Video Analysis with Deep Learning and without it

Pr. Jenny Benois-Pineau
LABRI UMR 5800/Université Bordeaux

Lecture 3

Incremental -Continual Learning

université de BORDEAUX

Chapter 3

Summary.

Incremental learning

1. Problem statement
2. Cases
3. « Move-to-data » method

Incremental Learning: Problem Statement

- Suppose model M_t was trained with training data S_t
- Model of ”Zero phases” - $M_{t\ ,\ t=0}$ pretrained on a bunch of available data
- Consider a new data set S_{t+1}
- How can we enrich the model M_t with S_{t+1} and get M_{t+1}?
- How can we adapt our decision boundary without a full re-training?

Different cases of incremental learning

- **Class-Incremental Learning (CIL)** : aims to learn a classi-
- fication model with the number of classes increasing phase-
- by-phase.

- **One class incremental learning** : adapted to the streaming of data or tracking in video sequences: binary classification task.

Dataset GIWT, https://api.nakala.fr/data/11280%2F24923973/d6c607da6cc647021d40cec45815436bd7a45053

B. Krawczyk, M. Wozniak, One-class classifiers with incremental learning and forgetting for data streams with concept drift, *Soft Computing* **volume 19**, pages 3387–3400 (2015)

Concept Drift

- Let us remind that we have trained a mapping

$$g(\mathbf{x},\alpha): X \to Y \qquad g(\mathbf{x}_n,\alpha) = \hat{y}_n$$

- We suppose that the mapping $g(\mathbf{x},\alpha): X \to Y$ trained on a bunch of data is good for all new data. But(!)

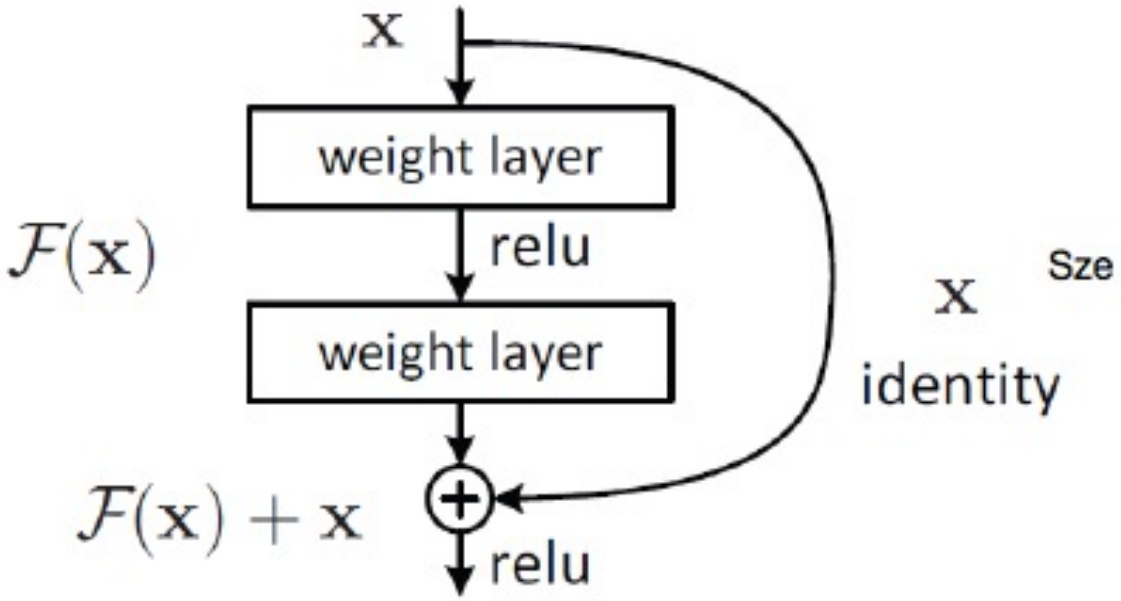
- **Concept drift** in machine learning refers to the change in the relationships between input and output data in the underlying problem over time.

- Otherwise called "*covariate shift*," "*dataset shift*"

- In the case of the Deep Learning for a network of a stable architecture, this means the necessity of adjusting the parameters.

