Paper Review Seminar # 1

(2022. 07. 12)

SimCLR & MoCo (v1 & v2)

통계데이터사이언스학과 통합과정 5학기 이승한

목차

1. Introduction of Contrastive Learning

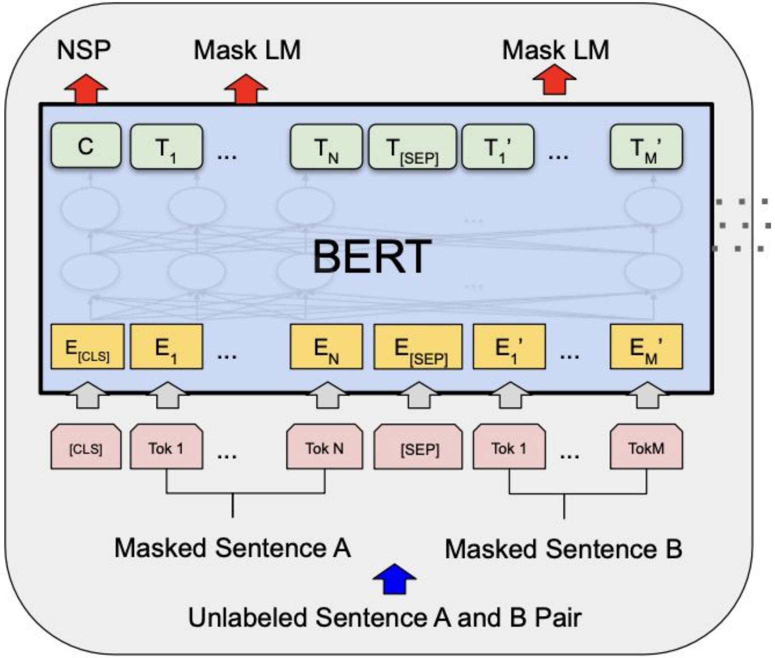
What is **Self-Supervised** Learning?

- Supervised Learning : label O
- Unsupervised Learning : label X
- Semi-supervised Learning : label O & X
- **Self-supervised** Learning : **get label from data itself !**

1. Introduction of Contrastive Learning

Self-Supervised Learning in NLP

- MLM (Masked Language Model)
- NSP (Next Sentence Prediction)



1. Introduction of Contrastive Learning

Self-Supervised Learning in CV

- Pretext Tasks
  - learn representation that could help downstream task
  - ex) Exemplar (2014), Context Prediction (2015), Jigsaw puzzle (2016), …
- Contrastive Learning
  - make similar/dissimilar data close/far

1. Introduction of Contrastive Learning

Self-Supervised Learning in CV - Pretext Tasks

Exemplar (2014)

- crop 32x32 patch, where gradient is significant & data augmentation with this patch
- 1 instance = 1 class
- cons) unsuitable for large dataset

$$L(X) = \sum_{\mathrm{x}_i \in X} \sum_{T \in T_i} l(i, T\mathrm{x}_i)$$

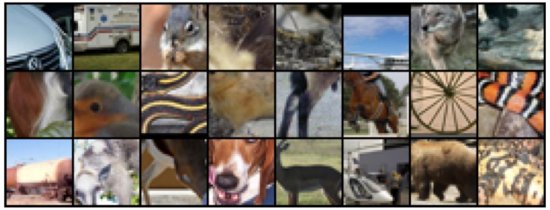


Fig. 1. Exemplary patches sampled from the STL unlabeled dataset which are later augmented by various transformations to obtain surrogate data for the CNN training.

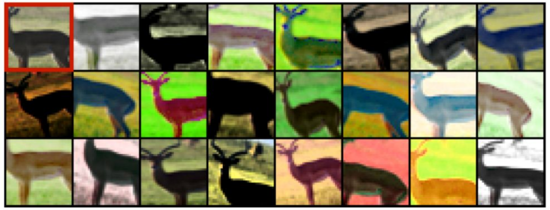


Fig. 2. Several random transformations applied to one of the patches extracted from the STL unlabeled dataset. The original ('seed') patch is in the top left corner.

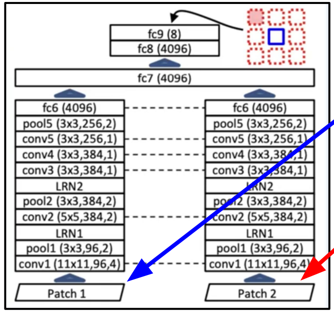
Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks (Dosoviskiy, NIPS 2014)

1. Introduction of Contrastive Learning

Self-Supervised Learning in CV - Pretext Tasks

Context Prediction (2015)

- guess the relative location (1~8)
- cons) cheating with texture/boundary



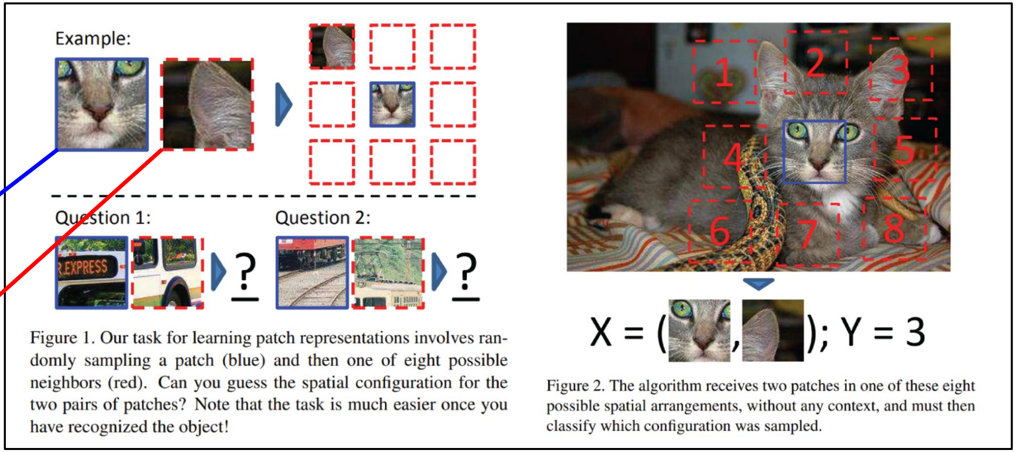


Figure 1. Our task for learning patch representations involves randomly sampling a patch (blue) and then one of eight possible neighbors (red). Can you guess the spatial configuration for the two pairs of patches? Note that the task is much easier once you have recognized the object!

Figure 2. The algorithm receives two patches in one of these eight possible spatial arrangements, without any context, and must then classify which configuration was sampled.

Unsupervised Visual Representation Learning by Context Prediction (Doersch et al., ICCV 2015)

1. Introduction of Contrastive Learning

Self-Supervised Learning in CV - Pretext Tasks

Jigsaw Puzzle (2016)

- solve jigsaw puzzle
- number of class :
    - original) 9! = 362,880
    - proposed) 100

      (remove similar permutation)

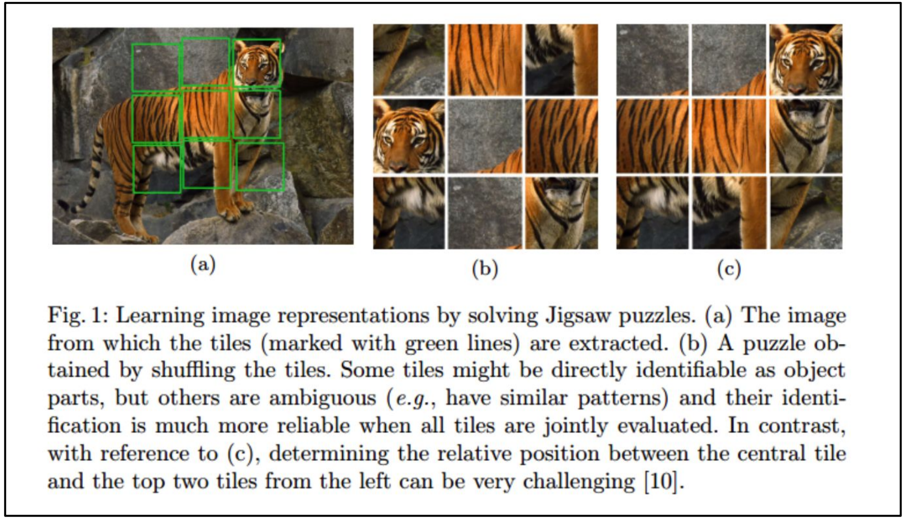

(a) (b) (c)

Fig. 1: Learning image representations by solving Jigsaw puzzles. (a) The image from which the tiles (marked with green lines) are extracted. (b) A puzzle obtained by shuffling the tiles. Some tiles might be directly identifiable as object parts, but others are ambiguous (*e.g.*, have similar patterns) and their identification is much more reliable when all tiles are jointly evaluated. In contrast, with reference to (c), determining the relative position between the central tile and the top two tiles from the left can be very challenging [10].

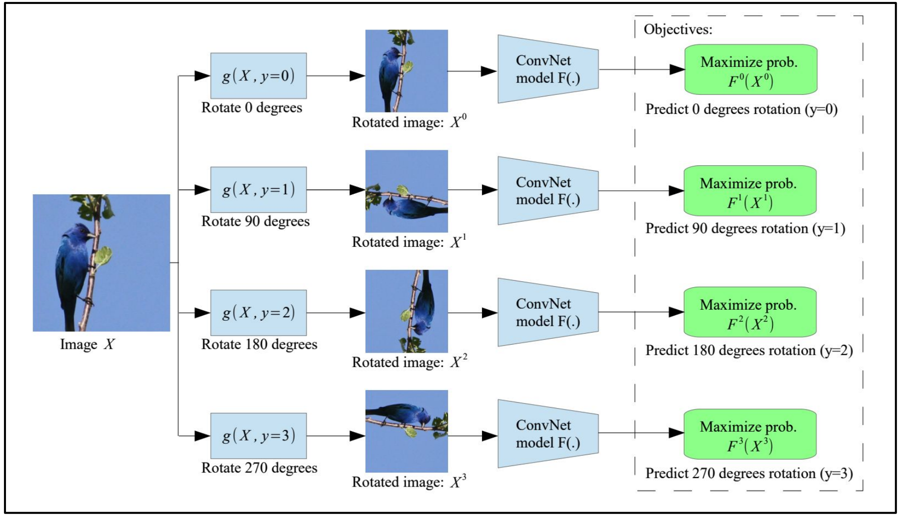
Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles (Noroozi and Favaro, ECCV 2016)

1. Introduction of Contrastive Learning

Self-Supervised Learning in CV - Pretext Tasks

Rotation Prediction (2016)

- rotate image & guess the rotation angle



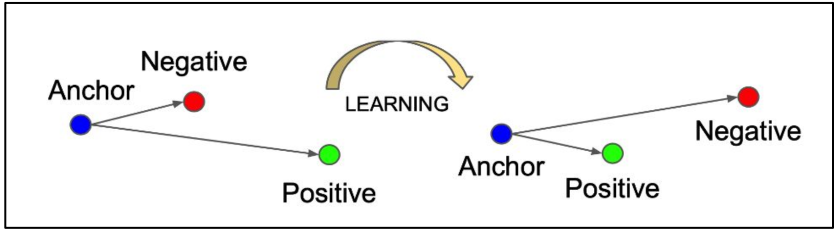
Unsupervised Representation Learning by Predicting Image Rotations (Gidaris et al., ICLR 2018)

