Self supervised visual representation learning : SimCLR, Deep Clustering

Verlon, Sorel, Samuael, Binta, Honorine and Mame Diarra Diop, Uriel

African Master's in Machine Intelligence (AMMI)

Supervised by prof Moustapha Cisse

April 25, 2023

AIMS African Institute for Mathematical Sciences SENEGAL

Overview

1. Motivation
2. What is Self Supervised learning(SSL)?
3. Pretraining methods in computer vision (CV)
4. Contrastive representative Learning
   - Intuition and formulation
   - A formulation of constrative learning
5. Deep Clustering
6. Applications of self supervised learning

AIMS African Institute for Mathematical Sciences SENEGAL

Motivation

- Supervised learning has demonstrated good results in many machine learning applications(image classification, spam detection,etc).

Motivation

- Supervised learning has demonstrated good results in many machine learning applications(image classification, spam detection,etc).
- Unfortunately, supervised learning usually require a large amount of labeled data which is not easy to obtain for many applications due to the cost and the domain limitations.

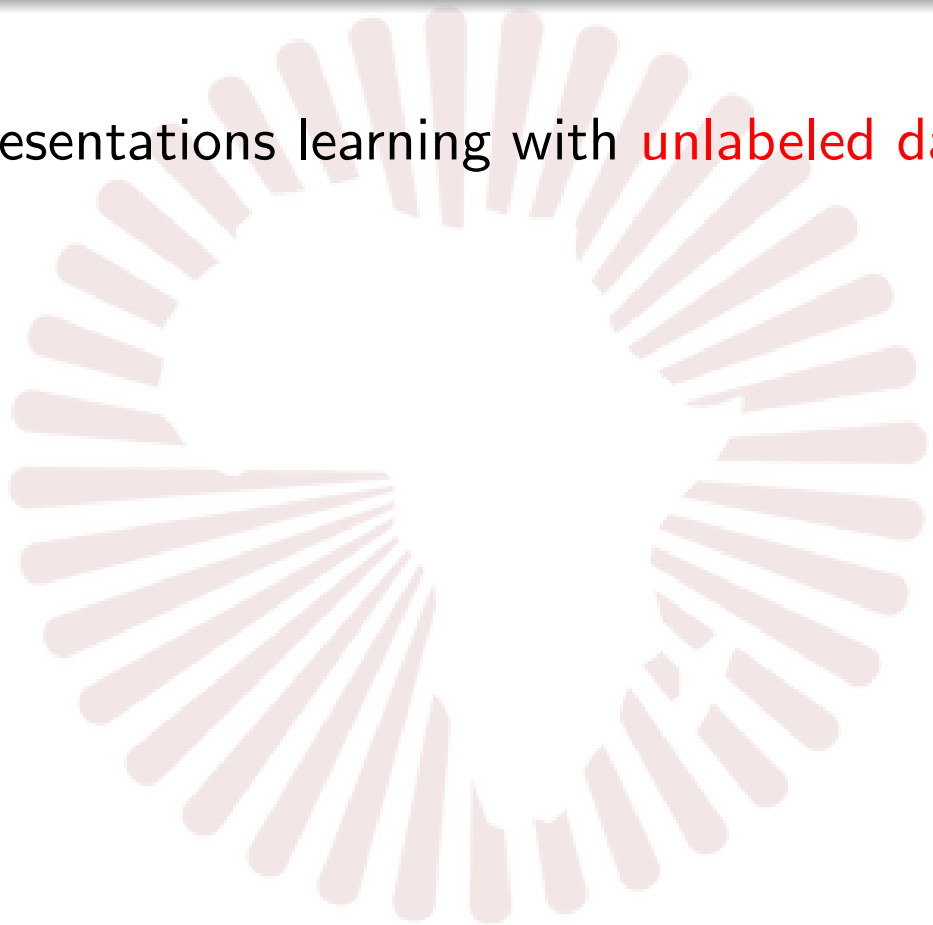
Motivation

- Supervised learning has demonstrated good results in many machine learning applications(image classification, spam detection,etc).
- Unfortunately, supervised learning usually require a large amount of labeled data which is not easy to obtain for many applications due to the cost and the domain limitations.
- With Self-Supervised Learning, we can used inexpensive unlabeled data to learn representations trough the pretext task for downstream task.

