Contrastive learning

Background

Presenter: Changyu Deng
February 2020

Motivation

Scope: self-supervised representation learning

- No labels
- To extract features (representations)

Why?

Labeled data

Unlabeled data

https://www.marketwatch.com/story/gold-heads-for-back-to-back-loss-amid-vaccine-hope-us-dollar-strength-2020-09-03
https://serbu.ca/different-types-of-rocks-stones-and-aggregates/

Motivation

How? Let us classify the following images into 2 categories

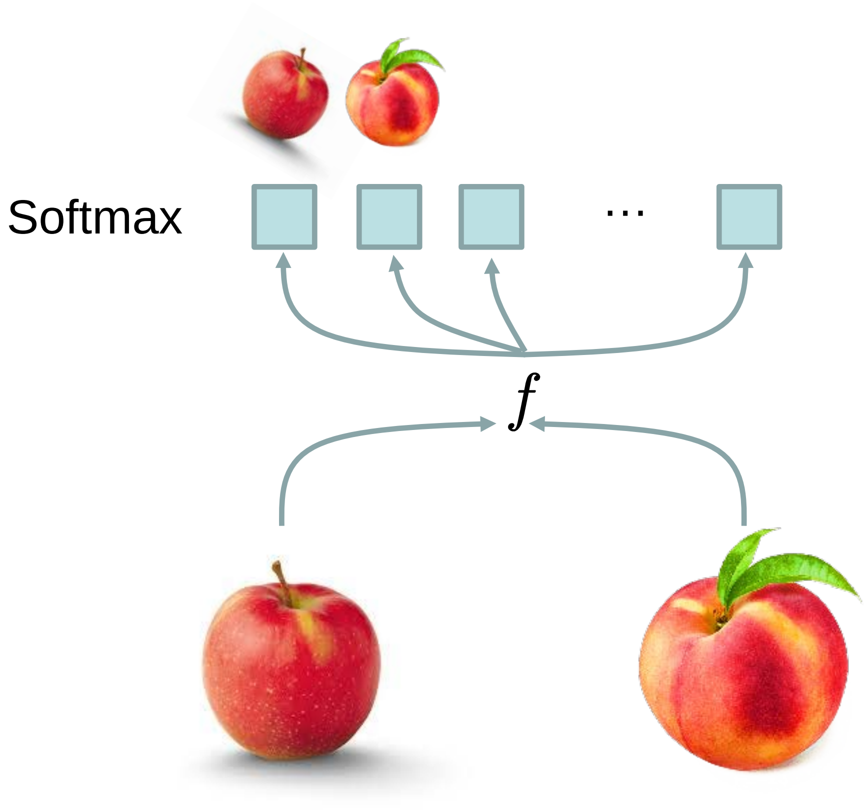
   

- Every instance image has its own feature/representation
- Can we learn representation by discriminating instances?

https://billsberryfarm.com/produce/peaches/
https://www.applesfromny.com/varieties/jonagold/
https://www.heb.com/static-page/apple-varieties
https://www.freshpoint.com/archive/commodity-stone-fruits-peach-yellow/