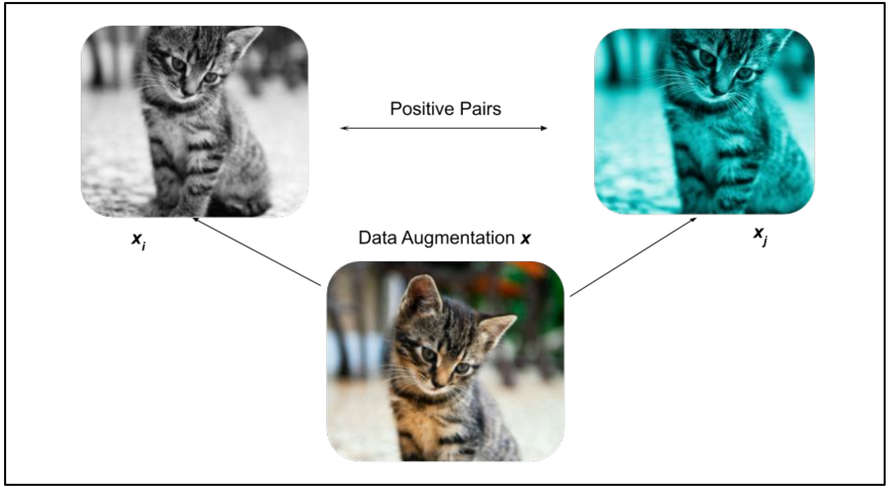
1. Introduction of Contrastive Learning

Self-Supervised Learning in CV - Contrastive Learning

- make **similar** data **close** to each other

  make **dissimilar** data **far** from each other
- various loss functions





1. Introduction of Contrastive Learning

Self-Supervised Learning in CV - Contrastive Learning

(1) Contrastive Loss

$$\mathcal{L}_{\text{cont}}(\mathbf{x}_i, \mathbf{x}_j, \theta) = \mathbb{1}[y_i = y_j] \| f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j) \|_2^2 + \mathbb{1}[y_i \neq y_j] \max(0, \epsilon - \| f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j) \|_2)^2$$

(2) Triplet Loss

$$\mathcal{L}_{\text{triplet}}(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-) = \sum_{\mathbf{x} \in \mathcal{X}} \max\left(0, \| f(\mathbf{x}) - f(\mathbf{x}^+) \|_2^2 - \| f(\mathbf{x}) - f(\mathbf{x}^-) \|_2^2 + \epsilon\right)$$

(3) Noise Contrastive Estimation

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[\log \sigma(\ell_\theta(\mathbf{x}_i)) + \log(1 - \sigma(\ell_\theta(\tilde{\mathbf{x}}_i))) \right]$$

$$\text{where } \sigma(\ell) = \frac{1}{1 + \exp(-\ell)} = \frac{p_\theta}{p_\theta + q}$$

target sample $\sim P(\mathbf{x} | C = 1; \theta) = p_\theta(\mathbf{x})$

noise sample $\sim P(\tilde{\mathbf{x}} | C = 0) = q(\tilde{\mathbf{x}})$

1. Introduction of Contrastive Learning

Self-Supervised Learning in CV - Contrastive Learning

(1) Contrastive Loss

$$\mathcal{L}_{\text{cont}}(\mathbf{x}_i,\mathbf{x}_j,\theta) = \mathbb{1}[y_i = y_j]\overset{\Downarrow}{\boxed{\|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|}}_2^2 + \mathbb{1}[y_i \neq y_j]\max(0, \epsilon - \overset{\Uparrow}{\boxed{\|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|_2}})^2$$

(2) Triplet Loss

$$\mathcal{L}_{\text{triplet}}(\mathbf{x},\mathbf{x}^+,\mathbf{x}^-) = \sum_{\mathbf{x}\in\mathcal{X}} \max\left(0, \overset{\Downarrow}{\boxed{\|f(\mathbf{x}) - f(\mathbf{x}^+)\|_2^2}} - \overset{\Uparrow}{\boxed{\|f(\mathbf{x}) - f(\mathbf{x}^-)\|_2^2}} + \epsilon\right)$$

🟥 : positive (= similar pair)

🟦 : negative (= dissimilar pair)

(3) Noise Contrastive Estimation

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\underset{\Uparrow}{\boxed{\log\sigma(\ell_\theta(\mathbf{x}_i))}} + \log(1 - \underset{\Downarrow}{\boxed{\sigma(\ell_\theta(\tilde{\mathbf{x}}_i))}})\right]$$

$$\text{where } \sigma(\ell) = \frac{1}{1+\exp(-\ell)} = \frac{p_\theta}{p_\theta + q}$$

target sample $\sim P(\mathbf{x}|C=1;\theta) = p_\theta(\mathbf{x})$

noise sample $\sim P(\tilde{\mathbf{x}}|C=0) = q(\tilde{\mathbf{x}})$

1. Introduction of Contrastive Learning

Self-Supervised Learning in CV - Contrastive Learning

(4) InfoNCE

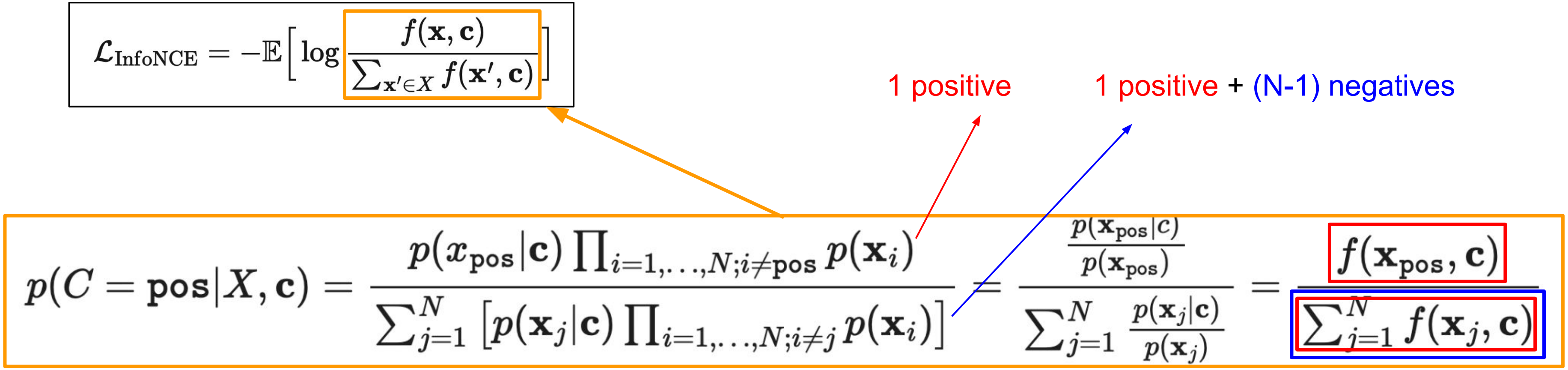
$$\boxed{\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \boxed{\frac{f(\mathbf{x},\mathbf{c})}{\sum_{\mathbf{x}'\in X} f(\mathbf{x}',\mathbf{c})}}\right]}$$

$$\boxed{p(C=\mathbf{pos}|X,\mathbf{c}) = \frac{p(x_{\mathbf{pos}}|\mathbf{c})\prod_{i=1,\ldots,N;i\neq\mathbf{pos}} p(\mathbf{x}_i)}{\sum_{j=1}^{N}\left[p(\mathbf{x}_j|\mathbf{c})\prod_{i=1,\ldots,N;i\neq j} p(\mathbf{x}_i)\right]} = \frac{\frac{p(\mathbf{x}_{\mathbf{pos}}|c)}{p(\mathbf{x}_{\mathbf{pos}})}}{\sum_{j=1}^{N}\frac{p(\mathbf{x}_j|\mathbf{c})}{p(\mathbf{x}_j)}} = \frac{f(\mathbf{x}_{\mathbf{pos}},\mathbf{c})}{\sum_{j=1}^{N} f(\mathbf{x}_j,\mathbf{c})}}$$

1. Introduction of Contrastive Learning

Self-Supervised Learning in CV - Contrastive Learning

(4) InfoNCE



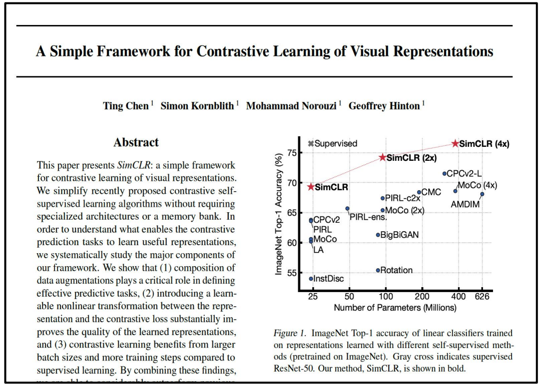
$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \boxed{\frac{f(\mathbf{x},\mathbf{c})}{\sum_{\mathbf{x}'\in X} f(\mathbf{x}',\mathbf{c})}}\right]$$

1 positive

1 positive + (N-1) negatives

$$p(C=\mathtt{pos}|X,\mathbf{c}) = \frac{p(x_{\mathtt{pos}}|\mathbf{c})\prod_{i=1,\ldots,N;i\neq\mathtt{pos}} p(\mathbf{x}_i)}{\sum_{j=1}^{N}\left[p(\mathbf{x}_j|\mathbf{c})\prod_{i=1,\ldots,N;i\neq j} p(\mathbf{x}_i)\right]} = \frac{\frac{p(\mathbf{x}_{\mathtt{pos}}|c)}{p(\mathbf{x}_{\mathtt{pos}})}}{\sum_{j=1}^{N}\frac{p(\mathbf{x}_j|\mathbf{c})}{p(\mathbf{x}_j)}} = \frac{\boxed{f(\mathbf{x}_{\mathtt{pos}},\mathbf{c})}}{\boxed{\boxed{\sum_{j=1}^{N} f(\mathbf{x}_j,\mathbf{c})}}}$$

SimCLR (2020)



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,

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.

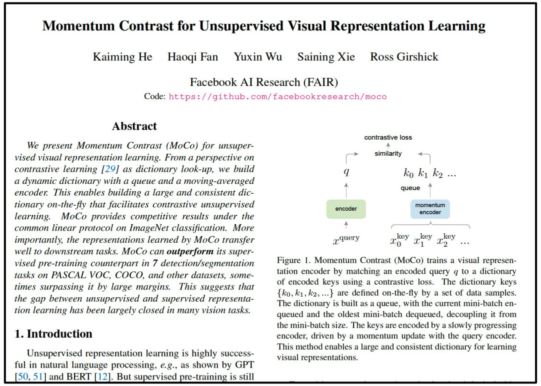
https://arxiv.org › cs ⋮

A Simple Framework for Contrastive Learning of Visual ... - arXiv

by T Chen · 2020 · Cited by 5112 — Abstract: This paper presents **SimCLR**: a simple framework for contrastive learning of visual representations. We simplify recently proposed ...

Cite as: arXiv:2002.05709

MoCo v1 (2019)



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

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 $\{k_0, k_1, k_2, ...\}$ 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.

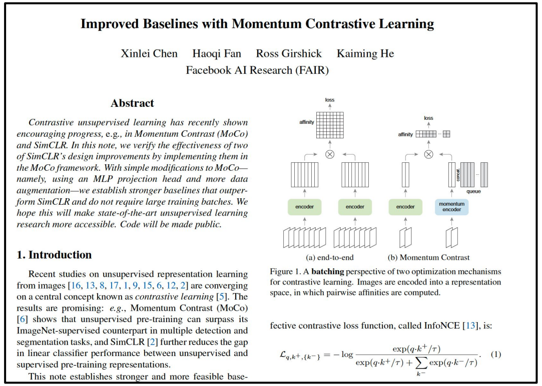
https://arxiv.org › cs ⋮

Momentum Contrast for Unsupervised Visual Representation ...

by K He · 2019 · Cited by 3735 — Abstract: We present Momentum Contrast (**MoCo**) for unsupervised visual representation learning. From a perspective on contrastive learning as ...

