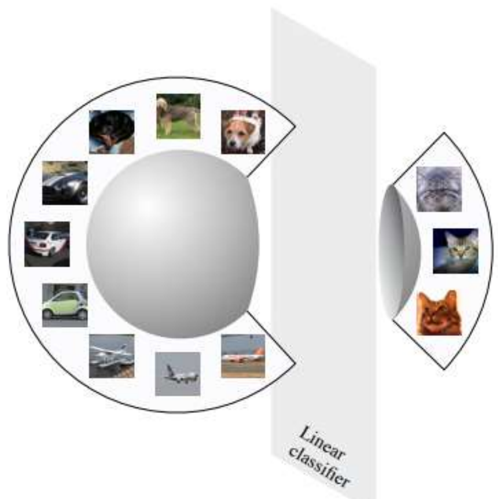
Understanding Contrastive Representation Learning through Geometry on the Hypersphere

Xiaobing Chen
May 11, 2021

New Metrics for Contrastive Learning

Representation on the Unit Hypersphere

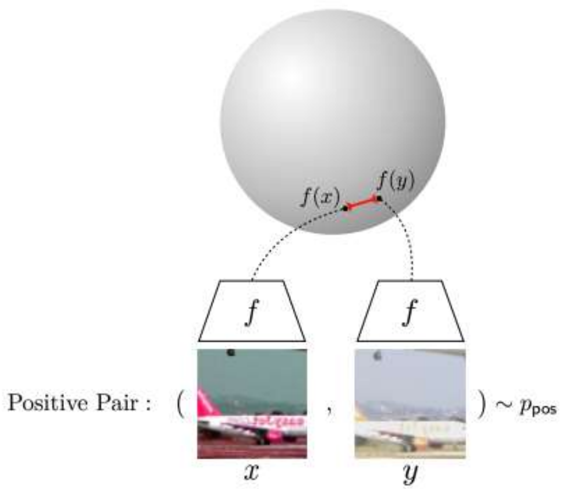
- Many work learn representations using l2 normalization, which restricts the latent space to the unit hypersphere
- Intuitively, high quality of representations can be linearly separable in the latent space.

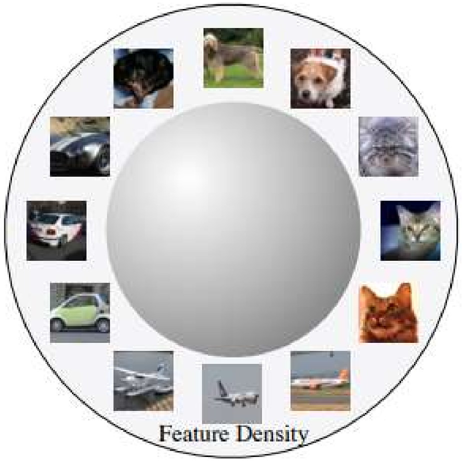


- Besides the clustering property, good representations should be invariant to unnecessary information. (InfoMax principle.)

Properties of Good Representations

- To explore what desirable properties of good representations, authors propose two metrics: alignment and uniformity, and use them to interpret the success of existing contrastive loss, InfoNCE.





Alignment: to measure the distance between positive representations.

representations of positive pairs are close in the latent space.

Uniformity: to measure how well representations are uniformly distributed.

uniform distribution would preserve maximal information.

Background-Contrastive Learning

Data:

$p_{data}(\cdot)$: data distribution in $\mathbb{R}^n$
$p_{pos}(\cdot)$: positive pair distribution over $\mathbb{R}^n \times \mathbb{R}^n$

Encoder:

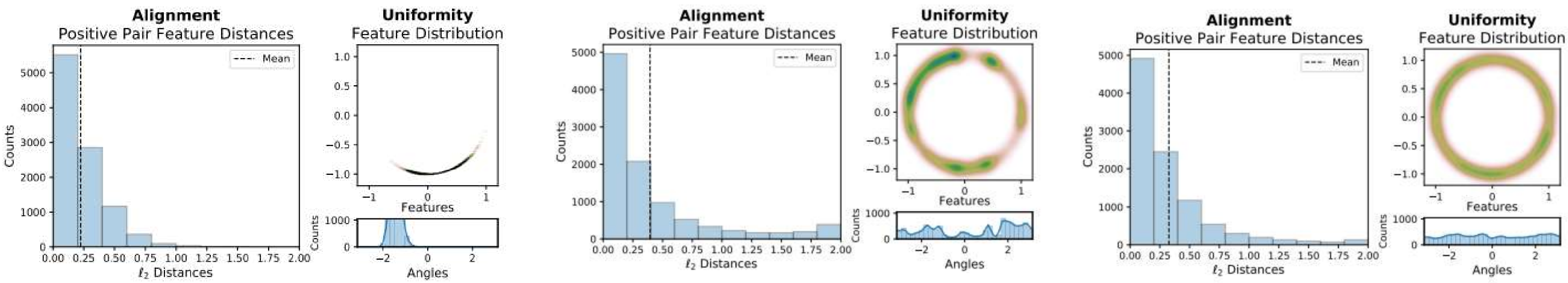
$$f : \mathbb{R}^n \to \mathbb{R}^d$$

InfoNCE loss:

$$\mathcal{L}_{\mathsf{contrastive}}(f; \tau, M) \triangleq \mathop{\mathbb{E}}_{\substack{(x,y)\sim p_{\mathsf{pos}} \\ \{x_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\mathsf{data}}}} \left[-\log \frac{e^{f(x)^\top f(y)/\tau}}{e^{f(x)^\top f(y)/\tau} + \sum_i e^{f(x_i^-)^\top f(y)/\tau}} \right]$$

Properties of Good Representations

- Toy example:
  - Train CIFAR-10 encoders
  - 2-dim representations, 1-sphere feature space (circle)
  - Computing alignment: l2 distance of positive pairs; uniformity: Gaussian kernel density estimation
  - Visualize feature distributions on the validation set.



Random Initialization:
Linear classification accuracy: 12.71%

Supervised Learning:
Linear classification accuracy: 57.19%

Contrastive Learning:
Linear classification accuracy: 28.60%

Features from contrastive learning is most uniformly distributed !

Quantifying Alignment and Uniformity

Alignment: expected distance between positive pairs

$$\mathcal{L}_{\text{align}}(f;\alpha) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left[\|f(x) - f(y)\|_2^{\alpha}\right], \quad \alpha > 0.$$

Uniformity: logarithm of the expected pairwise Gaussian potential

$$\mathcal{L}_{uniform}(f;t) = \log(\mathop{\mathbb{E}}_{\substack{x\sim p_{data}\\ y\sim p_{data}}}(G_t(f(x), f(y)))), t > 0$$

where $G_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is the radial basis kernel function, defined as follows.

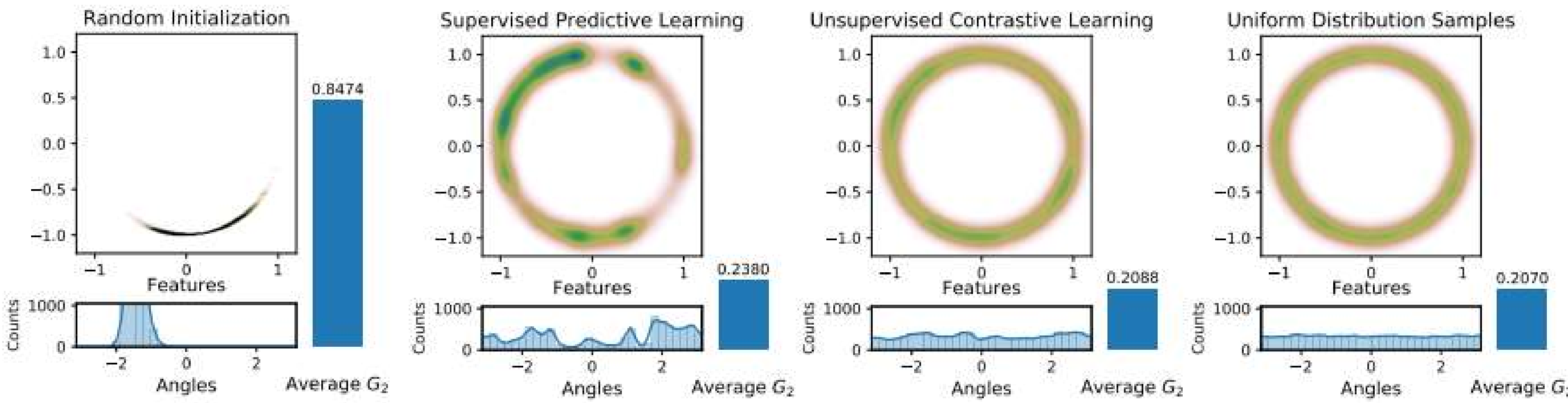
$$G_t(u, v) = e^{-t\|u-v\|^2} = e^{2tu^T v - 2t}, t > 0$$

where u and v are normalized vectors.