Motivation

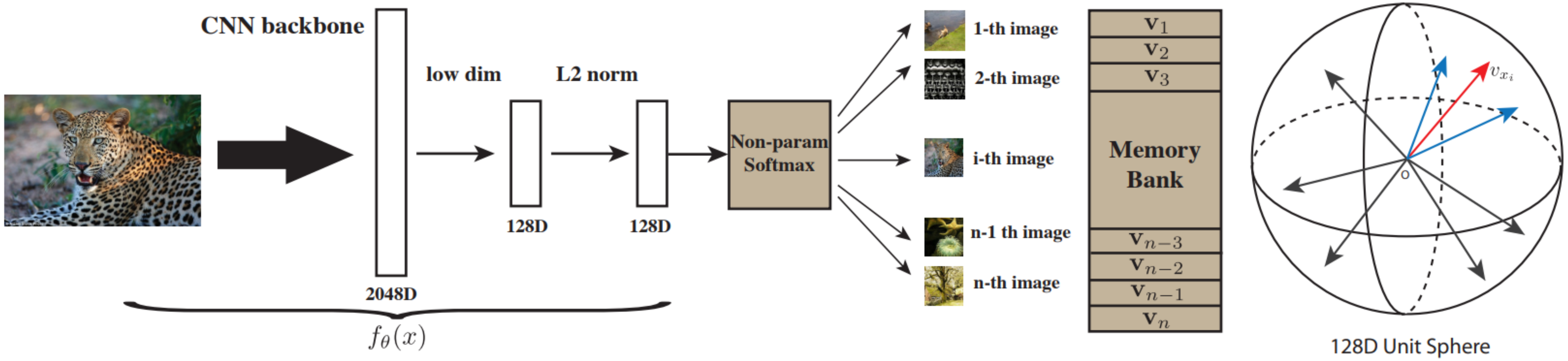
A naïve idea: train a classifier to classify N images into N categories.



Impractical, too many categories.

Memory bank

Non-Parametric Instance Discrimination
We reduce the dimension of classifications



We want to learn $\mathbf{v} = f_{\boldsymbol{\theta}}(x)$ subject to $\|\mathbf{v}\| = 1$

Non-parametric softmax $P(i|\mathbf{v}) = \dfrac{\exp\left(\mathbf{v}_i^T \mathbf{v}/\tau\right)}{\sum_{j=1}^{n} \exp\left(\mathbf{v}_j^T \mathbf{v}/\tau\right)}$

Loss $J(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log P(i|f_{\boldsymbol{\theta}}(x_i))$