Popular Backbones: ResNet

➔ Principle : instead of learning original mapping $F(X)$, a ResNet layer learns residual mapping $F(X)=F(X)-X$



Antipov, G. "Deep Learning for Semantic Description of Visual Human Traits, PhD 2017, TelecomParisTech
He, K., Zhang, X.,Ren, Sh., Sun, J. :Deeep Residual Learning for Image Recognition »,
in Proc. IEEE CVPR, Las Vegas , USA

How to adapt the learnt model to a concept drift (1)

➔ **1. Do Nothing (Static Model)**

develop a single “best” model once and use it on all future data

➔ **2. Periodically Re-Fit**

A good first-level intervention is to periodically update your static model with more recent historical data (CF!). A ned of back testing of the model to identify tghe amount of recent data to include.

➔ **3. Periodically Update**

This is an efficiency over the previous approach (periodically re-fit) where instead of discarding the static model completely, the existing state is used as the starting point for a fit process that updates the model fit using a sample of the most recent historical data.

4. Weight Data

➔ Use a weighting that is inversely proportional to the age of the data such that more attention is paid to the most recent data (higher weight) and less attention is paid to the least recent data (smaller weight).

https://machinelearningmastery.com/gentle-introduction-concept-drift-machine-learning/

How to adapt the learnt model to a concept drift (2)

➔ **5. Learn The Change**

An ensemble approach can be used where the static model is left untouched, but a new model learns to correct the predictions from the static model based on the relationships in more recent data.

➔ **6. Detect and Choose Model**

For some problem domains it may be possible to design systems to detect changes and choose a specific and different model to make predictions.

➔ **7. Data Preparation**

➔ In some domains, such as time series problems, the data may be expected to change over time. (Weatherforcast in winter and in Spring!)

In these types of problems, it is common to prepare the data in such a way as to remove the systematic changes to the data over time, such as trends and seasonality by differencing.

A « Catastrophic forgetting » phenomenon

- « Intelligent agents must demonstrate a capacity for **continual learning**: that is, the ability to learn consecutive tasks without forgetting how to perform previously trained tasks. » (J. Kirkpatrick, 2017 DeepMind)
- In DNNs : previously learnt tasks can be abruptly lost as information relevant to the current task (e.g. task B) is incorporated.

- This penomenon was called « **catastrophic forgetting** »*

- E.g. CNN inheretly contain catastrophic forgetting due to the optimsation of their parameters by SGD.
- Hence the problem is how we can adapt without forgetting too abruptly and too much.

*Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. The psychology of learning and motivation, 24(109-165):92, 1989.

Optimisation by SGD

- Simplest form

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \nabla \mathbf{L}\left(\mathbf{W}^{(t)}\right)$$

- (1)Initialize all parameters by random values
- (2) Randomly select a batch of *B* training data
- (3) Perform optimization with the batch of data.
- (4) Repeat 2 untill all data in the training set have been used
- (5) Repeat 2-4 Nbr Epochs

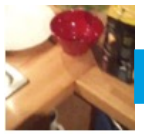
- Epoch use of the whole training dataset

- Nombre d'itérations of training :

$$NbrIt = NbrEpochs * \frac{NbrTData}{BatchSize}$$

Streaming Learning

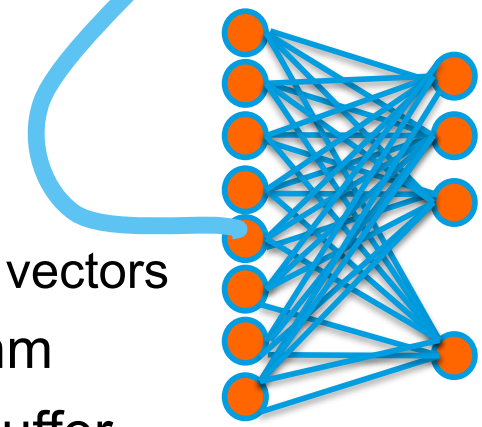
- It is a special case of incremental learning when the quantity of new data is just 1 data sample
- »Exstream » method [1]
  - - usage of a buffer



