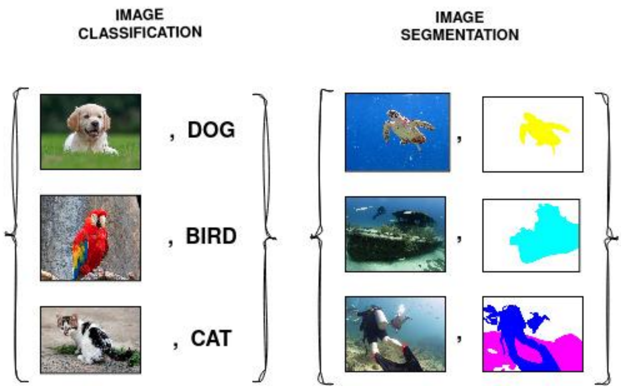
SELF-SUPERVISED LEARNING

FOR COMPUTER VISION APPLICATIONS

SONAL KUMAR, RESEARCH SCHOLAR, CSE DEPARTMENT

SUPERVISION IN DEEP LEARNING

- A DEEP LEARNING MODEL NEEDS SOME KIND OF SUPERVISION TO TRAIN ITSELF.
- IMAGE CLASSIFICATION: IMAGE AND CLASS LABEL PAIR
- SEMANTIC SEGMENTATION: IMAGE AND SEGMENTATION MASK PAIR



DEEP LEARNING METHODS

- SUPERVISED LEARNING —--------> LABELS AVAILABLE
    - SEMI-SUPERVISED LEARNING [SMALL AMOUNT OF LABEL DATA]
    - WEAKLY-SUPERVISED LEARNING [DATA WITH COARSE GRAINED OR INACCURATE LABELS]
- UNSUPERVISED LEARNING —----> LABELS NOT AVAILABLE —----> NO SUPERVISION
    - SELF-SUPERVISED LEARNING [GENERATE SUPERVISORY SIGNAL FROM UNLABELED DATA]

WHY SELF-SUPERVISED LEARNING ?

- COLLECTION AND ANNOTATION OF LARGE-SCALE DATASETS ARE TIME CONSUMING AND EXPENSIVE.
- IN REAL TIME APPLICATIONS LARGE-SCALE LABEL DATASET MAY NOT BE AVAILABLE. FOR EXAMPLE: DEEP SEA SPECIES CLASSIFICATION
- THE DARK MATTER OF ARTIFICIAL INTELLIGENCE

> "Supervised learning is a bottleneck for building more intelligent generalist models that can do multiple tasks and acquire new skills without massive amounts of labeled data."
>
> **-Yann LeCun, Ishan Misra (Facebook AI)**

SELF-SUPERVISED LEARNING (SSL)

- DEFINITION

An unsupervised way of training a deep learning model, which generate the supervisory single from the unlabeled image/video dataset itself.

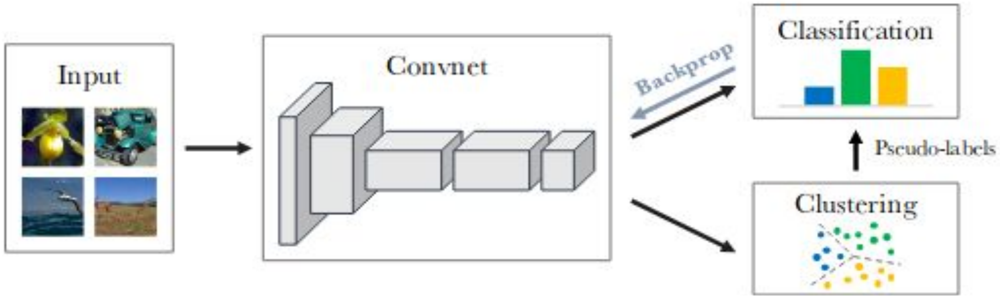
It considers the supervisory signal as one of the properties of the unlabeled dataset.