Cite as: arXiv:1911.05722

MoCo v2 (2020)



Improved Baselines with Momentum Contrastive Learning

Xinlei Chen Haoqi Fan Ross Girshick Kaiming He
Facebook AI Research (FAIR)

Abstract

Contrastive unsupervised learning has recently shown encouraging progress, e.g., *in Momentum Contrast (MoCo) and SimCLR. In this note, we verify the effectiveness of two of SimCLR's design improvements by implementing them in the MoCo framework. With simple modifications to MoCo—namely, using an MLP projection head and more data augmentation—we establish stronger baselines that outperform SimCLR and do not require large training batches. We hope this will make state-of-the-art unsupervised learning research more accessible. Code will be made public.*

1. Introduction

Recent studies on unsupervised representation learning from images [16, 13, 8, 17, 1, 9, 15, 6, 12, 2] are converging on a central concept known as *contrastive learning* [5]. The results are promising: *e.g.*, Momentum Contrast (MoCo) [6] shows that unsupervised pre-training can surpass its ImageNet-supervised counterpart in multiple detection and segmentation tasks, and SimCLR [2] further reduces the gap in linear classifier performance between unsupervised and supervised pre-training representations.

This note establishes stronger and more feasible base-

Figure 1. A **batching** perspective of two optimization mechanisms for contrastive learning. Images are encoded into a representation space, in which pairwise affinities are computed.

fective contrastive loss function, called InfoNCE [13], is:

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q{\cdot}k^+/\tau)}{\exp(q{\cdot}k^+/\tau) + \sum_{k^-} \exp(q{\cdot}k^-/\tau)}. \quad (1)$$

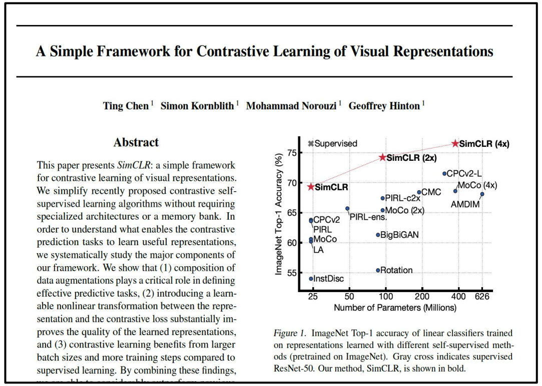
http://arxiv.org › cs ⋮

Improved Baselines with Momentum Contrastive Learning

by X Chen · 2020 · Cited by 1088 — Abstract: Contrastive unsupervised learning has recently shown encouraging progress, e.g., in Momentum Contrast (**MoCo**) and SimCLR.

Cite as: arXiv:2003.04297

SimCLR (2020)



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,

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.

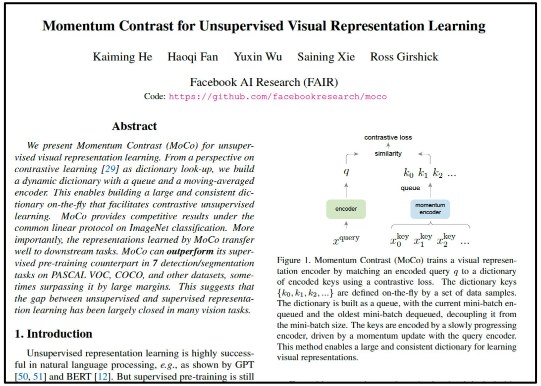
https://arxiv.org › cs ⋮

A Simple Framework for Contrastive Learning of Visual ... - arXiv

by T Chen · 2020 · Cited by 5112 — Abstract: This paper presents **SimCLR**: a simple framework for contrastive learning of visual representations. We simplify recently proposed ...

Cite as: arXiv:2002.05709

MoCo v1 (2019)



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

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 $\{k_0, k_1, k_2, ...\}$ 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.

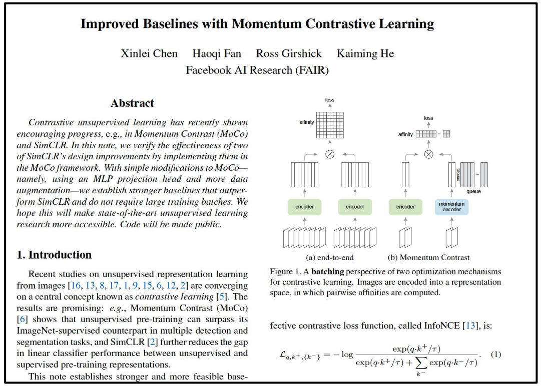
https://arxiv.org › cs ⋮

Momentum Contrast for Unsupervised Visual Representation ...

by K He · 2019 · Cited by 3735 — Abstract: We present Momentum Contrast (**MoCo**) for unsupervised visual representation learning. From a perspective on contrastive learning as ...

Cite as: arXiv:1911.05722

MoCo v2 (2020)



Improved Baselines with Momentum Contrastive Learning

Xinlei Chen Haoqi Fan Ross Girshick Kaiming He
Facebook AI Research (FAIR)

Abstract

Contrastive unsupervised learning has recently shown encouraging progress, e.g., *in Momentum Contrast (MoCo) and SimCLR. In this note, we verify the effectiveness of two of SimCLR's design improvements by implementing them in the MoCo framework. With simple modifications to MoCo—namely, using an MLP projection head and more data augmentation—we establish stronger baselines that outperform SimCLR and do not require large training batches. We hope this will make state-of-the-art unsupervised learning research more accessible. Code will be made public.*

1. Introduction

Recent studies on unsupervised representation learning from images [16, 13, 8, 17, 1, 9, 15, 6, 12, 2] are converging on a central concept known as *contrastive learning* [5]. The results are promising: *e.g.*, Momentum Contrast (MoCo) [6] shows that unsupervised pre-training can surpass its ImageNet-supervised counterpart in multiple detection and segmentation tasks, and SimCLR [2] further reduces the gap in linear classifier performance between unsupervised and supervised pre-training representations.

This note establishes stronger and more feasible base-

(a) end-to-end (b) Momentum Contrast

Figure 1. A **batching** perspective of two optimization mechanisms for contrastive learning. Images are encoded into a representation space, in which pairwise affinities are computed.

fective contrastive loss function, called InfoNCE [13], is:

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q{\cdot}k^+/\tau)}{\exp(q{\cdot}k^+/\tau) + \sum\limits_{k^-} \exp(q{\cdot}k^-/\tau)}. \quad (1)$$

http://arxiv.org › cs ⋮

Improved Baselines with Momentum Contrastive Learning

by X Chen · 2020 · Cited by 1088 — Abstract: Contrastive unsupervised learning has recently shown encouraging progress, e.g., in Momentum Contrast (**MoCo**) and SimCLR.

Cite as: arXiv:2003.04297

2. SimCLR

SimCLR = **Sim**ple Framework for **C**ontrastive **L**earning of Visual **R**epresentation

- contrastive **SELF-SUPERVISED** learning algorithm

Three Findings

- (1) **composition of data augmentations**
- (2) **learnable non-linear transformation** between representation & contrastive loss
- (3) contrastive learning benefits from ..
    - **larger batch sizes**
    - **more training steps**

2. SimCLR

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i,\boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i,\boldsymbol{z}_k)/\tau)}.$$

- where $\mathrm{sim}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/ \mid\mid \boldsymbol{u} \mid\mid\mid\mid \boldsymbol{v} \mid\mid$

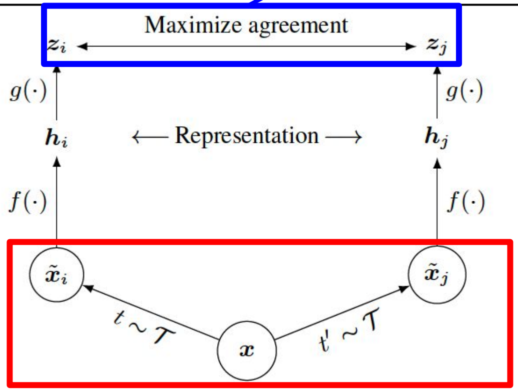


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

learns representation by …..

maximizing agreement between

differently augmented versions of **same data**,

with **contrastive loss**

2. SimCLR

4 Major Components

1. Stochastic Data Augmentation
2. Base Encoder
3. Projection head
4. Contrastive Loss Function

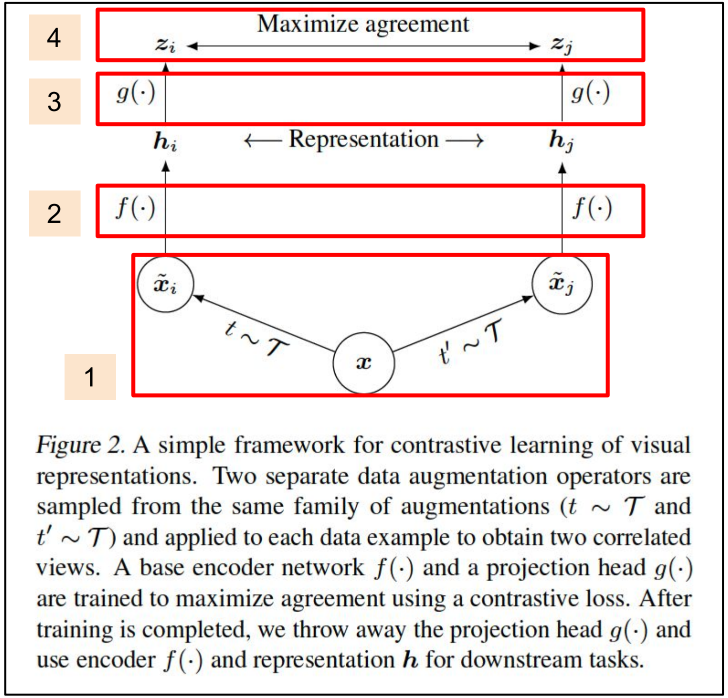


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

2. SimCLR

4 Major Components

1. **Stochastic Data Augmentation**

- positive pair $\tilde{x}_i$ & $\tilde{x}_j$
- apply 3 simple augmentations
  - (1) random cropping
  - (2) random color distortions
  - (3) random Gaussian blur

→ (1) + (2) : good performance!

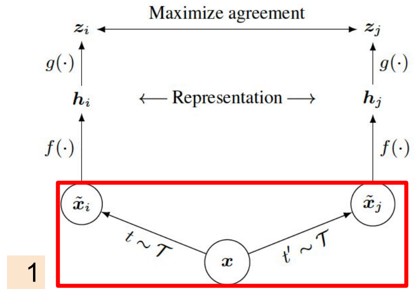


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

2. SimCLR

4 Major Components

2. **Base Encoder** $f(\cdot)$

- $\boldsymbol{h}_i = f(\tilde{\boldsymbol{x}}_i) = \mathrm{ResNet}(\tilde{\boldsymbol{x}}_i)$.
  - where $\boldsymbol{h}_i \in \mathbb{R}^d$ is the output after the GAP
  - extract representations from **augmented data samples**
- use ResNet

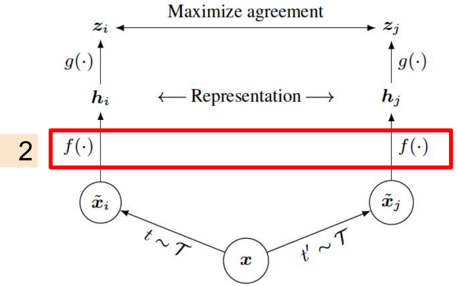


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

2. SimCLR

4 Major Components

3. Projection head $g(\cdot)$

- $\boldsymbol{z}_i = g(\boldsymbol{h}_i) = W^{(2)}\sigma(W^{(1)}\boldsymbol{h}_i).$
  - maps representations to the space **where contrastive loss is applied**
- use MLP with 1 hidden layer

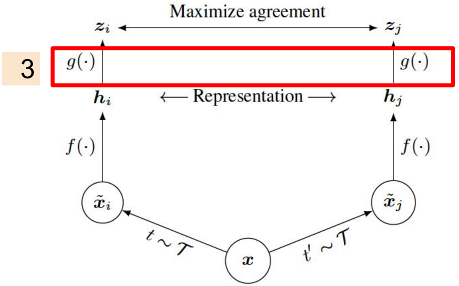


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

2. SimCLR

4 Major Components

4. Contrastive Loss Function

- Data : set $\{\tilde{\boldsymbol{x}}_k\}$ including a positive pair ($\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$)
- Task : aims to identify $\tilde{\boldsymbol{x}}_j$ in $\{\tilde{\boldsymbol{x}}_k\}_{k \neq i}$ for a given $\tilde{\boldsymbol{x}}_i$.