Cov. Layers of Phase 0

Flatten

Buffer



Pre-trained model – feature extractor
It does not change during learning

- Buffer

-for each of *C* classes store *s* feature vectors

- *s* is the parameter of the algorithm
- train classification layer on the buffer
- when full – update the buffer with a new data

Streaming Learning

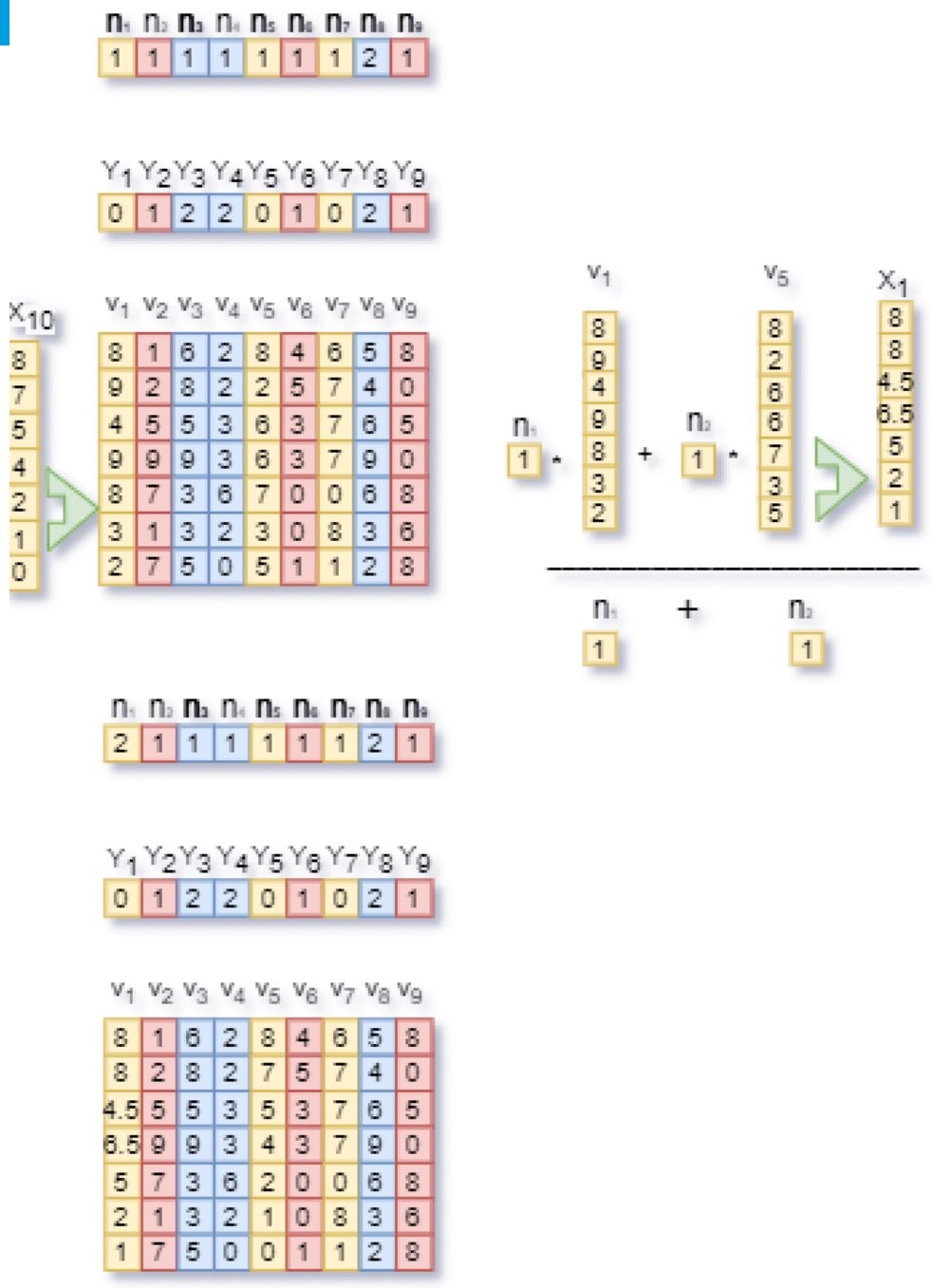
- Buffer update

$$v_i \leftarrow \frac{n_i v_i + n_j v_j}{n_i + n_j}, \quad n_i \leftarrow n_i + n_j, \quad v_j \leftarrow x_t, \quad n_j \leftarrow 1$$

