Complexity of Linear Regions in Self-supervised Deep ReLU Networks

Supplementary Material

Introduction

This supplementary section provides additional details about our experimental setup. We outline the full hyperparameter configurations used for each model across both benchmark datasets. We further elaborate on the high-dimensional projection strategy employed. We also expand on the high-dimensional projection strategy adopted. Finally, we include additional illustrations of the results from the FashionMNIST dataset and include supplementary visualisations of the geometric structure of linear regions at the final training epoch for all models.

Hyperparameters

Table 5. Training hyperparameter configurations for each model on MNIST dataset.

Models	Epochs	Batch Size	Learning Rate	Weight Decay	Temperature
Classifier	100 for each	256	0.01	0.0	-
Triplet loss		256	0.02	1.0e-5 for each	0.05
SupCon		256	0.01		0.05
SimCLR		256	0.01		0.2
MoCo		64	0.008		0.2
Simsiam		256	0.04		-
BYOL		256	0.04		-

Table 6. Training hyperparameter configurations for each model on FashionMNIST dataset.

Models	Epochs	Batch Size	Learning Rate	Weight Decay	Temperature
Classifier	100 for each	256	0.02	0.0	-
Triplet loss		256	0.04	1.0e-5 for each	0.05
SupCon		256	0.01		0.05
SimCLR		256	0.01		0.2
MoCo		64	0.01		0.2
Simsiam		256	0.04		-
BYOL		256	0.04		-

High-dimensional Projection

We describe the process used to project a two-dimensional point to a high-dimensional space [30, 33]. Let x_0, x_1 and $x_2 \in \mathbb{R}^d$ be three sampled input images. We construct:

1. the circumcenter of the triangle produced by the three points,
2. the orthonormal basis vector of the plane,

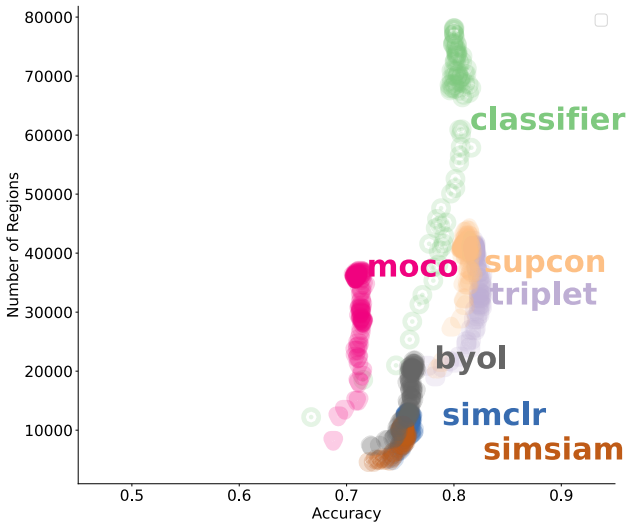

Figure 9. Number of regions and accuracy achieved by supervised and self-supervised methods on FashionMNIST dataset. The opacity reflects the progression of training over epochs. Low-opacity points depict results achieved early in training, while darker points represent results achieved later.

3. an affine mapping that transforms high-dimensional coordinates into the two-dimensional input space.

Let $\mathbf{v}_1 = x_1 - x_0$ and $\mathbf{v}_2 = x_2 - x_0$. The circumcenter C of the triangle formed by x_0, x_1 and x_2 can be expressed as:

$$C = x_0 + \beta_0 \mathbf{v}_1 + \beta_1 \mathbf{v}_2 \tag{9}$$

where $\beta_0, \beta_1 \in \mathbb{R}$ are the coefficients that satisfy the linear system

$$M \begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{2} b \tag{10}$$

such that

$$M = \begin{bmatrix} \mathbf{v}_1^T \mathbf{v}_1 & \mathbf{v}_1^T \mathbf{v}_2 \\ \mathbf{v}_1^T \mathbf{v}_2 & \mathbf{v}_2^T \mathbf{v}_2 \end{bmatrix}, \; b = \frac{1}{2} \begin{bmatrix} \|\mathbf{v}_1\|^2 \\ \|\mathbf{v}_2\|^2 \end{bmatrix} \tag{11}$$

