IPCL: ITERATIVE PSEUDO-SUPERVISED CONTRASTIVE LEARNING TO IMPROVE SELF-SUPERVISED FEATURE REPRESENTATION

Sonal Kumar
k.sonal@iitg.ac.in

Anurudh Phukan
aphukan@alumni@iitg.ac.in

Arijit Sur
arijit@iitg.ac.in

Multi Media Lab, Department of CSE, IIT Guwahati

IEEE ICASSP 2024, Seoul, Korea

Outline

Background

Self-supervised Learning (SSL)

- **Unsupervised way of training** a deep learning model.
- **Generate supervisory single** from unlabeled image/video dataset [1].
- **Supervisory signal** is one of the **properties of the unlabeled dataset**.
- **The dark matter of intelligence** [2].

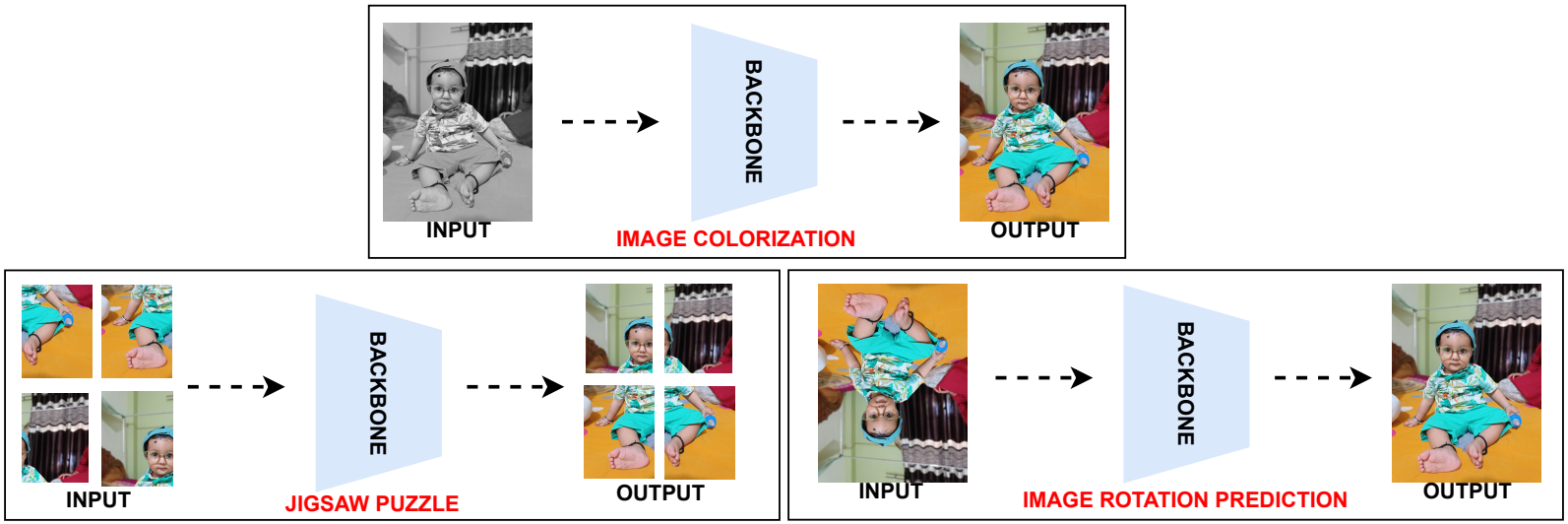


Figure 1: For **self-supervised training**, a series of **pretext tasks** are explored in the early stage.

SSL Framework

- Provides a strong baseline for various **downstream computer vision applications**.
- Performance of downstream tasks is **proportional** to the quality of the self-supervised feature representation [3].