- v – feature vectors, closes are fused
- x_t – new image feature vector at time t

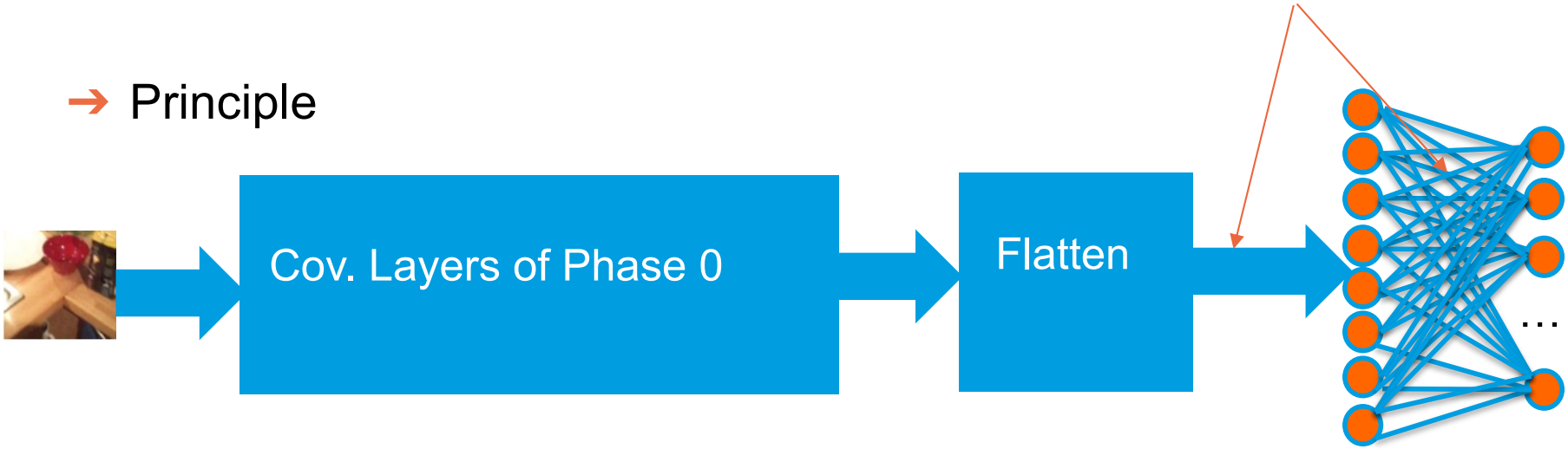
Example

- 3-class
- classification



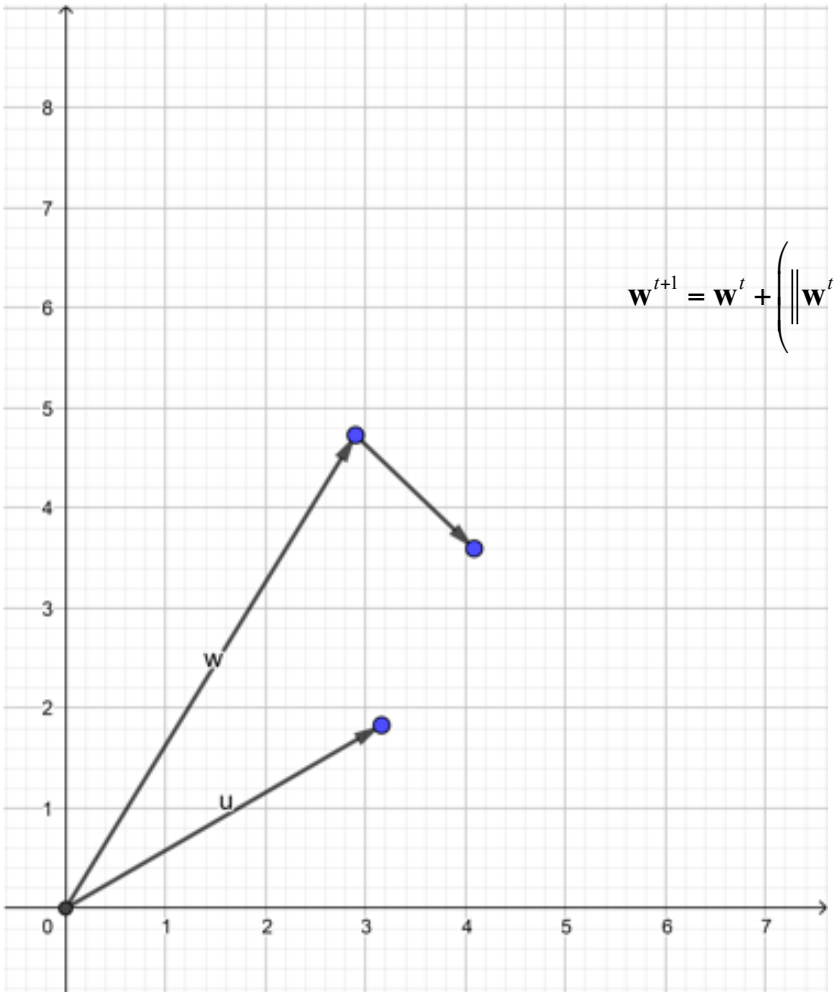
« Move-to-data »

- A method without retraining
- Uses only the last FC Layer of a CNN
- The core idea of the method is the adjustment of a weight of the neuron