ReduNet: White-Box Deep (Convolution) Networks from the Principle of Rate Reduction
(Lecture 21 & 22)

Yi Ma

EECS Department, UC Berkeley

November 16 & 18, 2021

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, Haozhi Qi, and John Wright

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

VIVE Center

"What I cannot create, I do not understand."
– Richard Feynman

High-Dim Data with Mixed Low-Dim Structures

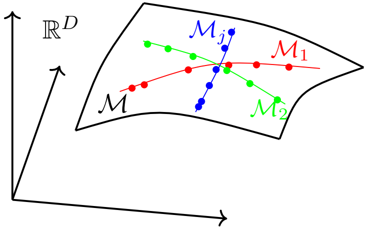


Figure: High-dimensional data $\boldsymbol{x} \in \mathbb{R}^D$ lying on a mixture of low-dimensional submanifolds $\{\mathcal{M}_j\}$.

Three Related Objectives of Learning from Data:

1. **Interpolation**: Identify which samples belong to the same structure.
2. **Extrapolation**: Determine to which structure a new sample belong.
3. **Representation**: Find most compact and discriminative representations.

Learning Lumped into a Black Box (Deep Learning)

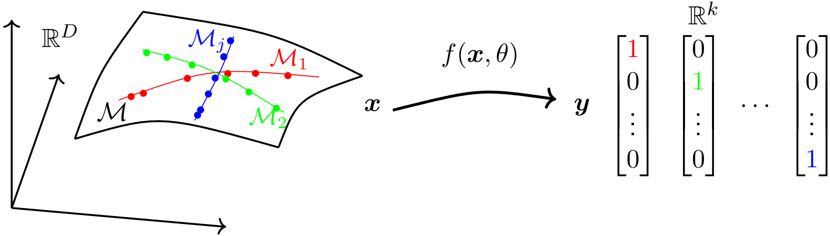


Figure: **Black Box Classification:** $\boldsymbol{y}$ is the class label of $\boldsymbol{x}$ represented as a "one-hot" vector in $\mathbb{R}^k$. To learn a nonlinear mapping $f(\cdot, \theta) : \boldsymbol{x} \mapsto \boldsymbol{y}$, say modeled by a deep network.

Fitting Class Labels via a Deep Network

In a supervised setting, using cross-entropy (CE) loss:

$$\min_{\theta \in \Theta} \mathsf{CE}(\theta, \boldsymbol{x}, \boldsymbol{y}) \doteq -\mathbb{E}[\langle \boldsymbol{y}, \log[f(\boldsymbol{x}, \theta)]\rangle] \approx -\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{y}_i, \log[f(\boldsymbol{x}_i, \theta)]\rangle. \quad (1)$$

Issues (an elephant in the room):

- A large deep neural networks can **fit arbitrary data and labels**.
- Statistical and geometric meaning of internal features **not clear**.
- Task/data-dependent and **not robust nor truly invariant**.

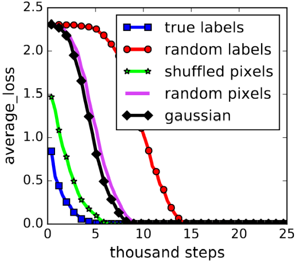


Figure: [Zhang et al, ICLR'17]

What did machines actually "learn" from doing this?

In terms of interpolating, extrapolating, or representing the data?

A Hypothesis: Information Bottleneck

[Tishby & Zaslavsky, 2015]

A feature mapping $f(\boldsymbol{x}, \theta)$ and a classifier $g(\boldsymbol{z})$ trained for downstream classification:

$$\boldsymbol{x} \xrightarrow{f(\boldsymbol{x},\theta)} \boldsymbol{z}(\theta) \xrightarrow{g(\boldsymbol{z})} \boldsymbol{y}.$$

The IB Hypothesis: Features learned in a deep network trying to

$$\max_{\theta \in \Theta} \mathsf{IB}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}(\theta)) \doteq I(\boldsymbol{z}(\theta), \boldsymbol{y}) - \beta I(\boldsymbol{x}, \boldsymbol{z}(\theta)), \quad \beta > 0, \tag{2}$$

where $I(\boldsymbol{z}, \boldsymbol{y}) \doteq H(\boldsymbol{z}) - H(\boldsymbol{z}|\boldsymbol{y})$ and $H(\boldsymbol{z})$ is the entropy of $\boldsymbol{z}$.

- **Minimal** informative features $\boldsymbol{z}$ that most correlate with the label $\boldsymbol{y}$
- Task and label-dependent, consequently sacrificing generalizability, robustness, or transferability

Gap between Theory and Practice (a Bigger Elephant)

For high-dimensional real data,
many statistical and information-theoretic concepts such as entropy, mutual information, K-L divergence, and maximum likelihood:

- curse of **dimensionality** for computation.
- ill-posed for **degenerate** distributions.
- lack guarantees with **finite** (or non-asymptotic) samples.

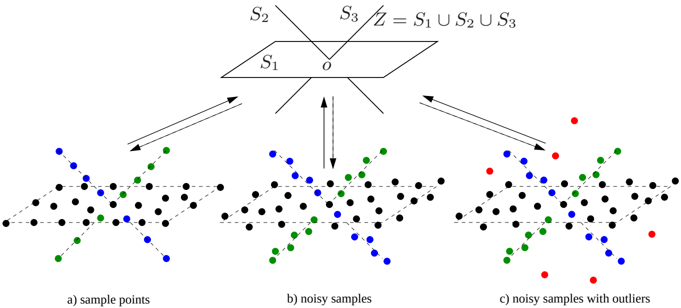
Reality check: principled formulations are replaced with approximate bounds, grossly simplifying assumptions, heuristics, even *ad hoc* tricks and hacks.

How to provide any performance guarantees at all?

A Principled Computational Approach

For high-dim data with mixed **low-dim** structures:

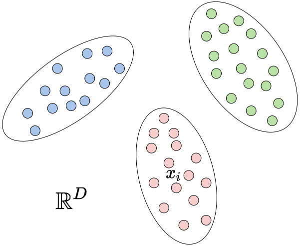
learn to compress, and compress to learn!



Generalized PCA for mixture of subspaces [Vidal, Ma, and Sastry, 2005]

1. Clustering Mixed Data (Interpolation)

Assume data $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$ are i.i.d. samples from a mixture of distributions: $p(\boldsymbol{x}, \theta) = \sum_{j=1}^{k} \pi_j p_j(\boldsymbol{x}, \theta)$.



Classic approaches to cluster the data: the maximum-likelihood (ML) estimate via Expectation Maximization (EM):

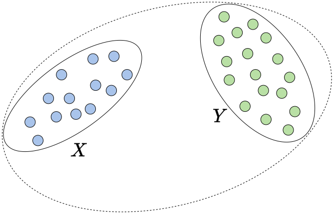
$$\max_{\theta,\pi} \mathbb{E}\Big[\log\Big(\sum_{j=1}^{k} \pi_j p_j(\boldsymbol{x}, \theta)\Big)\Big] \approx \max_{\theta,\pi} \frac{1}{m}\sum_{i=1}^{m} \log\Big(\sum_{j=1}^{k} \pi_j p_j(\boldsymbol{x}_i, \theta)\Big).$$

Difficulties: ML is not well-defined when distributions are degenerate.

Clustering via Compression

[Yi Ma, Harm Derksen, Wei Hong, and John Wright, TPAMI'07]

A Fundamental Idea:
Data belong to mixed low-dim structures should be compressible.



Cluster Criterion:
Whether the number of binary bits required to store the data:

$$\#\text{bits}(\boldsymbol{X} \cup \boldsymbol{Y}) \;\geq\; \#\text{bits}(\boldsymbol{X}) + \#\text{bits}(\boldsymbol{Y})?$$

"The whole is greater than the sum of the parts."
– Aristotle, 320 BC

Coding Length Function for Subspace-Like Data

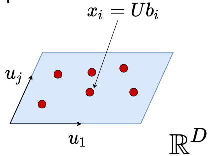
Theorem (Ma, TPAMI'07)

The number of bits needed to encode data $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m] \in \mathbb{R}^{D \times m}$ *up to a precision* $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \leq \epsilon$ *is bounded by:*

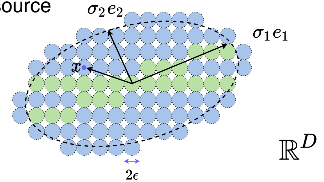
$$L(\boldsymbol{X}, \epsilon) \doteq \left(\frac{m+D}{2}\right) \log\det\left(\boldsymbol{I} + \frac{D}{m\epsilon^2}\boldsymbol{X}\boldsymbol{X}^\top\right).$$

This can be derived from constructively quantifying SVD of $\boldsymbol{X}$ or by sphere packing vol$(\boldsymbol{X})$ as samples of a noisy Gaussian source.

Linear subspace



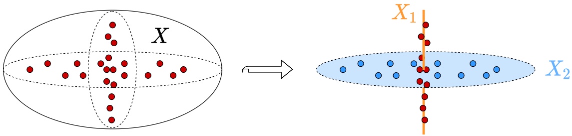
Gaussian source



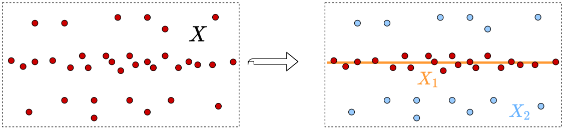
Cluster to Compress

$$L(\boldsymbol{X}) \;\geq\; L^c(\boldsymbol{X}) \doteq L(\boldsymbol{X}_1) + L(\boldsymbol{X}_2) + H(|\boldsymbol{X}_1|, |\boldsymbol{X}_2|)?$$

partitioning:



sifting:



A Greedy Algorithm

Seek a partition of the data $\boldsymbol{X} \to [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k]$ such that

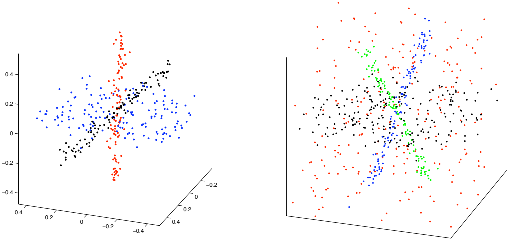
$$\min L^c(\boldsymbol{X}) \doteq L(\boldsymbol{X}_1) + \cdots + L(\boldsymbol{X}_k) + H(|\boldsymbol{X}_1|, \ldots, |\boldsymbol{X}_k|).$$

