[CVPR 2021 Tutorial] Leave Those Nets Alone: Advances in Self-Supervised Learning

# Introduction

### Motivation

Deep Learning works like this:

* Predefine set of visual concepts to be learned
* Collect diverse and large number of examples for each of them
* Train a deep model for several GPU hours or days

There are some drawbacks:

* Time consuming and expensive
* Require intense human labour (annotate and clean raw data)
* Error prone (human mistakes)

Moreover:

* Data distributions shift all the times
* Infeasible to launch large annotation campaigns each time
* Sensors specifics are frequently upgraded

To sum up: Deep Learning requires large amounts of carefully labelled data which is difficult to acquire and expensive to annotate. Still Supervised Learning there are several blind-spots in learning useful and rich representations.

### Definition

SSL is a form of **unsupervised learning** where data provides the supervision signal; it usually defines a **pretext task** for which the network is forced to learn what is the main goal; for most pretext tasks, **part of data is withheld** and the network has to **predict** it. The features/representations learned on the pretext task are subsequently used for a different **downstream task**, usually where some annotations are available. A common pretext task is the **Rotation Prediction**.

Multiple approaches can match the definition of SSL:

* **input or feature reconstruction**: [Hinton and Salakhutdinov (2006); Vincent et al. (2008); Gidaris et al. (2020), Grill et al. (2020)]
* **generating data**: [Goodfellow et al., (2014)]
* **training with paired signals**: [V. De Sa (1994); Arandjelovic and Zisserman (2017)]
* **hiding data from the networks**: [Doersch et al. (2015); Zhang et al. (2017)]
* **instance discrimination**: [Dosovitskiy et al. (2014); van der Ooord et al. (2018)] ...

### Scope

Focus on SSL methods that lead to useful representation obtained through invention of a pretext task and/or by hiding part or view of the original data

### SSL Methods

In most benchmarks the model is pretrained on ImageNet on a pretext task and subsequentially fine-tuned on other datasets or protocols.

The main three methods to compare SSL are:

* **LINEAR CLASSIFICATION**: (FC layer, linear SVM), typical dataset: ImageNet, Places205, Pascal V0C07. It’s hard to beat the SL on ImageNet, but they’re trying that.
* **ANNOTATION EFFICIENT CLASSIFICATION**: pre-trained on full dataset and then fine-tune on 5%-10%
* **TRANSFER LEARNING**: augment model with module and fine-tune partially or completely. Typical tasks and dataset:
* Object detection: VOC07, VOC12, COCO14
* Semantic segmentation: Cityscapes, ADE20K
* Other tasks: Surface Normal Estimation (NYUv2), Visual Navigation (Gibson)

# Contrastive Learning Approach

### What is Contrastive Learning

Sample a positive and a negative image and fit a scoring between them

### Contrastive Loss

While in Supervised Learning you have positive = same class, negative = random different class; in Self-Supervised Learning you have positive = sample PARTS of the same image, negative = sample parts of RANDOM image or SAME image at different location.

The Loss can be simple and privileging the positive and discarding the negative or be a Triplet loss which get nearer the positive while getting further the negative at same time. Example of Contrastive Loss is Sigmoid cross-Entropy used in word2vec, another common loss is the InfoNCE, it is the loglikelihood of predicting the positive sample correctly. This latter can be expanded into a version with a temperature feature.

# Clustering Learning Approach

Can we replace labels with clustering? First you perform a K-mean clustering dividing elements in the feature space and thus assigning pseudo-labels to instances. With backpropagation you train the NN as in Supervise Learning. In this way you are implicitly learning Invariance to Cropping via a random crop augmentation technique. How to evaluate this Pre-Training? Example: object-detection on Pascal V0C07 dataset where DeepCluster got results better than previous SSL techniques but still worse than Supervised Learning. Many limitations lead to use the SwAV method: how to produce the pseudo-labels? Comparing the point in the feature space to the different centroids using the similarity concept of finding the closest to the point in object; in this process you need a score for each cluster and the constraint is to impose the total score for each output must be the same. Sinkhorn adjust the scores to avoid a predominant cluster over the others.

# MultiModal Approaches to SSL

Usual training is based on:

* **Training** on large datasets with no human labels (e.g. IG65, Kinetics, YT8M, AudioSet, HowTo100M)
* **Evaluation** on smaller datasets with labels (e.g. HMDB51, Kinetics, UCF101, AudioSet, ESC-50)

We want to know how good is the representation, different procedures:

* **Nearest Neighbours** (classify by using the labels of the closest example in the feature space, is parameter free but also with lower performance than other methods)
* **Linear Evaluation** (one layer on top of frozen representation, computationally cheap, dependent of hyperparameters)
* **Fine tuning** (update weights of whole nNN to perform downstream task, computationally expensive, better performance, very dependant of hyperparameters)
* **Cross-Modal Retrieval** (retrieve the closest elements in a different modality e.g. video and text, requires multimodal paired datasets, parameter free)

**Classical methods in SSL:**

1. Contrastive Learning Methods:

The goal is **learning the semantic correspondence between synchronized modalities** (e.g. audio and video), using **contrastive approach**, learn **a joint multimodal space** where embeddings of modalities that occur together are close and far otherwise. Positive and negative samples can be created simply taking video of a source and audio of another source.

Immagine che contiene testo, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.

Example in details: MultiModalVersatile

Immagine che contiene testo, schermata, diagramma, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.Goal is to learn from multiple modalities present in videos: video, audio, text. How to embed the modalities? With a hierarchical space!

**Contrastive Learning: Main Tricks**

* **Backbones:** deeper and wider NN can improve performance a lot, more than with supervised methods
* **Number of negatives and batch size relation**: number of negatives grows quadratically with batch size, share your negatives across workers!
* **Sharding dataset across workers**: make sure each worker works on a different subset of the data to avoid false negatives
* **Preventing Shortcuts**: be careful when using BatchNorm! Sharing the statistics across workers can lead to strong performance gains

Contrastive Learning: Pros and Cons

* **Easy to understand**
* **As a by product**

**BUT**

* **Computationally intensive**
* **Negative choice matters**

1. **Teacher-Student in a Nutshell**

BraVe: an example in details with goal the representation from multiple modalities in videos (video and audio)

Key results:

* + broadening the temporal context is a good signal for self-supervision
  + additional augmentations on the complementary leadt to improved performance
  + negatives are not needed to learn SotA representation
  + SotA on SSL benchmarks

Immagine che contiene schermata, testo, diagramma, design

Il contenuto generato dall'IA potrebbe non essere corretto.

Pros and Cons

* No negatives needed
* Works very well

BUT

* Hard to understand: no notion of global/local optima just equilibria of the dynamic
* Sensitive to hyperparameters

Main tricks:

* Hyperparameters: start from an existing working model
* Augmentation: some have high, some no impact at all
* Small-scale vs Large-scale: try to stick to the largest possible scale of iterations/resolutions/data size
* Backbones: it is worth to use the final backbones as often as possible

1. Clustering Methods in a Nutshell

You want to learn semantic correspondence between synchronized modalities using clustering approach: cluster the feature, use cluster ids as “pseudo labels”

Immagine che contiene testo, linea, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Example in detail: XDC

Pros:

* + - No need for negatives definition
    - Standard classification problem once cluster ids are obtained
    - In offline version, small batch can be used

Cons: