much semantic information. By contrast, PIRL learns representations that are *invariant* to the transformation t and retain semantic information.

transformation, it encourages the construction of image representations that are *covariant* to the transformations. Although such covariance is beneficial for tasks such as predicting 3D correspondences [33, 57, 65], it is undesirable for most semantic recognition tasks. Representations ought to be *invariant* under image transformations to be useful for

A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]

Abstract

This paper presents *SimCLR*: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations, and (3) contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning. By combining these findings, we are able to considerably outperform previous methods for self-supervised and semi-supervised learning on ImageNet. A linear classifier trained on self-supervised representations learned by SimCLR achieves 76.5% top-1 accuracy, which is a 7% relative improvement over previous state-of-the-art, matching the performance of a supervised ResNet-50. When fine-tuned on only 1% of the labels, we achieve 85.8% top-5 accuracy, outperforming AlexNet with 100× fewer labels. [1]

1. Introduction

Learning effective visual representations without human

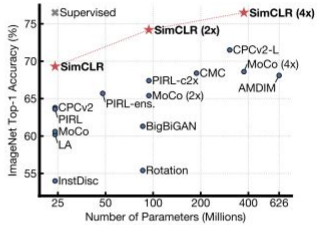
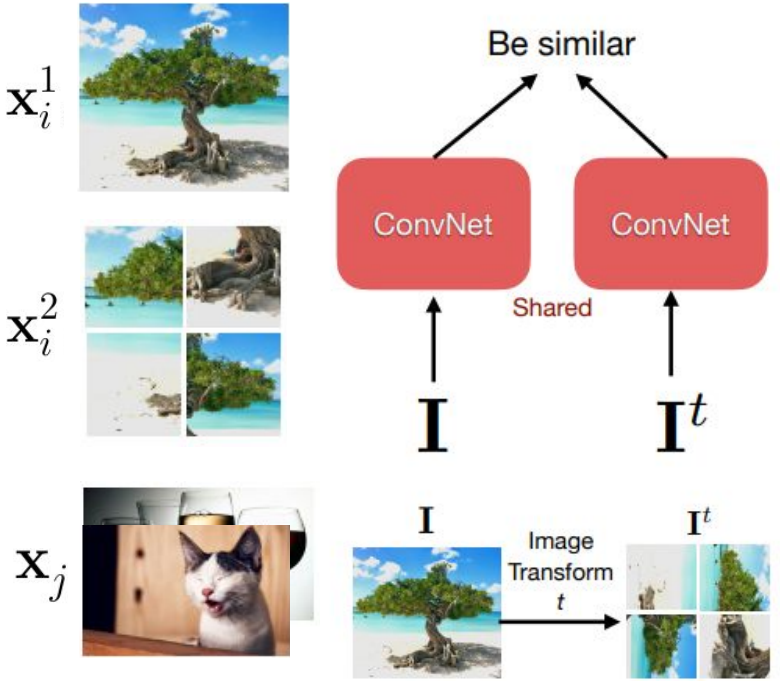


Figure 1. ImageNet Top-1 accuracy of linear classifiers trained on representations learned with different self-supervised methods (pretrained on ImageNet). Gray cross indicates supervised ResNet-50. Our method, SimCLR, is shown in bold.

However, pixel-level generation is computationally expensive and may not be necessary for representation learning. Discriminative approaches learn representations using objective functions similar to those used for supervised learning, but train networks to perform pretext tasks where both the inputs and labels are derived from an unlabeled dataset. Many such approaches have relied on heuristics to design pretext tasks (Doersch et al., 2015; Zhang et al., 2016; Noroozi & Favaro, 2016; Gidaris et al., 2018), which could limit the generality of the learned representations. Discriminative approaches based on contrastive learning in the latent space have recently shown great promise, achieving state-of-the-art results (Hadsell et al., 2006; Dosovitskiy et al., 2014; Oord et al., 2018; Bachman et al., 2019).

Pretext-Invariant Representation Learning

- Positive examples are transformed images
- Negative examples are unrelated images
- Model learns invariance to:
  - Data augmentations
  - Views created by pretext task

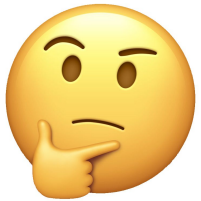


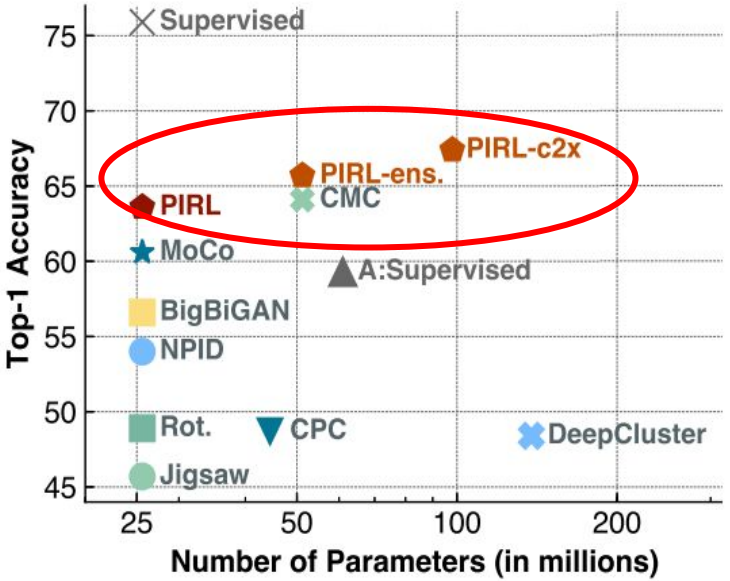
Misra et al. "Pretext-invariant representation learning." 2019