Optimize with a *bottom-up pair-wise* merging algorithm [Ma, TPAMI'07]:

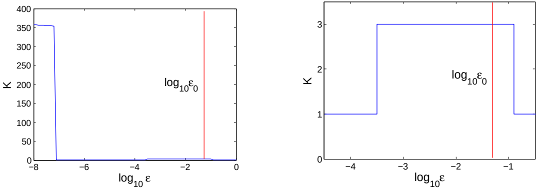
input: the data $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m] \in \mathbb{R}^{D \times m}$ and a distortion $\epsilon^2 > 0$.
initialize $\mathcal{S}$ as a set of sets with a single datum $\{S = \{\boldsymbol{x}\} \mid \boldsymbol{x} \in \boldsymbol{X}\}$.
while $|\mathcal{S}| > 1$ **do**
    choose distinct sets $S_1, S_2 \in \mathcal{S}$ such that
    $L^c(S_1 \cup S_2) - L^c(S_1, S_2)$ is minimal.
    if $L^c(S_1 \cup S_2) - L^c(S_1, S_2) \geq 0$ **then** break;
    else $\mathcal{S} := \big(\mathcal{S} \setminus \{S_1, S_2\}\big) \cup \{S_1 \cup S_2\}$.
end
output: $\mathcal{S}$

Surprisingly Good Performance

Empirically, **find global optimum and extremely robust to outliers**

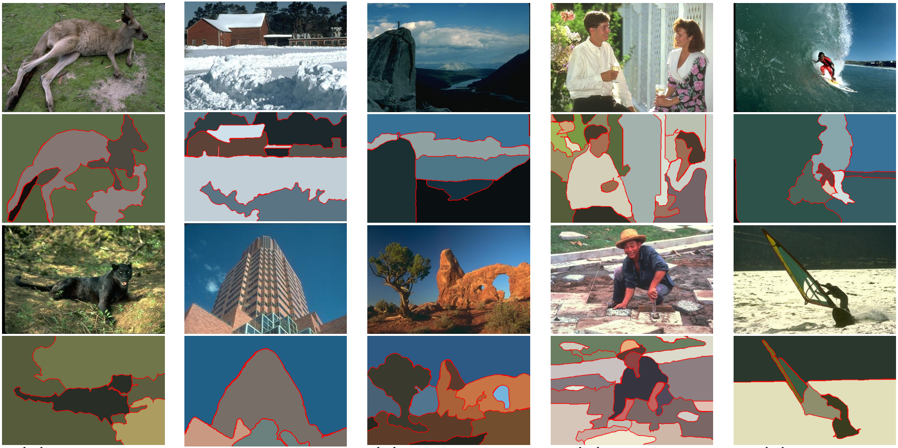


A strikingly sharp **phase transition** w.r.t. quantization ϵ



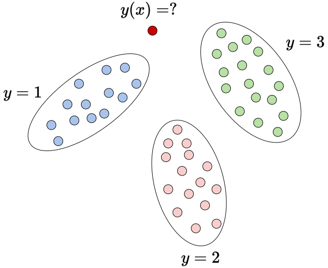
Natural Image Segmentation [Mobahi et.al., IJCV'09]

Compression alone, without any supervision, leads to **state of the art** segmentation on natural images (and many other types of data).



(a) Animals (b) Buildings (c) Landscape (d) People (e) Water

2. Classify Mixed Data (Extrapolation)



Assume data $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$ are i.i.d. samples from a mixture of distributions: $p(\boldsymbol{x}, \theta) = \sum_{j=1}^{k} \pi_j p_j(\boldsymbol{x}, \theta)$.

Classic approach to classify the data is via maximum a posteriori (MAP) classifier:

$$\hat{y}(\boldsymbol{x}) \;=\; \arg\max_j \log p_j(\boldsymbol{x}, \theta) + \log \pi_j.$$

Difficulties: distributions p_j are hard to estimate and log likelihood is not well-defined when distributions are degenerate.

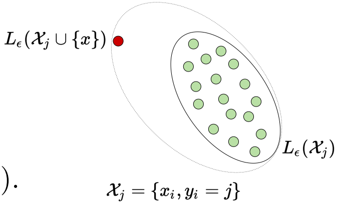
(**probably why SVMs or deep networks prevail instead...**)

Classify to Compress

[Wright, Tao, Lin, Shum, and Ma, NIPS'07]

A Fundamental Idea:
Count additional #bits needed to encode a query sample $\boldsymbol{x}$ with data in each class $\boldsymbol{X}_j$:

$$\delta L_\epsilon(\boldsymbol{x}, j) \doteq L_\epsilon(\boldsymbol{X}_j \cup \{\boldsymbol{x}\}) - L_\epsilon(\boldsymbol{X}_j) + L(j).$$



Classification Criterion: Minimum **I**ncremental **C**oding **L**ength (**MICL**):

$$\hat{y}(\boldsymbol{x}) \;=\; \arg\min_j \, \delta L_\epsilon(\boldsymbol{x}, j).$$

Law of Parsimony: *"Entities should not be multiplied without necessity."*
–William of Ockham

Asymptotic Property of MICL

Theorem (Wright, NIPS'07)

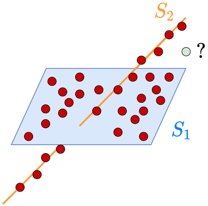
As the number of samples m goes to infinity, the MICL criterion converges at a rate of $O(m^{-1/2})$ to the following criterion:

$$\hat{y}_\epsilon(\boldsymbol{x}) = \arg\max_j \underbrace{L_G\left(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j + \frac{\epsilon^2}{D}\boldsymbol{I}\right) + \log \pi_j}_{\textit{Regularized MAP}} + \frac{1}{2} D_\epsilon(\boldsymbol{\Sigma}_j),$$

where $D_\epsilon(\boldsymbol{\Sigma}_j) \doteq \mathit{tr}\left(\boldsymbol{\Sigma}_j\left(\boldsymbol{\Sigma}_j + \frac{\epsilon^2}{D}\boldsymbol{I}\right)^{-1}\right)$ *is the effective dimension.*

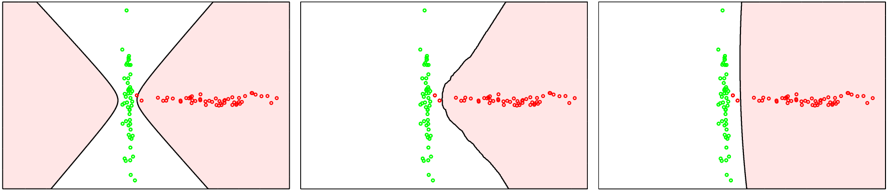
Everything else equal, MICL prefers a class with higher effective dimension.

Err on the side of caution!



Extrapolation of Low-Dim Structure for Classification

Figure: A truly extrapolating (nearest subspace) classifier!



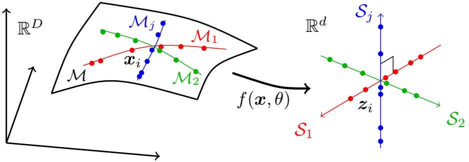
(a) **MICL** (b) k-Nearest Neighbors (c) SVM-RBF

Difficulty in practice: inference computationally costly (non-parametric) and possibly need a kernel (nonlinearity).

Go beyond (non-parametric) data interpolation and extrapolation?

3. Represent Mixed Data

Given samples $\boldsymbol{x} \in \mathbb{R}^D$ drawn from a mixture of k submanifolds $\mathcal{M} = \{\mathcal{M}_j\}_{j=1}^k$, we **seek a good representation** $\boldsymbol{z} \in \mathbb{R}^d$ through a continuous mapping: $f(\boldsymbol{x}, \theta) : \mathbb{R}^D \to \mathbb{R}^d$.



Goals of **"re-present"** the data $f(\boldsymbol{x}, \theta) : \boldsymbol{x} \mapsto \boldsymbol{z}$:

- from non-parametric (samples) to more compact (models).
- from nonlinear structures in $\boldsymbol{x}$ to linear in $\boldsymbol{z}$.
- from separable $\boldsymbol{x}$ to maximally discriminative $\boldsymbol{z}$.

What is a good representation? (Why a deep neural network?)

Seeking a Linear Discriminative Representation (LDR)

Desiderata: Representation $z = f(x, \theta)$ have the following properties:

1. *Within-Class Compressible:* Features of the same class/cluster should be highly compressed in a **low-dimensional** linear subspace.
2. *Between-Class Discriminative:* Features of different classes/clusters should be in highly **incoherent** linear subspaces.
3. *Maximally Informative Representation:* Dimension (or variance) of features for each class/cluster should be **as large as possible**.

Is there a principled measure for all such properties, together?

Why not cross entropy? Prevalence of **neural collapse** during the terminal phase of deep learning training, Papyan, Han, and Donoho, 2020.

Measure of Compactness for a Linear Representation

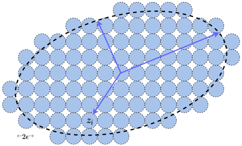
Consider a feature mapping:

$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m] \in \mathbb{R}^{D \times m} \xrightarrow{f(\boldsymbol{x},\theta)} \boldsymbol{Z}(\theta) = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m] \in \mathbb{R}^{d \times m}.$$

The average coding length per sample (rate) subject to a distortion ϵ:

$$R(\boldsymbol{Z}, \epsilon) \doteq \frac{1}{2} \log\det\left(\boldsymbol{I} + \frac{d}{m\epsilon^2} \boldsymbol{Z}\boldsymbol{Z}^\top\right). \quad (3)$$

Rate distortion is an intrinsic measure for the volume of all features.