Why choose RBF kernel?

Evaluating Uniformity

Empirical evaluation of toy example: evaluate the average pairwise potential of various finite point collections.



The average G2 decreases as the distribution becomes more uniform. It's a good metric for uniformity.

Asymptotics of InfoNCE

Theorem 1 is proposed to connect alignment and uniformity with InfoNCE.

Theorem 1 (Asymptotics of $\mathcal{L}_{\text{contrastive}}$). *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\begin{aligned}
&\lim_{M\to\infty} \mathcal{L}_{\text{contrastive}}(f;\tau,M) - \log M \\
&= \lim_{M\to\infty} \mathop{\mathbb{E}}_{\substack{(x,y)\sim p_{\text{pos}} \\ \{x_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[-\log \frac{e^{f(x)^\intercal f(y)/\tau}}{e^{f(x)^\intercal f(y)/\tau} + \sum_i e^{f(x_i^-)^\intercal f(y)/\tau}} \right] - \log M \\
&= -\frac{1}{\tau} \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left[f(x)^\intercal f(y) \right] + \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} \left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\text{data}}} \left[e^{f(x^-)^\intercal f(x)/\tau} \right] \right].
\end{aligned}$$

We have the following results:

1. *The first term is minimized iff f is perfectly aligned.*
2. *If perfectly uniform encoders exist, they form the exact minimizers of the second term.*

Minimizing InfoNCE loss indeed requires both alignment and uniformity.