Sample mini batches of size N

$\rightarrow$ 2 augmentations $\rightarrow$ $2N$ data points

(no negative samples ...only positive pairs)

- Just treat $2(N-1)$ augmented samples within a mini-batch as negative examples.

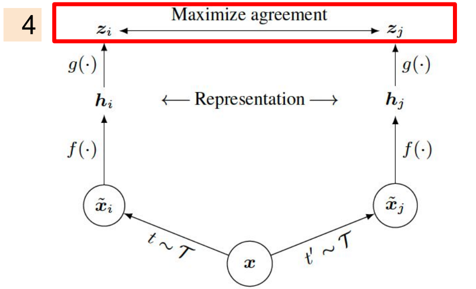


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

2. SimCLR

4 Major Components

4. Contrastive Loss Function

Loss Function for a positive pair of examples (NT-Xent)

(= Normalized Temperature scaled CE loss)

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i,\boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i,\boldsymbol{z}_k)/\tau)}.$$

- where $\mathrm{sim}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\|\|\boldsymbol{v}\|$

→ final loss : computed across all positive pairs

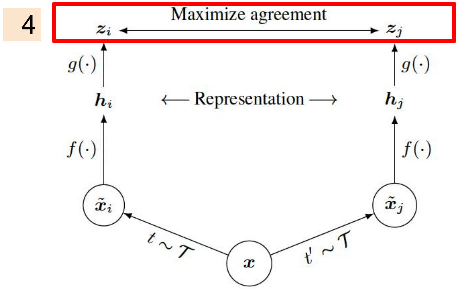
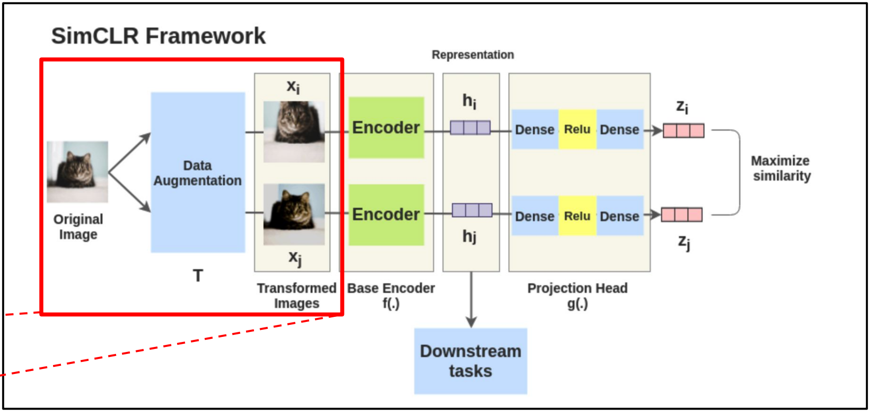
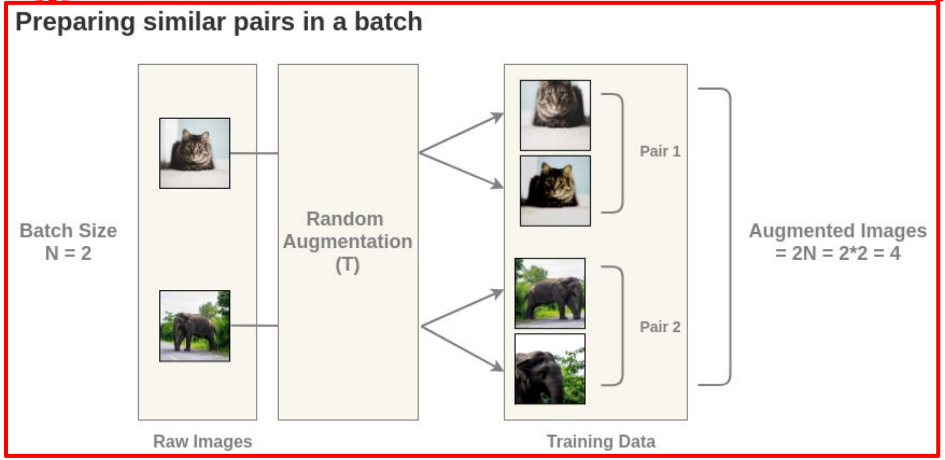


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

2. SimCLR

Example

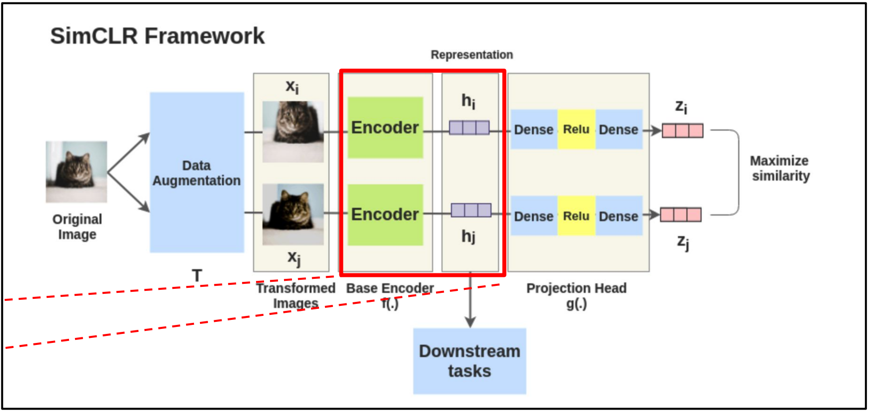


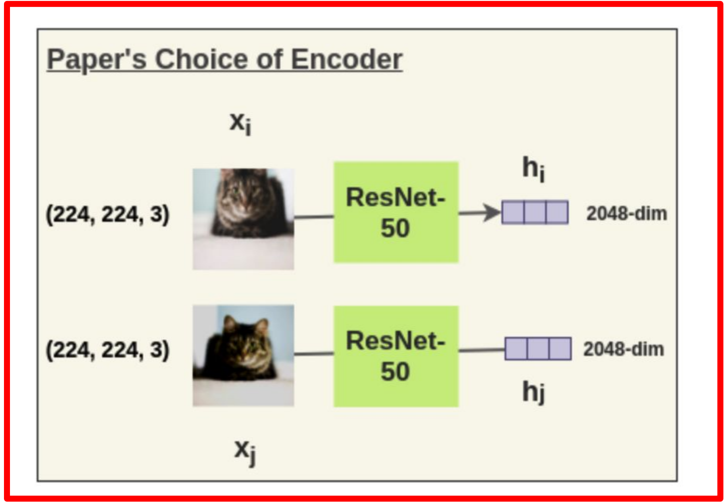


Step 1) Data Augmentation

2. SimCLR

Example

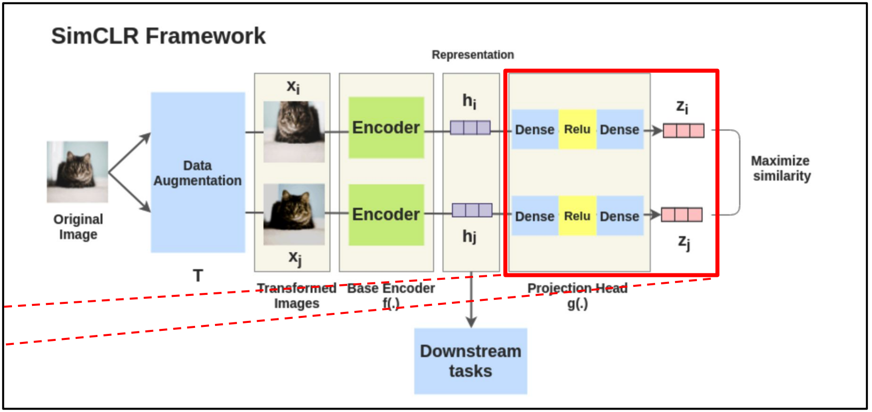


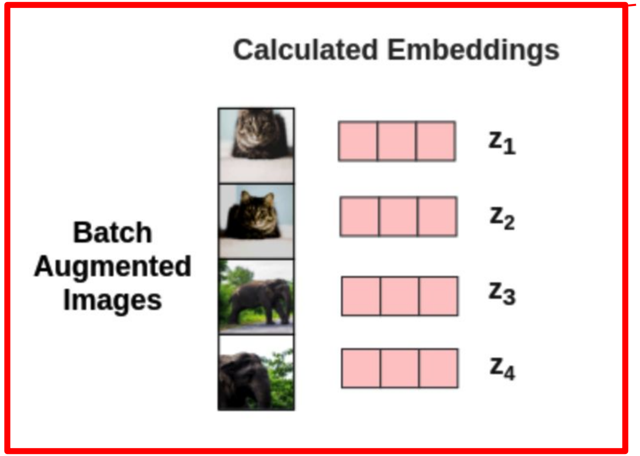


Step 2) Embedding with Base Encoder (= ResNet)