Pretext-Invariant Representation Learning

- Large improvement w.r.t handcrafted pretext tasks and AlexNet
- Still far from supervised

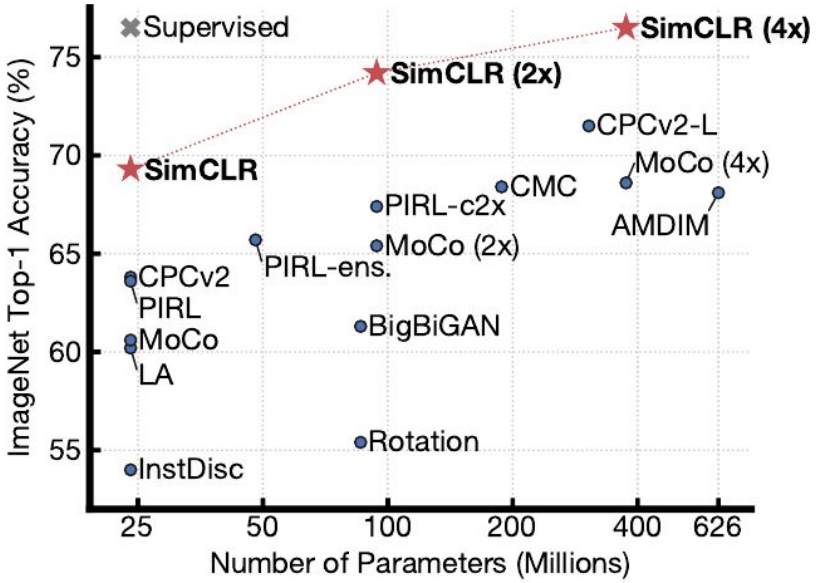
How to achieve supervised performance?





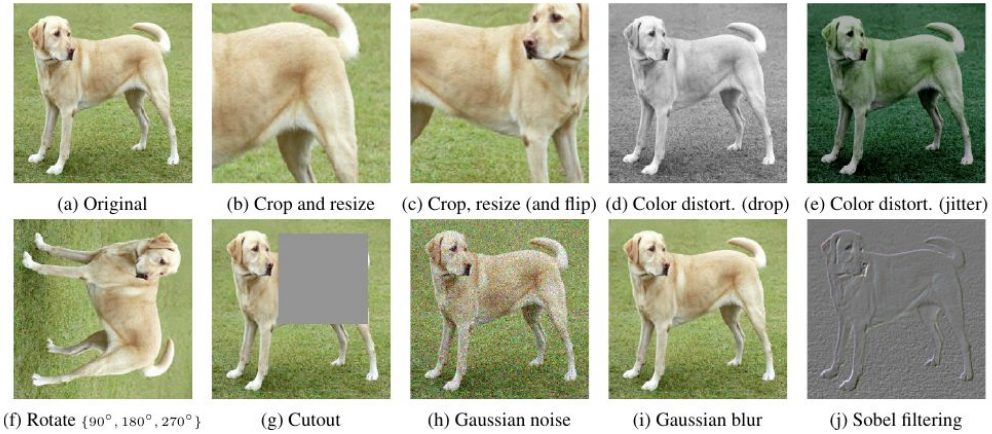
SimCLR

- Surpassed supervised performance
- Inspired a lot of the future SSL ideas
- Contributions:
  - Understanding of data augmentation
  - Introduced nonlinear projection head
  - Benefits of large batch sizes