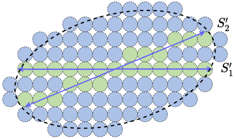


vol(Z)

Measure of Compactness for Mixed Representations

The features $\boldsymbol{Z}$ of multi-class data may be partitioned into multiple subsets:

$$\boldsymbol{Z} = \boldsymbol{Z}_1 \cup \boldsymbol{Z}_2 \cup \cdots \cup \boldsymbol{Z}_k.$$



W.r.t. this partition, the **average coding rate** is:

$$R^c(\boldsymbol{Z}, \epsilon \mid \boldsymbol{\Pi}) \doteq \sum_{j=1}^{k} \frac{\mathrm{tr}(\boldsymbol{\Pi}_j)}{2m} \log\det\left(\boldsymbol{I} + \frac{d}{\mathrm{tr}(\boldsymbol{\Pi}_j)\epsilon^2} \boldsymbol{Z}\boldsymbol{\Pi}_j\boldsymbol{Z}^\top\right), \tag{4}$$

where $\boldsymbol{\Pi} = \{\boldsymbol{\Pi}_j \in \mathbb{R}^{m\times m}\}_{j=1}^k$ encode the membership of the m samples in the k classes: the diagonal entry $\boldsymbol{\Pi}_j(i,i)$ of $\boldsymbol{\Pi}_j$ is the probability of sample i belonging to subset j. $\Omega \doteq \{\boldsymbol{\Pi} \mid \sum \boldsymbol{\Pi}_j = \boldsymbol{I}, \boldsymbol{\Pi}_j \geq \mathbf{0}.\}$

Learning Representation to Cluster and Classify

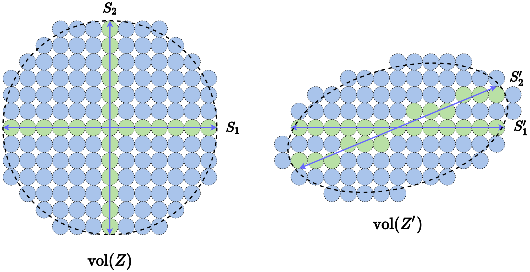
A Fundamental Idea: maximize the **difference** between the coding rate of all features and the average rate of features in the classes:

$$\Delta R(\boldsymbol{Z}, \boldsymbol{\Pi}, \epsilon) = \underbrace{\frac{1}{2} \log\det\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}\boldsymbol{Z}^\top\right)}_{R} - \underbrace{\sum_{j=1}^{k} \frac{\mathrm{tr}(\boldsymbol{\Pi}_j)}{2m} \log\det\left(\boldsymbol{I} + \frac{d}{\mathrm{tr}(\boldsymbol{\Pi}_j)\epsilon^2}\boldsymbol{Z}\boldsymbol{\Pi}_j\boldsymbol{Z}^\top\right)}_{R^c}.$$

- R: expand all features $\boldsymbol{Z}$ as **large** as possible.
- R^c: compress each class $\boldsymbol{Z}_j$ as **small** as possible.

Slogan: similarity contracts and dissimilarity contrasts!

Interpretation of MCR^2: Sphere Packing and Counting



Example: two subspaces S_1 and S_2 in $\mathbb{R}^2$.

- $\log \#$(green spheres + blue spheres) = rate of span of all samples R.
- $\log \#$(green spheres) = rate of the two subspaces R^c.
- $\log \#$(blue spheres) = rate reduction gain ΔR.

Maximal Coding Rate Reduction (MCR^2)

[Yu, Chan, You, Song, Ma, NeurIPS2020]

Learn a mapping $f(\boldsymbol{x}, \theta)$ (for a given partition $\boldsymbol{\Pi}$):

$$\boldsymbol{X} \xrightarrow{f(\boldsymbol{x},\theta)} \boldsymbol{Z}(\theta) \xrightarrow{\boldsymbol{\Pi},\epsilon} \Delta R(\boldsymbol{Z}(\theta), \boldsymbol{\Pi}, \epsilon) \tag{5}$$

so as to **M**aximize the **C**oding **R**ate **R**eduction (**MCR**^{2}):

$$\begin{aligned} \max_{\theta} \quad & \Delta R\big(\boldsymbol{Z}(\theta), \boldsymbol{\Pi}, \epsilon\big) = R(\boldsymbol{Z}(\theta), \epsilon) - R^c(\boldsymbol{Z}(\theta), \epsilon \mid \boldsymbol{\Pi}), \\ & \text{subject to} \quad \|\boldsymbol{Z}_j(\theta)\|_F^2 = m_j, \ \boldsymbol{\Pi} \in \Omega. \end{aligned} \tag{6}$$

Since ΔR is *monotonic* in the scale of $\boldsymbol{Z}$, one needs to:
normalize the features $z = f(x, \theta)$ so as to compare $Z(\theta)$ and $Z(\theta')$!

Batch normalization, Sergey Ioffe and Christian Szegedy, 2015.

Layer normalization'16, instance normalization'16; group normalization'18...

Theoretical Justification of the MCR^2 Principle

Theorem (Informal Statement [Yu et.al., NeurIPS2020])

Suppose $\boldsymbol{Z}^\star = \boldsymbol{Z}_1^\star \cup \cdots \cup \boldsymbol{Z}_k^\star$ is the optimal solution that maximizes the rate reduction (6). *We have:*

- Between-class Discriminative*: As long as the ambient space is adequately large ($d \geq \sum_{j=1}^k d_j$), the subspaces are all orthogonal to each other,* i.e. $(\boldsymbol{Z}_i^\star)^\top \boldsymbol{Z}_j^\star = \mathbf{0}$ *for* $i \neq j$.
- Maximally Informative Representation*: As long as the coding precision is adequately high, i.e.,* $\epsilon^4 < \min_j \left\{ \frac{m_j}{m} \frac{d^2}{d_j^2} \right\}$*, each subspace achieves its maximal dimension, i.e.* $\text{rank}(\boldsymbol{Z}_j^\star) = d_j$. *In addition, the largest* $d_j - 1$ *singular values of* $\boldsymbol{Z}_j^\star$ *are equal.*

A new slogan, beyond Aristotle:

The whole is **to be maximally** *greater than the sum of the parts!*

Comparison to Orthogonal Low-Rank Embedding (OLE) [Lezama, Qiu, Musé, and Sapiro, CVPR'18]

The following loss as a geometric regularizer for the cross entropy (1):

$$\max_{\theta} \mathsf{OLE}(\boldsymbol{Z}(\theta), \boldsymbol{\Pi}) \doteq \|\boldsymbol{Z}(\theta)\|_* - \sum_{j=1}^{k} \|\boldsymbol{Z}_j(\theta)\|_*.$$

The nuclear norm $\|\cdot\|_*$ in OLE is convex but non-smooth; the $\log\det(\cdot)$ in MCR^2 is concave but smooth.

OLE is always negative and reaches the maximum 0 iff the subspaces are orthogonal, regardless of the dimension.

OLE cannot avoid neural collapse whereas MCR^2 does.

Comparison to Contrastive Learning

[Hadsell, Chopra, and LeCun, CVPR'06]

When k is large, a randomly chosen **pair** $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is of high probability belonging to different classes. Minimize the **contrastive loss**:

$$\min -\log \frac{\exp(\langle \boldsymbol{z}_i, \boldsymbol{z}_i' \rangle)}{\sum_{j \neq i} \exp(\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle)}.$$

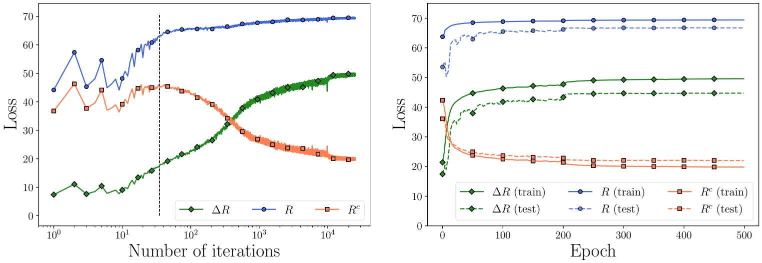
The learned features of such pairs of samples together with their augmentations $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$ should have large rate reduction:

$$\max \sum_{ij} \Delta R_{ij} \doteq R\big(\boldsymbol{Z}_i \cup \boldsymbol{Z}_j, \epsilon\big) - \frac{1}{2}\big(R(\boldsymbol{Z}_i, \epsilon) + R(\boldsymbol{Z}_j, \epsilon)\big).$$

MCR^2 contrasts triplets, quadruplets, or any number of sets.

Experiment I: Supervised Deep Learning

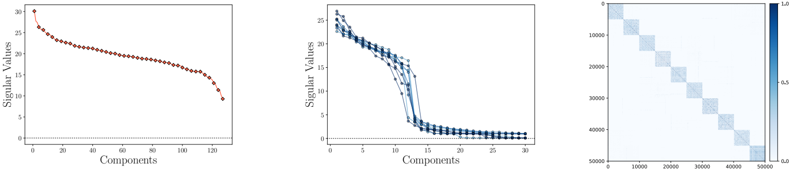
Experimental Setup: Train $f(\boldsymbol{x}, \theta)$ as ResNet18 on the CIFAR10 dataset, feature $\boldsymbol{z}$ dimension $d = 128$, precision $\epsilon^2 = 0.5$.



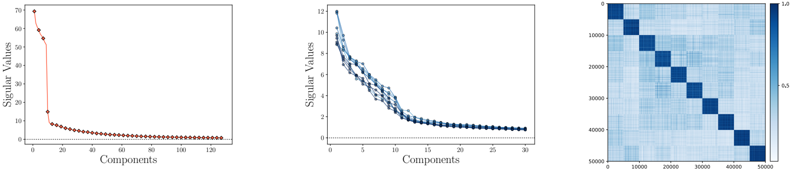
(a) (b)

Figure: (a). Evolution of $R, R^c, \Delta R$ during the training process; (b). Training loss versus testing loss.

Visualization of Learned Representations Z



(a) MCR^2 (overall) (b) MCR^2 (PCA of every class) (c) MCR^2 (cosine similarity)



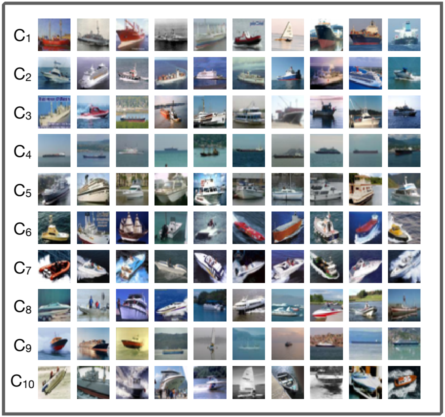
(d) CE (overall) (e) CE (PCA of every class) (f) CE (cosine similarity)

Figure: PCA of learned representations from MCR^2 and cross-entropy.

No neural collapse!

