Reviewing Research conducted on Self-Supervised Learning frameworks on image modality

Kareena Beniwal (20110095), Priya Gupta (20110147), Prajwal Singh, Prof Shanmuganathan Raman

CVIG Lab , Indian Institute of Technology Gandhinagar, India

Background and Motivation

- **Data Abundance but Label Scarcity**: Availability of large no of unlabeled image datasets has led to increased demand for more effective ways to process and understand this data.
- Inspired by breakthroughs in natural language processing, discriminative self-supervised learning methods have rapidly advanced the field of computer vision.
- The limitations of supervised learning became increasingly evident as the incremental improvements traditionally gained through these methods began to plateau, underscoring the need for innovative approaches like self-supervised learning in computer vision.
- Idea behind **Self Supervised Learning (SSL)**: Devise an experimental setting in which the task that provides the supervisory signal can be solved without human annotation and then train DNNs to solve it.

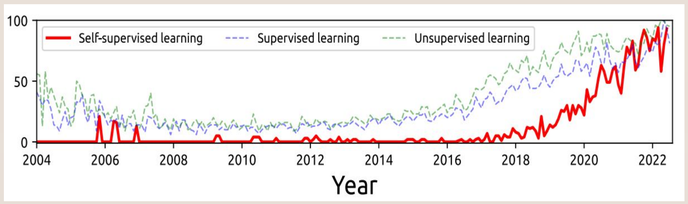
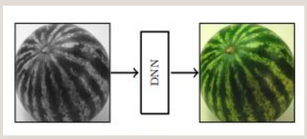


Fig: Interest over time for different learning approaches

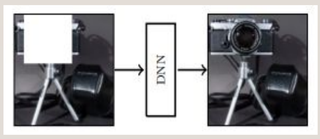
Generative & Discriminative Approaches

Pretext tasks: auxiliary tasks or designed to generate supervisory signals or labels from the input data itself.

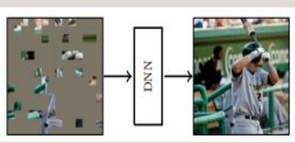
- Encourage models to learn rich and meaningful feature representations from raw data.
- Transfer Learning and Adaptability.



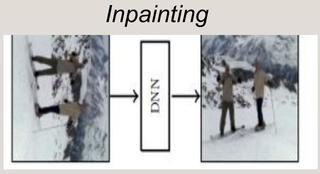
Colorization



Inpainting



Masked Image Modeling



Geometric transformation

Fig: various image-based pretext tasks for self-supervised learning

Generative SSL methods:
Raw data is directly processed and transformed at the pixel level during training. Learn the underlying patterns in the data by generating new data samples that closely resemble the original dataset
Advantages: They learn rich representations and handle complex data distributions effectively.
Ex: Autoencoders, Generative Adversarial Networks (GANs).

Discriminative SSL methods:
learn to discriminate between different classes or features.
Advantages: They improve performance in downstream tasks by learning task-specific features.
Ex: Contrastive learning, Triplet loss, Siamese networks.

Dimensionality Reduction

The approach highlights four key attributes:

- The method learns a globally coherent ***nonlinear function*** that maps high-dimensional data to a low-dimensional manifold. Focuses on preserving neighborhood relationships without relying on a pre-defined distance metric in the input space.
- Learn mappings that are invariant to complex transformations of the inputs such as lighting changes and geometric distortions.
- Utilizes a *contrastive loss function* that encourages similar points to come closer and dissimilar points to move apart.
- The learned function can effectively map new samples that were not part of the training set, without requiring prior information.

```
layer1: Conv2d(1, 15, kernel_size=(6, 6), stride=(1, 1), padding=(2, 2))
pool: MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
layer2: Conv2d(15, 15, kernel_size=(5, 5), stride=(1, 1))
layer3: Conv2d(15, 30, kernel_size=(5, 5), stride=(1, 1))
fc: Linear(in_features=30, out_features=2, bias=True)
```