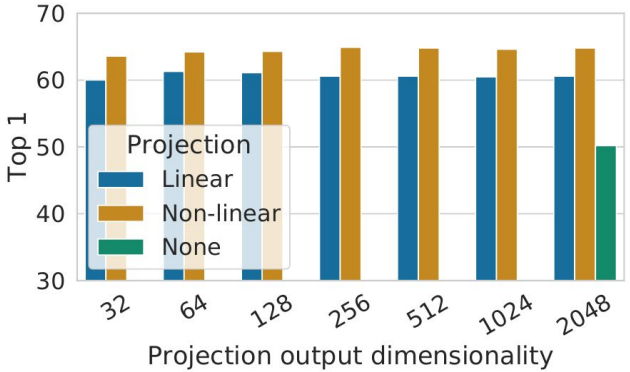
Data Augmentation in SSL

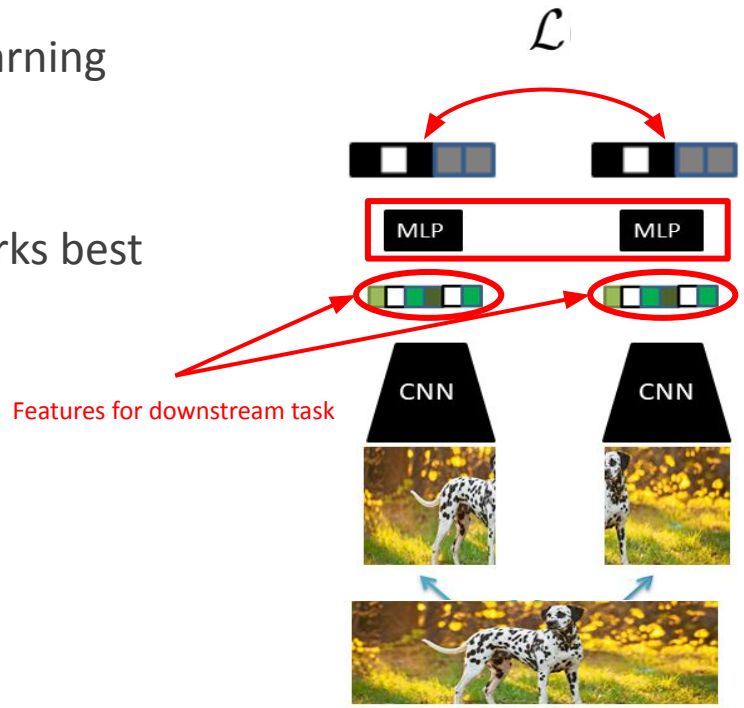
- Selection of data augmentations is important for model performance
- Random cropping & color jittering achieve best performance
  - Cropping enforces learning object parts
  - Color jittering prevents learning color histograms



SimCLR Projection Head

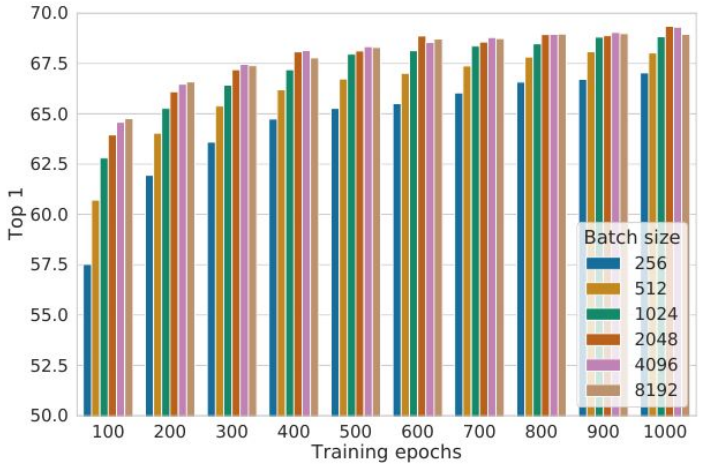
- Adding projection head for similarity learning
- Not used for downstream task
- Two-layer MLP with ReLU activation works best





Benefits of Large Batch Sizes

- Good negatives are important in contrastive learning
- Large batch sizes provide more negative examples
- Large differences during first training epochs

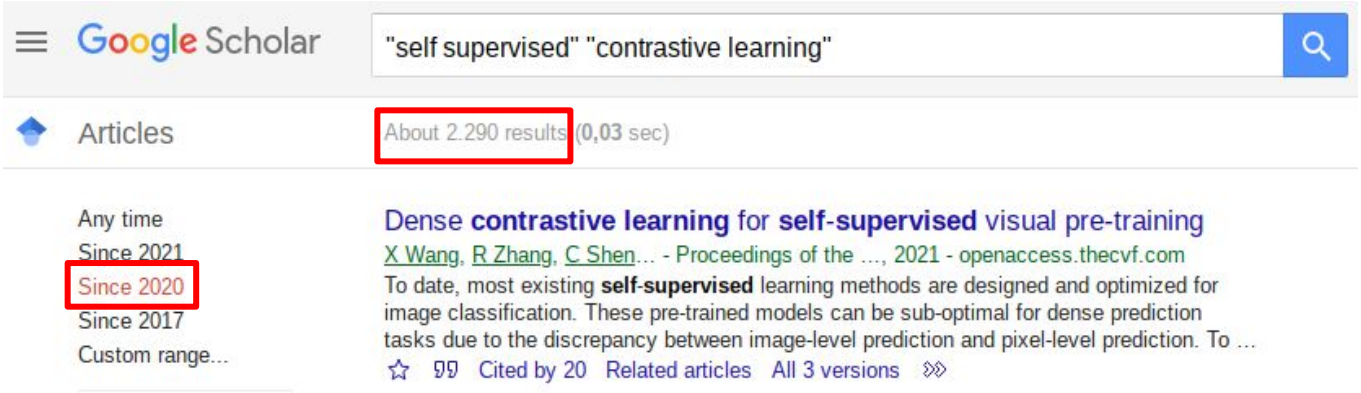


Cons: Not easy to train with batch size of 2048 if you are not Google

Illustrated SimCLR

Outlook

Hot-Topic in DL Research



Other Contrastive SSL Works

Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He Haoqi Fan Yuxin Wu Saining Xie Ross Girshick