Wu, Zhirong, et al. "Unsupervised feature learning via non-parametric instance discrimination." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Motivation

Can we incorporate more semantic info? Let us classify the following images into 2 categories

What are the attributes to distinguish between these two types of objects

- Shape? Yes
- Texture? Yes
- Color? Maybe
- Orientation? No
- Image size? No
- Location? No

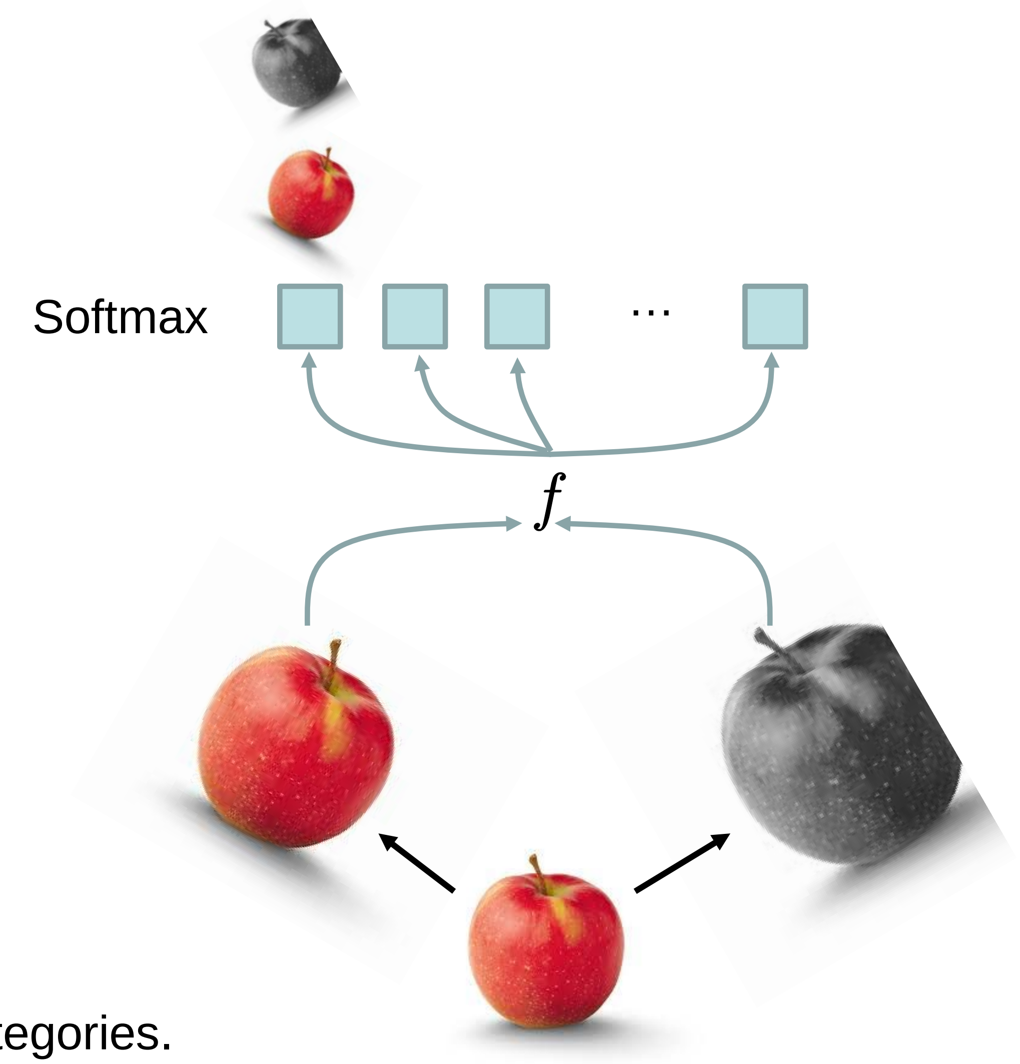
The *views* are generated from 2 *images* by **data augmentation**