Visualization - Samples along Principal Components

(a) Bird



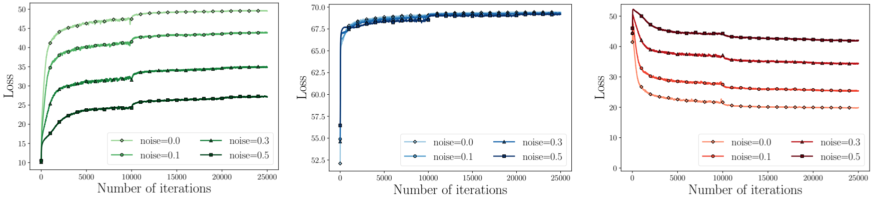
(b) Ship

Figure: Top-10 "principal" images for class - "Bird" and "Ship" in the CIFAR10.

Robustness to Label Noise

Table 1: Classification results with features learned with labels corrupted at different levels.

	RATIO=0.1	RATIO=0.2	RATIO=0.3	RATIO=0.4	RATIO=0.5
CE TRAINING	90.91%	86.12%	79.15%	72.45%	60.37%
MCR^2 TRAINING	**91.16%**	**89.70%**	**88.18%**	**86.66%**	**84.30%**



(a) $\Delta R\big(\boldsymbol{Z}(\theta), \boldsymbol{\Pi}, \epsilon\big)$ (b) $R(\boldsymbol{Z}(\theta), \epsilon)$ (c) $R^c(\boldsymbol{Z}(\theta), \epsilon \mid \boldsymbol{\Pi})$

Figure: Evolution of $R, R^c, \Delta R$ of MCR^2 during training with corrupted labels.

Represent only what can be jointly compressed.

Experiment II: Self-supervised Learning

Without label information, we use the MCR^2 principle (6) to learn representations $\boldsymbol{Z}$ that are *invariant* to certain class of transformations/ augmentations, say $\mathcal{T}$ with a distribution $P_{\mathcal{T}}$.

1. Given a mini-batch of data $\{\boldsymbol{x}_j\}_{j=1}^k$, augment each sample $\boldsymbol{x}_j$ with n augmentations $\{\tau_i(\cdot)\}_{i=1}^n$ randomly drawn from $P_{\mathcal{T}}$.
2. Label all the augmented samples $\boldsymbol{X}_j = [\tau_1(\boldsymbol{x}_j), \ldots, \tau_n(\boldsymbol{x}_j)]$ of $\boldsymbol{x}_j$ as the j-th class, and $\boldsymbol{Z}_j$ the corresponding learned features.
3. Using this self-labeled data, train feature mapping $f(\cdot, \theta)$ via maximizing the MCR^2 in Eq. (6).

Experimental Setup: Train ResNet18 on CIFAR10 dataset, feature dimension $d = 128$, precision $\epsilon = 0.5$. For every mini-batch, the total number of samples for training is $m = kn$.

Clustering of Real Datasets

Figure: Clustering results on CIFAR10, CIFAR100, and STL10 datasets.

Dataset	Metric	K-Means	JULE	RTM	DEC	DAC	DCCM	MCR^2-Ctrl
CIFAR10	NMI	0.087	0.192	0.197	0.257	0.395	0.496	**0.630**
	ACC	0.229	0.272	0.309	0.301	0.521	0.623	**0.684**
	ARI	0.049	0.138	0.115	0.161	0.305	0.408	**0.508**
CIFAR100	NMI	0.084	0.103	-	0.136	0.185	0.285	**0.387**
	ACC	0.130	0.137	-	0.185	0.237	0.327	**0.375**
	ARI	0.028	0.033	-	0.050	0.087	0.173	**0.178**
STL10	NMI	0.124	0.182	-	0.276	0.365	0.376	**0.446**
	ACC	0.192	0.182	-	0.359	0.470	0.482	**0.491**
	ARI	0.061	0.164	-	0.186	0.256	0.262	**0.290**

Set the mini-batch size as $k = 20$, number of augmentations for each sample as $n = 50$ and the precision parameter as $\epsilon^2 = 0.5$. Compare with JULE, RTM [Nina'19], DEC [Xie'16], DAC [Chang'17], and DCCM [Wu'19].

Deep Networks from Optimizing Rate Reduction

$$\boldsymbol{X} \xrightarrow{f(\boldsymbol{x},\theta)} \boldsymbol{Z}(\theta); \quad \max_{\theta} \Delta R(\boldsymbol{Z}(\theta), \boldsymbol{\Pi}, \epsilon).$$

Final features learned by MCR^2 are more interpretable and robust, **but**:

- The borrowed deep network (e.g. ResNet) is still a "black box"!
- Why is a "deep" architecture necessary, and how wide and deep?
- What are the roles of the "linear and nonlinear" operators?
- Why "multi-channel" convolutions?
- ...

Replace black box networks with entirely "white box" networks?

Projected Gradient Ascent for Rate Reduction

Recall the rate reduction objective:

$$\max_{\boldsymbol{Z}} \Delta R(\boldsymbol{Z}) \doteq \underbrace{\frac{1}{2}\log\det\left(\boldsymbol{I} + \alpha \boldsymbol{Z}\boldsymbol{Z}^*\right)}_{R(\boldsymbol{Z})} - \underbrace{\sum_{j=1}^{k} \frac{\gamma_j}{2}\log\det\left(\boldsymbol{I} + \alpha_j \boldsymbol{Z}\boldsymbol{\Pi}^j \boldsymbol{Z}^*\right)}_{R_c(\boldsymbol{Z},\boldsymbol{\Pi})}, \quad (7)$$

where $\alpha = d/(m\epsilon^2)$, $\alpha_j = d/(\mathrm{tr}(\boldsymbol{\Pi}^j)\epsilon^2)$, $\gamma_j = \mathrm{tr}(\boldsymbol{\Pi}^j)/m$ for $j = 1, \ldots, k$.

Consider directly maximizing ΔR with **projected gradient ascent** (PGA):

$$\boldsymbol{Z}_{\ell+1} \;\propto\; \boldsymbol{Z}_\ell + \eta \cdot \left.\frac{\partial \Delta R}{\partial \boldsymbol{Z}}\right|_{\boldsymbol{Z}_\ell} \quad \text{subject to} \quad \boldsymbol{Z}_{\ell+1} \subset \mathbb{S}^{d-1}. \quad (8)$$

Gradients of the Two Terms

The derivatives $\frac{\partial R(\boldsymbol{Z})}{\partial \boldsymbol{Z}}$ and $\frac{\partial R_c(\boldsymbol{Z},\boldsymbol{\Pi})}{\partial \boldsymbol{Z}}$ are:

$$\frac{1}{2}\frac{\partial \log\det(\boldsymbol{I}+\alpha \boldsymbol{Z}\boldsymbol{Z}^*)}{\partial \boldsymbol{Z}}\bigg|_{\boldsymbol{Z}_\ell} = \underbrace{\alpha(\boldsymbol{I}+\alpha \boldsymbol{Z}_\ell \boldsymbol{Z}_\ell^*)^{-1}}_{\boldsymbol{E}_\ell\ \in \mathbb{R}^{d\times d}} \boldsymbol{Z}_\ell, \tag{9}$$

$$\frac{1}{2}\frac{\partial\left(\gamma_j \log\det(\boldsymbol{I}+\alpha_j \boldsymbol{Z}\boldsymbol{\Pi}^j \boldsymbol{Z}^*)\right)}{\partial \boldsymbol{Z}}\bigg|_{\boldsymbol{Z}_\ell} = \gamma_j \underbrace{\alpha_j(\boldsymbol{I}+\alpha_j \boldsymbol{Z}_\ell \boldsymbol{\Pi}^j \boldsymbol{Z}_\ell^*)^{-1}}_{\boldsymbol{C}_\ell^j\ \in \mathbb{R}^{d\times d}} \boldsymbol{Z}_\ell \boldsymbol{\Pi}^j. \tag{10}$$

Hence the gradient $\frac{\partial \Delta R(\boldsymbol{Z})}{\partial \boldsymbol{Z}}$ is:

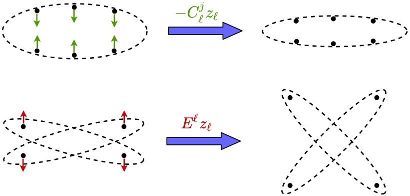
$$\frac{\partial \Delta R}{\partial \boldsymbol{Z}}\bigg|_{\boldsymbol{Z}_\ell} = \underbrace{\boldsymbol{E}_\ell}_{\text{Expansion}} \boldsymbol{Z}_\ell - \sum_{j=1}^{k} \gamma_j \underbrace{\boldsymbol{C}_\ell^j}_{\text{Compression}} \boldsymbol{Z}_\ell \boldsymbol{\Pi}^j \quad \in \mathbb{R}^{d\times m}. \tag{11}$$

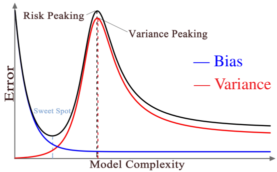
Interpretation of the Linear Operators $\boldsymbol{E}$ and $\boldsymbol{C}^j$

For any $\boldsymbol{z}_\ell \in \mathbb{R}^d$, we have

$$\boldsymbol{E}_\ell \boldsymbol{z}_\ell = \alpha(\boldsymbol{z}_\ell - \boldsymbol{Z}_\ell \boldsymbol{q}_\ell^*) \quad \text{with} \quad \boldsymbol{q}_\ell^* \doteq \arg\min_{\boldsymbol{q}_\ell} \alpha \|\boldsymbol{z}_\ell - \boldsymbol{Z}_\ell \boldsymbol{q}_\ell\|_2^2 + \|\boldsymbol{q}_\ell\|_2^2.$$

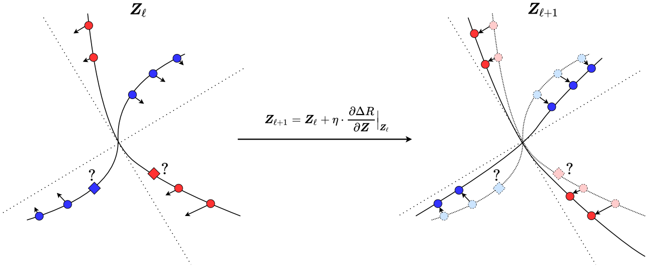
$\boldsymbol{E}_\ell \boldsymbol{z}_\ell$ and $\boldsymbol{C}_\ell^j \boldsymbol{z}_\ell$ are the "residuals" of $\boldsymbol{z}_\ell$ against the subspaces spanned by columns of $\boldsymbol{Z}_\ell$ and $\boldsymbol{Z}_\ell^j$, respectively.