Facebook AI Research (FAIR)

Code: https://github.com/facebookresearch/moco

Abstract

We present Momentum Contrast (MoCo) for unsupervised visual representation learning. From a perspective on contrastive learning [29] as dictionary look-up, we build a dynamic dictionary with a queue and a moving-averaged encoder. This enables building a large and consistent dictionary on-the-fly that facilitates contrastive unsupervised learning. MoCo provides competitive results under the common linear protocol on ImageNet classification. More importantly, the representations learned by MoCo transfer well to downstream tasks. MoCo can ***outperform*** *its supervised pre-training counterpart in* **7** *detection/segmentation tasks on PASCAL VOC, COCO, and other datasets, sometimes surpassing it by large margins. This suggests that the gap between unsupervised and supervised representation learning has been largely closed in many vision tasks.*

1. Introduction

Unsupervised representation learning is highly successful in natural language processing, *e.g.*, as shown by GPT [50, 51] and BERT [12]. But supervised pre-training is still dominant in computer vision, where unsupervised methods generally lag behind. The reason may stem from differences in their respective signal spaces. Language tasks have discrete signal spaces (words, sub-word units, *etc.*) for building tokenized *dictionaries*, on which unsupervised learning can be based. Computer vision, in contrast, further concerns dictionary building [54, 9, 5], as the raw signal is in a continuous, high-dimensional space and is not structured for human communication (*e.g.*, unlike words).

Several recent studies [61, 46, 36, 66, 35, 56, 2] present promising results on unsupervised visual representation learning using approaches related to the *contrastive loss* [29]. Though driven by various motivations, these methods can be thought of as building *dynamic dictionaries*. The

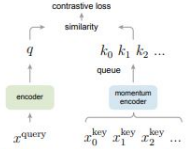


Figure 1. Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss. The dictionary keys $\{k_0, k_1, k_2, \ldots\}$ are defined on-the-fly by a set of data samples. The dictionary is built as a queue, with the current mini-batch enqueued and the oldest mini-batch dequeued, decoupling it from the mini-batch size. The keys are encoded by a slowly progressing encoder, driven by a momentum update with the query encoder. This method enables a large and consistent dictionary for learning visual representations.

From this perspective, we hypothesize that it is desirable to build dictionaries that are: (i) large *and* (ii) consistent as they evolve during training. Intuitively, a larger dictionary may better sample the underlying continuous, high-dimensional visual space, while the keys in the dictionary should be represented by the same or similar encoder so that their comparisons to the query are consistent. However, existing methods that use contrastive losses can be limited in one of these two aspects (discussed later in context).

We present *Momentum Contrast* (MoCo) as a way of building large and consistent dictionaries for unsupervised learning with a contrastive loss (Figure 1). We maintain the dictionary as a *queue* of data samples: the encoded representations of the current mini-batch are enqueued, and the

Big Self-Supervised Models are Strong Semi-Supervised Learners

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, Geoffrey Hinton
Google Research, Brain Team

Abstract

One paradigm for learning from few labeled examples while making best use of a large amount of unlabeled data is *unsupervised pretraining* followed by *supervised fine-tuning*. Although this paradigm uses unlabeled data in a *task-agnostic* way, in contrast to common approaches to semi-supervised learning for computer vision, we show that it is surprisingly effective for semi-supervised learning on ImageNet. A key ingredient of our approach is the use of big (deep and wide) networks during pretraining and fine-tuning. We find that, the fewer the labels, the more this approach (task-agnostic use of unlabeled data) benefits from a bigger network. After fine-tuning, the big network can be further improved and distilled into a much smaller one with little loss in classification accuracy by using the unlabeled examples for a second time, but in a *task-specific* way. The proposed semi-supervised learning algorithm can be summarized in three steps: *unsupervised pretraining* of a big ResNet model using SimCLRv2, *supervised fine-tuning* on a few labeled examples, and *distillation with unlabeled examples* for refining and transferring the task-specific knowledge. This procedure achieves 73.9% ImageNet top-1 accuracy with just 1% of the labels ($\leq$13 labeled images per class) using ResNet-50, a $10\times$ improvement in label efficiency over the previous state-of-the-art. With 10% of labels, ResNet-50 trained with our method achieves 77.5% top-1 accuracy, outperforming standard supervised training with all of the labels. [1]

1 Introduction

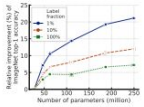


Figure 1: Bigger models yield larger gains when fine-tuning with fewer labeled examples.

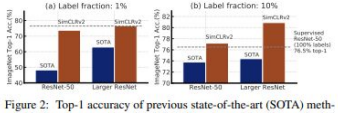


Figure 2: Top-1 accuracy of previous state-of-the-art (SOTA) methods [1, 2] and our method (SimCLRv2) on ImageNet using only 1% or 10% of the labels. Dashed line denotes fully supervised ResNet-50 trained with 100% of labels. Full comparisons in Table 3.