SSL CATEGORIZATION

- BASED ON THE TRAINING SCHEME
    - END-TO-END SSL —-------------------------> TASK-SPECIFIC LEARNING
    - GENERALISED / TWO-STEP SSL—------------------------> TRANSFER LEARNING

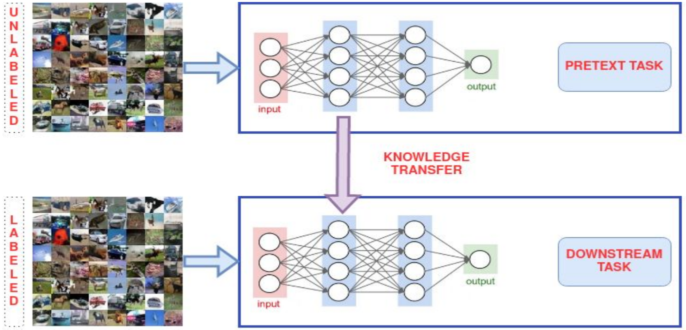
END-TO-END SSL

- LEARNS FEATURE REPRESENTATIONS AND LABELS SIMULTANEOUSLY.
  - IMAGE CLASSIFICATION TASK:
    - EXAMPLE: DEEP CLUSTERING ——-----> JOINTLY LEARNS NETWORK PARAMETERS AND CLUSTER ASSIGNMENTS.
  - SEMANTIC SEGMENTATION TASK:
    - EXAMPLE: SegSort ——------------------------>JOINTLY LEARNS PIXEL-WISE EMBEDDING AND CLUSTERING



GENERALIZED SSL

- SOLVE A PREDEFINED TASK (CALLED PRETEXT TASK) TO LEARN VISUAL FEATURES (TRAIN A CNN BACKBONE) ——------—> REPRESENTATION LEARNING



- PRETEXT TASKS CAN BE PREDICTIVE, GENERATIVE, CONTRASTIVE OR COMBINATION OF THEM.
- DOWNSTREAM TASKS ARE COMPUTER VISION APPLICATIONS LIKE IMAGE CLASSIFICATION, IMAGE CLUSTERING, SEMANTIC SEGMENTATION, OBJECT DETECTION, IMAGE RETRIEVAL, DEPTH ESTIMATION, KEY POINT DETECTION. ETC.

- NEXT, TRANSFER LEARNED VISUAL FEATURES AS PRE-TRAINED MODEL TO DOWNSTREAM TASKS ——-----------------------------—> TRANSFER LEARNING

FEW EXAMPLES OF PRETEXT TASK

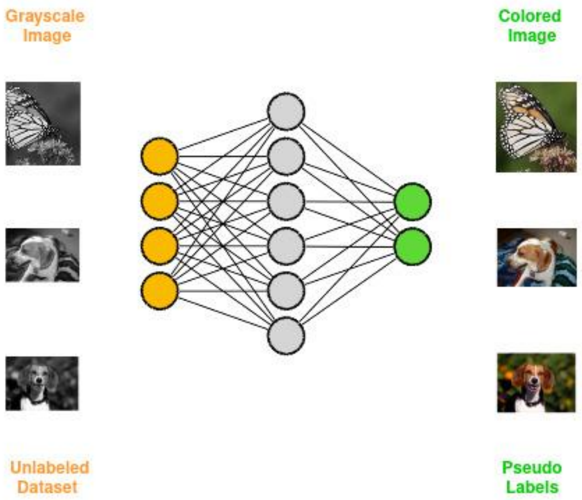
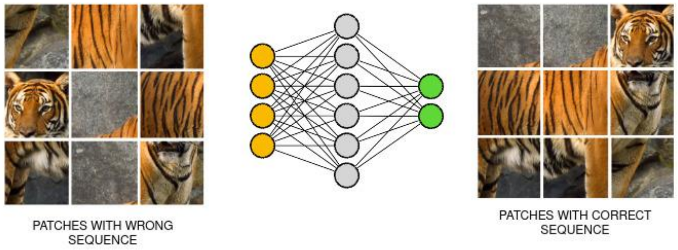
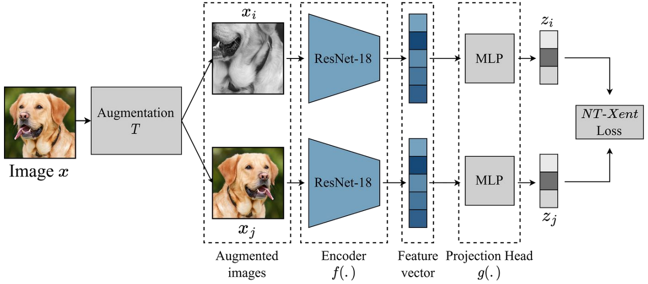