AIMS African Institute for Mathematical Sciences SENEGAL

What is Self Supervised learning(SSL)?

SSL is the representations learning with unlabeled data



What is Self Supervised learning(SSL)?

SSL is the representations learning with unlabeled data

- Learn useful feature representations from unlabeled data through pretext task to solve downstream tasks (classification, regression, clustering). In computer vision, the pretext task can be (Image rotation, image colorization,filling missing piece in the image)

What is Self Supervised learning(SSL)?

SSL is the representations learning with unlabeled data

- Learn useful feature representations from unlabeled data through pretext task to solve downstream tasks (classification, regression, clustering). In computer vision, the pretext task can be (Image rotation, image colorization,filling missing piece in the image)

- The term " Self supervised" refers to creating its own supervision(i.e., without supervision, without lablels)

What is Self Supervised learning(SSL)?

SSL is the representations learning with unlabeled data

- Learn useful feature representations from unlabeled data through pretext task to solve downstream tasks (classification, regression, clustering). In computer vision, the pretext task can be (Image rotation, image colorization,filling missing piece in the image)

- The term " Self supervised" refers to creating its own supervision(i.e., without supervision, without lablels)

- Self-supervised learning is one category of unsupervised learning.

AIMS African Institute for Mathematical Sciences SENEGAL

What is Self Supervised learning(SSL)?

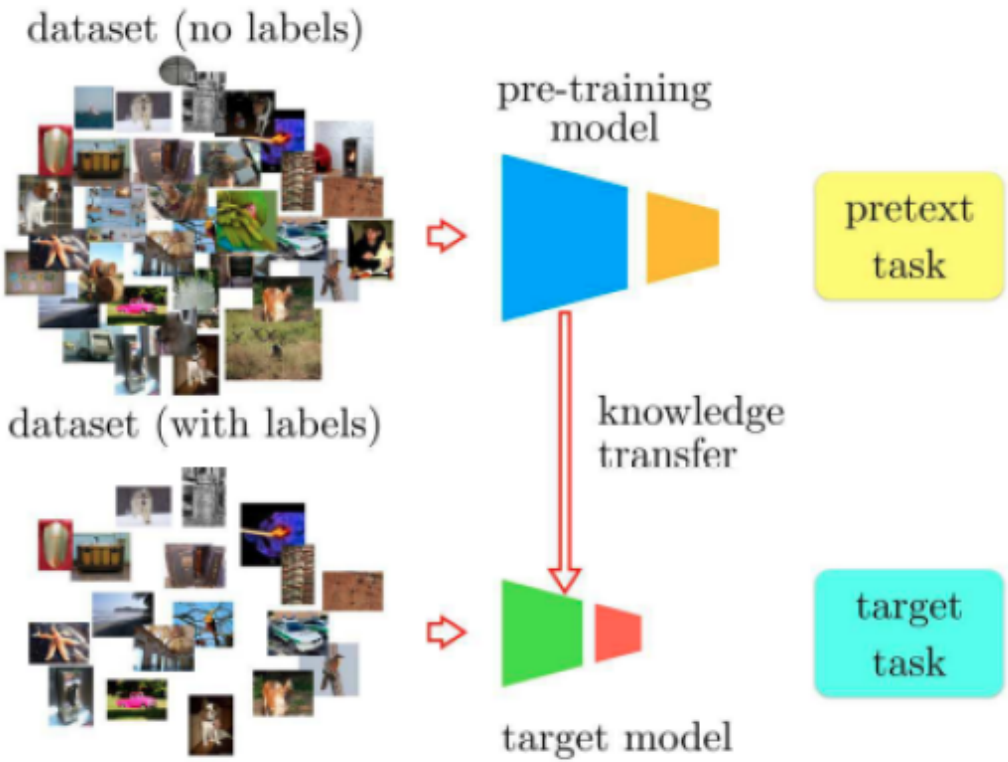


image source

Pretraining methods in computer vision (CV)

