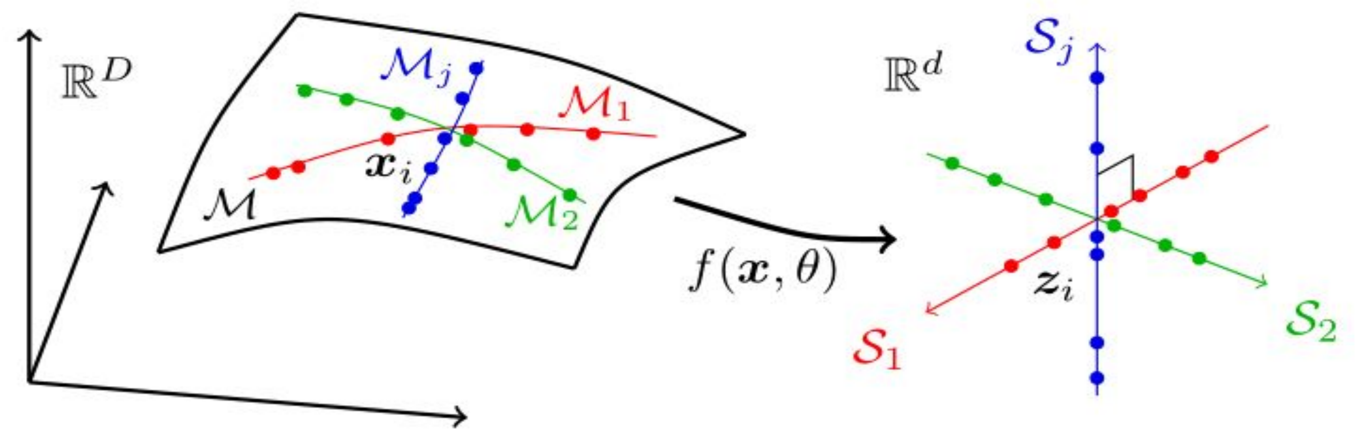
Learning Diverse and Discriminative Representations via the Principle of Maximal Coding Rate Reduction

Yaodong Yu*[1], Kwan Ho Ryan Chan*[1], Chong You[1], Chaobing Song[1,2], Yi Ma[1]
University of California, Berkeley[1], Tsinghua University[2]

Desired Properties in Learned Features

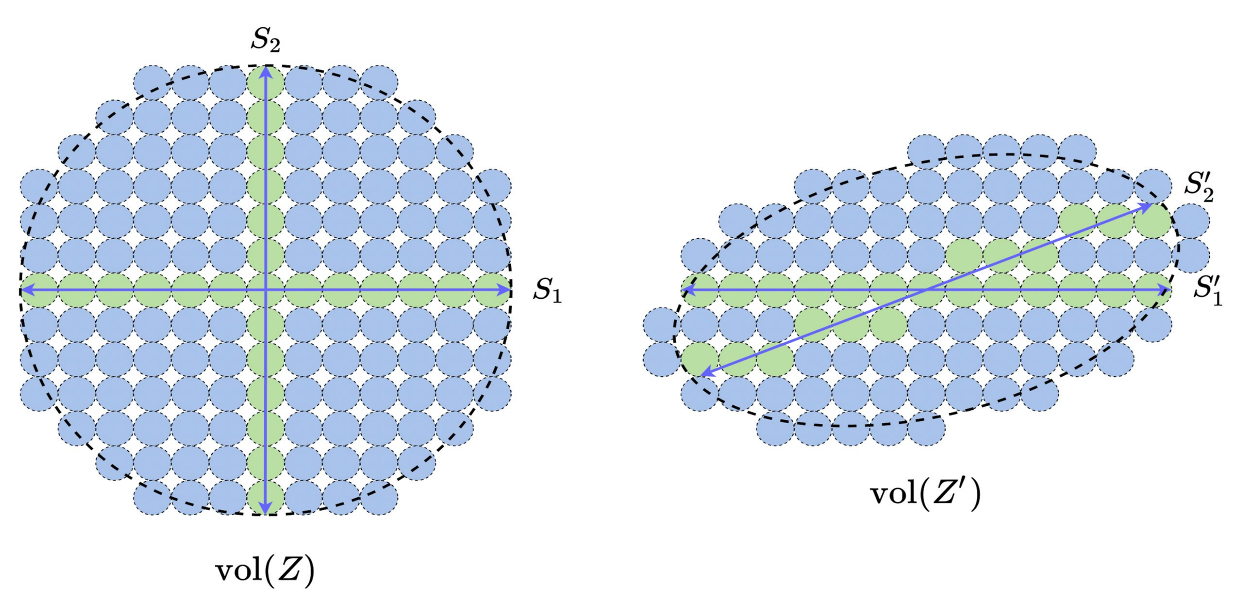
Goal: Given a random vector $x \in \mathbb{R}^D$ drawn from a mixture of k distributions, $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^k$ we **seek a good representation** $z \in \mathbb{R}^d$ through a continuous mapping, $f(x, \theta) = \mathbb{R}^D \to \mathbb{R}^d$, such that:



