南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Introduction to SimCLR

Yiming Zhang

July 20, 2023

Contents

How to get representations without supervision?

Generative approach

- modelling the joint distribution
- converge more quickly, can deal with latent features
- computationally expensive and may not be necessary
- can switch to discriminative approach
- Naive Bayes, GAN, GPT

Discriminative approach

- modelling the boundary, distribution-free
- can't switch to generative approach
- Linear Regression, SVM, KNN

Examplar Learning

- Main idea: treat each instance as a class represented by a feature vector
- Apparent similarity is from the visual data themselves.

Dosovitskiy et al. (2014)

- applying transformations to a 'seed' image
- better performance in object classification and descriptor matching

Wu et al. (2018)

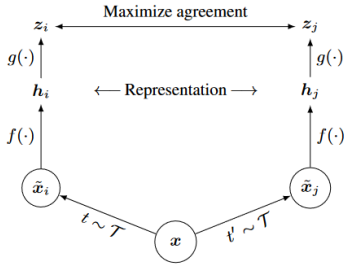
- treat the feature space as an unit sphere
- use Non-Parametric Softmax Classifier and Memory Bank

Contents

The Contrastive Learning Framework

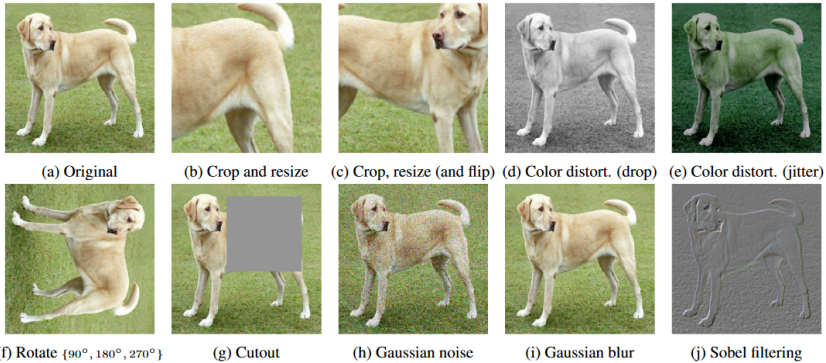
Idea:maximizing agreement between differently augmented views of the same data example

- stochastic data augmentation
- neural network: base encoder
- small neural network: projection head
- contrastive loss function



demo

Stochastic Data Augmentation Module

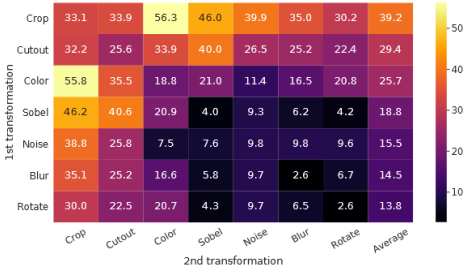


Data augmentation operations are crucial!

Stochastic Data Augmentation Module

Result of linear evaluation(ImageNet top-1 accuracy):

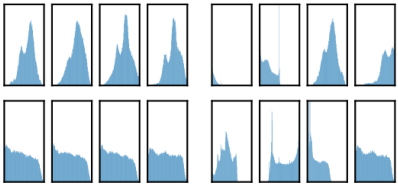
- No single transformation suffices to learn good representations
- Random cropping and random color distortion stands out.
- Why color distortion is crucial?



Effect of Color Distortion

From experiments:

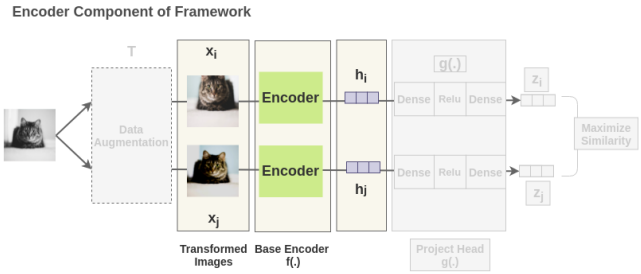
- It is critical to compose cropping with color distortion in order to learn generalizable features.
- Contrastive learning needs stronger data augmentation than supervised learning.



(a) Without color distortion. (b) With color distortion.

Methods	Color distortion strength					AutoAug
	1/8	1/4	1/2	1	1 (+Blur)	
SimCLR	59.6	61.0	62.6	63.2	64.5	61.1
Supervised	77.0	76.7	76.5	75.7	75.4	77.1

Architectures for Encoder and Head

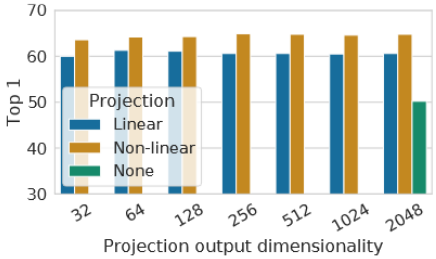


- $h_i = f(x_i) = \text{ResNet}(x_i)$
- output: 2048-dim vector $\boldsymbol{h}$

Architectures for Encoder and Head

Existence of projection head?