2. SimCLR

Example

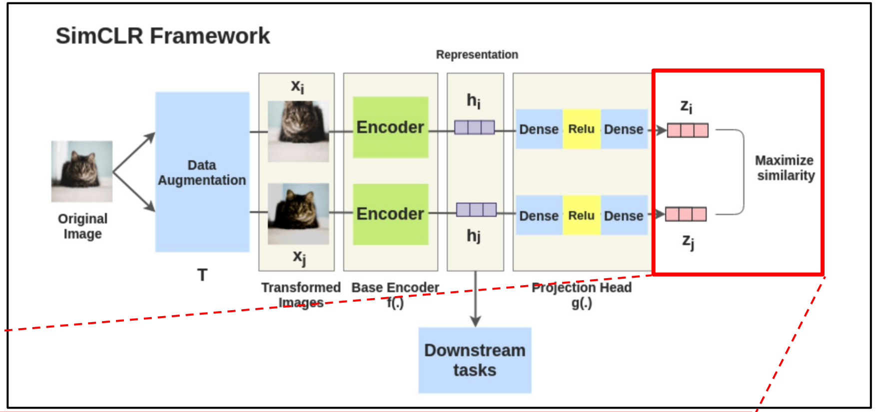




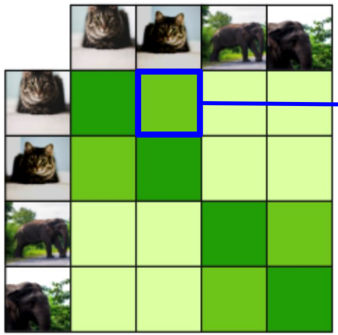
Step 3) Embedding with Projection Head

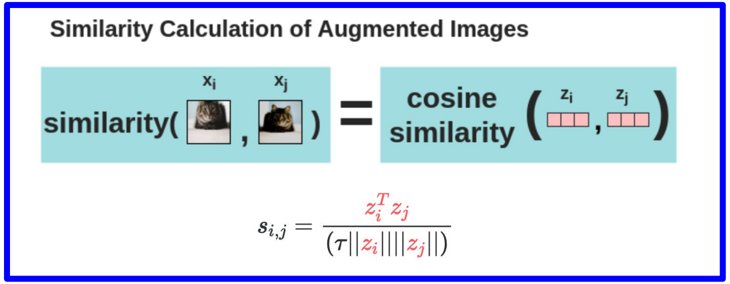
2. SimCLR

Example



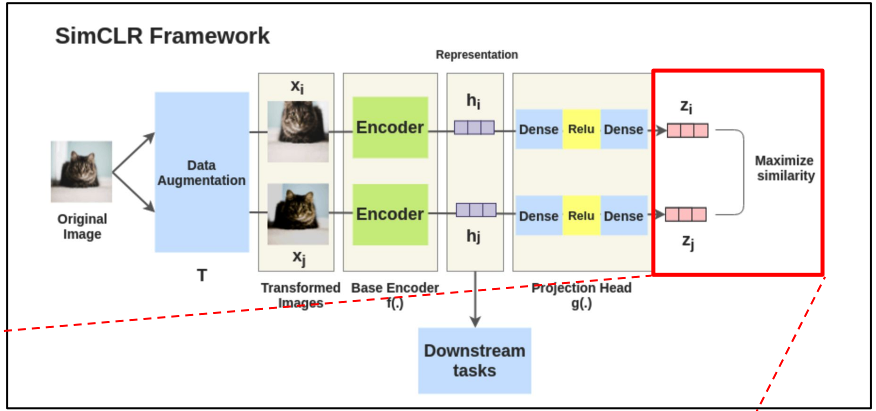
Pairwise cosine similarity



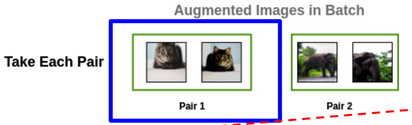


$$s_{i,j} = \frac{z_i^T z_j}{(\tau||z_i||||z_j||)}$$

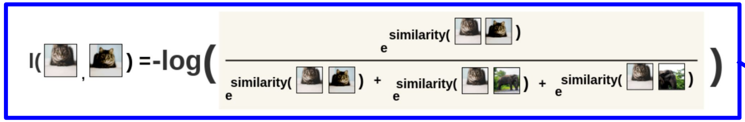
2. SimCLR

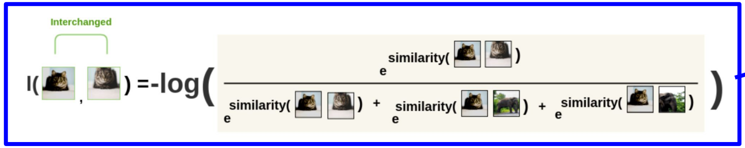


Example

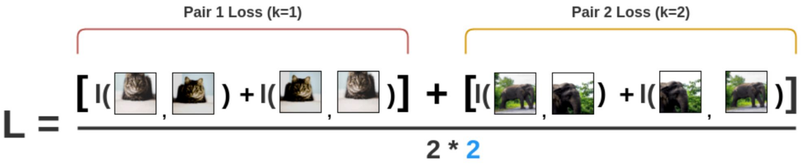


$$l(i,j) = -log\frac{exp(s_{i,j})}{\sum_{k=1}^{2N} l_{[k!=i]}exp(s_{i,k})}$$

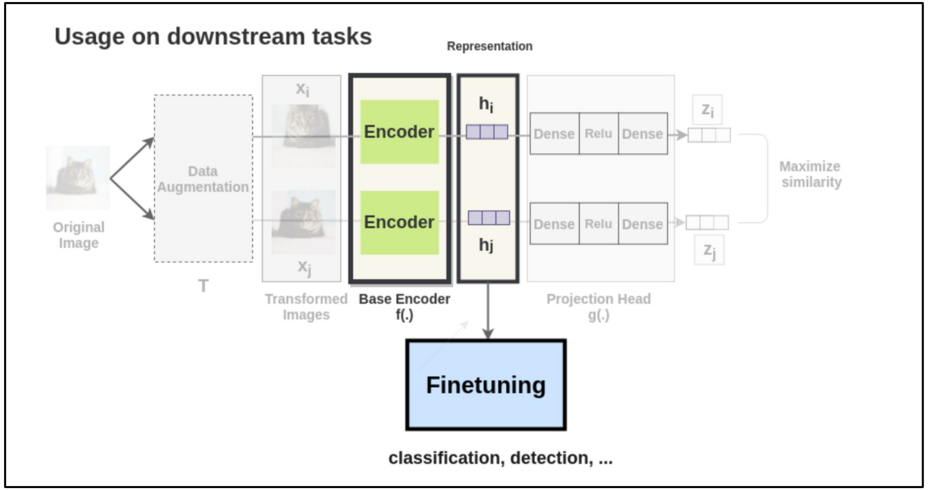




$$L = \frac{1}{2N}\sum_{k=1}^{N}[l(2k-1,2k)+l(2k,2k-1)]$$



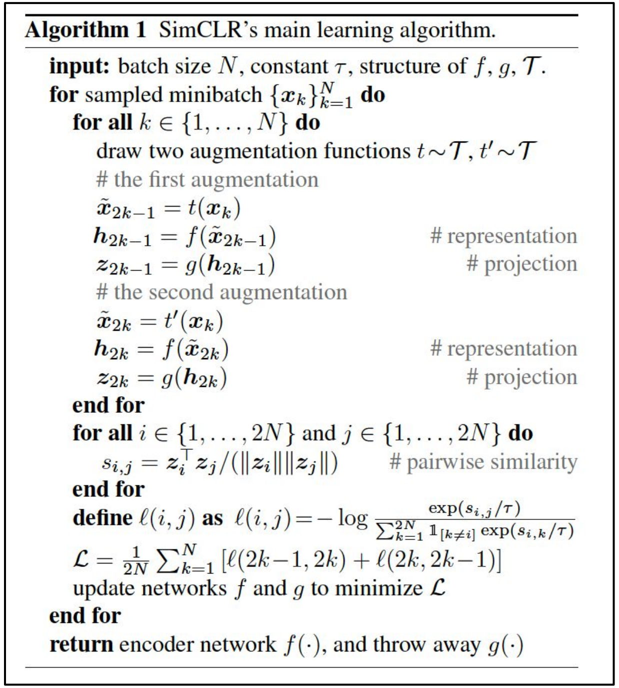
2. SimCLR



Representation obtained from **base encoder (not projection head)** can be used for other tasks!
(will be discussed in the Experiment Section)

2. SimCLR

PseudoCode



```
Algorithm 1 SimCLR's main learning algorithm.
input: batch size $N$, constant $\tau$, structure of $f$, $g$, $\mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ do
    for all $k \in \{1, \ldots, N\}$ do
        draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
        # the first augmentation
        $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
        $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$                # representation
        $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$                  # projection
        # the second augmentation
        $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
        $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$                    # representation
        $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$                      # projection
    end for
    for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ do
        $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$    # pairwise similarity
    end for
    define $\ell(i,j)$ as $\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
    $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
    update networks $f$ and $g$ to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$
```

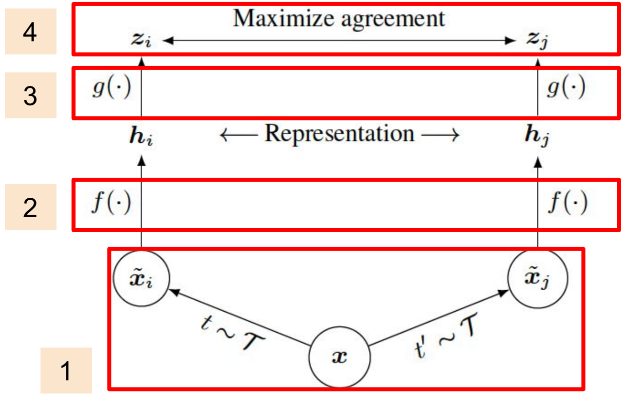
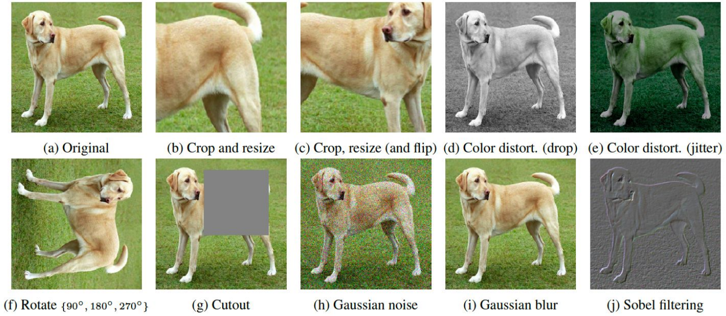


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

2. SimCLR

Experiments


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

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

Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

Data Augmentations

	Crop	Cutout	Color	Sobel	Noise	Blur	Rotate	Average
Crop	33.1	33.9	56.3	46.0	39.9	35.0	30.2	39.2
Cutout	32.2	25.6	33.9	40.0	26.5	25.2	22.4	29.4
Color	55.8	35.5	18.8	21.0	11.4	16.5	20.8	25.7
Sobel	46.2	40.6	20.9	4.0	9.3	6.2	4.2	18.8
Noise	38.8	25.8	7.5	7.6	9.8	9.8	9.6	15.5
Blur	35.1	25.2	16.6	5.8	9.7	2.6	6.7	14.5
Rotate	30.0	22.5	20.7	4.3	9.7	6.5	2.6	13.8

Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.

Data Augmentations Pairs Results

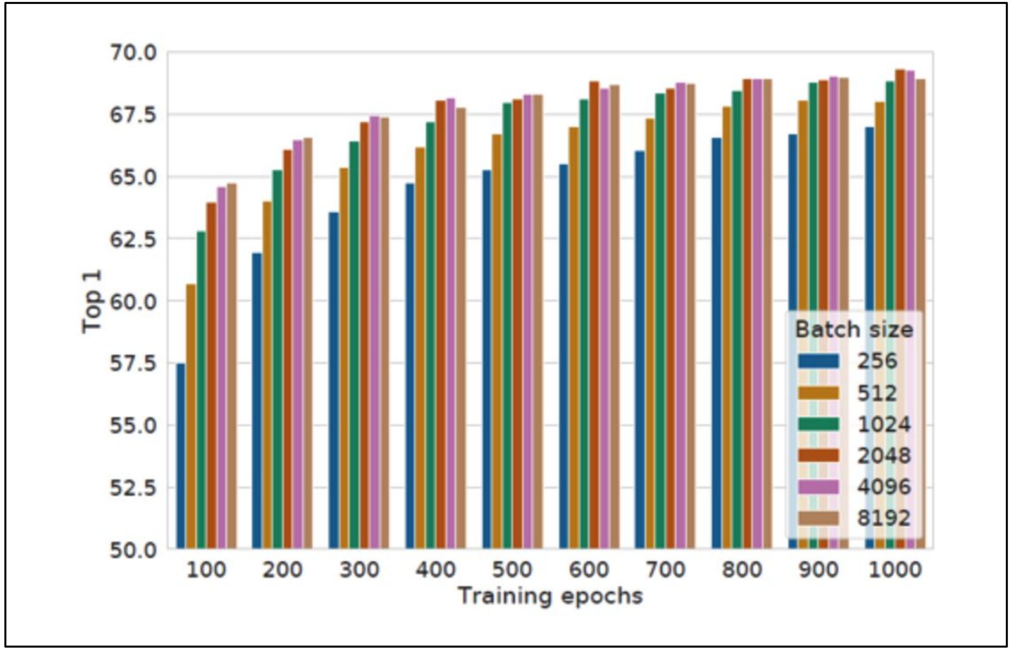
2. SimCLR

Experiments

contrastive learning benefits from ..