They are two concrete pretraining methods for CV :

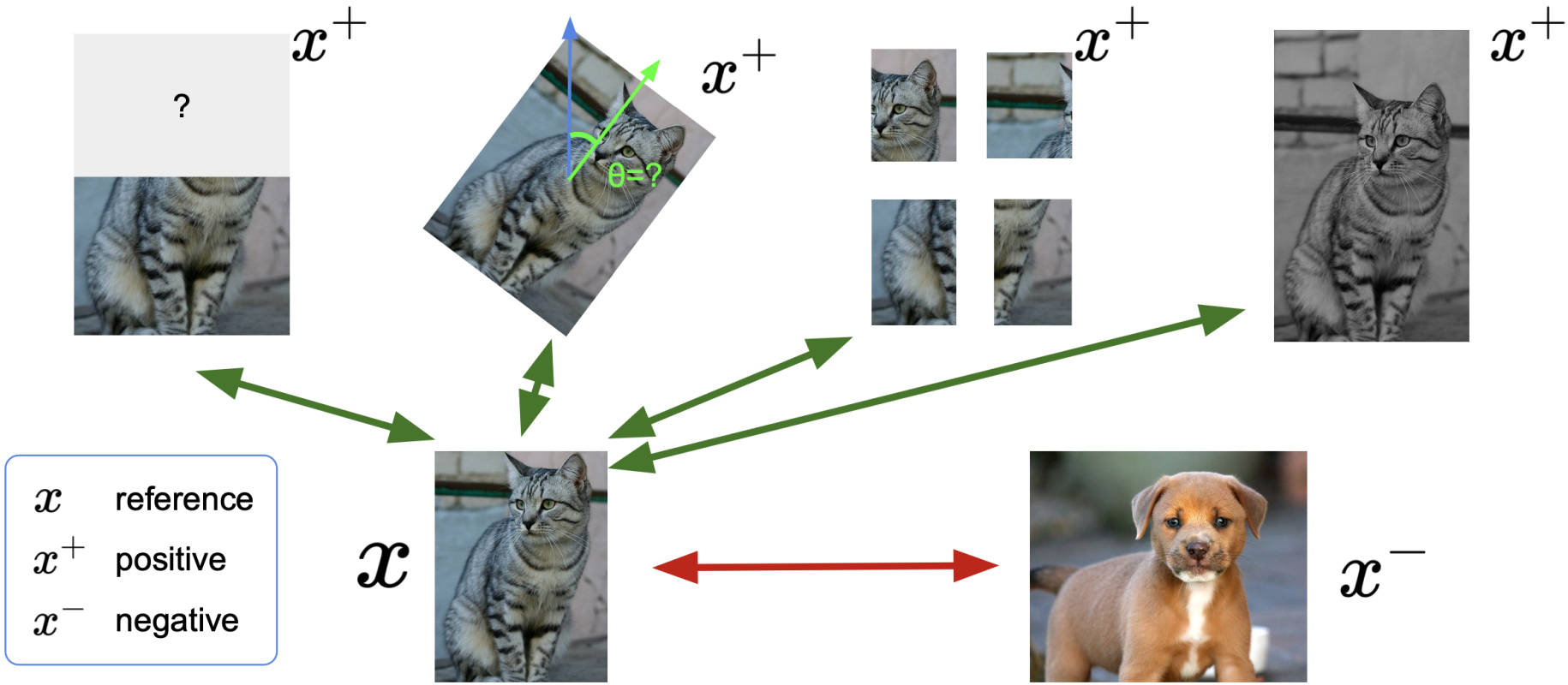
- Contrastive representative learning
- Supervised pretraining

In our work we will focus in Contrastive representative learning and in particular on a Simple Framework for Contrastive Learning(SimCLR)

Overview

1. Motivation
2. What is Self Supervised learning(SSL)?
3. Pretraining methods in computer vision (CV)
4. Contrastive representative Learning
   - Intuition and formulation
   - A formulation of constrative learning
5. Deep Clustering
6. Applications of self supervised learning

Intuition and formulation



source : Chen, Ting, et al.