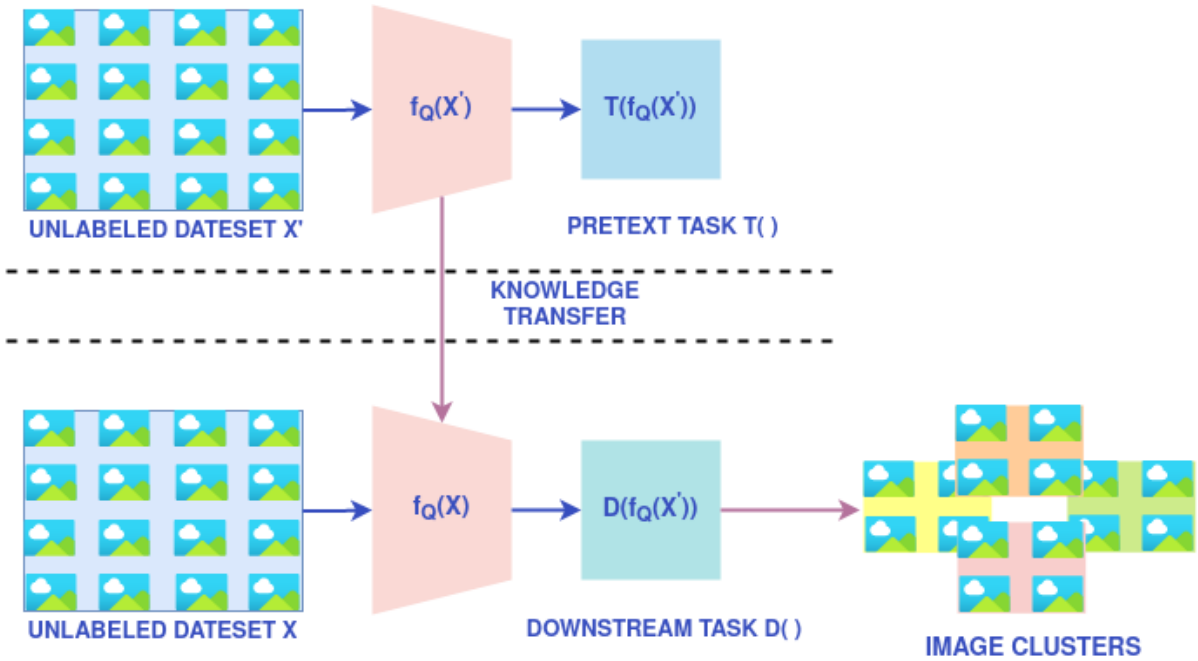


Figure 2: General Overview of SSL Framework.

Existing Works

- **Photometric and geometric prediction-based** pretext tasks **fail to learn** discriminative feature representations.
- Quality of feature representation is **further improved** by introducing a batch contrastive approach [4-6].