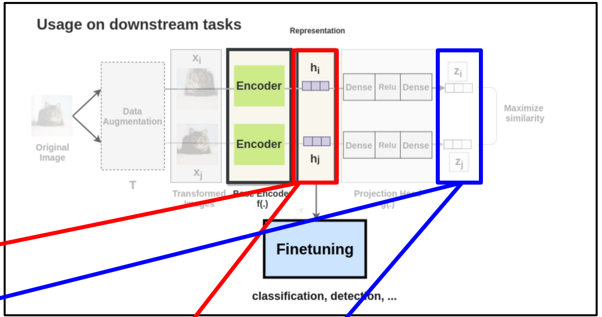
- **larger batch sizes**
- **more training steps**

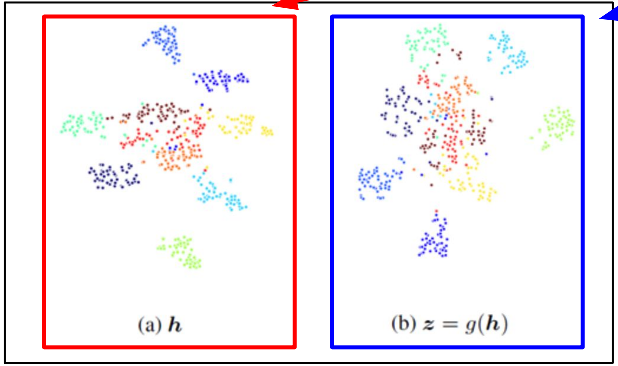
than supervised learning



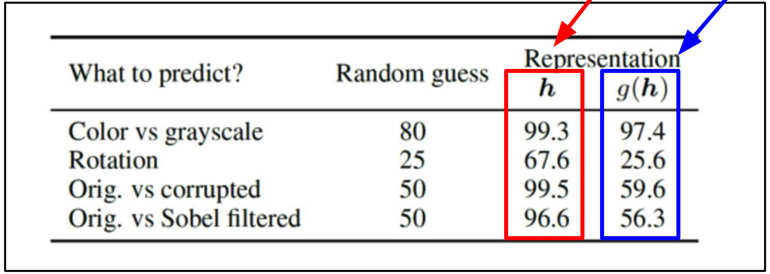
2. SimCLR

Experiments




(a) $\boldsymbol{h}$

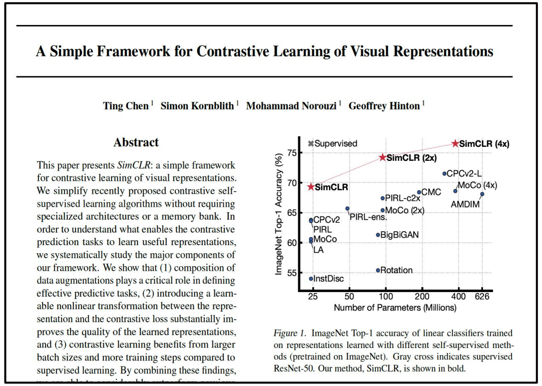
(b) $\boldsymbol{z} = g(\boldsymbol{h})$



What to predict?	Random guess	Representation	
		$\boldsymbol{h}$	$g(\boldsymbol{h})$
Color vs grayscale	80	99.3	97.4
Rotation	25	67.6	25.6
Orig. vs corrupted	50	99.5	59.6
Orig. vs Sobel filtered	50	96.6	56.3

Representation obtained from **base encoder (not projection head)** can be used for other tasks!

SimCLR (2020)



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,

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.

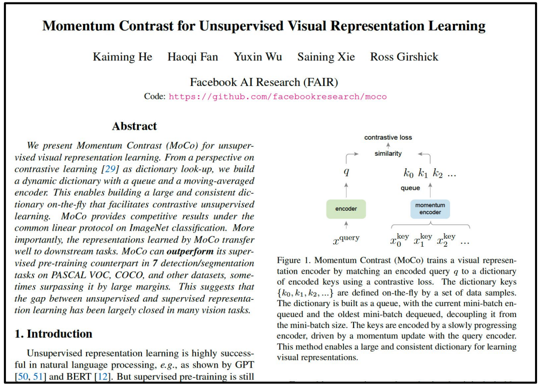
https://arxiv.org › cs ⋮

A Simple Framework for Contrastive Learning of Visual ... - arXiv

by T Chen · 2020 · Cited by 5112 — Abstract: This paper presents **SimCLR**: a simple framework for contrastive learning of visual representations. We simplify recently proposed ...

Cite as: arXiv:2002.05709

MoCo v1 (2019)



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

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 $\{k_0, k_1, k_2, ...\}$ 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.

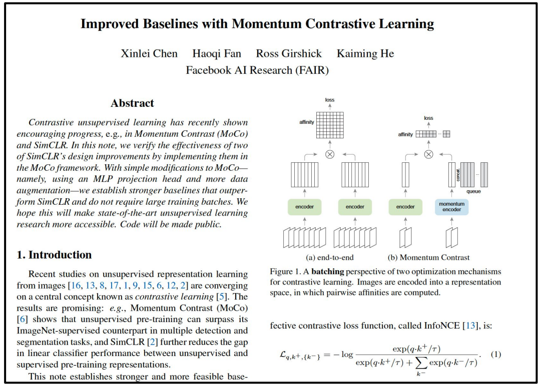
https://arxiv.org › cs ⋮

Momentum Contrast for Unsupervised Visual Representation ...

by K He · 2019 · Cited by 3735 — Abstract: We present Momentum Contrast (**MoCo**) for unsupervised visual representation learning. From a perspective on contrastive learning as ...

Cite as: arXiv:1911.05722

MoCo v2 (2020)



Improved Baselines with Momentum Contrastive Learning

Xinlei Chen Haoqi Fan Ross Girshick Kaiming He
Facebook AI Research (FAIR)

Abstract

Contrastive unsupervised learning has recently shown encouraging progress, e.g., *in Momentum Contrast (MoCo) and SimCLR. In this note, we verify the effectiveness of two of SimCLR's design improvements by implementing them in the MoCo framework. With simple modifications to MoCo—namely, using an MLP projection head and more data augmentation—we establish stronger baselines that outperform SimCLR and do not require large training batches. We hope this will make state-of-the-art unsupervised learning research more accessible. Code will be made public.*

1. Introduction

Recent studies on unsupervised representation learning from images [16, 13, 8, 17, 1, 9, 15, 6, 12, 2] are converging on a central concept known as *contrastive learning* [5]. The results are promising: *e.g.*, Momentum Contrast (MoCo) [6] shows that unsupervised pre-training can surpass its ImageNet-supervised counterpart in multiple detection and segmentation tasks, and SimCLR [2] further reduces the gap in linear classifier performance between unsupervised and supervised pre-training representations.

This note establishes stronger and more feasible base-

(a) end-to-end (b) Momentum Contrast

Figure 1. A **batching** perspective of two optimization mechanisms for contrastive learning. Images are encoded into a representation space, in which pairwise affinities are computed.

fective contrastive loss function, called InfoNCE [13], is:

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q{\cdot}k^+/\tau)}{\exp(q{\cdot}k^+/\tau) + \sum_{k^-} \exp(q{\cdot}k^-/\tau)}. \quad (1)$$

http://arxiv.org › cs ⋮

Improved Baselines with Momentum Contrastive Learning

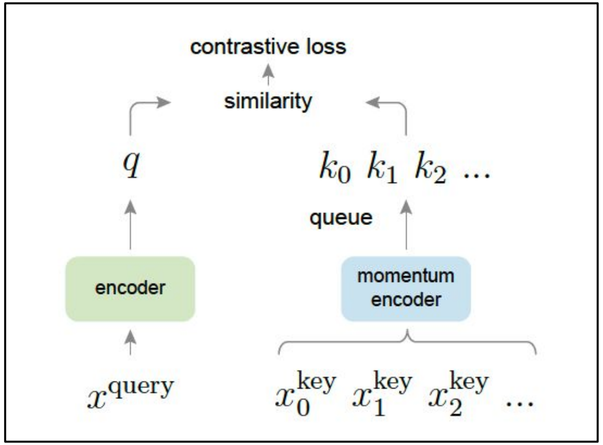
by X Chen · 2020 · Cited by 1088 — Abstract: Contrastive unsupervised learning has recently shown encouraging progress, e.g., in Momentum Contrast (**MoCo**) and SimCLR.

Cite as: arXiv:2003.04297

3. MoCo v1

MoCo = **Mo**mentum **Co**ntrast

- **UN-SUPERVISED** visual representation learning algorithm
- (1) **Dictionary** look-up perspective
    - build a **Dynamic dictionary** (with a queue (FIFO))
- (2) **Moving Average Encoder**

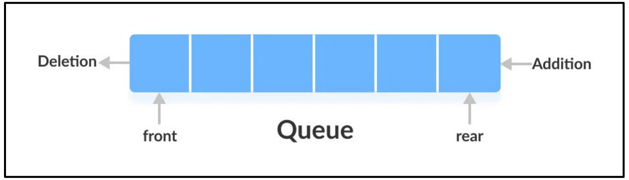


3. MoCo v1

(1) Dynamic Dictionary

- (SimCLR) positive pair & negative pair in ONE BATCH
- (MoCo) define a Dictionary
  - **match O** with query -> **positive** key
  - **match X** with query -> **negative** key

Use “**Dynamic**” Dictionary, by using **queue (FIFO)**



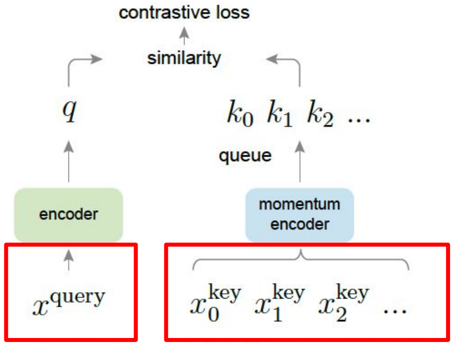


Figure 1. Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss. The dictionary keys $\{k_0, k_1, k_2, \ldots\}$ are defined on-the-fly by a set of data samples. The dictionary is built as a queue, with 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.

3. MoCo v1

(1) Dynamic Dictionary

Notation

- encoded query : q
- keys of a dictionary :
  - set of encoded samples : $\{k_0, k_1, k_2, \cdots\}$
  - positive key : k_+
  - negative key : k_-

Contrastive Loss : low , when...

- q is similar to its positive key k_+
- dissimilar to other key (= negative keys)

(similarity : measured by **dot product**)

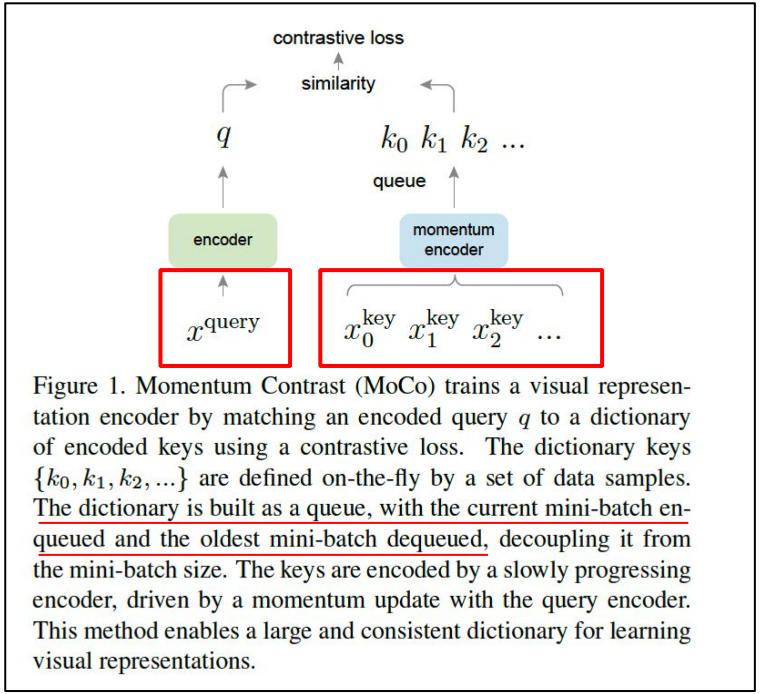


Figure 1. Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss. The dictionary keys $\{k_0, k_1, k_2, ...\}$ are defined on-the-fly by a set of data samples. The dictionary is built as a queue, with the 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.