Dense Contrastive Learning for Self-Supervised Visual Pre-Training

Xinlong Wang[1], Rufeng Zhang[2], Chunhua Shen[1]*, Tao Kong[3], Lei Li[3]

[1]The University of Adelaide, Australia [2]Tongji University, China [3]ByteDance AI Lab

Abstract

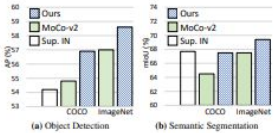
To date, most existing self-supervised learning methods are designed and optimized for image classification. These pre-trained models can be sub-optimal for dense prediction tasks due to the discrepancy between image-level prediction and pixel-level prediction. To fill this gap, we aim to design an effective, dense self-supervised learning method that directly works at the level of pixels (or local features) by taking into account the correspondence between local features. We present dense contrastive learning (DenseCL), which implements self-supervised learning by optimizing a pairwise contrastive (dis)similarity loss at the pixel level between two views of input images.

Compared to the baseline method MoCo-v2, our method introduces negligible computation overhead (only <1% slower), but demonstrates consistently superior performance when transferring to downstream dense prediction tasks including object detection, semantic segmentation and instance segmentation; and outperforms the state-of-the-art methods by a large margin. Specifically, over the strong MoCo-v2 baseline, our method achieves significant improvements of 2.0% AP on PASCAL VOC object detection, 1.1% AP on COCO object detection, 0.9% AP on COCO instance segmentation, 3.0% mIoU on PASCAL VOC semantic segmentation and 1.8% mIoU on Cityscapes semantic segmentation.

Code and models are available at: https://git.io/DenseCL

1. Introduction

Pre-training has become a well-established paradigm in many computer vision tasks. In a typical pre-training paradigm, models are first pre-trained on large-scale datasets and then fine-tuned on target tasks with less training data. Specifically, the supervised ImageNet pre-training has been dominant for years, where the models are pre-



(a) Object Detection (b) Semantic Segmentation

Figure 1 – Comparisons of pre-trained models by fine-tuning on object detection and semantic segmentation datasets. 'Sup. IN' denotes the supervised pre-training on ImageNet. 'COCO' and 'ImageNet' indicate the pre-training models trained on COCO and ImageNet respectively. (a): The object detection results of a Faster R-CNN detector fine-tuned on VOC `trainval07+12` for 24k iterations and evaluated on VOC `test2007`; (b): The semantic segmentation results of an FCN model fine-tuned on VOC `train_aug2012` for 20k iterations and evaluated on `val2012`. The results are averaged over 5 independent trials.

age classification pre-training and target dense prediction tasks, such as object detection [9, 25] and semantic segmentation [5]. The former focuses on assigning a category to an input image, while the latter needs to perform dense classification or regression over the whole image. For example, semantic segmentation aims to assign a category for each pixel, and object detection aims to predict the categories and bounding boxes for all object instances of interest. A straightforward solution would be to pre-train on dense prediction tasks directly. However, these tasks' annotation is notoriously time-consuming compared to the image-level labeling, making it hard to collect data at a massive scale to pre-train a universal feature representation.

Recently, unsupervised visual pre-training has attracted much research attention, which aims to learn a proper visual representation from a large set of unlabeled images. A

Clustering SSL

Deep Clustering for Unsupervised Learning of Visual Features

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze

Facebook AI Research
{mathilde,bojanowski,ajoulin,matthijs}@fb.com

Abstract. Clustering is a class of unsupervised learning methods that has been extensively applied and studied in computer vision. Little work has been done to adapt it to the end-to-end training of visual features on large-scale datasets. In this work, we present DeepCluster, a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. DeepCluster iteratively groups the features with a standard clustering algorithm, k-means, and uses the subsequent assignments as supervision to update the weights of the network. We apply DeepCluster to the unsupervised training of convolutional neural networks on large datasets like ImageNet and YFCC100M. The resulting model outperforms the current state of the art by a significant margin on all the standard benchmarks.

Keywords: unsupervised learning, clustering

1 Introduction

Pre-trained convolutional neural networks, or convnets, have become the building blocks in most computer vision applications [8, 9, 50, 65]. They produce excellent general-purpose features that can be used to improve the generalization of models learned on a limited amount of data [53]. The existence of ImageNet [12], a large fully-supervised dataset, has been fueling advances in pre-training of convnets. However, Stock and Cisse [57] have recently presented empirical evidence that the performance of state-of-the-art classifiers on ImageNet is largely underestimated, and little error is left unresolved. This explains in part why the performance has been saturating despite the numerous novel architectures proposed in recent years [9, 21, 23]. As a matter of fact, ImageNet is relatively small by today's standards; it "only" contains a million images that cover the specific domain of object classification. A natural way to move forward is to build a bigger and more diverse dataset, potentially consisting of billions of images. This,

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

Mathilde Caron[1,2] **Ishan Misra**[2] **Julien Mairal**[1]

Priya Goyal[2] **Piotr Bojanowski**[2] **Armand Joulin**[2]