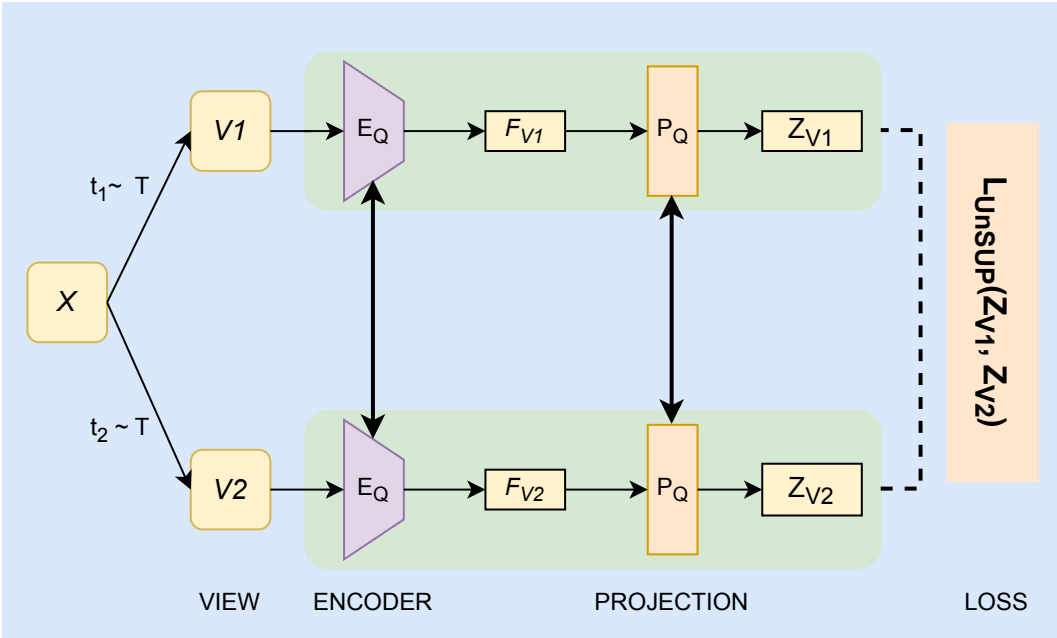


Figure 3: A General Batch Contrastive approach-based framework for Self-supervised Training (SimCLR).

Limitations

- Existing unsupervised contrastive batch approaches heavily depend on **data augmentation strategy**.

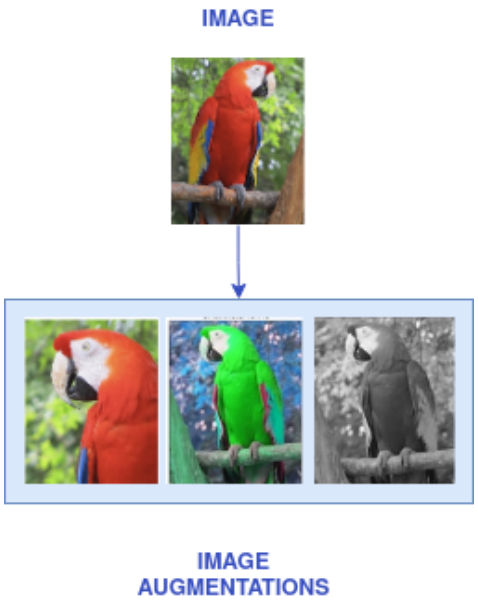


Figure 4: Generate **augmented versions** (positive pairs) of an image.

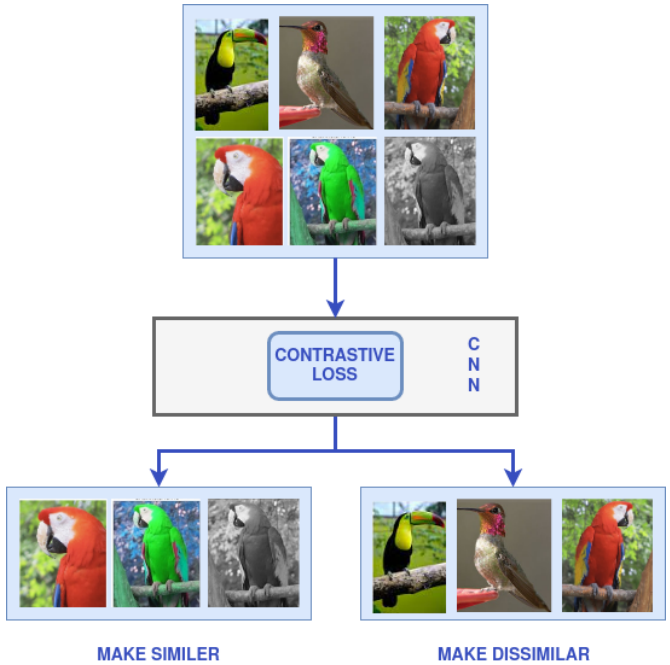
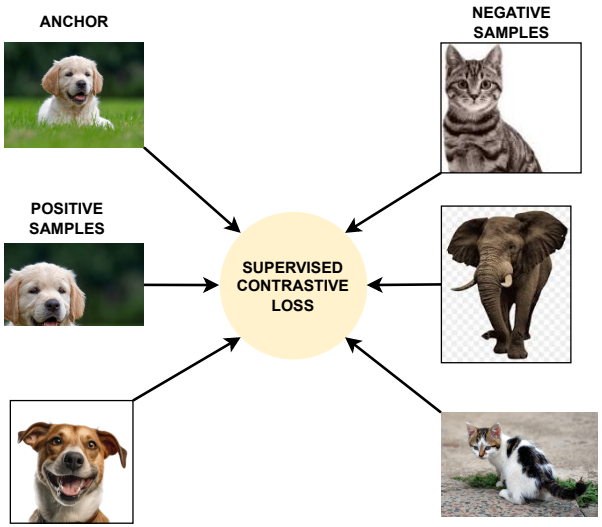


Figure 5: Attract positive pairs and repel negative pairs

Motivation

- Paper [7] study the **effect of replacing image augmentations** with **sample images from same class** in supervised setting.