Such "auto" ridge regressions **do not overfit** even with redundant random regressors, due to a "double descent" risk [Yang, ICML'20]!

Incremental Deformation via Gradient Flow



Extrapolate the gradient $\frac{\partial \Delta R(\boldsymbol{Z})}{\partial \boldsymbol{Z}}$ from training samples $\boldsymbol{Z}$ to all $\boldsymbol{z} \in \mathbb{R}^d$:

$$\left.\frac{\partial \Delta R}{\partial \boldsymbol{Z}}\right|_{\boldsymbol{Z}_\ell} = \boldsymbol{E}_\ell \boldsymbol{Z}_\ell - \sum_{j=1}^{k} \gamma_j \boldsymbol{C}_\ell^j \boldsymbol{Z}_\ell \underbrace{\boldsymbol{\Pi}^j}_{\text{known}} \in \mathbb{R}^{d \times m}, \tag{12}$$

$$g(\boldsymbol{z}_\ell, \boldsymbol{\theta}_\ell) \doteq \boldsymbol{E}_\ell \boldsymbol{z}_\ell - \sum_{j=1}^{k} \gamma_j \boldsymbol{C}_\ell^j \boldsymbol{z}_\ell \underbrace{\boldsymbol{\pi}^j(\boldsymbol{z}_\ell)}_{\text{unknown}} \in \mathbb{R}^{d}. \tag{13}$$

Estimate of the Membership $\boldsymbol{\pi}^j(\boldsymbol{z}_\ell)$

Estimate the membership $\boldsymbol{\pi}^j(\boldsymbol{z}_\ell)$ with "softmax" on the residuals $\|\boldsymbol{C}_\ell^j \boldsymbol{z}_\ell\|$:

$$\boldsymbol{\pi}^j(\boldsymbol{z}_\ell) \approx \widehat{\boldsymbol{\pi}}^j(\boldsymbol{z}_\ell) \doteq \frac{\exp\left(-\lambda \|\boldsymbol{C}_\ell^j \boldsymbol{z}_\ell\|\right)}{\sum_{j=1}^k \exp\left(-\lambda \|\boldsymbol{C}_\ell^j \boldsymbol{z}_\ell\|\right)} \quad \in [0,1]. \tag{14}$$

Thus the weighted residuals for contracting:

$$\boldsymbol{\sigma}\Big([\boldsymbol{C}_\ell^1 \boldsymbol{z}_\ell, \ldots, \boldsymbol{C}_\ell^k \boldsymbol{z}_\ell]\Big) \doteq \sum_{j=1}^k \gamma_j \boldsymbol{C}_\ell^j \boldsymbol{z}_\ell \cdot \widehat{\boldsymbol{\pi}}^j(\boldsymbol{z}_\ell) \quad \in \mathbb{R}^d. \tag{15}$$

Many alternatives, e.g. enforcing all features to be in the first quadrant:

$$\boldsymbol{\sigma}(\boldsymbol{z}_\ell) \approx \boldsymbol{z}_\ell - \sum_{j=1}^k \mathrm{ReLU}\big(\boldsymbol{P}_\ell^j \boldsymbol{z}_\ell\big), \tag{16}$$

The ReduNet for Optimizing Rate **Reduction**

Iterative projected gradient ascent (PGA) :

$$\boldsymbol{z}_{\ell+1} \;\propto\; \boldsymbol{z}_\ell + \eta \cdot \underbrace{\Big[\boldsymbol{E}_\ell \boldsymbol{z}_\ell + \boldsymbol{\sigma}\big([\boldsymbol{C}_\ell^1 \boldsymbol{z}_\ell, \dots, \boldsymbol{C}_\ell^k \boldsymbol{z}_\ell]\big)\Big]}_{g(\boldsymbol{z}_\ell, \boldsymbol{\theta}_\ell)} \quad \text{s.t.} \quad \boldsymbol{z}_{\ell+1} \in \mathbb{S}^{d-1}, \qquad (17)$$

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \phi^L \circ \phi^{L-1} \circ \cdots \circ \phi^0(\boldsymbol{x}), \text{ with } \phi^\ell(\boldsymbol{z}_\ell, \boldsymbol{\theta}_\ell) \doteq \mathcal{P}_{\mathbb{S}^{d-1}}[\boldsymbol{z}_\ell + \eta \cdot g(\boldsymbol{z}_\ell, \boldsymbol{\theta}_\ell)].$$

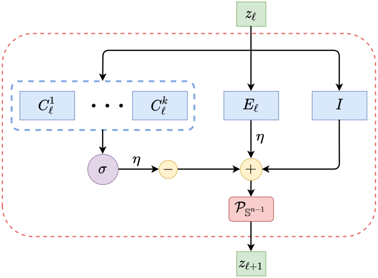


Figure: One layer of the **ReduNet**: one PGA iteration.

The ReduNet versus ResNet or ResNeXt

Iterative projected gradient ascent (PGA):

$$\boldsymbol{z}_{\ell+1} \propto \boldsymbol{z}_\ell + \eta \cdot \underbrace{\left[\boldsymbol{E}_\ell \boldsymbol{z}_\ell + \boldsymbol{\sigma}\left([\boldsymbol{C}_\ell^1 \boldsymbol{z}_\ell, \ldots, \boldsymbol{C}_\ell^k \boldsymbol{z}_\ell]\right)\right]}_{g(\boldsymbol{z}_\ell, \boldsymbol{\theta}_\ell)} \quad \text{s.t.} \quad \boldsymbol{z}_{\ell+1} \in \mathbb{S}^{d-1}, \qquad (18)$$

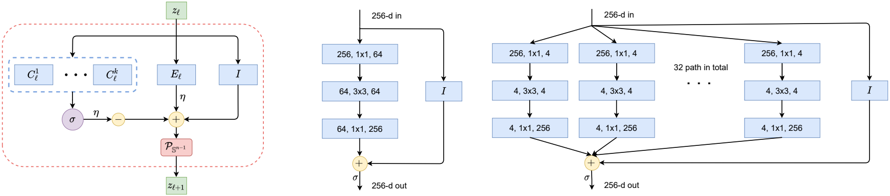


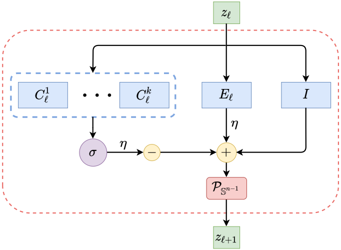
Figure: Left: **ReduNet**. Middle and Right: **ResNet** [He et. al. 2016] and **ResNeXt** [Xie et. al. 2017] (hundreds of layers).

Forward construction instead of back propagation!

The ReduNet versus Mixture of Experts

Approximate iterative projected gradient ascent (PGA) :

$$\boldsymbol{z}_{\ell+1} \;\propto\; \boldsymbol{z}_\ell + \eta \cdot \underbrace{\Big[\boldsymbol{E}_\ell \boldsymbol{z}_\ell + \boldsymbol{\sigma}\big([\boldsymbol{C}_\ell^1 \boldsymbol{z}_\ell, \ldots, \boldsymbol{C}_\ell^k \boldsymbol{z}_\ell]\big)\Big]}_{g(\boldsymbol{z}_\ell, \boldsymbol{\theta}_\ell)} \quad \text{s.t.} \quad \boldsymbol{z}_{\ell+1} \in \mathbb{S}^{d-1}, \qquad (19)$$



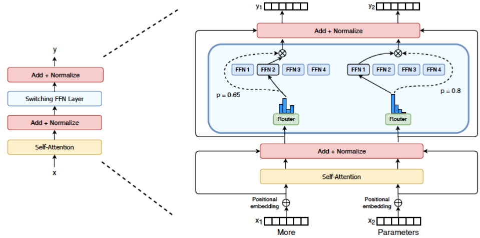


Figure: Left: **ReduNet** layer. Right: **Mixture of Experts** [Shazeer et. al. 2017] or **Switched Transformer** [Fedus et. al. 2021] (1.7 trillion parameters).

Forward construction instead of back propagation!

ReduNet Features for Mixture of Gaussians

$L =$**2000**-Layers ReduNet: $m = 500, \eta = 0.5, \epsilon = 0.1.$

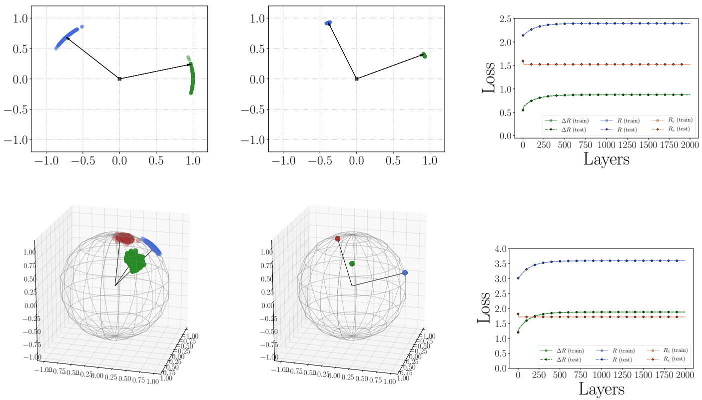


Figure: Left: original samples $\boldsymbol{X}$ and ReduNet features $\boldsymbol{Z} = f(\boldsymbol{Z}, \boldsymbol{\theta})$ for 2D and 3D Mixture of Gaussians. Right: plots for the progression of values of the rates.

Group Invariant Classification

Feature mapping $f(\boldsymbol{x}, \boldsymbol{\theta})$ is invariant to a group of transformations:

$$\textit{Group Invariance:} \quad f(\boldsymbol{x} \circ \mathfrak{g}, \boldsymbol{\theta}) \sim f(\boldsymbol{x}, \boldsymbol{\theta}), \quad \forall \mathfrak{g} \in \mathbb{G}, \tag{20}$$

where "$\sim$" indicates two features belonging to the same equivalent class.

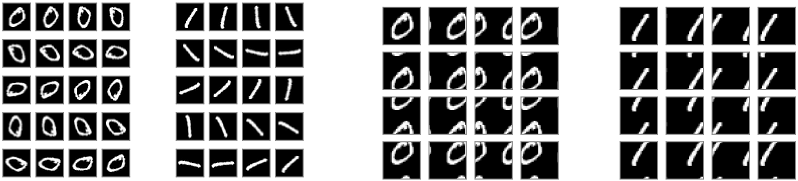


Figure: Left: 1D rotation $\mathbb{S}^1$; Right: 2D cyclic translation $\mathcal{T}^2$.