solving the linear system gives the circumcenter. Next we use the Gram–Schmidt process to obtain the orthonormal basis vectors. Given the center vectors $w_1 = x_1 - C$ and $w_2 = x_2 - C$. If $u_1 = \frac{w_1}{\|w_1\|}$ is the normalized vector then:

$$w_2^{\perp} = w_2 - \left(\frac{w_2^{\top} u_1}{u_1^{\top} u^1} \right) u_1 \tag{12}$$

points in the perpendicular direction. Let $u_2 = \frac{w_2^{\perp}}{w_2^{\perp}}$ then the orthonomal basis vector is defined as $\{u_1, u_2\} \subseteq \mathbb{R}^d$.

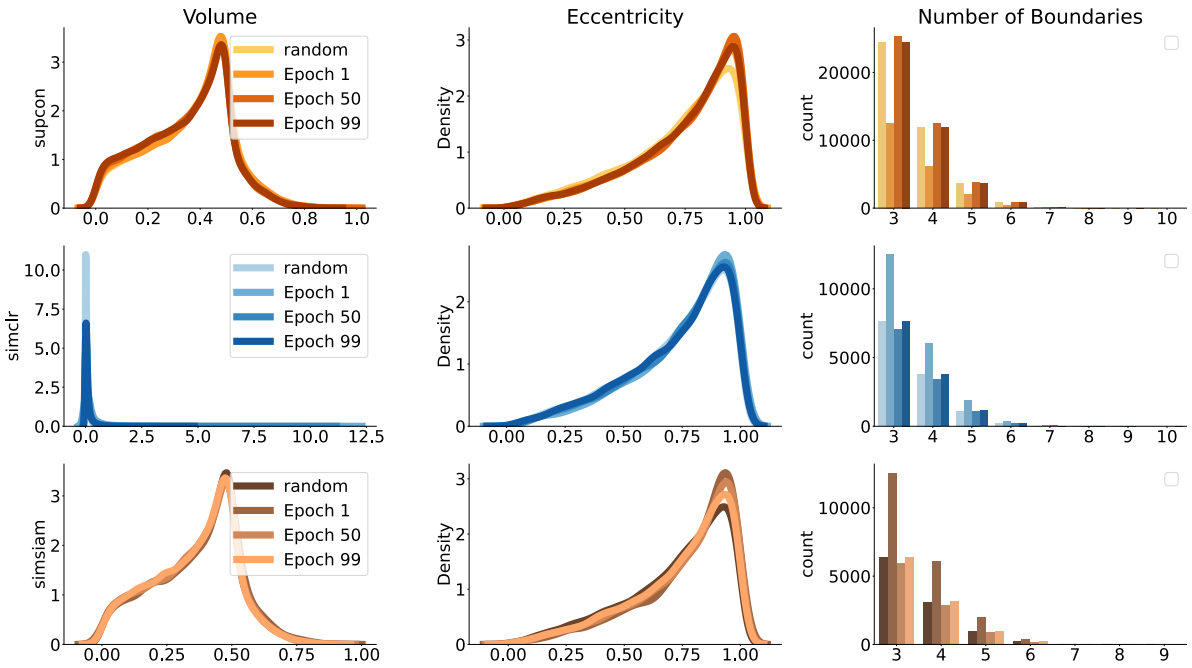

Figure 10. Evolution of region area, eccentricity, and boundary count distributions across epochs for SupCon, SimCLR, and SimSiam on FashionMNIST dataset.

The input x can be obtained from the point $[a, b]$ in the two-dimensional space by setting

$$x(a, b) = C + au_1 + bu_2 \tag{13}$$

similarly

$$x = \begin{bmatrix} u_1 & u_2 & C \end{bmatrix} \begin{bmatrix} a \\ b \\ 1 \end{bmatrix} \tag{14}$$

such that $T = \begin{bmatrix} u_1 & u_2 & C \end{bmatrix}$ is the transformation matrix from a two-dimensional space to a high-dimensional space.