AIMS African Institute for Mathematical Sciences SENEGAL

Overview

1 Motivation
2 What is Self Supervised learning(SSL)?
3 Pretraining methods in computer vision (CV)
4 Contrastive representative Learning
- Intuition and formulation
- A formulation of constrative learning

5 Deep Clustering
6 Applications of self supervised learning

AIMS African Institute for Mathematical Sciences SENEGAL

A formulation of constrative learning

We want :

$$score(f(x), f(x^+)) >> score(f(x), f(x^-))$$

where,
x: reference sample;
x^+ : positive sample;
x^-: negative sample.
So, given a choosen score function, we aim to learn an **encoder function** f that yields high score for positive (x, x^+) and the low scores for negative pairs (x, x^-).

A formulation of constrative learning

Loss function given one positive and N-1 negative samples:

$$L = -\mathbb{E}_X \left[\log \frac{\exp(s(f(x), f(x^+)))}{\exp(s(f(x), f(x^+))) + \sum_{j=1}^{N-1} \exp(s(f(x), f(x_j^-)))} \right], \tag{4.1}$$

where,

L: The contrastive loss or Noise-Contrastive Estimation (NCE) loss

$\exp(s(f(x), f(x^+)))$: score for the positive pair

$\sum_{j=1}^{N-1} \exp(s(f(x), f(x_j^-)))$: score for the N-1 negative pairs.

$$s(u, v) = \frac{u^T v}{\|u\|\|v\|}$$

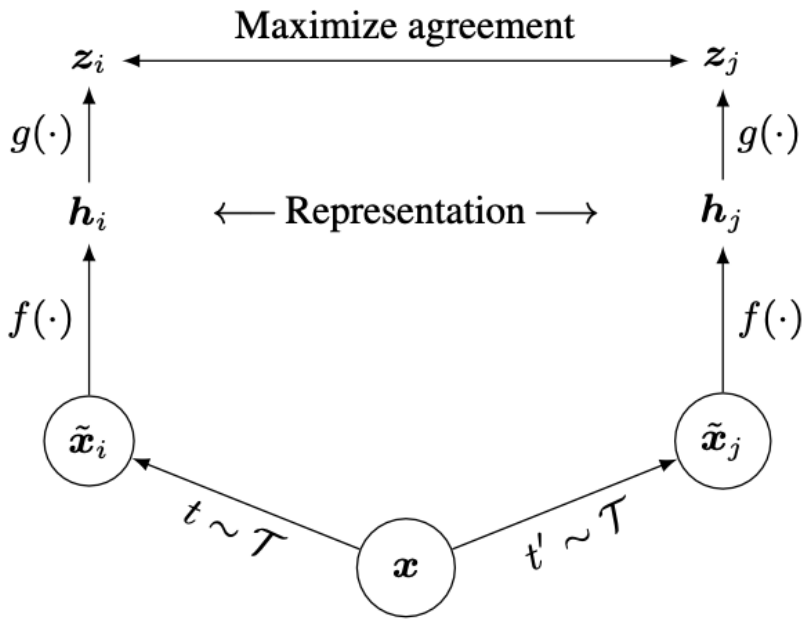
AIMS African Institute for Mathematical Sciences SENEGAL

SimCLR architecture

- Data augmentation operations t and t'
- Base encoder $f(\bullet)$ to obtain the code representation h_i and h_j.
- Prediction head encoder or neural network projection head $g(\bullet)$