[1] Inria* [2] Facebook AI Research

Abstract

Unsupervised image representations have significantly reduced the gap with supervised pretraining, notably with the recent achievements of contrastive learning methods. These contrastive methods typically work online and rely on a large number of explicit pairwise feature comparisons, which is computationally challenging. In this paper, we propose an online algorithm, SwAV, that takes advantage of contrastive methods without requiring to compute pairwise comparisons. Specifically, our method simultaneously clusters the data while enforcing consistency between cluster assignments produced for different augmentations (or "views") of the same image, instead of comparing features directly as in contrastive learning. Simply put, we use a "swapped" prediction mechanism where we predict the code of a view from the representation of another view. Our method can be trained with large and small batches and can scale to unlimited amounts of data. Compared to previous contrastive methods, our method is more memory efficient since it does not require a large memory bank or a special momentum network. In addition, we also propose a new data augmentation strategy, `multi-crop`, that uses a mix of views with different resolutions in place of two full-resolution views, without increasing the memory or compute requirements. We validate our findings by achieving 75.3% top-1 accuracy on ImageNet with ResNet-50, as well as surpassing supervised pretraining on all the considered transfer tasks.

1 Introduction

Unsupervised visual representation learning, or self-supervised learning, aims at obtaining features without using manual annotations and is rapidly closing the performance gap with supervised pre-

Prototypical Contrastive Learning of Unsupervised Representations

Junnan Li, Pan Zhou, Caiming Xiong, Steven C.H. Hoi
Salesforce Research
{junnan.li,pzhou,cxiong,shoi}@salesforce.com

Abstract

This paper presents Prototypical Contrastive Learning (PCL), an unsupervised representation learning method that bridges contrastive learning with clustering. PCL not only learns low-level features for the task of instance discrimination, but more importantly, it encodes semantic structures discovered by clustering into the learned embedding space. Specifically, we introduce prototypes as latent variables to help find the maximum-likelihood estimation of the network parameters in an Expectation-Maximization framework. We iteratively perform E-step as finding the distribution of prototypes via clustering and M-step as optimizing the network via contrastive learning. We propose ProtoNCE loss, a generalized version of the InfoNCE loss for contrastive learning, which encourages representations to be closer to their assigned prototypes. PCL outperforms state-of-the-art instance-wise contrastive learning methods on multiple benchmarks with substantial improvement in low-resource transfer learning. Code and pretrained models are available at https://github.com/salesforce/PCL

1 Introduction

Unsupervised visual representation learning aims to learn image representations from pixels themselves without relying on semantic annotations, and recent advances are largely driven by instance discrimination tasks (Wu et al., 2018; Ye et al., 2019; He et al., 2020; Misra & van der Maaten, 2020; Hjelm et al., 2019; Oord et al., 2018; Tian et al., 2019). These methods usually consist of two key components: image transformation and contrastive loss. Image transformation aims to generate multiple embeddings that represent the same image, by data augmentation (Ye et al., 2019; Bachman et al., 2019; Chen et al., 2020a), patch perturbation (Misra & van der Maaten, 2020), or using momentum features (He et al., 2020). The contrastive loss, in the form of a noise contrastive estimator (Gutmann & Hyvärinen, 2010), aims to bring closer samples from the same instance and separate samples from different instances. Essentially, instance-wise contrastive learning leads to an embedding space where all instances are well-separated, and each instance is locally smooth (*i.e.* input with perturbations have similar representations).

Distillation SSL

Bootstrap Your Own Latent
A New Approach to Self-Supervised Learning

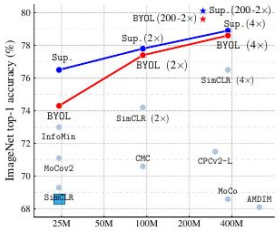
Jean-Bastien Grill[*,1] Florian Strub[*,1] Florent Altché[*,1] Corentin Tallec[*,1] Pierre H. Richemond[*,1,2]
Elena Buchatskaya[1] Carl Doersch[1] Bernardo Avila Pires[1] Zhaohan Daniel Guo[1]
Mohammad Gheshlaghi Azar[1] Bilal Piot[1] Koray Kavukcuoglu[1] Rémi Munos[1] Michal Valko[1]
[1]DeepMind [2]Imperial College
[jbgrill,fstrub,altche,corentint,richemond]@google.com

Abstract

We introduce **B**ootstrap **Y**our **O**wn **L**atent (BYOL), a new approach to self-supervised image representation learning. BYOL relies on two neural networks, referred to as *online* and *target* networks, that interact and learn from each other. From an augmented view of an image, we train the online network to predict the target network representation of the same image under a different augmented view. At the same time, we update the target network with a slow-moving average of the online network. While state-of-the art methods rely on negative pairs, BYOL achieves a new state of the art *without them*. BYOL reaches 74.3% top-1 classification accuracy on ImageNet using a linear evaluation with a ResNet-50 architecture and 79.6% with a larger ResNet. We show that BYOL performs on par or better than the current state of the art on both transfer and semi-supervised benchmarks. Our implementation and pretrained models are given on GitHub.