- Know the "category" of views without label
- Train the network to discard unneeded attributes

https://billsberryfarm.com/produce/peaches/ https://www.applesfromny.com/varieties/jonagold/

Motivation

Another naïve idea: train a classifier to classify N images (unlimited views) into N categories.

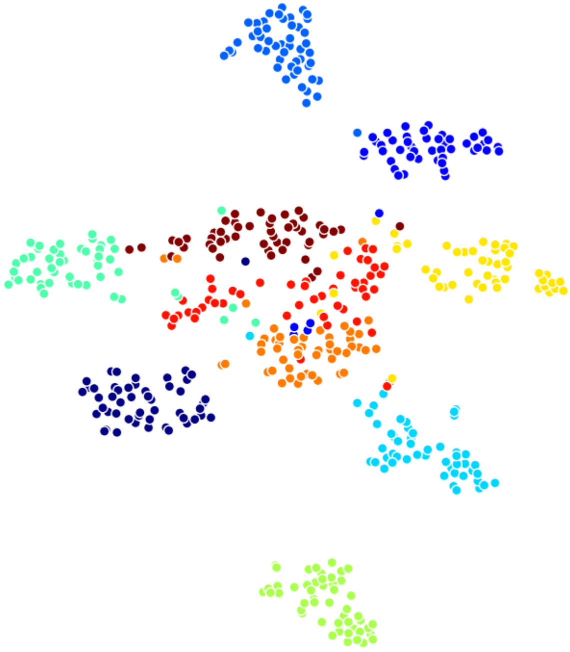


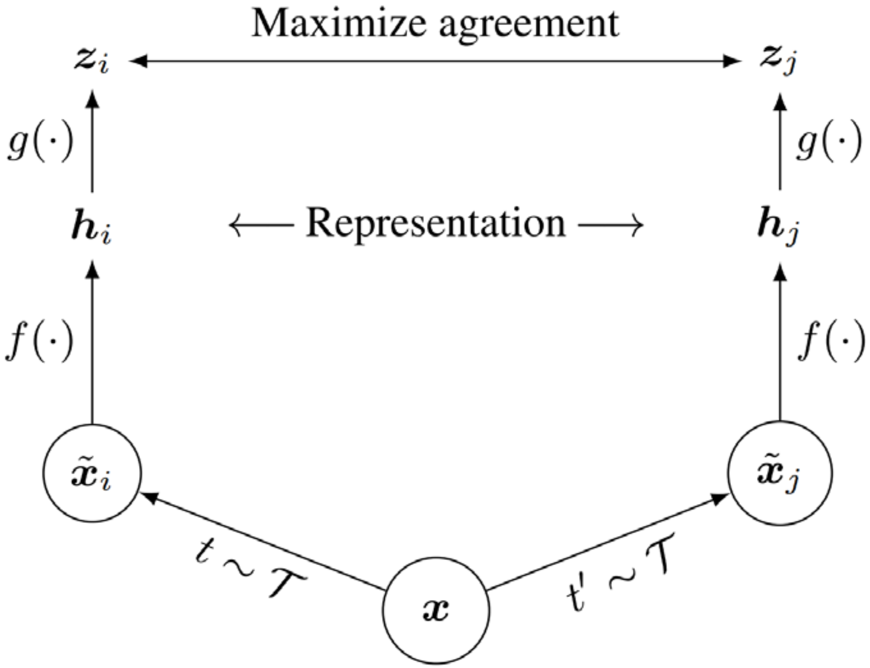
Impractical, too many categories.

SimCLR

SimCLR

- Use data augmentation to generate views
- Measure the distance between representations
- Cluster representations of views





Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

SimCLR

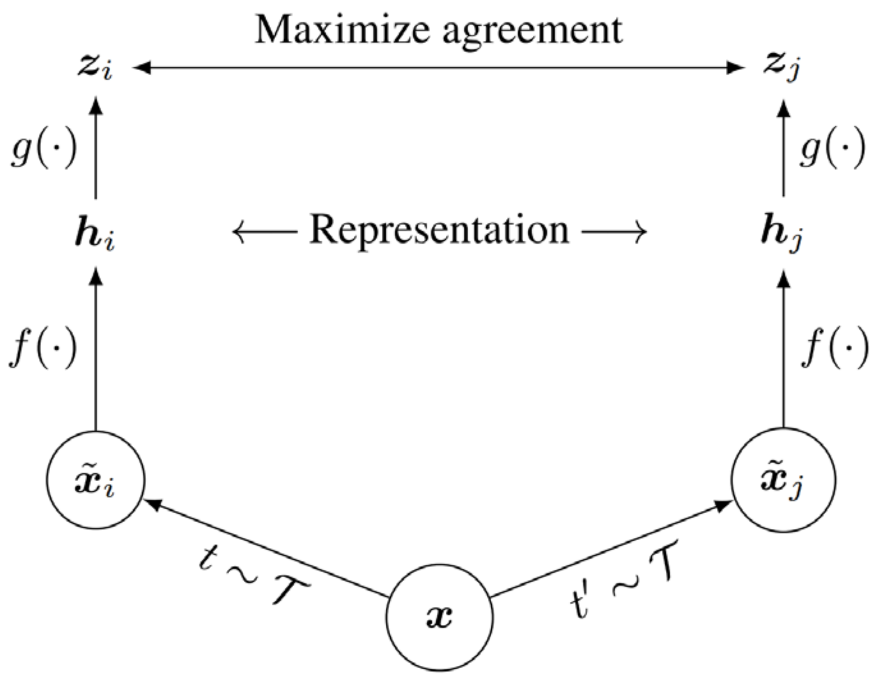
First define distance by inner product

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

Then define loss function by distance

For a positive pair of views i,j (from the same image)