responding to the class of the training example coming sequentially "on the fly".
- Principle

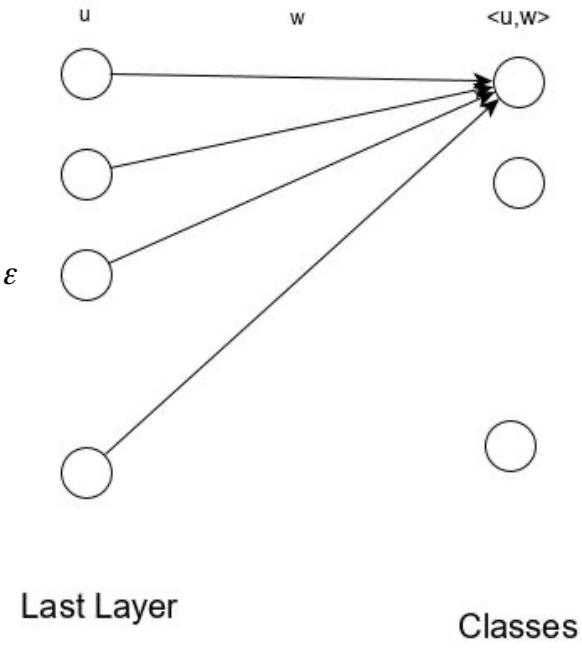


Incremental learning: Move to data

➔ Consider last layer of NN



$$\mathbf{w}^{t+1} = \mathbf{w}^{t} + \left(\left\| \mathbf{w}^{t} \right\| \frac{\mathbf{u}}{\left\| \mathbf{u} \right\|} - \mathbf{w}^{t} \right) \cdot \varepsilon$$