Fig: Implemented CNN architecture

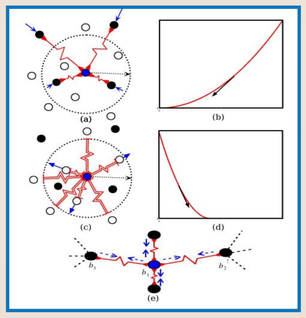


Fig: (a) Shows the points connected to similar points with attract-only springs. (b) The loss function and gradient of similar pairs. (c) The point connected to dissimilar points with m-repulse-only springs. (d) The loss function and gradient of dissimilar pairs. (e) A point is pulled by other points in different directions.

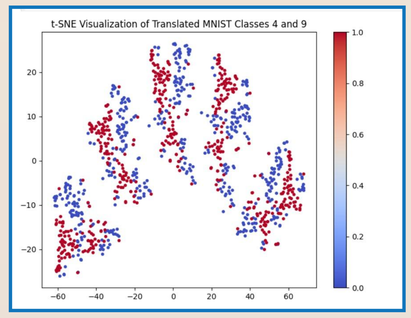


Fig: This fig shows the impact of a basic distance-based mapping technique on MNIST data, incorporating horizontal translations of -6, -3, +3, and +6 pixels. Due to the translations, the samples are spread apart, resulting in five distinct clusters representing each translation. It's noteworthy that while the clusters are separated, they maintain internal organization. These findings are based on test samples.

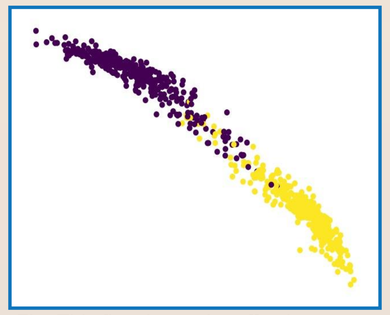


Fig: This fig. shows the effectiveness of DrLIM in learning a mapping from high-dimensional, shifted digit images to a 2D manifold. The results presented focus on test samples, where the relationships among neighboring samples are not pre-defined. Despite variations in shift, similar characters are consistently positioned in close proximity.

Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks

Aim: Generic feature learning and present an approach for training a convolutional network using only unlabeled data.

Outcomes of this training approach are a set of features that:

- are invariant to transformations.
- retain discriminative power, allowing them to distinguish between different images and their transformed versions.
- are applicable in class-independent scenarios, such as descriptor matching tasks.

Methodology	Accuracy (%)
Exemplar CNN (paper)	74.2 ± 0.4
Exemplar CNN (ours)	42.10

Name	Description	Output size
input	32 × 32 RGB image	3 × 32 × 32
conv1	92 filters, 5 × 5, pad = 2, ReLU	92 × 32 × 32
pool1	Maxpool 3 × 3 pixels, stride = 2	92 × 15 × 15
drop1	Dropout, $p = 0.25$	92 × 15 × 15
conv2	256 filters, 5 × 5, pad = 2, ReLU	256 × 15 × 15
pool2	Maxpool 3 × 3 pixels, stride = 2	256 × 7 × 7
drop2	Dropout, $p = 0.25$	256 × 7 × 7
conv3	512 filters, 5 × 5, pad = 2, ReLU	512 × 7 × 7
drop3	Dropout, $p = 0.25$	512 × 7 × 7
	Flatten 512 × 7 × 7 → 25088	25088
dense1	Fully connected 25088 → 1024	1024
dropDense1	Dropout, $p = 0.5$	1024
dense2	Fully connected 1024 → 16000	16000
output	Softmax	16000

Fig: Architecture of the Exemplar CNN

Limitations of the Exemplar CNN:
- Computational Overhead
- Scalability with Respect to Number of Classes
- Dependency on Patch Extraction Strategy

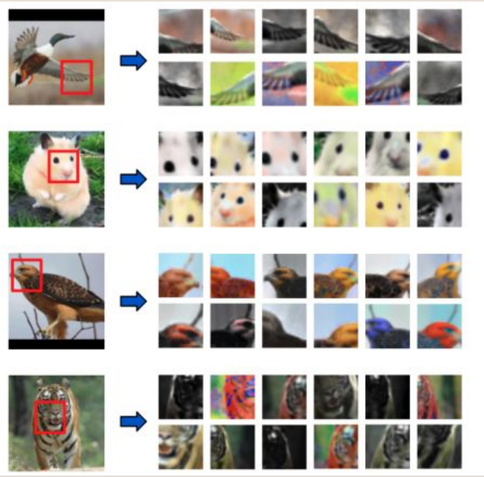


Fig: Examples of random transformations applied to the patches (seed) extracted from the unlabeled set of STL-10.

SimCLR : A Simple Framework for Contrastive Learning of Visual Representations

The framework shows that:

- Unsupervised contrastive learning benefits from stronger data augmentation than supervised learning.
- Applying a simple nonlinear transformation to representations before calculating the contrastive loss significantly enhances the quality of the learned representations.
- Contrastive learning benefits from larger batch sizes and longer training compared to its supervised counterpart. Like supervised learning, contrastive learning benefits from deeper and wider networks.

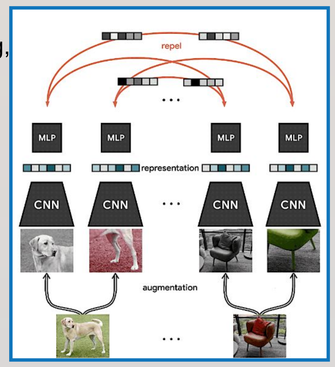


Fig: A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations (t ~ T and t 0 ~ T) and applied to each data example to obtain two correlated views. A base encoder network f(·) and a projection head g(·) are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head g(·) and use encoder f(·) and representation h for downstream tasks.

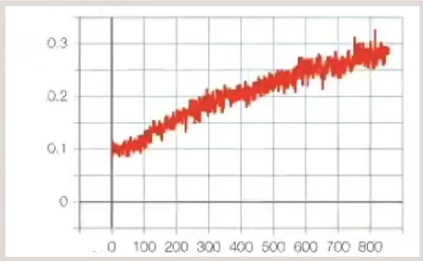


Fig: MLP Accuracy

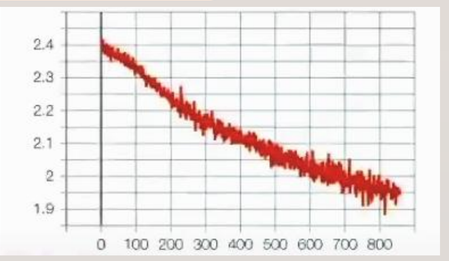


Fig: MLP Loss

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

Fig: Loss function for a positive pair of examples (i, j)

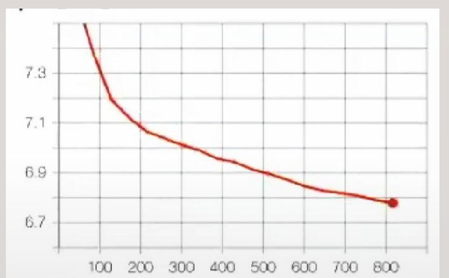


Fig: Training Loss

References

[1] Özbulak, U., Lee, H. J., Boga, B., Anzaku, E. T., Park, H., Van Messem, A., De Neve, W., & Vankerschaver, J. (2023). Know your Self-supervised Learning: A survey on image-based Generative and Discriminative training. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.13689.
[2] R. Hadsell, S. Chopra and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 2006, pp. 1735-1742, doi: 10.1109/CVPR.2006.100.
[3] Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., & Brox, T. (2014). Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1406.6909.
[4] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2002.05709.

github repo