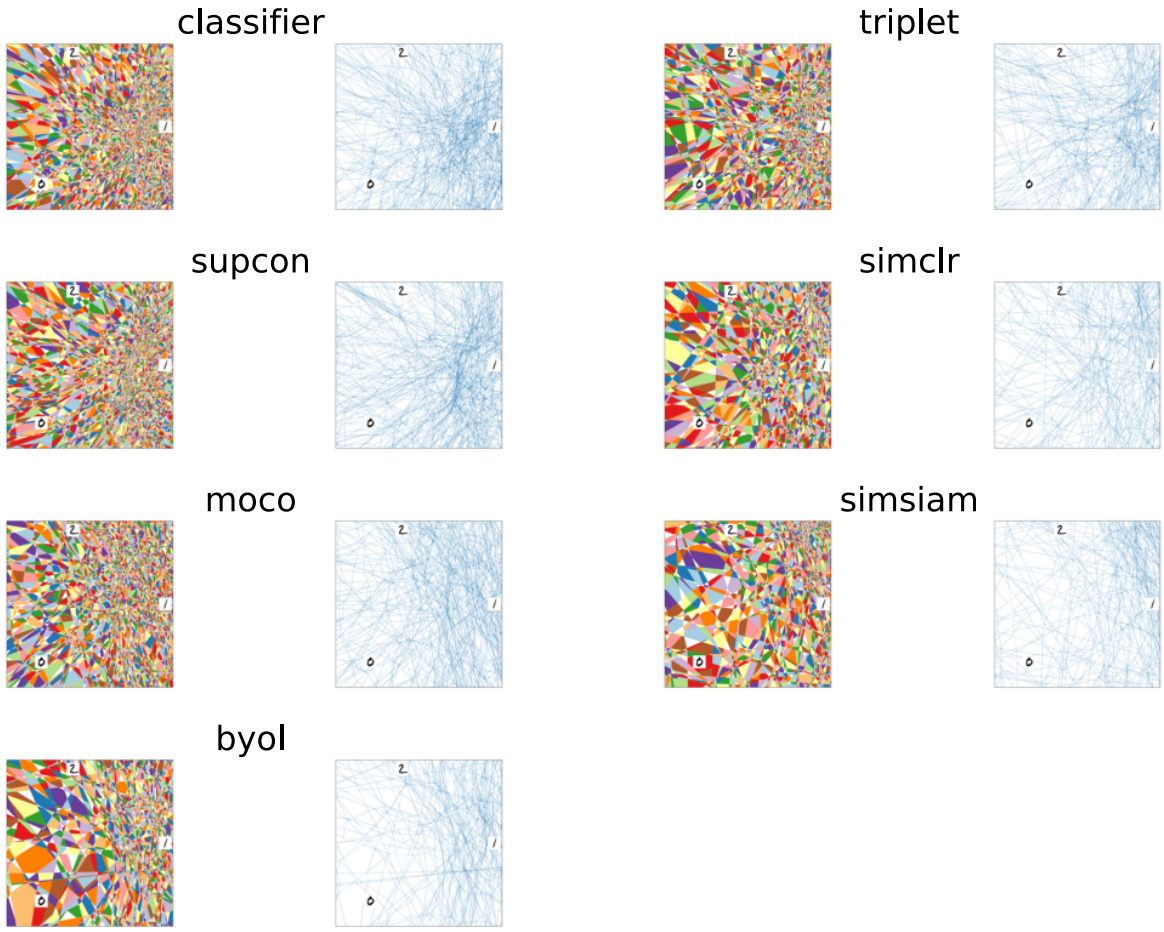

Figure 11. Final geometry of regions at the 100th epoch for each model on the MNIST dataset. Supervised methods (Classifier, Triplet and SupCon) produce highly dense and smaller regions. Self-supervised methods (SimCLR, MoCo, SimSiam, and BYOL) produce less dense regions with larger volumes. Contrastive self-supervised methods (MoCo and SimCLR) have linear regions that are more dense than the self-distillation methods (Simsiam and BYOL). The use of negatives as a repulsive force within the embedding space partitions the geometric space further.