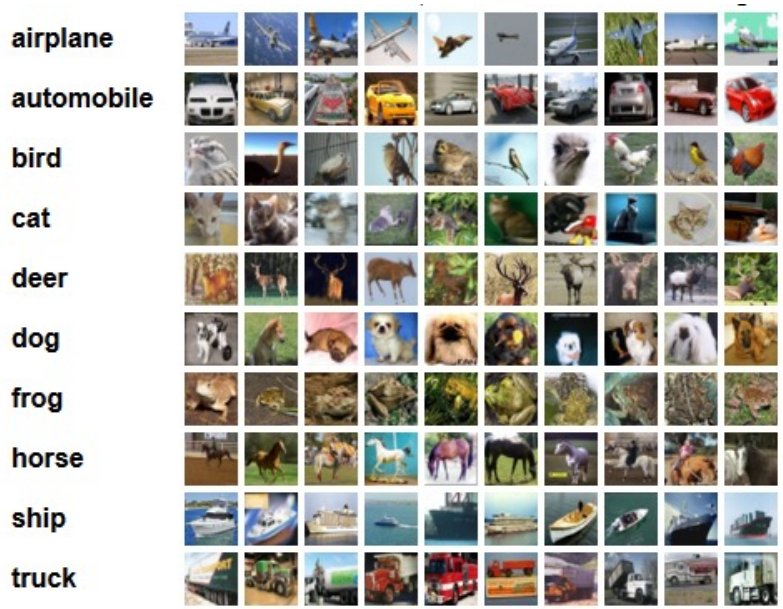
Cauchy Formula:

$\langle u, w \rangle \leq \|u\| \|w\|, \forall u, w \in \mathbb{R}^d$

Equality when co-linear

Testing of incremental principle on CIFAR-10

- 10 classes, 60.000 images, 6.000 images per class
- 50.000 training images, 10.000 test images
- LeNet architecture



https://www.cs.toronto.edu/~kriz/cifar.html

- Three methods: gradient descent on all layers, gradient descent on last layer and moving to data
  - On gradient descent: momentum, Learning rate: $2*10^{-4}$
- 50 images per class for incremental phase