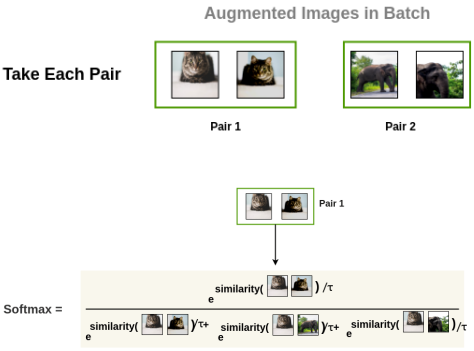
- A nonlinear projection head improves the representation quality
- $g(\boldsymbol{h})$ may remove information that may be useful for the downstream task
- In SimCLR, set $z_i = g(\boldsymbol{h_i}) = W^{(2)}\sigma(W^{(1)}\boldsymbol{h_i})$, where σ is a ReLU nonlinearity.

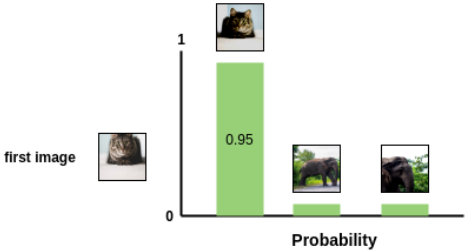


Loss Functions: NT-Xent function

Name	Negative loss function	Gradient w.r.t. $\boldsymbol{u}$
NT-Xent	$\boldsymbol{u}^T\boldsymbol{v}^+/\tau - \log \sum_{\boldsymbol{v}\in\{\boldsymbol{v}^+,\boldsymbol{v}^-\}} \exp(\boldsymbol{u}^T\boldsymbol{v}/\tau)$	$(1 - \frac{\exp(\boldsymbol{u}^T\boldsymbol{v}^+/\tau)}{Z(\boldsymbol{u})})/\tau\boldsymbol{v}^+ - \sum_{\boldsymbol{v}^-} \frac{\exp(\boldsymbol{u}^T\boldsymbol{v}^-/\tau)}{Z(\boldsymbol{u})}/\tau\boldsymbol{v}^-$
NT-Logistic	$\log\sigma(\boldsymbol{u}^T\boldsymbol{v}^+/\tau) + \log\sigma(-\boldsymbol{u}^T\boldsymbol{v}^-/\tau)$	$(\sigma(-\boldsymbol{u}^T\boldsymbol{v}^+/\tau))/\tau\boldsymbol{v}^+ - \sigma(\boldsymbol{u}^T\boldsymbol{v}^-/\tau)/\tau\boldsymbol{v}^-$
Margin Triplet	$-\max(\boldsymbol{u}^T\boldsymbol{v}^- - \boldsymbol{u}^T\boldsymbol{v}^+ + m, 0)$	$\boldsymbol{v}^+ - \boldsymbol{v}^-$ if $\boldsymbol{u}^T\boldsymbol{v}^+ - \boldsymbol{u}^T\boldsymbol{v}^- < m$ else $\boldsymbol{0}$

Loss Functions: NT-Xent function

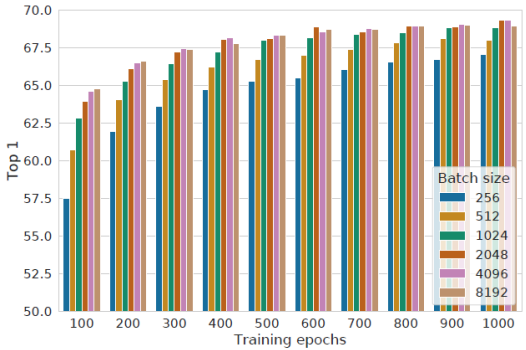




Batch Sizes and Training Time

Contrastive learning benefits more from larger batch sizes and longer training.

- When the number of training epochs is small (e.g. 100 epochs), larger batch sizes have a significant advantage over the smaller ones.
- With more training steps/epochs, the gaps between different batch sizes decrease or disappear, provided the batches are randomly resampled.



Contents

Summary

- Composition of data augmentations plays a critical role in defining effective predictive tasks.
- a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations.
- Contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning.

Thanks for listening!