IMAGE COLORIZATION



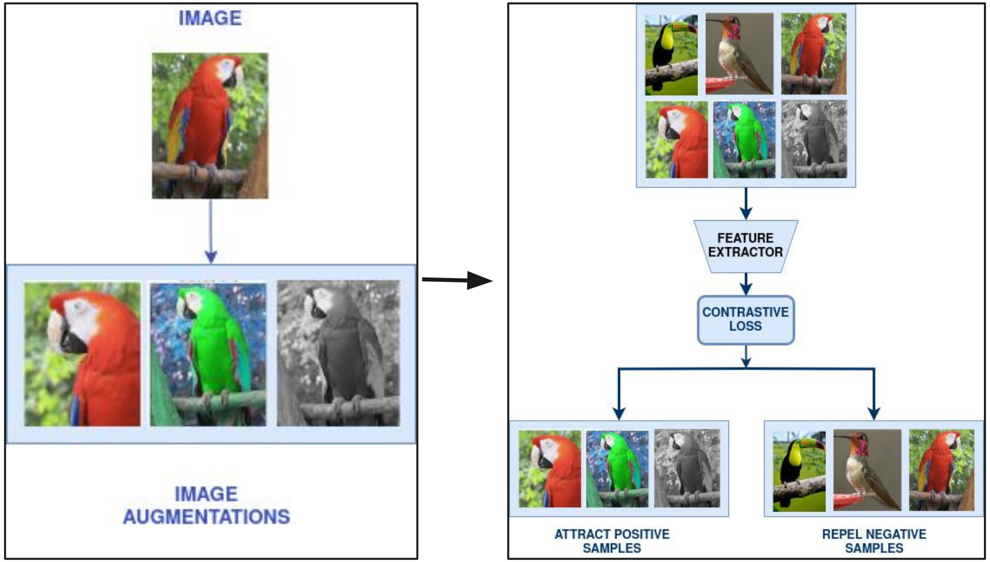
JIGSAW PUZZLE



CONTRASTIVE LEARNING

CONTRASTIVE LEARNING

- WHAT IS CONTRASTIVE LEARNING ?



SELF-SUPERVISED CONTRASTIVE LOSS

$i \in I := \{1, \ldots, 2N\} \rightarrow$ index of an arbitrary augmented sample

$j(i) \rightarrow$ index of the other augmented sample originating from the same source sample

$$\mathcal{L}^{self} = \sum_{i \in I} \mathcal{L}_i^{self} = -\sum_{i \in I} \log \frac{\exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_{j(i)} / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \bullet \boldsymbol{z}_a / \tau\right)}$$