- We argue that introducing intra-class variance in a batch contrastive approach can **further improve the quality** of self-supervised feature representation.
- Challenges in unsupervised setting:
  1. **Class-level information is not available.**
  2. **How to utilize the class-level information to converge the training?**

Proposed Architecture

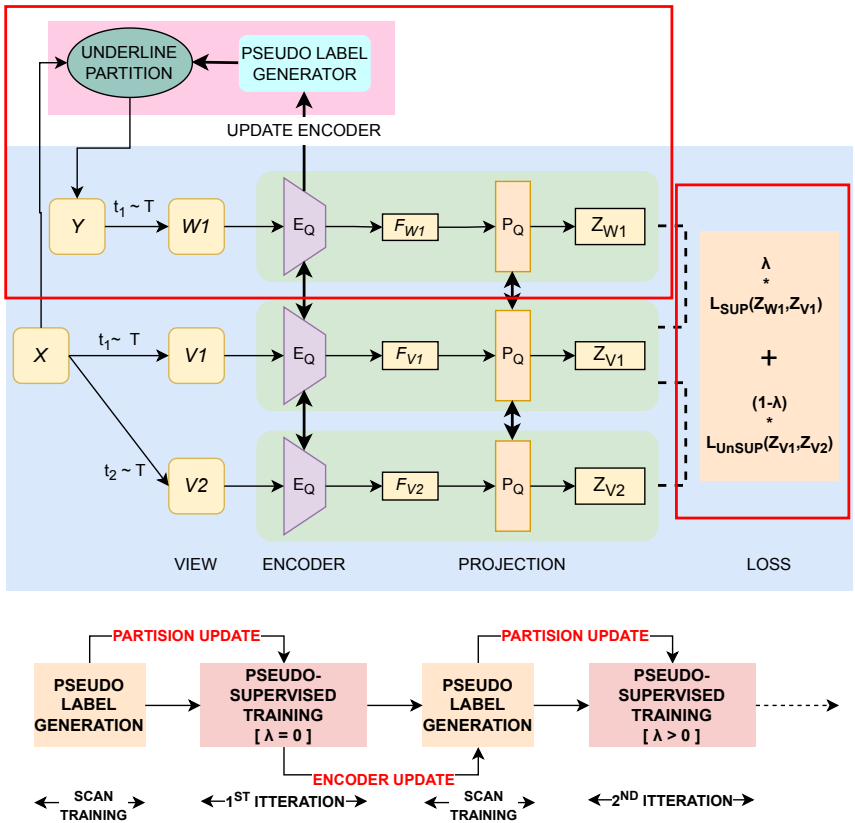


Figure 6: A high-level view of the proposed method, i.e., IPCL.

Major Contributions

- Pseudo-label Generator: Iteratively discovers the underlying partition of the unlabeled dataset in an unsupervised setting.
  - Utilizes **Semantic Clustering by Adopting Nearest Neighbors** (SCAN) [9] to generate pseudo labels.
  - Encoder's parameter of the SCAN framework is **updated with** the trained encoder from the previous training iteration.
- Pseudo-supervised Contrastive Loss ($\mathcal{L}_{\text{IPCL}}$): a convex combination of self-supervised and supervised contrastive loss to capture the benefits of both augmentations and pseudo-class information.

$$\mathcal{L}_{\text{IPCL}} = (1-\lambda)\mathcal{L}_{\text{UnSup}} + \lambda\mathcal{L}_{\text{Sup}} \tag{1}$$

  - λ is a weight factor determining the relative importance of the supervised and unsupervised losses

$\mathcal{L}_{\text{UnSup}}$ & $\mathcal{L}_{\text{Sup}}$

- **Self-supervised Contrastive Loss** ($\mathcal{L}_{\text{UnSup}}$) [4]: Utilizes the image augmentations to sample positive and negative pairs.

$$\mathcal{L}_{\text{UnSup}} = -\sum_{i=1}^{2N} \log \frac{\exp\left(\text{sim}(Z_i, Z_{i+N}) \,/\, \tau\right)}{\sum_{j=1, j\neq i}^{2N} \exp\left(\text{sim}(Z_i, Z_j) \,/\, \tau\right)} \tag{2}$$

  - Z_i and Z_{i+N} are feature representations of augmented views of an image, $2N$ is the multi-view batch size, τ is a temperature parameter, $\text{sim}(.,.)$ is a similarity function.