1. Fooling CNNs with simple transformations, Engstrom et.al., 2017.

2. Why do deep convolutional networks generalize so poorly to small image transformations? Azulay & Weiss, 2018.

Group Invariant Classification

Feature mapping $f(\boldsymbol{x}, \boldsymbol{\theta})$ is invariant to a group of transformations:

$$\textit{Group Invariance:} \quad f(\boldsymbol{x} \circ \mathfrak{g}, \boldsymbol{\theta}) \sim f(\boldsymbol{x}, \boldsymbol{\theta}), \quad \forall \mathfrak{g} \in \mathbb{G}, \tag{21}$$

where "$\sim$" indicates two features belonging to the same equivalent class.

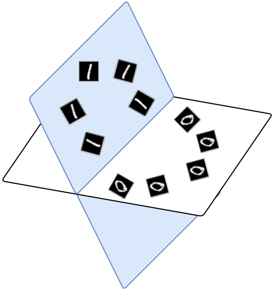
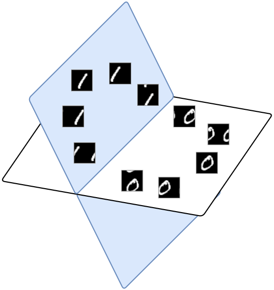



Figure: **Embed all equivariant samples to the same subspace.**

Circulant Matrix and Convolution

Given a vector $\boldsymbol{z} = [z_0, z_1, \ldots, z_{n-1}]^* \in \mathbb{R}^n$, we may arrange all its circular shifted versions in a circulant matrix form as

$$\text{circ}(\boldsymbol{z}) \quad \doteq \quad \begin{bmatrix} z_0 & z_{n-1} & \cdots & z_2 & z_1 \\ z_1 & z_0 & z_{n-1} & \cdots & z_2 \\ \vdots & z_1 & z_0 & \ddots & \vdots \\ z_{n-2} & \vdots & \ddots & \ddots & z_{n-1} \\ z_{n-1} & z_{n-2} & \cdots & z_1 & z_0 \end{bmatrix} \in \mathbb{R}^{n \times n}. \tag{22}$$

A circular (or cyclic) convolution:

$$\text{circ}(\boldsymbol{z}) \cdot \boldsymbol{x} = \boldsymbol{z} \circledast \boldsymbol{x}, \quad \text{where} \quad (\boldsymbol{z} \circledast \boldsymbol{x})_i = \sum_{j=0}^{n-1} x_j z_{i+n-j \bmod n}. \tag{23}$$

Convolutions from Cyclic Shift Invariance

Given a set of sample vectors $\boldsymbol{Z} = [\boldsymbol{z}^1, \ldots, \boldsymbol{z}^m]$, construct the ReduNet from cyclic-shift augmented families $\boldsymbol{Z} = [\text{circ}(\boldsymbol{z}^1), \ldots, \text{circ}(\boldsymbol{z}^m)]$.

Proposition (Convolution Structures of $\boldsymbol{E}$ and $\boldsymbol{C}^j$)

The linear operator in the ReduNet:

$$\boldsymbol{E} = \alpha\Big(\boldsymbol{I} + \alpha \sum_{i=1}^{m} \text{circ}(\boldsymbol{z}^i)\text{circ}(\boldsymbol{z}^i)^*\Big)^{-1}$$

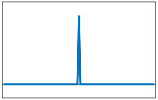
is a circulant matrix and represents a circular convolution:

$$\boldsymbol{E}\boldsymbol{z} = \boldsymbol{e} \circledast \boldsymbol{z},$$

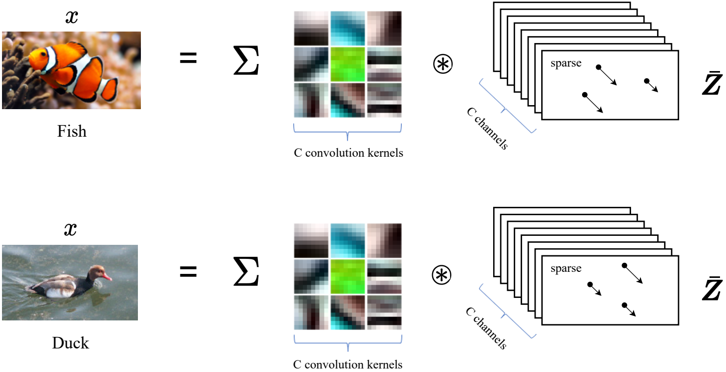
where $\boldsymbol{e}$ is the first column vector of $\boldsymbol{E}$. Similarly, the operators $\boldsymbol{C}^j$ associated with subsets $\boldsymbol{Z}^j$ are also circular convolutions.

Tradeoff between Invariance and Separability

A problem with separability: superposition of shifted "delta" functions can generate any other signals: $\text{span}[\text{circ}(\boldsymbol{x})] = \mathbb{R}^n$!



A necessary assumption: $\boldsymbol{x}$ is **sparsely generated** from incoherent dictionaries for different classes: $\boldsymbol{x} = [\text{circ}(\mathcal{D}_1), \text{circ}(\mathcal{D}_2), \ldots, \text{circ}(\mathcal{D}_k)]\bar{\boldsymbol{z}}$.

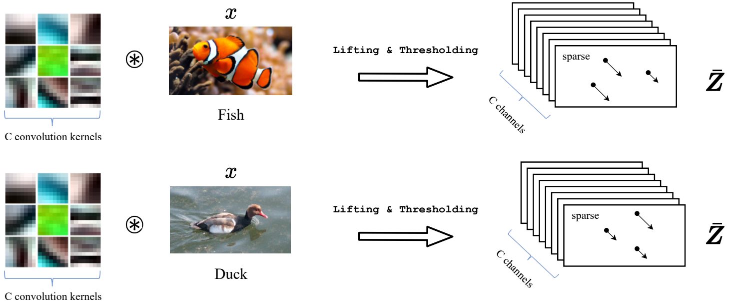


Tradeoff between Invariance and Separability

A basic idea: estimate sparse codes $\bar{\boldsymbol{z}}$ by taking their responses to multiple analysis filters $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_C \in \mathbb{R}^n$ [Rubinstein & Elad 2014]:

$$\bar{\boldsymbol{z}} = \boldsymbol{\tau}\left[\boldsymbol{k}_1 \circledast \boldsymbol{x}, \ldots, \boldsymbol{k}_C \circledast \boldsymbol{x}\right]^* \quad \in \mathbb{R}^{C \times n}. \tag{24}$$

for some entry-wise "sparsity-promoting" operator $\boldsymbol{\tau}(\cdot)$.



Multi-Channel Convolutions

Given a set of multi-channel sparse codes $\bar{\boldsymbol{Z}} = [\bar{\boldsymbol{z}}^1, \ldots, \bar{\boldsymbol{z}}^m]$, construct the ReduNet from their circulant families $\bar{\boldsymbol{Z}} = [\mathrm{circ}(\bar{\boldsymbol{z}}^1), \ldots, \mathrm{circ}(\bar{\boldsymbol{z}}^m)]$.

Proposition (Convolution Structures of $\bar{\boldsymbol{E}}$ and $\bar{\boldsymbol{C}}^j$)

The linear operator in the ReduNet:

$$\bar{\boldsymbol{E}} = \alpha\Big(\boldsymbol{I} + \alpha \sum_{i=1}^{m} \mathrm{circ}(\bar{\boldsymbol{z}}^i)\mathrm{circ}(\bar{\boldsymbol{z}}^i)^*\Big)^{-1} \in \mathbb{R}^{Cn \times Cn}$$

is a block circulant matrix and represents a multi-channel convolution:

$$\bar{\boldsymbol{E}}(\bar{\boldsymbol{z}}) = \bar{\boldsymbol{e}} \circledast \bar{\boldsymbol{z}} \ \in \mathbb{R}^{Cn},$$

where $\bar{e}$ is the first slice of $\bar{\boldsymbol{E}}$. Similarly, the operators $\bar{\boldsymbol{C}}^j$ associated with subsets $\bar{\boldsymbol{Z}}^j$ are also multi-channel circular convolutions.

Multi-Channel Convolutions

$$\bar{\boldsymbol{E}}(\bar{\boldsymbol{z}}) = \bar{\boldsymbol{e}} \circledast \bar{\boldsymbol{z}} \;\in \mathbb{R}^{Cn}, \quad \bar{\boldsymbol{C}}^j(\bar{\boldsymbol{z}}) = \bar{\boldsymbol{c}}^j \circledast \bar{\boldsymbol{z}} \;\in \mathbb{R}^{Cn}:$$

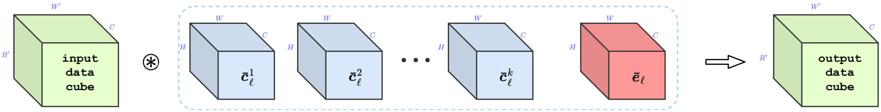


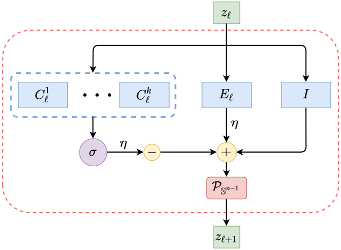
Figure: $\bar{\boldsymbol{E}}$ **and** $\bar{\boldsymbol{C}}^j$ **are automatically multi-channel convolutions!**

The Convolution ReduNet versus Scattering Network

Iterative projected gradient ascent (PGA) for invariant rate reduction:

$$\bar{z}_{\ell+1} \;\propto\; \bar{z}_\ell + \eta \cdot \underbrace{\left[\bar{E}_\ell \bar{z}_\ell + \boldsymbol{\sigma}\left([\bar{C}_\ell^1 \bar{z}_\ell, \ldots, \bar{C}_\ell^k \bar{z}_\ell]\right)\right]}_{g(\bar{z}_\ell, \theta_\ell)}, \tag{25}$$

with each layer being a fixed number of multi-channel convolutions!



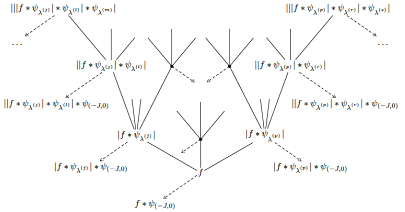


Fig. 2: Scattering network architecture based on wavelet filters and the modulus non-linearity. The elements of the feature vector $\Phi_W(f)$ in (1) are indicated at the tips of the arrows.