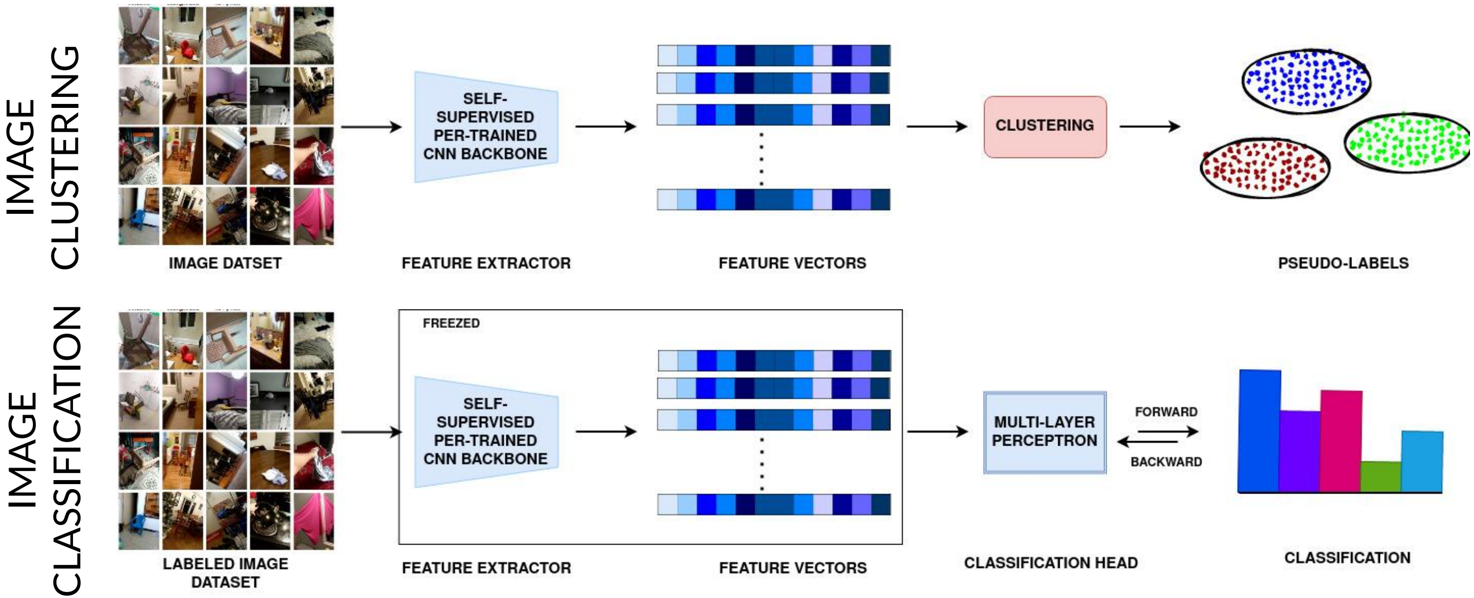
$$z_\ell = \mathrm{Proj}(\mathrm{Enc}(\tilde{x}_\ell)) \in \mathbb{R}^{D_P}$$

$$A(i) = I \setminus \{i\}$$

$i \rightarrow$ anchor, $j(i) \rightarrow$ positive, $k \in A(i) \setminus \{j(i)\} \rightarrow$ negatives

one positive and $2(N-1)$ negatives

FEW EXAMPLES OF DOWNSTREAM TASK

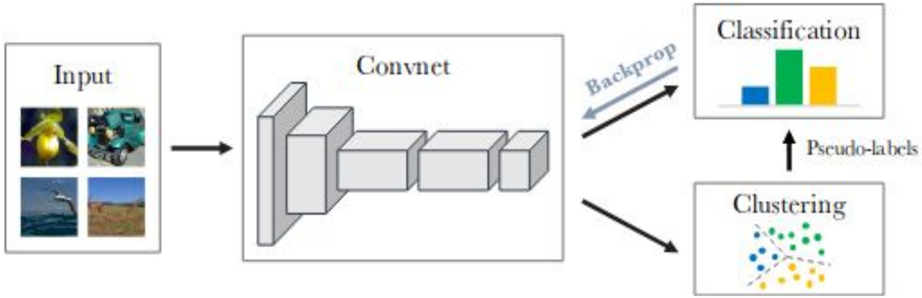


EVALUATION OF SELF-SUPERVISED LEARNING

- NEAREST NEIGHBOUR RETRIEVAL
- PERFORMANCE OF DOWNSTREAM TASK
  - IMAGE CLASSIFICATION: CLASSIFICATION ACCURACY
  - IMAGE CLUSTERING: CLASSIFICATION ACCURACY, CLUSTERING QUALITY (ARI, NMI), ETC.
  - SEMANTIC SEGMENTATION: MiOU, PIXEL ACCURACY, F1 SCORE, ETC.