- **Supervised Contrastive Loss** ($\mathcal{L}_{\text{Sup}}$) [7]: Utilizes the pseudo-level information to sample positive and negative pairs.

$$\mathcal{L}_{\text{Sup}} = \sum_{i=1}^{2N} \frac{-1}{|A(i)|} \sum_{a\in A(i)} \log \frac{\exp\left(\text{sim}(Z_i, Z_a) \,/\, \tau\right)}{\sum_{b\in B(i)} \exp\left(\text{sim}(Z_i, Z_b) \,/\, \tau\right)} \tag{3}$$

  - A(i) and B(i) are the set of view indices from the same and different pseudo-class as i_{th} view in a multi-view batch of size 2N, respectively.

IPCL Algorithm.

Input D: dataset, E_Q: encoder, totalEpochs: total number of epochs, pretrainCount: number of epochs for pre-training, stepLength: number of epochs for an iteration, λ: weight factor, T: augmentation set
Initialize Encoder with random weights, nextIteration = pretrainCount
for $e = 0$ to totalEpochs **do**
    Sample mini-batch of size N from D
    Sample two transformations from T and obtain corresponding views of each image(2N total)
    if $e <$ *pretrainCount* **then**
        Train E_Q by setting $\lambda = 0$ in equation 1
    else
        if e == (nextIteration - 1) **then**
            Update encoder of pseudo-label generator
            Set nextIteration = nextIteration + stepLength
        end if
        Generate pseudo-labels for 2N.
        Train E_Q by setting λ to pre-defined non-zero value in equation 1
    end if
end for
Output Encoder

Experimental Settings

- Datasets:
  - **CIFAR-10** [9]: 50,000 training and 10,000 testing 32x32 color images, 10 classes, 6,000 images per category.
  - **STL-10** [10]: 100,000 training images and 8,000 test 96x96 color images, 10 classes.
- Architecture:
  - ResNet18 base encoder as a feature extractor.
  - A 2-layer MLP projection head Yield 128-dimensional feature vectors.
- Hyper-parameters:
  - Temperature coefficient of 0.1.
  - Stochastic gradient descent (SGD) optimizer (momentum $= 0.9$ and learning rate $= 0.4$) and a cosine learning rate scheduler.
  - Batch size of 512.

Metrics & Results

- Evaluate the effectiveness of the learned representations through:
  - K-nearest neighbors (KNN) assessment, and
  - Unsupervised classification (UnCls) [8] as a **downstream task**.

Table 1: Comparison of IPCL and SimCLR for Top-K NN precision over CIFAR-10 and STL-10 datasets

Dataset	Method	Top-1	Top-5	Top-20
CIFAR-10	SimCLR	83.33	79.98	75.23
	IPCL[100, 0.05]	**84.77**	**82.20**	**78.84**
STL-10	SimCLR	85.02	81.92	77.39
	IPCL[-, 0.025]	**85.55**	**82.62**	**78.73**

Here, IPCL[X, Y], with X representing the parameter *stepLength* and Y representing the parameter λ.

Downstream Results

- Compute the **cluster quality** using **Cluster Accuracy (Acc)**, **Normalized Random Index (NMI)**, and **Adjusted Random Index (ARI)** metrics.

Table 2: Results for downstream unsupervised classification(UnCls) task. Here, 'SL' denotes the method followed by the self-labeling (SL) step.