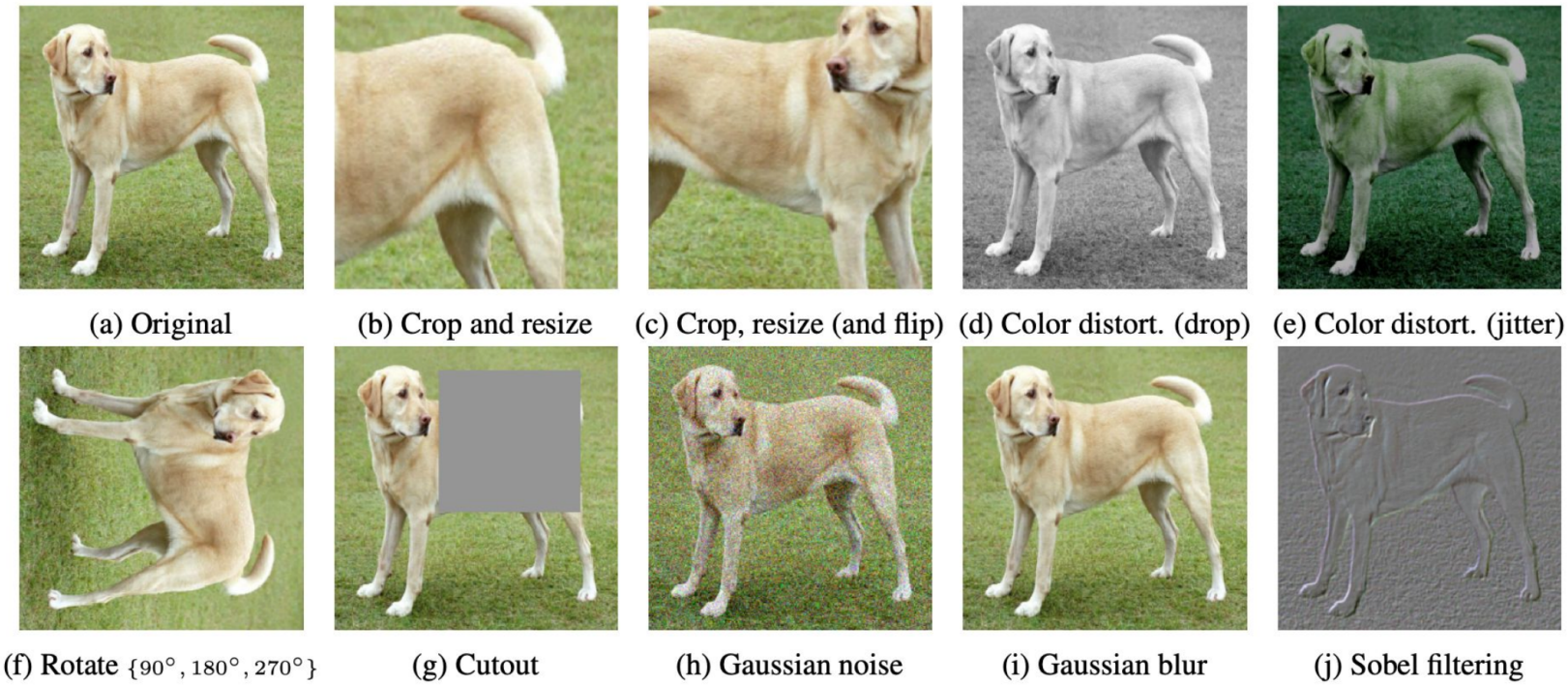
We use Multi Layer Perceptron(MLP) with one hidden layer to obtain

$$z_i = g(h_i) = W^2\sigma(W^1 h_i) \ .$$



source :Chen, Ting, et al.

SIMCLR : generating positive samples from data augmentation



(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

source : Chen, Ting, et al.

SimCLR's algorithm

SimCLR

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, 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 $g(\cdot)$

Generate a positive pair by sampling data augmentation functions

Iterate through and use each of the 2N sample as reference, compute average loss

InfoNCE loss: Use all non-positive samples in the batch as x^-

source: Chen, Ting, et al.

SSL Limitations

- Computing ressources:It takes large amount of data and computation resources to pre-train the model.
- Domain-specific limitations: It may not always be suitable for all domains or applications. In some cases, labeled data may be readily available and may be more efficient to use for training instead.
- Limited types of data: It typically works best with certain types of data, such as images, text, or audio, which have a clear underlying structure. However, it may not be as effective with other types of data, such as graphs or time-series data.