1. ***Between-Class Discriminative***: Features of samples from different classes/clusters should be highly uncorrelated and belong to different low-dimensional linear subspaces.
2. **Within-Class Compressible**: Features of samples from the same class/cluster should be relatively correlated in a sense that they belong to a low-dimensional linear subspace.
3. **Maximally Diverse Representation**: Dimension (or variance) of features for each class/cluster should be as large as possible as long as they stay uncorrelated from the other classes.

Intuitive Visualization

Here we offer a informational geometric perspective on coding rate reduction using sphere packing:



We consider 2-dimensional case, with two two distributions S_1 and S_2.

- ∑ [vol(green spheres)] = sum of coding rate of subspace R^c.
- ∑ [vol(green spheres + blue spheres)] = sum of coding rate of all samples R.
- ∑ [vol(blue spheres)] = loss

MCR²

MCR² loss aims to maximize the reduction *between* the coding rate of all features and that of the sum of features w.r.t. their classes:

$$\max_{\theta} \ \Delta R(\theta) = \underbrace{\frac{1}{2}\log\det\left(I + \frac{d}{m\epsilon^2}ZZ^\top\right)}_{R} - \underbrace{\sum_{j=1}^{k}\frac{\mathrm{tr}(\Pi_j)}{2m}\log\det\left(I + \frac{d}{\mathrm{tr}(\Pi_j)\epsilon^2}Z\Pi_j Z^\top\right)}_{R^c}$$

where $\|Z_j(\theta)\|_F^2 = m_j, j \in [k]$

- Rate distortion of data with a mixed distribution: The features of Z of multi-class data may belong to multiple low-dimensional subspaces, and we may partition the data Z into multiple subsets: $Z = Z_1 \cup Z_2 \cup \cdots \cup Z_k$. With respect to this partition, the average number of bits per sample (the coding rate) is:

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

where $\Pi = \{\Pi_j \in \mathbb{R}^{m\times m}\}_{j=1}^k$ be a set of diagonal matrices whose diagonal entries encode the membership of the m samples in the k classes and the diagonal entry $\Pi_j(i,i)$ of Π_j indicates the probability of sample i belonging to subset j.

- Nonasymptotic rate distortion for finite samples: The average coding length per sample (as the sample m is large) subject to the distortion ϵ:

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

Each sample should be as decorrelated as possible to encourages diversity across all learned representations z.