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

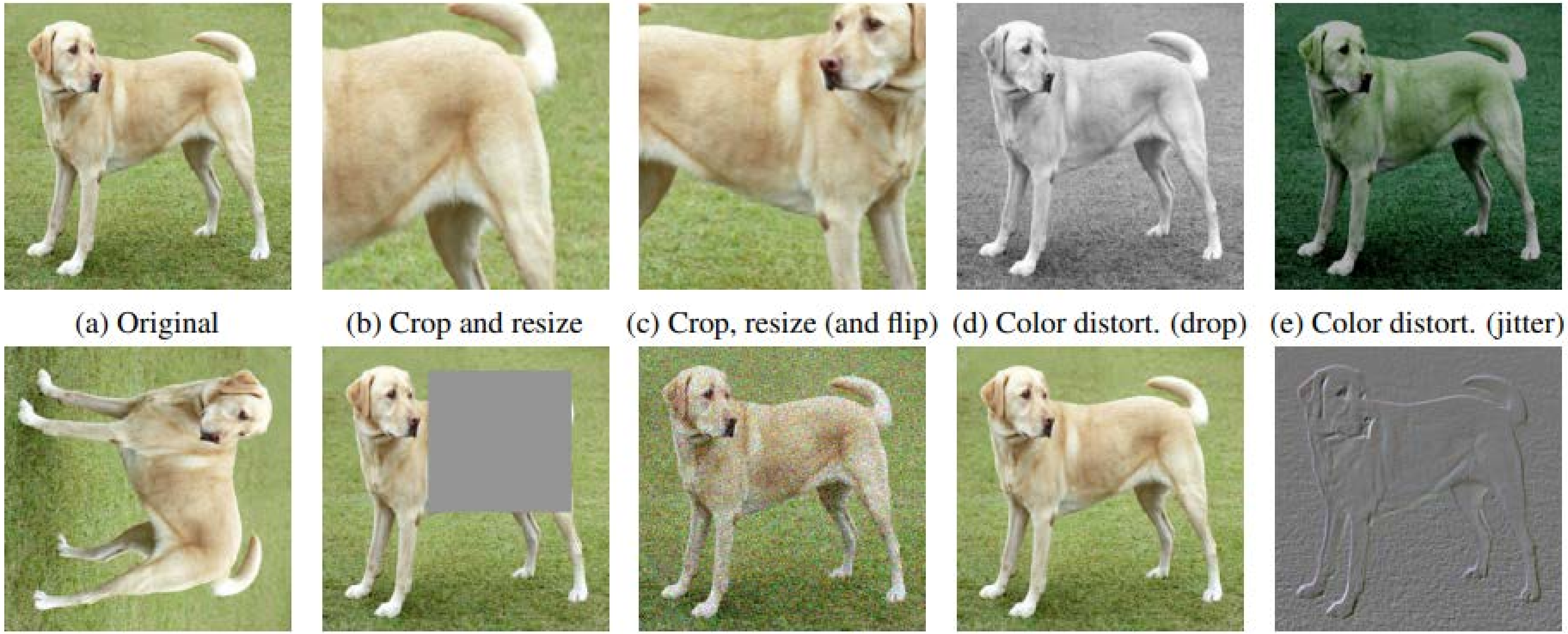


To make this work, we need
- Large batch size (256-8192 in the paper)
- Various data augmentation techniques

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

SimCLR

Data augmentation used in SimCLR



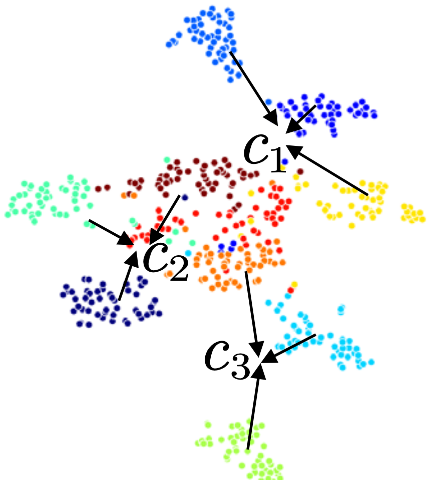
(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