Dataset	Method	UnCls	Acc	NMI	ARI
CIFAR-10	SimCLR	k-means	65.9	59.8	50.9
	SimCLR	SCAN	81.86	72.64	66.66
	IPCL	SCAN	**84.44**	**74.24**	**69.78**
	SimCLR	SCAN+SL	87.96	79.11	76.44
	IPCL	SCAN+SL	**88.81**	**81.02**	**77.95**
STL-10	SimCLR	k-means	65.8	60.4	50.6
	SimCLR	SCAN	79.04	67.84	62.66
	IPCL	SCAN	**79.33**	**69.01**	**63.84**
	SimCLR	SCAN+SL	80.43	69.86	64.82
	IPCL	SCAN+SL	**80.91**	**70.97**	**66.12**

Ablation Study

Table 3: Ablation results with CIFER-10 dataset.

Method	Top-1	Top-5	Top-20
SimCLR	83.33	79.98	75.23
IPCL[-, 1.0]	72.74	72.69	72.53
IPCL[-, 0.1]	82.92	80.61	**77.73**
IPCL[-, 0.05]	**83.55**	**80.86**	77.56

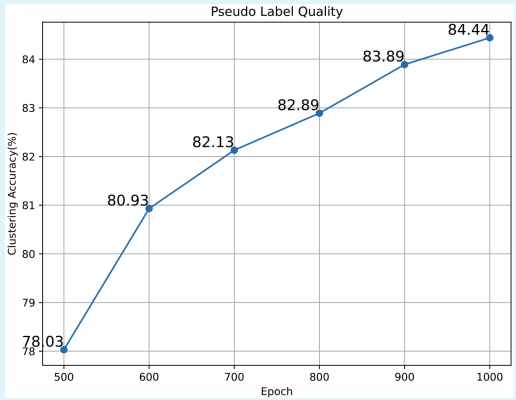
(a) **Effect of λ** on **Top-K NN precision**

Method	Top-1	Top-5	Top-20
SimCLR	83.33	79.98	75.23
IPCL[-, 0.1]	82.92	80.61	77.73
IPCL[250, 0.1]	83.76	81.42	78.32
IPCL[-, 0.05]	83.55	80.86	77.56
IPCL[100, 0.05]	**84.77**	**82.20**	**78.84**

(b) **Effect of stepLength** on **Top-K NN precision**

Method	Top-1	Top-5	Top-20
SimCLR	83.33	79.98	75.23
IPCL[-, 0.05](10 epochs)	**83.98**	80.84	77.27
IPCL[-, 0.05](50 epochs)	83.55	**80.86**	**77.56**

(c) **Effect of SCAN training time** on **Top-K NN precision**



(d) **Evolution of Clustering Accuracy (Acc.)** with the **number of epochs**.

Conclusion

- Proposes a novel self-supervised learning method, IPCL, to improve visual representation learning.
- Iteratively enhances the visual representation with a pseudo-level generator and novel pseudo-supervised contrastive loss function.
- Experiment on multiple datasets to show that the proposed method outperforms the baseline self-supervised method.
- Perform a detailed ablation study to understand the influence of parameters and sub-modules.

References

1. Jing, Longlong, and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey." IEEE transactions on pattern analysis and machine intelligence 43.11 (2020): 4037-4058.
2. LeCun, Yann, and Ishan Misra. "Self-Supervised Learning: The Dark Matter of Intelligence." AI Meta, ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/, Accessed 8 April 2024.
3. E. Zheltonozhskii, C. Baskin, A. M. Bronstein, and A. Mendelson, "Self-supervised learning for large-scale unsupervised image clustering," arXiv preprint arXiv:2008.10312, 2020.
4. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in International con- ference on machine learning. PMLR, 2020, pp. 1597–1607.
5. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.
6. Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," Advances in Neural Information Processing Systems, vol. 33, pp. 9912–9924, 2020.
7. Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," Advances in Neural Information Processing Systems, vol. 33, pp. 18661–18673, 2020.
8. Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool, "Scan: Learning to classify images without labels," in European Conference on Computer Vision. Springer, 2020, pp. 268–285.
9. Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.
10. Adam Coates, Andrew Ng, and Honglak Lee, "An analysis of single-layer networks in unsupervised feature learning," in Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.

Q&A

THANK YOU

GitHub Link