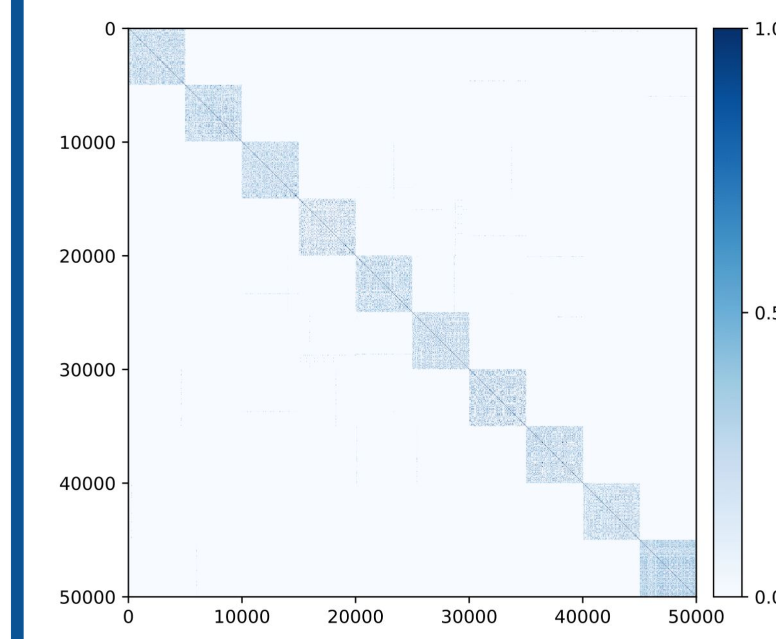
Theorem

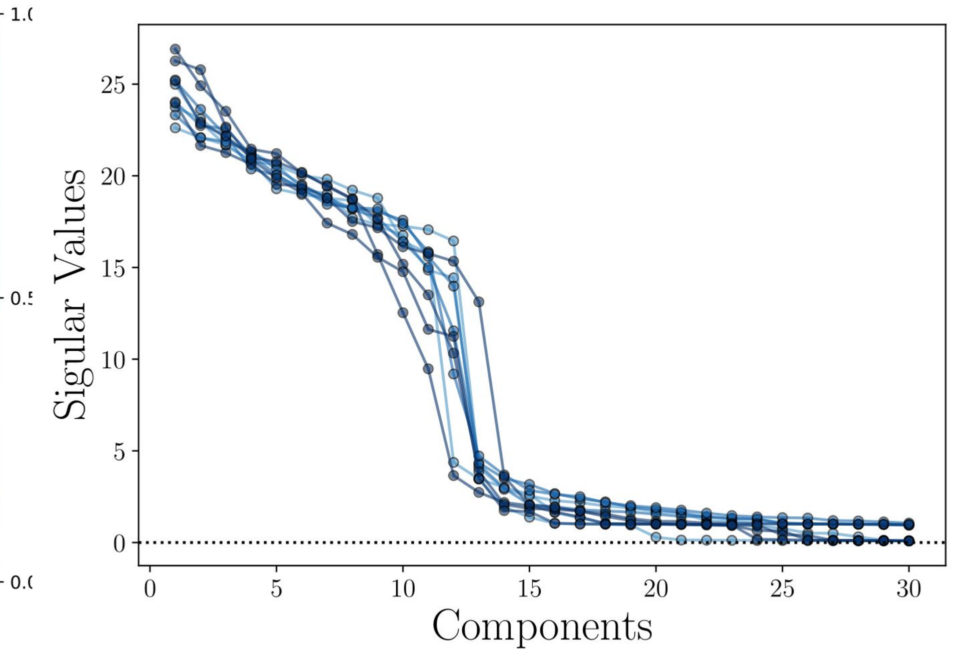
Here we provide theoretical guarantee to the properties of learned representations: (Theorem 2.1)

Since $Z^* = Z_1^* \cup \cdots Z_k^*$ is the optimal solution that maximizes the rate reduction. 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. $(Z_i^*)^\top Z_j^* = 0$ for $i \neq j$.
- *Maximal Diverse Representations:* As long as the coding precision is adequately high, i.e., $\epsilon^4 < \min_j\{\frac{m_j}{m}\frac{d^2}{d_j^2}\}$, each subspace achieves its maximal dimension, i.e. $\mathrm{rank}(Z_j^*) = d_j$. In addition, the largest $d_j - 1$ singular values of Z_j^* are equal.

Desired Properties in Learned Features





- **Setup:** Image classification task on CIFAR10 using ResNet18, with $d = 128$.
- **Left:** A heatmap of the cosine similarity score between features. Each class has 5,000 samples and their features span a subspace of over 10 dimensions. Here we can see that *between-class features are discriminative* and *in-class features are highly correlated*.
- **Right:** Singular values after performing Principal Component Analysis on each class of features. We can see that each subspace spans approximately 12-13 dimensions, which in total spans the whole 128 dimensional representation space.

Robustness to Label Corruption: Classification Results for features learned with labels at different corruption level. CE training means Cross-Entropy Training. As we can see, features learned using MCR² are more robust to label corruption:

	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%
MCR² Training	**91.16%**	**89.70%**	**88.18%**	**86.66%**	**84.30%**

Clustering: Results based on features learned using self-supervised learning. MCR² also has superior results over multiple datasets:

Dataset	Metric	JULE	RTM	DEC	DAC	DCCM	MCR²-Ctrl
CIFAR10	NMI	0.192	0.197	0.257	0.395	0.496	**0.630**
	ACC	0.272	0.309	0.301	0.521	0.623	**0.684**
	ARI	0.138	0.115	0.161	0.305	0.408	**0.508**
CIFAR100	NMI	0.103	-	0.136	0.185	0.285	**0.387**
	ACC	0.137	-	0.185	0.237	0.327	**0.375**
	ARI	0.033	-	0.050	0.087	0.173	**0.178**
STL10	NMI	0.182	-	0.276	0.365	0.376	**0.446**
	ACC	0.182	-	0.359	0.470	0.482	**0.491**
	ARI	0.164	-	0.186	0.256	0.262	**0.290**