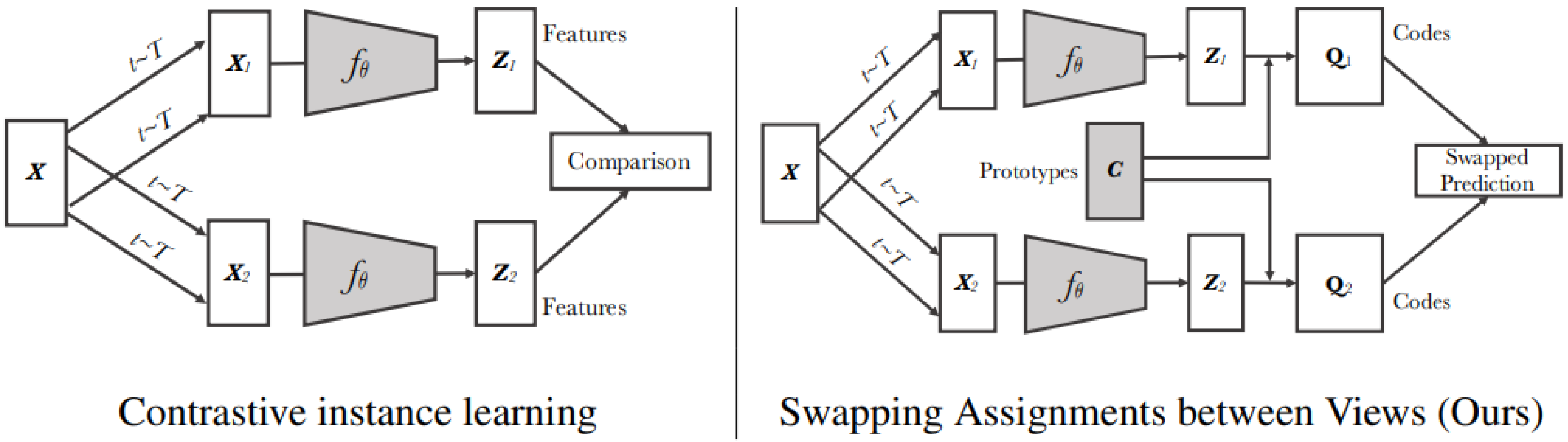
(f) Rotate $\{90^\circ, 180^\circ, 270^\circ\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

SwAV

SwAV: use given number of clusters/categories





Loss

$$-\frac{1}{N}\sum_{n=1}^{N}\sum_{s,t\sim\mathcal{T}}\left[\frac{1}{\tau}\mathbf{z}_{nt}^{\top}\mathbf{C}\mathbf{q}_{ns}+\frac{1}{\tau}\mathbf{z}_{ns}^{\top}\mathbf{C}\mathbf{q}_{nt}-\log\sum_{k=1}^{K}\exp\left(\frac{\mathbf{z}_{nt}^{\top}\mathbf{c}_k}{\tau}\right)-\log\sum_{k=1}^{K}\exp\left(\frac{\mathbf{z}_{ns}^{\top}\mathbf{c}_k}{\tau}\right)\right]$$

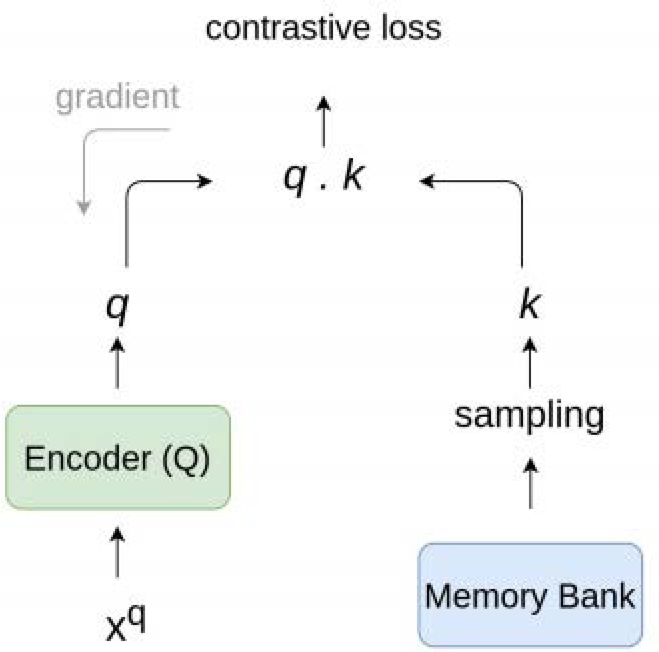
To avoid trivial solution, Q is regularized by complicated constraints