Figure: Left: **ReduNet** layer. Right: **Scatterting Network** [J. Bruna and S. Mallat, 2013] [T. Wiatowski and H. Blcskei, 2018] (**only 2-3 layers**).

Fast Computation in Spectral Domain

Fact: all circulant matrices can be simultaneously diagonalized by the *discrete Fourier transform* $\boldsymbol{F}$: $\text{circ}(\boldsymbol{z}) = \boldsymbol{F}^*\boldsymbol{D}\boldsymbol{F}$.

$$\left(\boldsymbol{I} + \sum_{i=1}^{m} \text{circ}(\bar{\boldsymbol{z}}^i)\text{circ}(\bar{\boldsymbol{z}}^i)^*\right)^{-1} = \left(\boldsymbol{I} + \begin{bmatrix} \boldsymbol{F}^* & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{F}^* \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_{11} & \cdots & \boldsymbol{D}_{1C} \\ \vdots & \ddots & \vdots \\ \boldsymbol{D}_{C1} & \cdots & \boldsymbol{D}_{CC} \end{bmatrix} \begin{bmatrix} \boldsymbol{F} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{F} \end{bmatrix}\right)^{-1} \in \mathbb{R}^{nC \times nC}$$

where $\boldsymbol{D}_{ij}$ are all diagonal of size n.

Computing the inverse is $O(C^3 n)$ in the spectral domain, instead of $O(C^3 n^3)$! *Learning convolutional networks for invariant classification is* **naturally far more efficient in the spectral domain***!*

Nature: In visual cortex, neurons encode and transmit information in frequency, hence called "spiking neurons" [Softky & Koch, 1993; Eliasmith & Anderson, 2003].

A "White Box" Deep Convolutional ReduNet by Construction (Spectral Domain)

Require: $\bar{\boldsymbol{Z}} \in \mathbb{R}^{C\times T\times m}$, $\boldsymbol{\Pi}$, $\epsilon > 0$, λ, and a learning rate η.

Set $\alpha = \frac{C}{m\epsilon^2}$, $\{\alpha_j = \frac{C}{\mathrm{tr}(\boldsymbol{\Pi}^j)\epsilon^2}\}_{j=1}^k$, $\{\gamma_j = \frac{\mathrm{tr}(\boldsymbol{\Pi}^j)}{m}\}_{j=1}^k$.

Set $\bar{\boldsymbol{V}}_0 = \{\bar{\boldsymbol{v}}_0^i(p) \in \mathbb{C}^C\}_{p=0,i=1}^{T-1,m} \doteq \mathrm{DFT}(\bar{\boldsymbol{Z}}) \in \mathbb{C}^{C\times T\times m}$.

for $\ell = 1, 2, \ldots, L$ **do**

  for $p = 0, 1, \ldots, T-1$ **do**

    Compute $\bar{\mathcal{E}}_\ell(p) \in \mathbb{C}^{C\times C}$ and $\{\bar{\mathcal{C}}_\ell^j(p) \in \mathbb{C}^{C\times C}\}_{j=1}^k$ as

    $\bar{\mathcal{E}}_\ell(p) \doteq \alpha \cdot \left[\boldsymbol{I} + \alpha \cdot \bar{\boldsymbol{V}}_{\ell-1}(p) \cdot \bar{\boldsymbol{V}}_{\ell-1}(p)^*\right]^{-1}$,

    $\bar{\mathcal{C}}_\ell^j(p) \doteq \alpha_j \cdot \left[\boldsymbol{I} + \alpha_j \cdot \bar{\boldsymbol{V}}_{\ell-1}(p) \cdot \boldsymbol{\Pi}^j \cdot \bar{\boldsymbol{V}}_{\ell-1}(p)^*\right]^{-1}$;

  end for

  for $i = 1, \ldots, m$ **do**

    for $p = 0, 1, \ldots, T-1$ **do**

      Compute $\{\bar{\boldsymbol{p}}_\ell^{ij}(p) \doteq \bar{\mathcal{C}}_\ell^j(p) \cdot \bar{\boldsymbol{v}}_\ell^i(p) \in \mathbb{C}^{C\times 1}\}_{j=1}^k$;

    end for

    Let $\{\bar{\boldsymbol{P}}_\ell^{ij} = [\bar{\boldsymbol{p}}_\ell^{ij}(0), \ldots, \bar{\boldsymbol{p}}_\ell^{ij}(T-1)] \in \mathbb{C}^{C\times T}\}_{j=1}^k$;

    Compute $\left\{\widehat{\boldsymbol{\pi}}_\ell^{ij} = \frac{\exp(-\lambda \|\bar{\boldsymbol{P}}_\ell^{ij}\|_F)}{\sum_{j=1}^k \exp(-\lambda \|\bar{\boldsymbol{P}}_\ell^{ij}\|_F)}\right\}_{j=1}^k$;

    for $p = 0, 1, \ldots, T-1$ **do**

      $\bar{\boldsymbol{v}}_\ell^i(p) = \bar{\boldsymbol{v}}_{\ell-1}^i(p) + \eta\left(\bar{\mathcal{E}}_\ell(p)\bar{\boldsymbol{v}}_\ell^i(p) - \sum_{j=1}^k \gamma_j \cdot \widehat{\boldsymbol{\pi}}_\ell^{ij} \cdot \bar{\mathcal{C}}_\ell^j(p) \cdot \bar{\boldsymbol{v}}_\ell^i(p)\right)$;

    end for

    $\bar{\boldsymbol{v}}_\ell^i = \bar{\boldsymbol{v}}_\ell^i \;/\; \|\bar{\boldsymbol{v}}_\ell^i\|_F$;

  end for

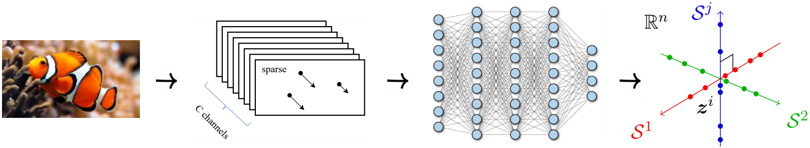
  Set $\bar{\boldsymbol{Z}}_\ell = \mathrm{IDFT}(\bar{\boldsymbol{V}}_\ell)$ as the feature at the ℓ-th layer;

  $\frac{1}{2T}\sum_{p=0}^{T-1}\left(\log\det[\boldsymbol{I} + \alpha\bar{\boldsymbol{V}}_\ell(p) \cdot \bar{\boldsymbol{V}}_\ell(p)^*] - \frac{\mathrm{tr}\left(\boldsymbol{\Pi}^j\right)}{m}\log\det[\boldsymbol{I} + \alpha_j\bar{\boldsymbol{V}}_\ell(p) \cdot \boldsymbol{\Pi}^j \cdot \bar{\boldsymbol{V}}_\ell(p)^*]\right)$;

end for

Ensure: features $\bar{\boldsymbol{Z}}_L$, the learned filters $\{\bar{\mathcal{E}}_\ell(p)\}_{\ell,p}$ and $\{\bar{\mathcal{C}}_\ell^j(p)\}_{j,\ell,p}$.

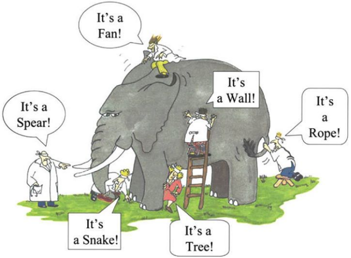
Overall Process (the Elephant)



Multi-Class Signals → Lifting & Sparse Coding → Invariant Rate Reduction → Incoherent Subspaces

Necessary components:

- **sparse coding** for class separability;
- **deep networks** maximize rate reduction;
- **spectral computing** for shift-invariance;
- convolution, normalization, nonlinearity...



Experiment: 1D Cyclic Shift Invariance of 0 and 1

2000 training samples, 1980 testing, $C = 5$, $L =$ **3500**-layers ReduNet.

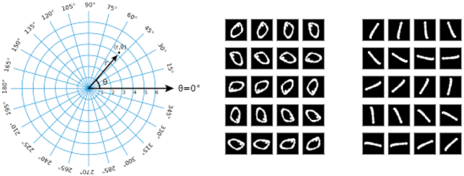


Figure: Left: Multi-channel feature representation of an image in polar coordinates. Right: Example of training/testing samples.

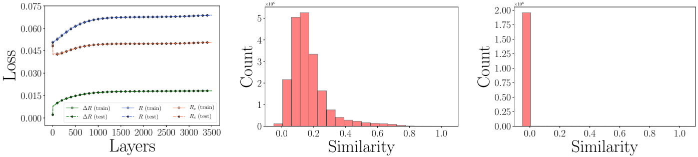


Figure: Left: Rates along the layers; Middle: cross-class cosine similarity among trainings; Right: similarity among testings.

Experiment: 1D Cyclic Shift Invariance of 0 and 1

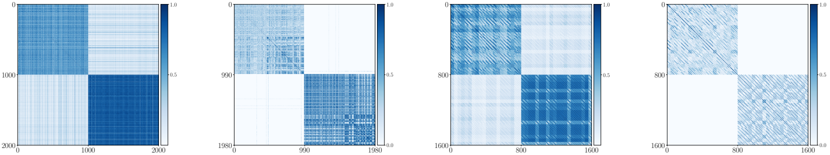


Figure: Left two: heat maps for training and testing. Right two: heat maps for one pair of samples at every possible shift.

Table: Network performance on digits with **all rotations**.

	ReduNet	ReduNet (invariant)
Acc (Original Test Data)	0.983	0.996
Acc (Test with All Shifts)	0.707	0.993

1. Fooling CNNs with simple transformations, Engstrom et.al., 2017.

2. Why do deep convolutional networks generalize so poorly to small image transformations? Azulay & Weiss, 2018.

Experiment: 1D Cyclic Shift Invariance of All 10 Digits

100 training samples, 100 testing, $C = 20$, $L =$**40**-layers ReduNet.

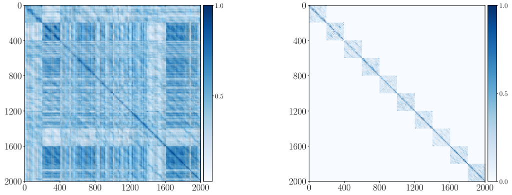


Figure: Heatmaps of cosine similarity among shifted training data $\boldsymbol{X}_{\text{shift}}$ (left) and learned features $\boldsymbol{Z}_{\text{shift}}$ (right).