1 Introduction

Learning good image representations is a key challenge in computer vision [1, 2, 3] as it allows for efficient training on downstream tasks [4, 5, 6, 7]. Many different training approaches have been proposed to learn such representations, usually relying on visual pretext tasks. Among them, state-of-the-art contrastive methods [8, 9, 10, 11, 12] are trained by reducing the distance between representations of different augmented views of the same image ('positive pairs'), and increasing the distance between representations of augmented views from different images ('negative pairs'). These methods need careful treatment of negative pairs [13] by either relying on large batch sizes [8, 12], memory banks [9] or customized mining strategies [14, 15] to retrieve the negative pairs. In addition, their performance critically depends on the choice of image augmentations [8, 12].



Exploring Simple Siamese Representation Learning

Xinlei Chen Kaiming He
Facebook AI Research (FAIR)

Abstract

Siamese networks have become a common structure in various recent models for unsupervised visual representation learning. These models maximize the similarity between two augmentations of one image, subject to certain conditions for avoiding collapsing solutions. In this paper, we report surprising empirical results that ***simple Siamese*** *networks can learn meaningful representations even using* ***none*** *of the following: (i) negative sample pairs, (ii) large batches, (iii) momentum encoders. Our experiments show that collapsing solutions do exist for the loss and structure, but a stop-gradient operation plays an essential role in preventing collapsing. We provide a hypothesis on the implication of stop-gradient, and further show proof-of-concept experiments verifying it. Our "SimSiam" method achieves competitive results on ImageNet and downstream tasks. We hope this simple baseline will motivate people to rethink the roles of Siamese architectures for unsupervised representation learning. Code will be made available.*

1. Introduction

Recently there has been steady progress in un-/self-supervised representation learning, with encouraging results on multiple visual tasks (*e.g.*, [2, 17, 8, 15, 7]). Despite various original motivations, these methods generally involve certain forms of Siamese networks [4]. Siamese networks are weight-sharing neural networks applied on two or more inputs. They are natural tools for *comparing* (including but not limited to "*contrasting*") entities. Recent methods define the inputs as two augmentations of one image, and maximize the similarity subject to different conditions.

An undesired trivial solution to Siamese networks is

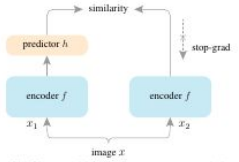


Figure 1. **SimSiam architecture**. Two augmented views of one image are processed by the same encoder network f (a backbone plus a projection MLP). Then a prediction MLP h is applied on one side, and a stop-gradient operation is applied on the other side. The model maximizes the similarity between both sides. It uses neither negative pairs nor a momentum encoder.

clustering, BYOL [15] relies only on positive pairs but it does not collapse in case a momentum encoder is used.

In this paper, we report that simple Siamese networks can work surprisingly well with *none* of the above strategies for preventing collapsing. Our model directly maximizes the similarity of one image's two views, using *neither* negative pairs *nor* a momentum encoder. It works with typical batch sizes and does not rely on large-batch training. We illustrate this "SimSiam" method in Figure 1.

Thanks to the conceptual simplicity, SimSiam can serve as a hub that relates several existing methods. In a nutshell, our method can be thought of as "BYOL *without* the momentum encoder". Unlike BYOL but like SimCLR and SwAV, our method directly shares the weights between the two branches, so it can also be thought of as "SimCLR *without* negative pairs", and "SwAV *without* online clustering". Interestingly, SimSiam is related to each method by removing one of its core components. Even so, SimSiam

And many more ideas...

Barlow Twins: Self-Supervised Learning via Redundancy Reduction

Jure Zbontar [*1] Li Jing [*1] Ishan Misra [1] Yann LeCun [1 2] Stéphane Deny [1]

Abstract

Self-supervised learning (SSL) is rapidly closing the gap with supervised methods on large computer vision benchmarks. A successful approach to SSL is to learn embeddings which are invariant to distortions of the input sample. However, a recurring issue with this approach is the existence of trivial constant solutions. Most current methods avoid such solutions by careful implementation details. We propose an objective function that naturally avoids collapse by measuring the cross-correlation matrix between the outputs of two identical networks fed with distorted versions of a sample, and making it as close to the identity matrix as possible. This causes the embedding vectors of distorted versions of a sample to be similar, while minimizing the redundancy between the components of these vectors. The method is called BARLOW TWINS, owing to neuroscientist H. Barlow's *redundancy-reduction principle* ap-

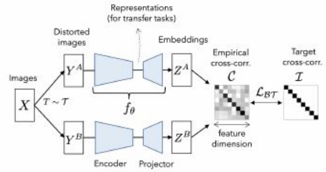


Figure 1. BARLOW TWINS's objective function measures the cross-correlation matrix between the embeddings of two identical networks fed with distorted versions of a batch of samples, and tries to make this matrix close to the identity. This causes the embedding vectors of distorted versions of a sample to be similar, while minimizing the redundancy between the components of these vectors. BARLOW TWINS is competitive with state-of-the-art methods for self-supervised learning while being conceptually simpler, naturally avoiding trivial constant (i.e. collapsed) embeddings, and being robust to the training batch size.