Deep Clustering

- Simple clustering is an unsupervised learning method that groups datapoints together based on their similarities.

Deep Clustering

- Simple clustering is an unsupervised learning method that groups datapoints together based on their similarities.

- Deep supervised learning : work well when we have large amounts labeled data.

Deep Clustering

• Simple clustering is an unsupervised learning method that groups datapoints together based on their similarities.

• Deep supervised learning : work well when we have large amounts labeled data.

• DeepCluster is a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features.

Deep Clustering

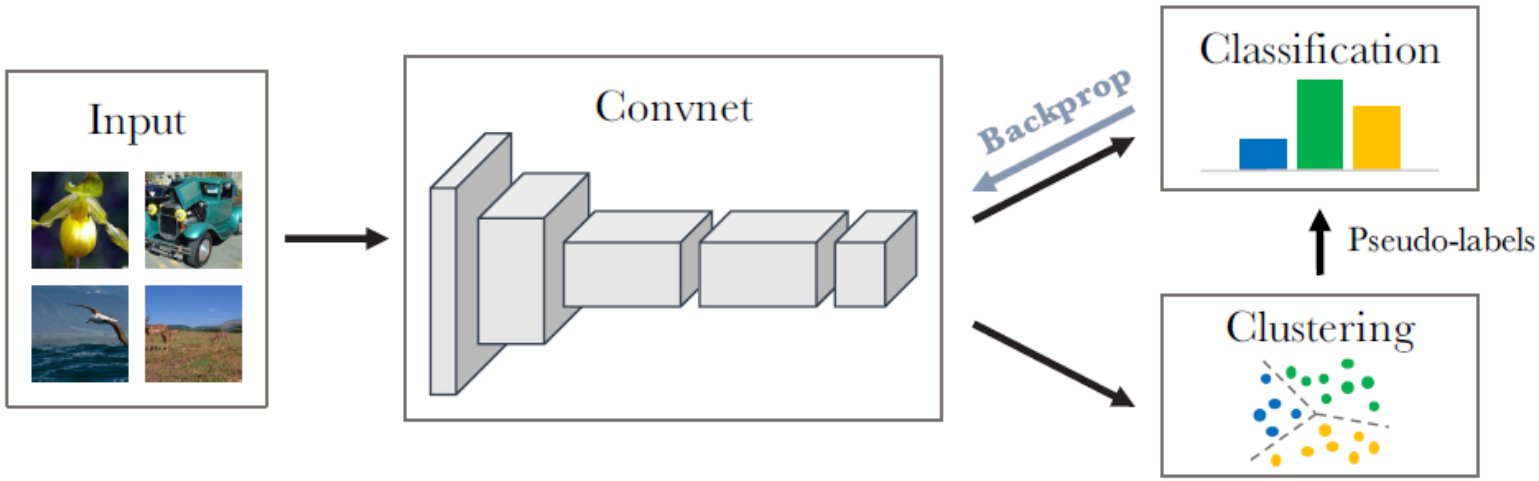


Figure: Illustration of the proposed method: we iteratively cluster deep features and use the cluster assignments as pseudo-labels to learn the parameters of the convnet. [1]

Deep Clustering

$\{x_1, ..., x_N\}$ is the training set of N images.
K-means clusters the features $f_\theta(x_n)$ into k distinct groups.
$C \in^{d \times k} \longrightarrow$ centroid matrix
$y_n \longrightarrow$ cluster assignment of each image

$$\min_{C \in^{d \times k}} \frac{1}{N} \sum_{n=1}^{N} \min_{y_n \in \{0,1\}^*} \|f_\theta(x_n) - C y_n\|_2^2 \text{, such that } y_n^T 1_k = 1 \quad (5.1)$$

$g_W \longrightarrow$ parametrized classifier on top of the features $f_\theta(x_n)$

$$\min_{\theta, W} \frac{1}{N} \sum_{n=1}^{N} l(g_W(f_\theta(x_n)), y_n) \quad (5.2)$$

l is the negative log softmax.