DEEP CLUSTERING

- OBJECTIVE: To show that it is possible to obtain useful general purpose visual features with a clustering framework.
- KEY OBSERVATION: A multilayer perceptron classifier on top of the last convolutional layer of a random AlexNet achieves 12% in accuracy on ImageNet while the chance is at 0.1%.
- PROBLEM STATEMENT: Given {x1,x2,x3...,xN } as training set of N images, find parameters θ^* such that the convnet mapping f_{θ^*} produces good general-purpose features representations.



- K-means clustering: Groups the features f_{θ^*} into k distinct groups.

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

  - $C \in R^{d \times k}$: Centroid matrix
  - y_n : cluster assignment of each image.

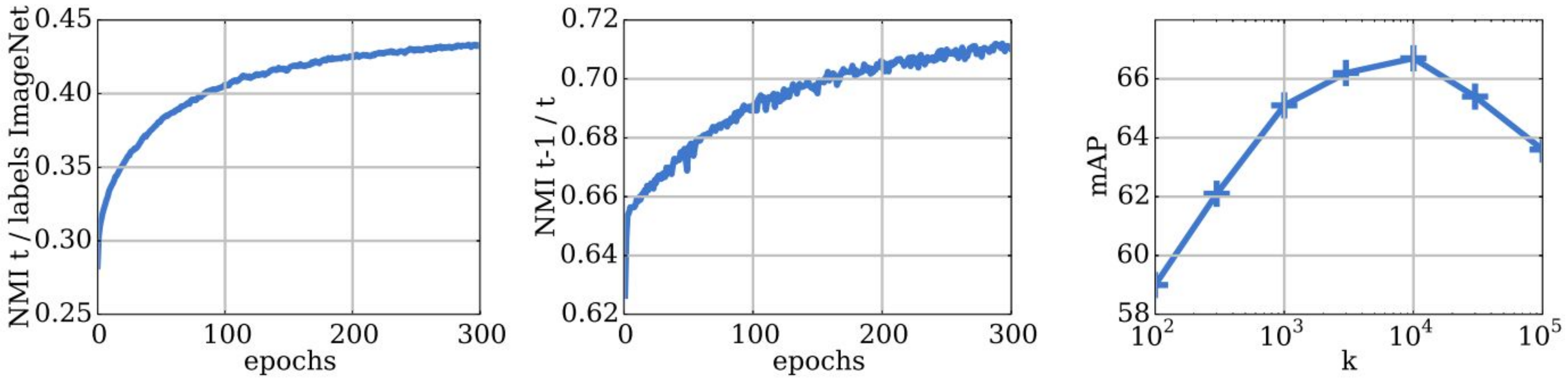
- Network parameters (θ^*) and classifier parameters (***W***) update:

$$\min_{\theta, W} \frac{1}{N} \sum_{n=1}^{N} \ell \left(g_W \left(f_\theta(x_n) \right), y_n \right) \qquad (1)$$

- Alternates between clustering the features to produce pseudo-labels using Eq. (2) and updating the parameters by predicting pseudo-labels using Eq. (1)
- ISSUES: Empty Clusters, Unbalanced Cluster

Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

DEEP CLUSTERING PERFORMANCE



(a) Clustering quality (b) Cluster reassignment (c) Influence of k

Pascal VOC transfer tasks		Classification		Detection		Segmentation	
Method	Training set	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

Pascal VOC 2007 object detection

Method	AlexNet	VGG-16
ImageNet labels	56.8	67.3
Random	47.8	39.7
Doersch *et al.* [13]	51.1	61.5
Wang and Gupta [63]	47.2	60.2
Wang *et al.* [64]	–	63.2
DeepCluster	**55.4**	**65.9**

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

Method	Oxford5K	Paris6K
ImageNet labels	72.4	81.5
Random	6.9	22.0
Doersch *et al.* [13]	35.4	53.1
Wang *et al.* [64]	42.3	58.0
DeepCluster	**61.0**	**72.0**

REFERENCES

- Jing, L. and Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *43*(11), pp.4037-4058.
- Caron, M., Bojanowski, P., Joulin, A. and Douze, M., 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 132-149).
- Self-supervised learning: The dark matter of intelligence

THANK YOU