3. MoCo v1

(2) Loss Function : **InfoNCE**

$$\mathcal{L}_q = -\log \frac{\exp(q{\cdot}k_+/\tau)}{\sum_{i=0}^{K}\exp(q{\cdot}k_i/\tau)}.$$

- τ : temperature

→ sum over **one** positive & **K** negatives

(= log loss of $(K+1)$ way softmax classifier , that tries to **classify** q **as** k_+)

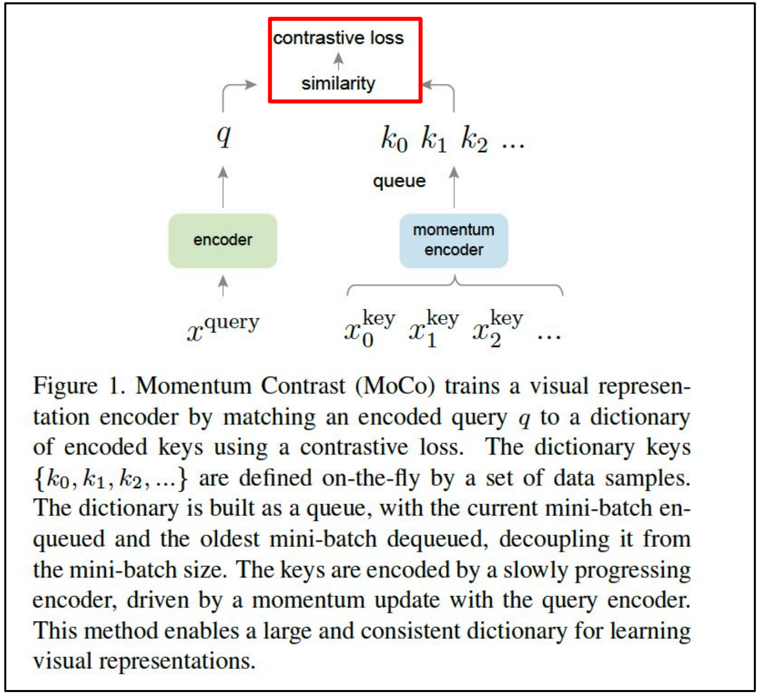


Figure 1. Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss. The dictionary keys $\{k_0, k_1, k_2, ...\}$ are defined on-the-fly by a set of data samples. The dictionary is built as a queue, with the 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.

Model Notation

- query : $q = f_{\mathrm{q}}(x^q)$
    - f_q : encoder network
    - x^q : query
- key : $k = f_{\mathrm{k}}(x^k)$
    - f_k : encoder network
    - x^k : key

3. MoCo v1

(3) Moving Average Encoder

- **slowly progressing** key encoder

  = **momentum-based MA of query encoder**
- to maintain “**consistency**”

$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}.$$

- $m \in [0,1)$: momentum coefficient
- Only the parameters θ_{q} are updated!!

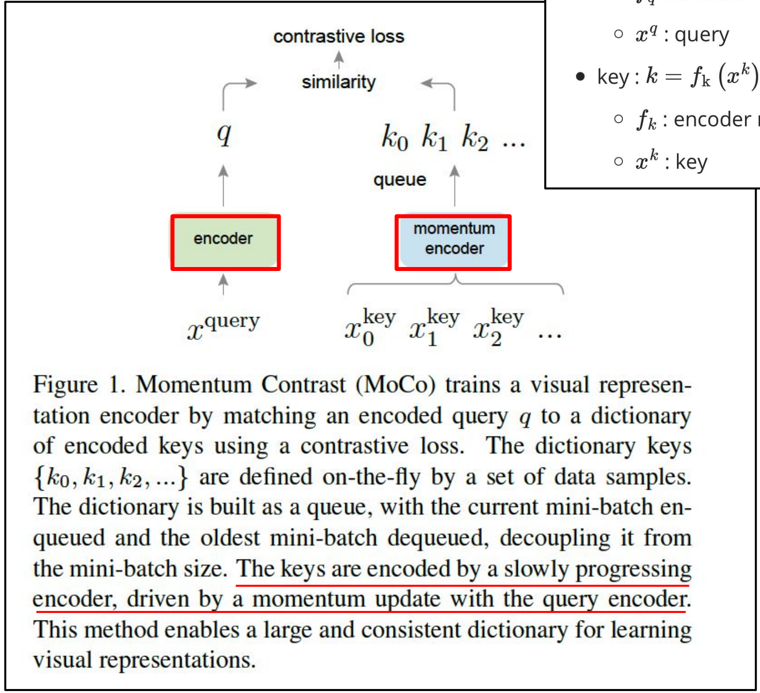
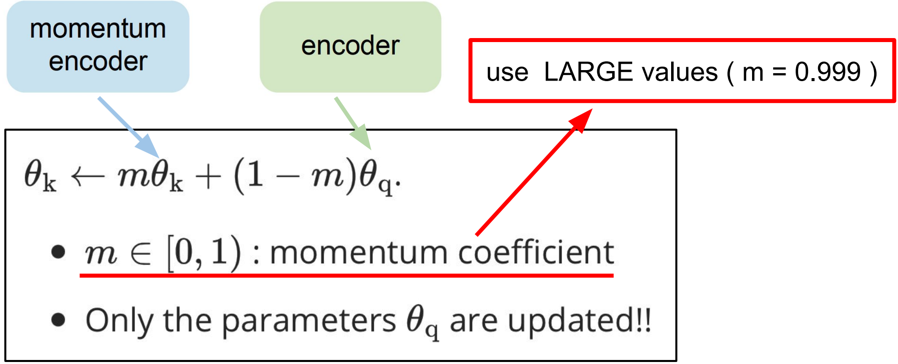


Figure 1. Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss. The dictionary keys $\{k_0, k_1, k_2, ...\}$ are defined on-the-fly by a set of data samples. The dictionary is built as a queue, with the 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.

3. MoCo v1

(3) Moving Average Encoder

- **slowly progressing** key encoder
  = **momentum-based MA of query encoder**
- to maintain “**consistency**”



momentum encoder

encoder

use LARGE values (m = 0.999)

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q.$$

- $m \in [0,1)$: momentum coefficient
- Only the parameters θ_q are updated!!

Model Notation

- query : $q = f_q(x^q)$
  - f_q : encoder network
  - x^q : query
- key : $k = f_k(x^k)$
  - f_k : encoder network
  - x^k : key

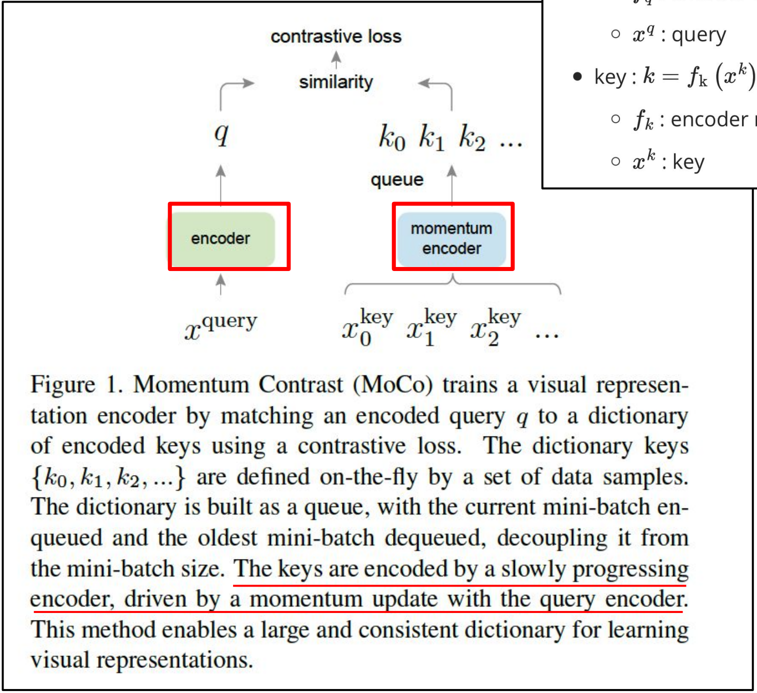


Figure 1. Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss. The dictionary keys $\{k_0, k_1, k_2, ...\}$ are defined on-the-fly by a set of data samples. The dictionary is built as a queue, with the 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.

3. MoCo v1

(3) Moving Average Encoder

- **slowly progressing** key encoder

  = **momentum-based MA of query encoder**

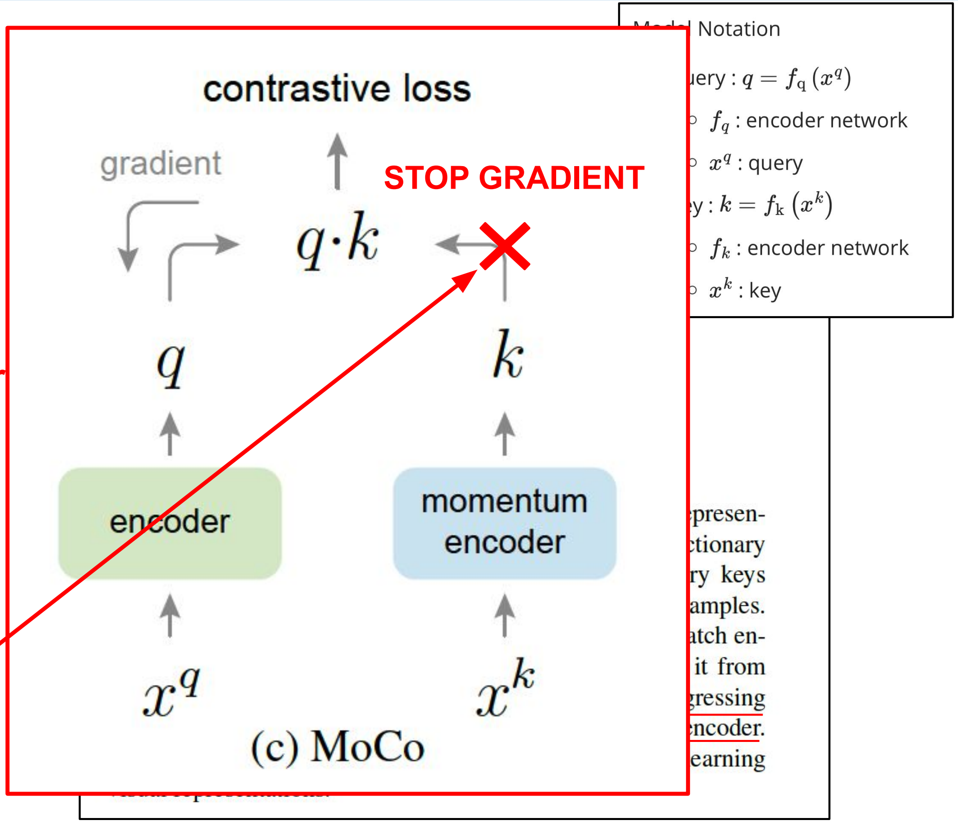
- to maintain “**consistency**”

momentum encoder

encoder

$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}.$$

- $m \in [0,1)$: momentum coefficient
- Only the parameters θ_{q} are updated!!