Deep Clustering

Trivial solutions

- **Empty clusters**: Because of the absence of mechanisms to prevent from empty cluster, we may have during a clustering some clusters with no object while normally an optimal separation should assign all of the inputs to a single cluster.

- **Trivial parametrization**: When the number of images per class is highly unbalanced we can meet a trivial parametrization.

Deep Clustering

Method	Training set	Classification		Detection		Segmentation	
		FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
Best competitor	ImageNet	63.0	67.7	43.4†	53.2	35.8†	37.7
DeepCluster	ImageNet	72.0	73.7	51.4	55.4	43.2	45.1
DeepCluster	YFCC100M	67.3	69.3	45.6	53.0	39.2	42.2

Figure: Illustration of DeepCluster process

- Impact of the training set on the performance of DeepCluster measured on the Pascal VOC transfer tasks. We compare ImageNet with a subset of 1M images from YFCC100M.
- Regardless of the training set, DeepCluster outperforms the best published results on most tasks.

AIMS African Institute for Mathematical Sciences SENEGAL

Deep Clustering

Method	AlexNet	VGG-16
ImageNet labels	56.8	67.3
Random	47.8	39.7
Doersch *et al.* [25]	51.1	61.5
Wang and Gupta [29]	47.2	60.2
Wang *et al.* [46]	–	63.2
DeepCluster	**55.4**	**65.9**

(a) Pascal VOC 2007 object detection with AlexNet and VGG- 16.

Method	Oxford5K	Paris6K
ImageNet labels	72.4	81.5
Random	6.9	22.0
Doersch *et al.* [25]	35.4	53.1
Wang *et al.* [46]	42.3	58.0
DeepCluster	**61.0**	**72.0**

(b) mAP on instance-level image retrieval on Oxford and Paris dataset with a VGG-16

- Compares a VGG-16 and AlexNet trained with DeepCluster and tested on the Pascal VOC 2007 object detection task.
- Apply VGG-16 architecture with Sobel filter on Oxford Buildings and Paris datasets.

AIMS African Institute for Mathematical Sciences SENEGAL

Applications of SSL

- **Computer Vision**: Self-supervised learning has been used extensively in computer vision tasks such as image and video analysis, object detection, segmentation, and tracking.
- **Naturel Language Processing**: Self-supervised learning has also been applied to natural language processing tasks such as language modeling, sentiment analysis, and machine translation.
- **Robotics**: Self-supervised learning has been used in robotics applications such as robot navigation, manipulation, and control.
- **Healthcare**: Self-supervised learning has been applied to healthcare applications such as medical image analysis, patient diagnosis, and drug discovery.

AIMS African Institute for Mathematical Sciences SENEGAL

Conclusion

In this work, we talked about self supervised learning which is a kind of unsupervised learning where we build models with unlabelled data. There are many types of self supervised learning algorithms but we have just focused on SimCLR and Deep Clustering. Despite some limitations, those algorithms are very useful in solving problem in different areas such as healthcare, robotics, natural language processing, etc...

References

Chen, Ting, et al.

A simple framework for contrastive learning of visual representations.

International conference on machine learning, 1597–1607 2020 .

Caron, Mathilde et al.

Deep clustering for unsupervised learning of visual features.

Proceedings of the European conference on computer vision (ECCV), pages=132–149, year=2018.

Xu, Linli et al.

Maximum Margin Clustering.

Proceedings of the European conference on computer vision (ECCV), volume = 17, year = 2004.

AIMS African Institute for Mathematical Sciences SENEGAL

References

Philbin, Noroozi et al.
Boosting self-supervised learning via knowledge transfer.
Proceedings of the IEEE conference on computer vision and pattern recognition, 9359–9367, 2018.

Philbin, James et al.
Object retrieval with large vocabularies and fast spatial matching.
2007 IEEE Conference on Computer Vision and Pattern Recognition, year=2007, volume=, number=, pages=1-8.

AIMS African Institute for Mathematical Sciences SENEGAL

Acknowledgements

African Institute for Mathematical Sciences
SENEGAL