1. Introduction

Understanding Self-Supervised Learning Dynamics without Contrastive Pairs

Yuandong Tian [1] Xinlei Chen [1] Surya Ganguli [1 2]

Abstract

While contrastive approaches of self-supervised learning (SSL) learn representations by minimizing the distance between two augmented views of the same data point (positive pairs) and maximizing views from different data points (negative pairs), recent *non-contrastive* SSL (e.g., BYOL and SimSiam) show remarkable performance *without* negative pairs, with an extra learnable predictor and a stop-gradient operation. A fundamental question arises: why do these methods not collapse into trivial representations? We answer this question via a simple theoretical study and propose a novel approach, **DirectPred**, that *directly* sets the linear predictor based on the statistics of its inputs, without gradient training. On ImageNet, it performs comparably with more complex two-layer non-linear predictors that employ BatchNorm and outperforms a linear predictor by 2.5% in 300-epoch training (and 5% in 60-epoch). **DirectPred** is motivated by our theoretical study of the nonlinear learning dynamics of non-contrastive SSL in simple linear networks. Our study yields conceptual insights into how non-contrastive SSL methods learn, how they avoid representational

quiring expensive target labels (Devlin et al., 2018). Many state-of-the-art SSL methods in computer vision employ the principle of contrastive learning (Oord et al., 2018; Tian et al., 2019; He et al., 2020; Chen et al., 2020a; Bachman et al., 2019) whereby the hidden representations of two augmented views of the same object (positive pairs) are brought closer together, while those of different objects (negative pairs) are encouraged to be further apart. Minimizing differences between positive pairs encourages modeling invariances, while contrasting negative pairs is thought to be required to prevent representational collapse (i.e., mapping all data to the same representation).

However, some recent SSL work, notably BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2020), have shown the remarkable capacity to learn powerful representations using only positive pairs, *without* ever contrasting negative pairs. These methods employ a dual pair of Siamese networks (Bromley et al., 1994) (Fig. 1): the representation of two views are trained to match, one obtained by the composition of an online and predictor network, and the other by a target network. The target network is *not* trained via gradient descent; and either employs a direct copy of the online network (e.g., SimSiam (Chen & He, 2020)), or a momentum encoder that slowly follows the online network in a delayed fashion through an exponential moving average (EMA) (e.g., MoCo (He et al., 2020; Chen et al.,

VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning

Adrien Bardes [1,2] Jean Ponce [2,4] Yann LeCun [1,3,4]
[1] Facebook AI Research
[2] Inria, École normale supérieure, CNRS, PSL Research University
[3] Courant Institute, New York University
[4] Center for Data Science, New York University

Abstract

Recent self-supervised methods for image representation learning are based on maximizing the agreement between embedding vectors from different views of the same image. A trivial solution is obtained when the encoder outputs constant vectors. This *collapse problem* is often avoided through implicit biases in the learning architecture, that often lack a clear justification or interpretation. In this paper, we introduce VICReg (Variance-Invariance-Covariance Regularization), a method that explicitly avoids the collapse problem with a simple regularization term on the variance of the embeddings along each dimension individually. VICReg combines the variance term with a decorrelation mechanism based on redundancy reduction and covariance regularization, and achieves results on par with the state of the art on several downstream tasks. In addition, we show that incorporating our new variance term into other methods helps stabilize the training and leads to performance improvements.

1 Introduction

Self-supervised representation learning has made significant progress over the last years, almost reaching the performance of supervised baselines on many downstream tasks [2, 33, 22, 40, 7, 21, 11, 49]. Several recent approaches rely on joint embedding learning with siamese networks [4], trained by maximising the agreement between different views of the same image. Contrastive methods [25, 22, 8] use a negative term that explicitly encourages the representations of different images

References

1. https://towardsdatascience.com/all-you-want-to-know-about-deep-learning-8d68dcffc258
2. Goodfellow, Ian, et al. Deep learning. Vol. 1. No. 2. Cambridge: MIT press, 2016.
3. https://www.youtube.com/watch?v=8L10w1KoOU8
4. Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." arXiv preprint arXiv:1803.07728 (2018).
5. Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised visual representation learning by context prediction." Proceedings of the IEEE international conference on computer vision. 2015.
6. Lee, Hsin-Ying, et al. "Unsupervised representation learning by sorting sequences." Proceedings of the IEEE International Conference on Computer Vision. 2017.
7. Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.
8. He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

References

8. Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." European conference on computer vision. Springer, Cham, 2016.
9. Vincent, Pascal, et al. "Extracting and composing robust features with denoising autoencoders." Proceedings of the 25th international conference on Machine learning. 2008.
10. Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Colorful image colorization." European conference on computer vision. Springer, Cham, 2016.
11. https://lilianweng.github.io/lil-log/2019/11/10/self-supervised-learning.html
12. Arandjelovic, Relja, and Andrew Zisserman. "Objects that sound." Proceedings of the European conference on computer vision (ECCV). 2018.
13. Jenni, Simon, and Paolo Favaro. "Self-supervised feature learning by learning to spot artifacts." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
14. Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).