(c) MoCo

Notation

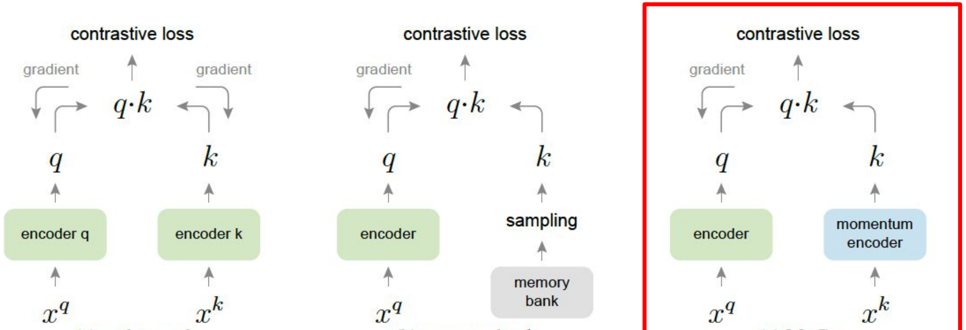
uery : $q = f_q(x^q)$
  - f_q : encoder network
  - x^q : query

y : $k = f_k(x^k)$
  - f_k : encoder network
  - x^k : key

epresen-
ctionary
ry keys
amples.
atch en-
it from
gressing
ncoder.
earning

3. MoCo v1

Relations to Previous Mechanisms (end-to-end & memory bank)

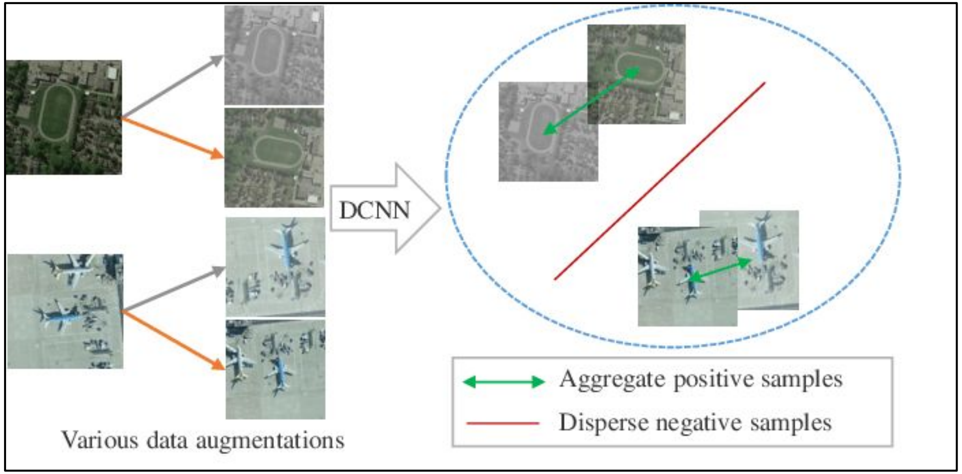


(a) end-to-end (b) memory bank (c) MoCo

Figure 2. **Conceptual comparison of three contrastive loss mechanisms** (empirical comparisons are in Figure 3 and Table 3). Here we illustrate one pair of query and key. The three mechanisms differ in how the keys are maintained and how the key encoder is updated. **(a)**: The encoders for computing the query and key representations are updated ***end-to-end*** by back-propagation (the two encoders can be different). **(b)**: The key representations are sampled from a ***memory bank*** [61]. **(c)**: ***MoCo*** encodes the new keys on-the-fly by a momentum-updated encoder, and maintains a queue (not illustrated in this figure) of keys.

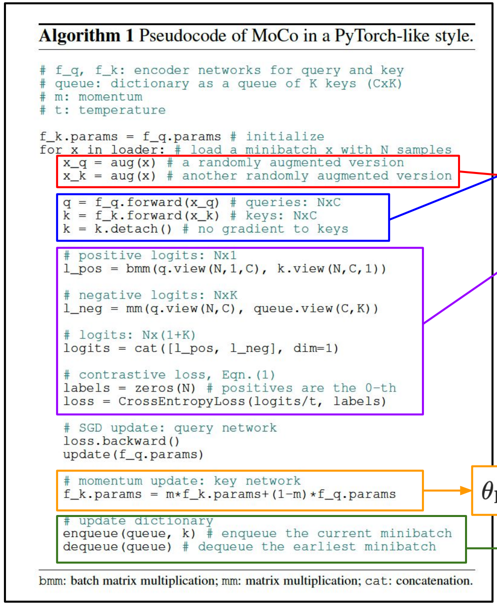
3. MoCo v1

Pre-text Task : Instance Discrimination



3. MoCo v1

Pseudo-code



Algorithm 1 Pseudocode of MoCo in a PyTorch-like style.

```
# f_q, f_k: encoder networks for query and key
# queue: dictionary as a queue of K keys (CxK)
# m: momentum
# t: temperature

f_k.params = f_q.params # initialize
for x in loader: # load a minibatch x with N samples
   x_q = aug(x) # a randomly augmented version
   x_k = aug(x) # another randomly augmented version

   q = f_q.forward(x_q) # queries: NxC
   k = f_k.forward(x_k) # keys: NxC
   k = k.detach() # no gradient to keys

   # positive logits: Nx1
   l_pos = bmm(q.view(N,1,C), k.view(N,C,1))

   # negative logits: NxK
   l_neg = mm(q.view(N,C), queue.view(C,K))

   # logits: Nx(1+K)
   logits = cat([l_pos, l_neg], dim=1)

   # contrastive loss, Eqn.(1)
   labels = zeros(N) # positives are the 0-th
   loss = CrossEntropyLoss(logits/t, labels)

   # SGD update: query network
   loss.backward()
   update(f_q.params)

   # momentum update: key network
   f_k.params = m*f_k.params+(1-m)*f_q.params

   # update dictionary
   enqueue(queue, k) # enqueue the current minibatch
   dequeue(queue) # dequeue the earliest minibatch
```

bmm: batch matrix multiplication; mm: matrix multiplication; cat: concatenation.

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q$$

Dynamic Dictionary

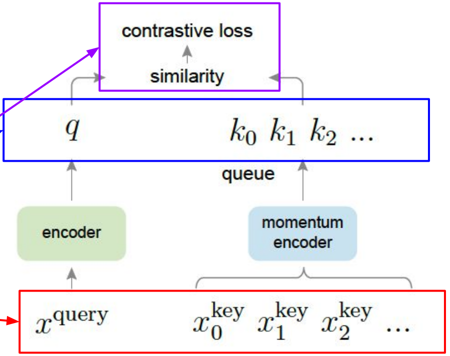
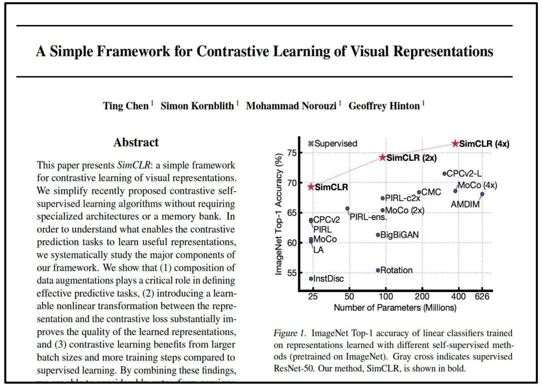


Figure 1. Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss. The dictionary keys $\{k_0, k_1, k_2, ...\}$ are defined on-the-fly by a set of data samples. The dictionary is built as a queue, with the 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.

SimCLR (2020)



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,

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.

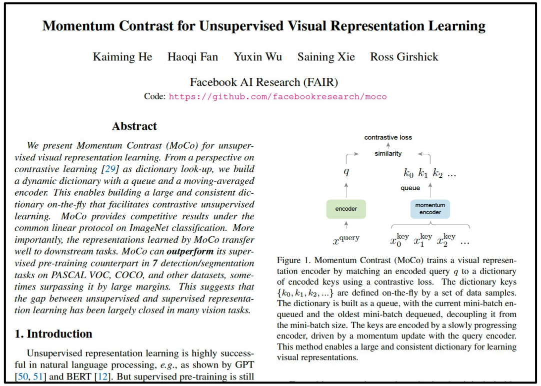
https://arxiv.org › cs ⋮

A Simple Framework for Contrastive Learning of Visual ... - arXiv

by T Chen · 2020 · Cited by 5112 — Abstract: This paper presents **SimCLR**: a simple framework for contrastive learning of visual representations. We simplify recently proposed ...

Cite as: arXiv:2002.05709

MoCo v1 (2019)



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

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 $\{k_0, k_1, k_2, ...\}$ 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.

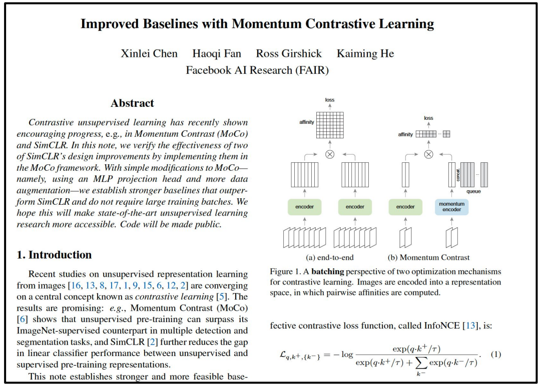
https://arxiv.org › cs ⋮

Momentum Contrast for Unsupervised Visual Representation ...

by K He · 2019 · Cited by 3735 — Abstract: We present Momentum Contrast (**MoCo**) for unsupervised visual representation learning. From a perspective on contrastive learning as ...

Cite as: arXiv:1911.05722

MoCo v2 (2020)



Improved Baselines with Momentum Contrastive Learning

Xinlei Chen Haoqi Fan Ross Girshick Kaiming He
Facebook AI Research (FAIR)

Abstract

Contrastive unsupervised learning has recently shown encouraging progress, e.g., *in Momentum Contrast (MoCo) and SimCLR. In this note, we verify the effectiveness of two of SimCLR's design improvements by implementing them in the MoCo framework. With simple modifications to MoCo—namely, using an MLP projection head and more data augmentation—we establish stronger baselines that outperform SimCLR and do not require large training batches. We hope this will make state-of-the-art unsupervised learning research more accessible. Code will be made public.*

1. Introduction

Recent studies on unsupervised representation learning from images [16, 13, 8, 17, 1, 9, 15, 6, 12, 2] are converging on a central concept known as *contrastive learning* [5]. The results are promising: *e.g.*, Momentum Contrast (MoCo) [6] shows that unsupervised pre-training can surpass its ImageNet-supervised counterpart in multiple detection and segmentation tasks, and SimCLR [2] further reduces the gap in linear classifier performance between unsupervised and supervised pre-training representations.

This note establishes stronger and more feasible base-

(a) end-to-end (b) Momentum Contrast

Figure 1. A **batching** perspective of two optimization mechanisms for contrastive learning. Images are encoded into a representation space, in which pairwise affinities are computed.

fective contrastive loss function, called InfoNCE [13], is:

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q{\cdot}k^+/\tau)}{\exp(q{\cdot}k^+/\tau) + \sum\limits_{k^-} \exp(q{\cdot}k^-/\tau)}. \quad (1)$$

http://arxiv.org › cs ⋮

Improved Baselines with Momentum Contrastive Learning

by X Chen · 2020 · Cited by 1088 — Abstract: Contrastive unsupervised learning has recently shown encouraging progress, e.g., in Momentum Contrast (**MoCo**) and SimCLR.

Cite as: arXiv:2003.04297

4. MoCo v2

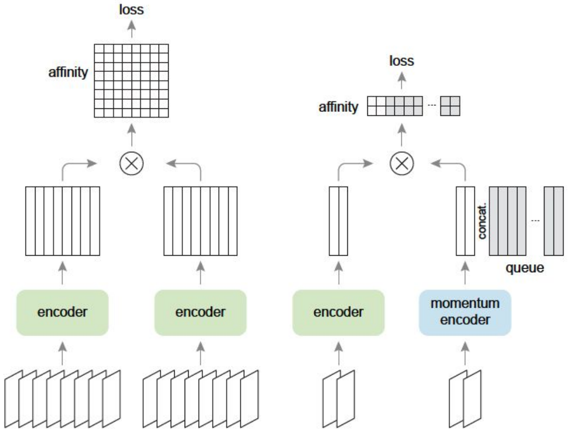
(1) **MoCo v1** :

- (a) Dynamic Dictionary
- (b) Moving-Averaged Encoder

(2) **SimCLR**

- (a) larger batch size for lots of negative samples
- **(b) stronger augmentation**
- **(c) MLP Projection head**

MoCo v2 = **MoCo v1** + **SimCLR ((b) + (c))**



(a) end-to-end (b) Momentum Contrast

Figure 1. A **batching** perspective of two optimization mechanisms for contrastive learning. Images are encoded into a representation space, in which pairwise affinities are computed.

5. Reference

- Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks (Dosoviskiy, NIPS 2014)
- Unsupervised Visual Representation Learning by Context Prediction (Doerscch et al., ICCV 2015)
- Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles (Noroozi and Favaro, ECCV 2016)
- Unsupervised Representation Learning by Predicting Image Rotations (Gidaris et al., ICLR 2018)
- A Simple Framework for Contrastive Learning of Visual Representations (Chen et al., ICML 2020)
- Momentum Contrast for Unsupervised Visual Representation Learning (He et al., CVPR 2020)
- Improved Baselines with Momentum Contrastive Learning (Chen et al., arxiv 2020)
- https://lilianweng.github.io/posts/2021-05-31-contrastive/
- https://seunghan96.github.io/categories/cl/

Thank You