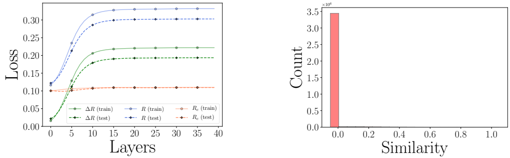
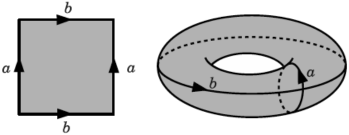


Figure: Left: Rates evolution with iterations; Right: histograms of the cosine similarity (in absolute value) between all pairs of features across different classes.

Experiment: 2D Cyclic Translation Invariance

1000 for training, 500 for testing, $C = 5$, $L =$**2000**-layers ReduNet.



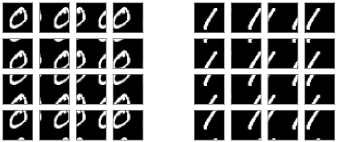


Table: Network performance on digits with **all translations**.

	ReduNet	ReduNet (invariant)
Acc (Original Test Data)	0.980	0.975
Acc (Test with All Shifts)	0.540	0.909

1. Fooling CNNs with simple transformations, Engstrom et.al., 2017.
2. Why do deep convolutional networks generalize so poorly to small image transformations? Azulay & Weiss, 2018.

Experiment: 2D Cyclic Trans. Invariance of All 10 Digits

100 training samples, 100 testing, $C = 75$, $L =$**25**-layers ReduNet.

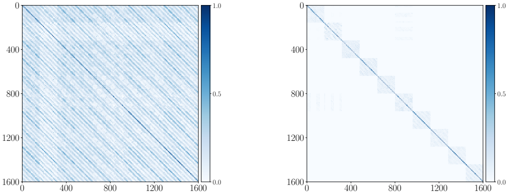


Figure: Heatmaps of cosine similarity among shifted training data $\boldsymbol{X}_{\text{shift}}$ (left) and learned features $\boldsymbol{Z}_{\text{shift}}$ (right).

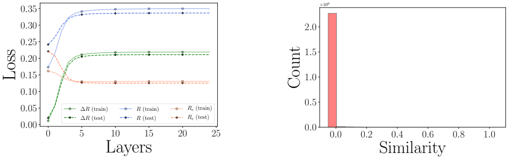


Figure: Left: Rates evolution with iterations; Right: histograms of the cosine similarity (in absolute value) between all pairs of features across different classes.

Experiment: Back Propagation of ReduNet

2D cyclic trans. of 10 digits, 500 training samples, all testing, $C = 16$, $L =$**30**-layers invariant ReduNet.

Initialization	Backpropagation	Test Accuracy
✓	✗	89.8%
✗	✓	93.2%
✓	✓	97.8%

Table: Test accuracy of 2D translation-invariant ReduNet, ReduNet-bp (without initialization), and ReduNet-bp (with initialization) on the MNIST dataset.

- **Backprop:** the ReduNet architecture *can* be fine-tuned by SGD and achieves better standard accuracy after back propagation;
- **Initialization:** using ReduNet for initialization can achieve better performance than the same architecture with random initialization.

Conclusions: Learn to Compress and Compress to Learn!

Principles of Parsimony:

- Clustering via compression: $\min_{\mathbf{\Pi}} R^c(\boldsymbol{X}, \mathbf{\Pi})$
- Classification via compression: $\min_{\boldsymbol{\pi}} \delta R^c(\boldsymbol{x}, \boldsymbol{\pi})$
- Representation via maximizing rate reduction: $\max_{\boldsymbol{Z}} \Delta R(\boldsymbol{Z}, \mathbf{\Pi})$
- Deep networks via optimizing rate reduction: $\dot{\boldsymbol{Z}} = \eta \cdot \frac{\partial \Delta R}{\partial \boldsymbol{Z}}$

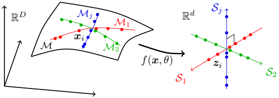
A Unified Framework:

- A principled objective for all settings of learning: **compression**
- A principled approach to interpret deep networks: **optimization**

"Everything should be made as simple as possible, but not simpler."
– Albert Einstein

Conclusions: Learn Linear Discriminative Representations

Compared to conventional deep neural networks:



	Conventional DNNs	Compression ReduNets
Objectives	label fitting	rate reduction
Deep architectures	trial & error	iterative optimization
Layer operators	empirical	projected gradient
Shift invariance	CNNs+augmentation	invariant ReduNets
Initializations	random/pre-design	forward computed
Training/fine-tuning	back prop	forward/back prop
Interpretability	black box	white box
Representations	unknown	incoherent subspaces

Open Problems: Theory

$$\textbf{MCR}^2\text{:}\quad \max_{\boldsymbol{Z}\subset\mathbb{S}^{d-1},\boldsymbol{\Pi}\in\Omega} \Delta R\big(\boldsymbol{Z},\boldsymbol{\Pi},\epsilon\big) = R(\boldsymbol{Z},\epsilon) - R^c(\boldsymbol{Z},\epsilon \mid \boldsymbol{\Pi}).$$

- **Phase transition** phenomenon in clustering via compression?
- Statistical justification for **robustness** of MCR^2 to label noise?
- **Optimal configurations** for broader conditions and distributions?
- Fundamental **tradeoff** between sparsity and invariance?
- **Jointly optimizing** both representation $\boldsymbol{Z}$ and clustering $\boldsymbol{\Pi}$?

$$\textbf{Joint Dynamics}\text{:}\quad \dot{\boldsymbol{Z}} = \eta \cdot \frac{\partial \Delta R}{\partial \boldsymbol{Z}}, \quad \dot{\boldsymbol{\Pi}} = \gamma \cdot \frac{\partial \Delta R}{\partial \boldsymbol{\Pi}}.$$

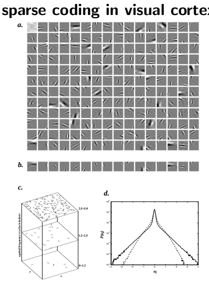
Open Problems: Architectures and Algorithms

ReduNet: $\bar{\boldsymbol{z}}_{\ell+1} \ \propto \ \bar{\boldsymbol{z}}_{\ell} + \eta \cdot \Big[\bar{\boldsymbol{e}}_{\ell} \circledast \bar{\boldsymbol{z}}_{\ell} + \boldsymbol{\sigma}\big([\bar{\boldsymbol{c}}_{\ell}^{1} \circledast \bar{\boldsymbol{z}}_{\ell}, \ldots, \bar{\boldsymbol{c}}_{\ell}^{k} \circledast \bar{\boldsymbol{z}}_{\ell}]\big) \Big] \in \mathbb{S}^{d-1}.$

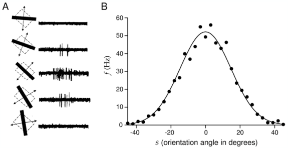
- New architectures from **accelerated** gradient schemes?
- Conditions for channel-wise **separable and short** convolutions?
- Architectures from invariant rate reduction for **other groups**?
- **Transformer**: $\frac{1}{2} \log\det \Big(\boldsymbol{I} + \alpha \boldsymbol{Z}\boldsymbol{Z}^{*} \Big) = \frac{1}{2} \log\det \Big(\boldsymbol{I} + \alpha \underbrace{\boldsymbol{Z}^{*}\boldsymbol{Z}}_{\boldsymbol{z}_i^{*}\boldsymbol{U}^{*}\boldsymbol{U}\boldsymbol{z}_j?} \Big)$?
- Algorithmic architectures (or networks) for optimizing $\boldsymbol{\Pi}$?

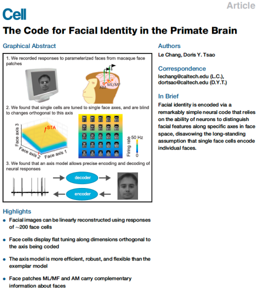
Open Directions: Extensions

- Data with other **dynamical or graphical** structures.
- Better transferability and robustness w.r.t. **low-dim structures**.
- Combine with a **generative model** (a generator or decoder).
- Sparse coding, spectral computing, subspace embedding in **Nature**.[1]



Rate coding hypothesis: the signal conveyed by a neuron is in the *rate* of spiking. Spiking irregularity is largely due to noise and does not convey information.




Article

Cell

The Code for Facial Identity in the Primate Brain

Graphical Abstract

Authors

Le Chang, Doris Y. Tsao

Correspondence

lechang@caltech.edu (L.C.),
dortsao@caltech.edu (D.Y.T.)

In Brief

Facial identity is encoded via a remarkably simple neural code that relies on the ability of neurons to distinguish facial features along specific axes in face space, disavowing the long-standing assumption that single face cells encode individual faces.

Highlights

- Facial images can be linearly reconstructed using responses of ~200 face cells
- Face cells display flat tuning along dimensions orthogonal to the axis being coded
- The axis model is more efficient, robust, and flexible than the exemplar model
- Face patches ML/MF and AM carry complementary information about faces

[1] figures from Bruno Olshausen of Neuroscience Dept., UC Berkeley.

References: Learning via Compression and Rate Reduction

1. **Clustering** via Lossy Coding and Compression (TPAMI 2007): http://people.eecs.berkeley.edu/~yima/psfile/Ma-PAMI07.pdf
2. **Classification** via Minimal Incremental Coding Length (NIPS 2007): http://people.eecs.berkeley.edu/~yima/psfile/MICL_SJIS.pdf
3. **Representation** via Maximal Coding Rate Reduction (NeurIPS 2020): https://arxiv.org/abs/2006.08558
4. **ReduNet:** A Whitebox Deep Network from Maximizing Rate Reduction: https://arxiv.org/abs/2105.10446
5. **A New Textbook**: *High-Dim Data Analysis with Low-Dim Models* https://book-wright-ma.github.io

Source Code: Whitebox ReduNet

1. **Github Link**: https://github.com/Ma-Lab-Berkeley/ReduNet
2. **Google Colab**: https://colab.research.google.com/github/ryanchankh/redunet_demo/blob/master/gaussian3d.ipynb
3. **Jupyter Notebook**: https://github.com/ryanchankh/redunet_demo/blob/master/gaussian3d.ipynb

"What I cannot create, I do not understand."
– Richard Feynman

Deep (Convolution) Network Architectures are Iterative Optimization Schemes for Compression!

Thank you!

Questions, please?