- High entropy
- Equal partition of images by prototypes (clusters)

A large batch size or memory bank is needed

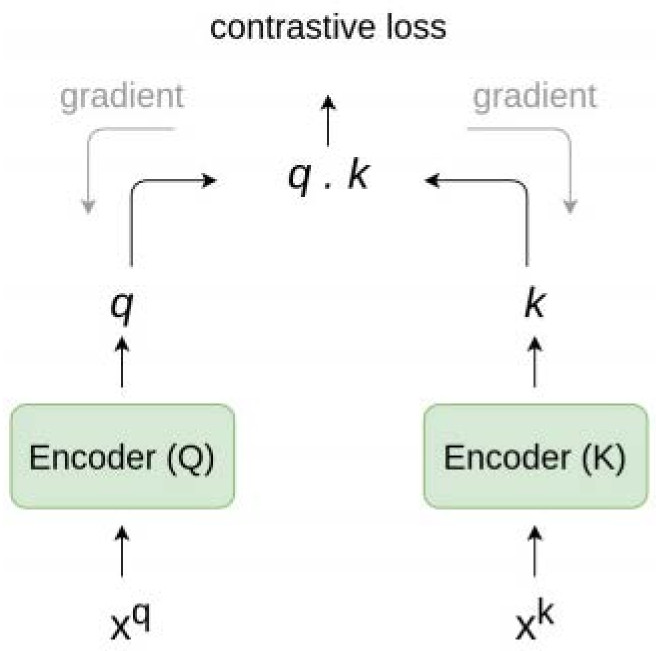
Caron, Mathilde, et al. "Unsupervised learning of visual features by contrasting cluster assignments." *arXiv preprint arXiv:2006.09882* (2020).

Brief summary



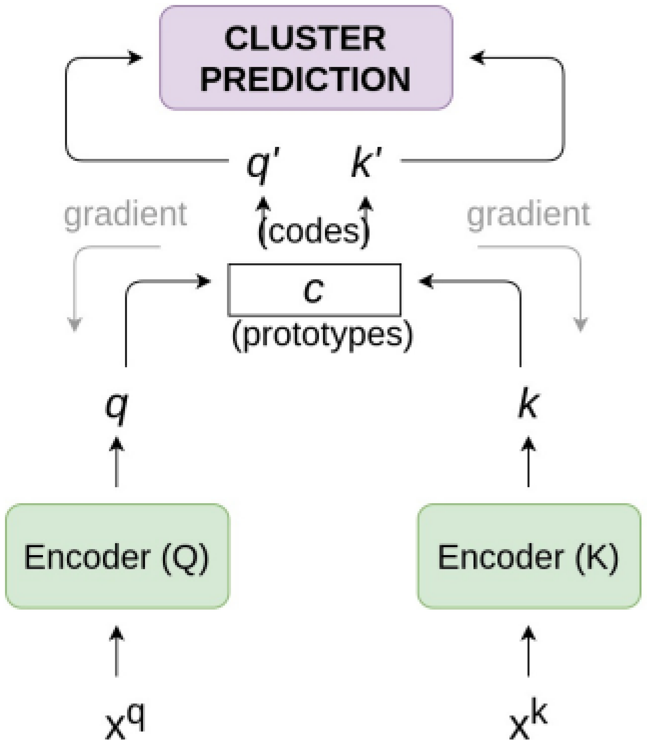
Memory bank

Needs to store all features



End-to-end

Needs a large batch size



Clustering

Needs a large batch size or memory bank

High GPU memory

Jaiswal, Ashish, et al. "A survey on contrastive self-supervised learning." *Technologies* 9.1 (2021): 2.

Thank you