Results on CIFAR

➔ Example: Class Plane

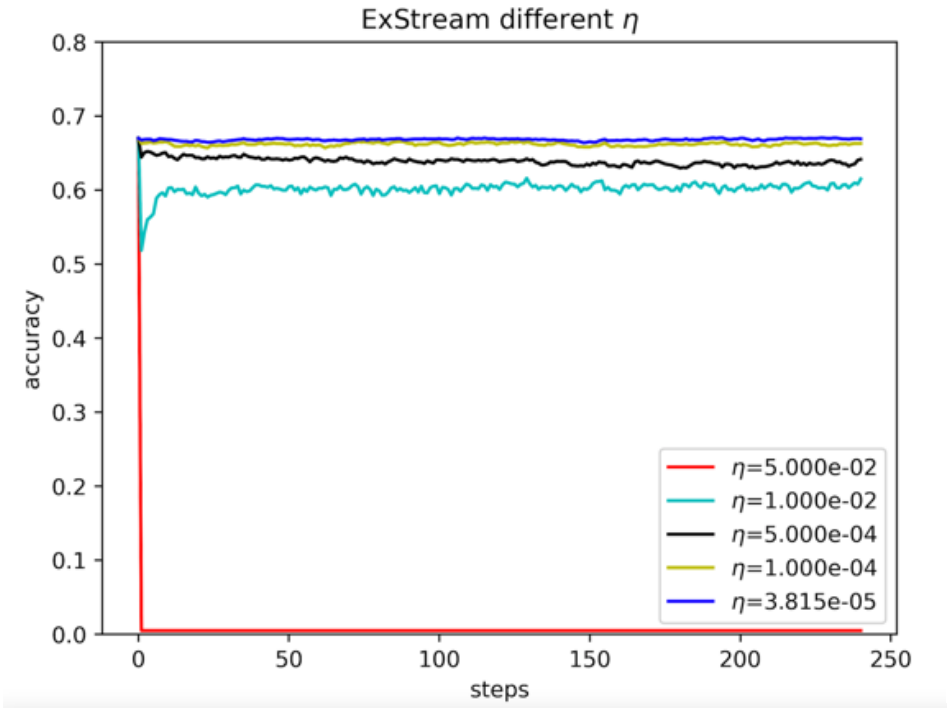
➔ Global Accuracy stays the same

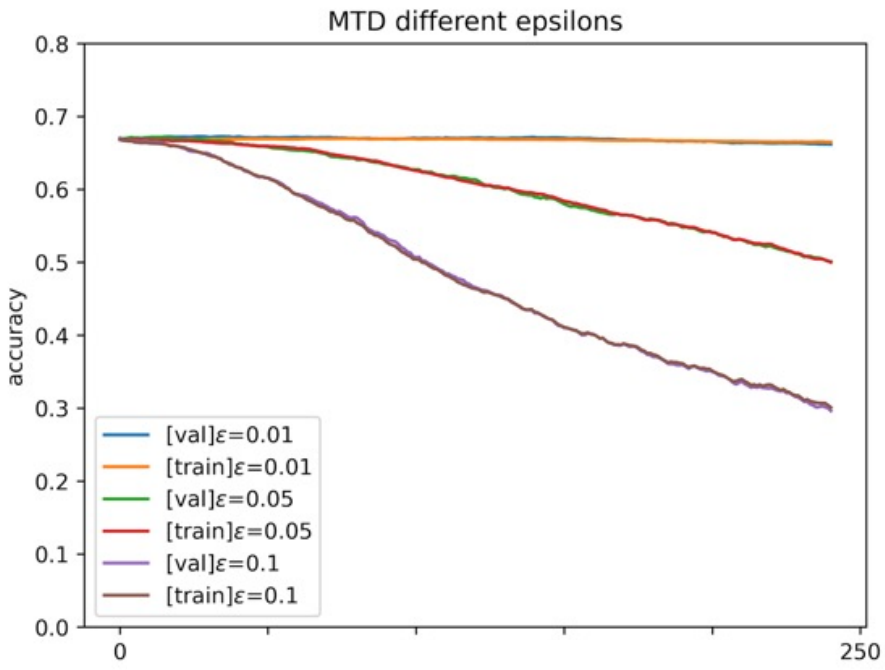
	overall accuracy
Backbone	53%
gradient descent last layer	52%
gradient descent all layers	53%
moving weights	53%

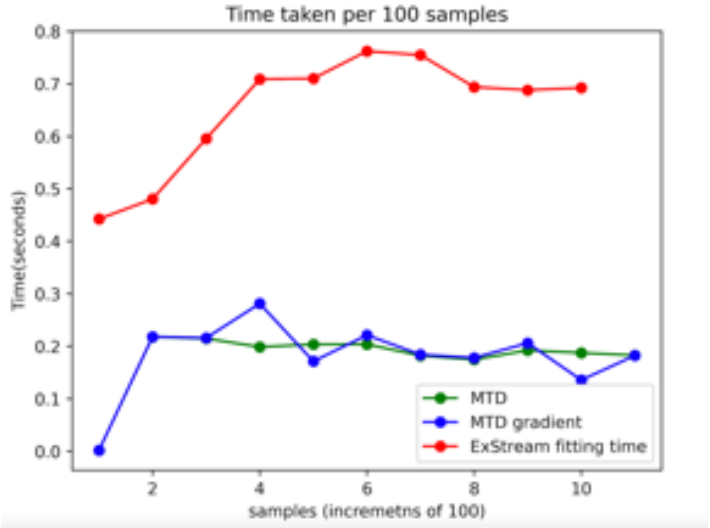
➔ Moving to data has similar precision and recall to applying gradient descent

	Recall		Precision	
Backbone	61%		62%	
gradient descent last	75%	14%	51%	-11%
gradient descent all	70%	9%	55%	-7%
moving weights	78%	17%	47%	-15%

Comparison of Exstream and MTD







Database : ImageNet200

Phase 0: 50K
Phase S: 50K
Validation 10K

Bibliography

- [1].T. L. Hayes, N. D. Cahill, and C. Kanan, « Memory efficient experience replay for streaming learning, » in 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 9769–9776.
- [2].A. K.Gebreslassie, J. Benois-Pineau, A. Zemmari, « Streaming learning with Move-to-Data approach for image classification ». CBMI 2022: 167-173
- [2].M. Poursanidis, J. Benois-Pineau, A. Zemmari, B. Mansencal, A. de Rugy, « Move-to-Data: A new Continual Learning approach with Deep CNNs, Application for image-class recognition ». CoRR abs/2006.07152 (2020)