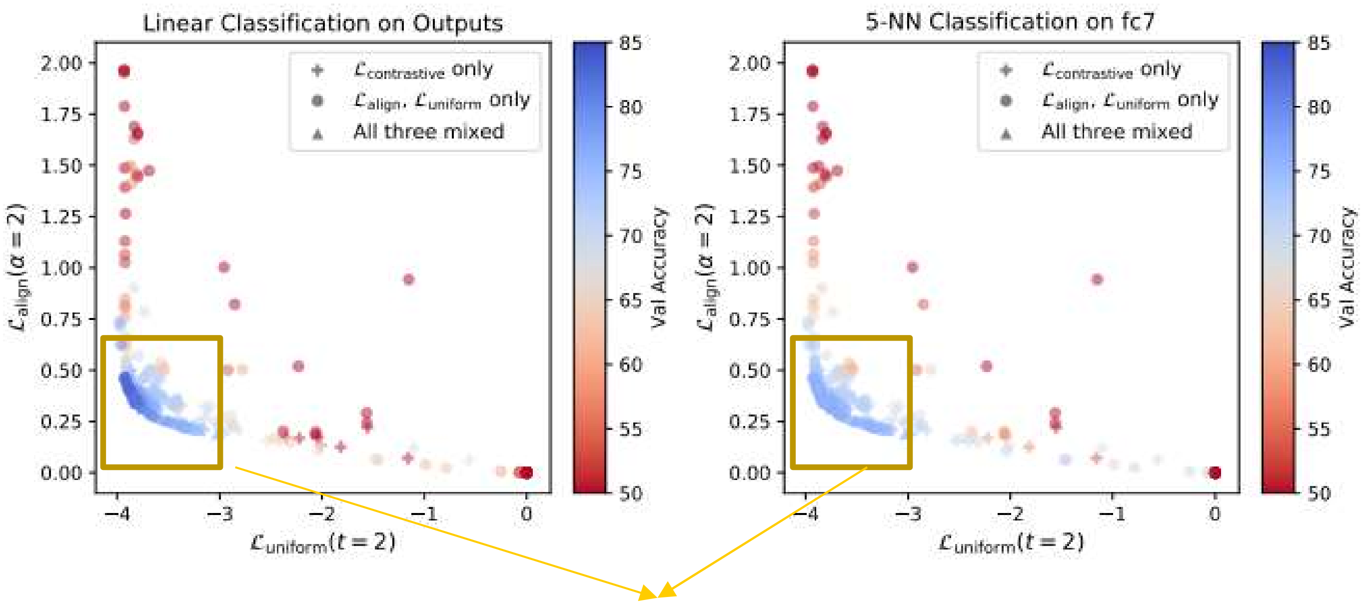
Experiments

LSU

Experimental Settings

- **Encoder**: multiple encoders based on CNN and RNN
- **Downstream task & datasets**:
  - STL-10: classification on AlexNet based encoder outputs or intermediate activations with a linear or k-nearest neighbor (k-NN) classifier.
  - NYU-DEPTH-V2: depth prediction on CNN encoder intermediate activations after convolution layers.
  - IMAGENET and IMAGENET-100 (random 100-class subset of IMAGENET): classification on CNN encoder penultimate layer activations with a linear classifier.
  - BOOKCORPUS: RNN sentence encoder outputs used for Moview Review Sentence Polarity (MR) (Pang & Lee, 2005) and Customer Product Review Sentiment (CR) (Wang & Manning, 2012) binary classification tasks with logisitic classifiers.
- **Positive pair construction**:
  - two augmented images from the same image for image-relevant tasks
  - neighboring sentences for NLP
- Experiments with varying:
  - Objectives: weighted combinations of $\mathcal{L}_{\text{contrastive}}$, $\mathcal{L}_{\text{align}}$, and/or $\mathcal{L}_{\text{uniform}}$
  - Hyperparameters: temperatures, batch size, representation dim, …

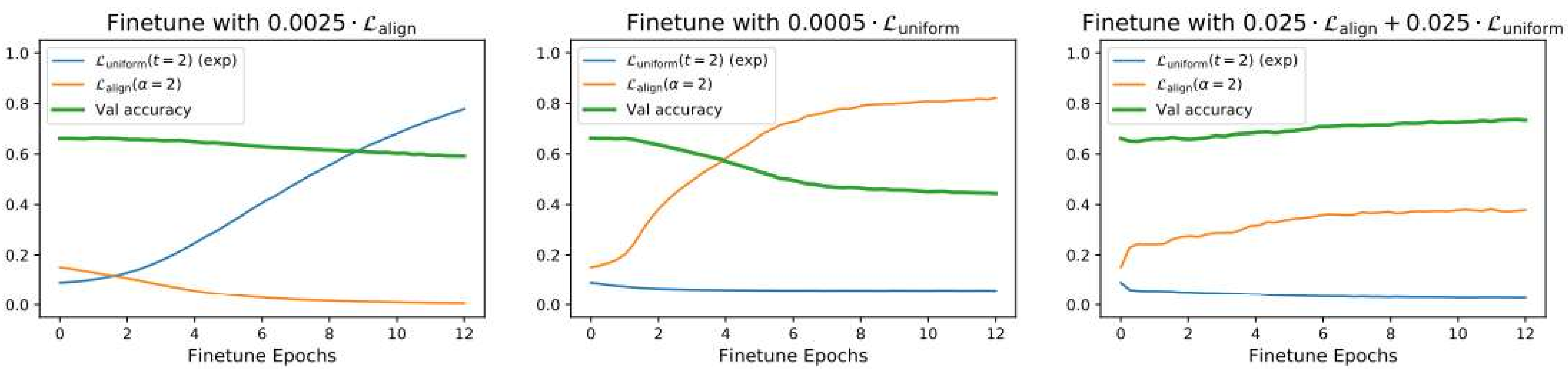
Results-STL-10 Classification



High performance region

Best downstream performance give lowest alignment and uniformity.

Results-STL-10 Classification



Both alignment and uniformity are necessary for a good representation.

Conclusions & Discussions

Conclusions & Discussions

- Alignment and uniformity are two important goals of contrastive learning and good metrics to evaluate the quality of representation.
- It provides theoretical proof that uniform distribution on sphere is the optimal solution for the Radial Basis Function (RBF) kernel.
- Why unit hypersphere benefits the representation learning remains explored.

Thanks !

Q & A