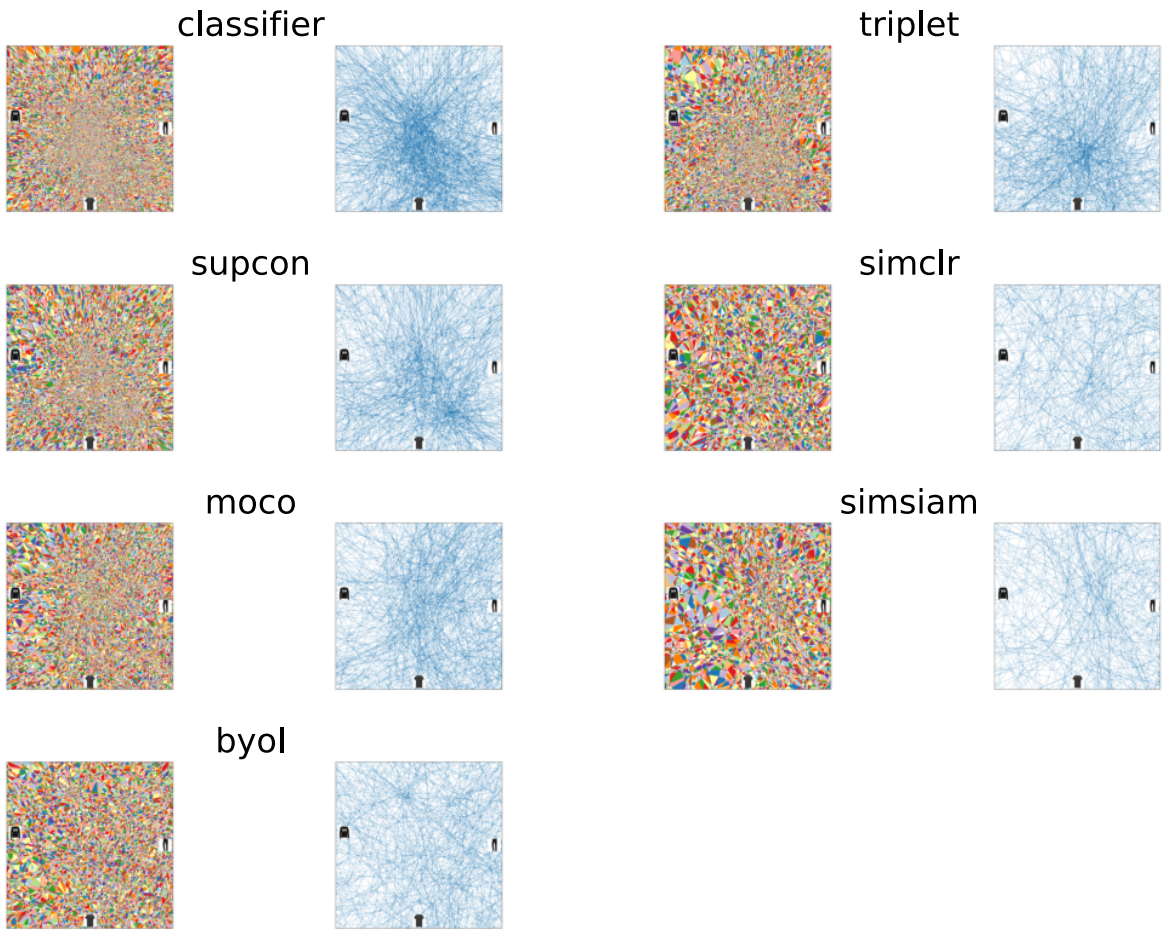

Figure 12. Final geometry of regions at the 100th epoch for each model on the FashionMNIST dataset. The same qualitative trends observed on MNIST are reproduced here. Regions consolidate in self-distillation methods and finer partitioning in supervised and self-supervised contrastive models.