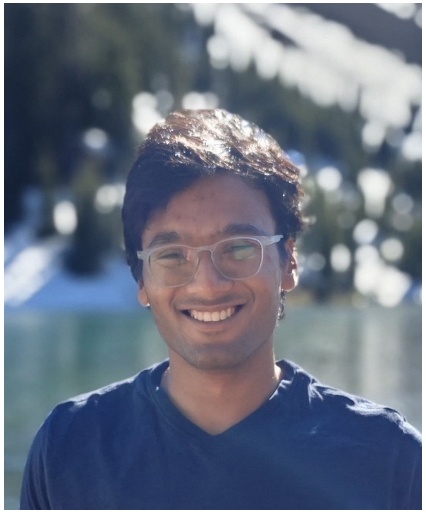
Understanding Contrastive Learning via Gaussian Mixture Models



Parikshit Bansal, Ali Kavis, Sujay Sanghavi

University of Texas, Austin

Understanding Model Retraining using its own outputs

Rudrajit Das

Just another guy working towards a PhD in Machine Learning

w/ Rudrajit Das
University of Texas, Austin

Understanding Model Retraining using its own outputs

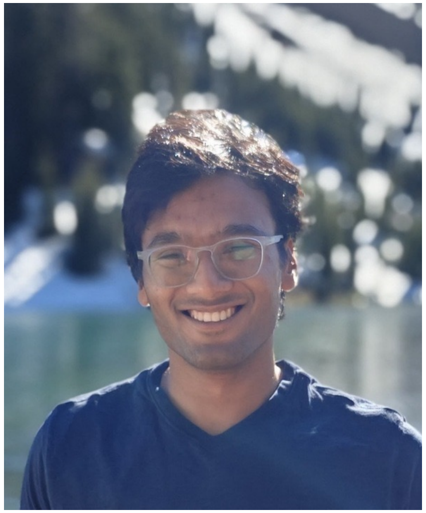
Rudrajit Das

Just another guy working towards a PhD in Machine Learning

w/ Rudrajit Das

University of Texas, Austin

Understanding Contrastive Learning via Gaussian Mixture Models



Parikshit Bansal, Ali Kavis, Sujay Sanghavi

University of Texas, Austin

Contrastive Learning

Unsupervised way to learn representations

Built on a simple idea:

- **Pairs of points :** For every point, we are given a “partner” point
    - Partner may be same modality, or a different modality
- **Contrastive Loss :** Embedding of a point should be **close to its “partner”** while being **far from everything else**

Contrastive Learning

Unsupervised way to learn representations

Built on a simple idea:

- **Pairs of points :** For every point, we are given a “partner” point
    - Partner may be same modality, or a different modality
- **Contrastive Loss :** Embedding of a point should be **close to its “partner”** while being **far from everything else**

Common recent theme:

Train (very large) models on a (very large) corpus of unsupervised data using contrastive loss

Fine-tune / further train / adapt this model to downstream task (e.g. classification)

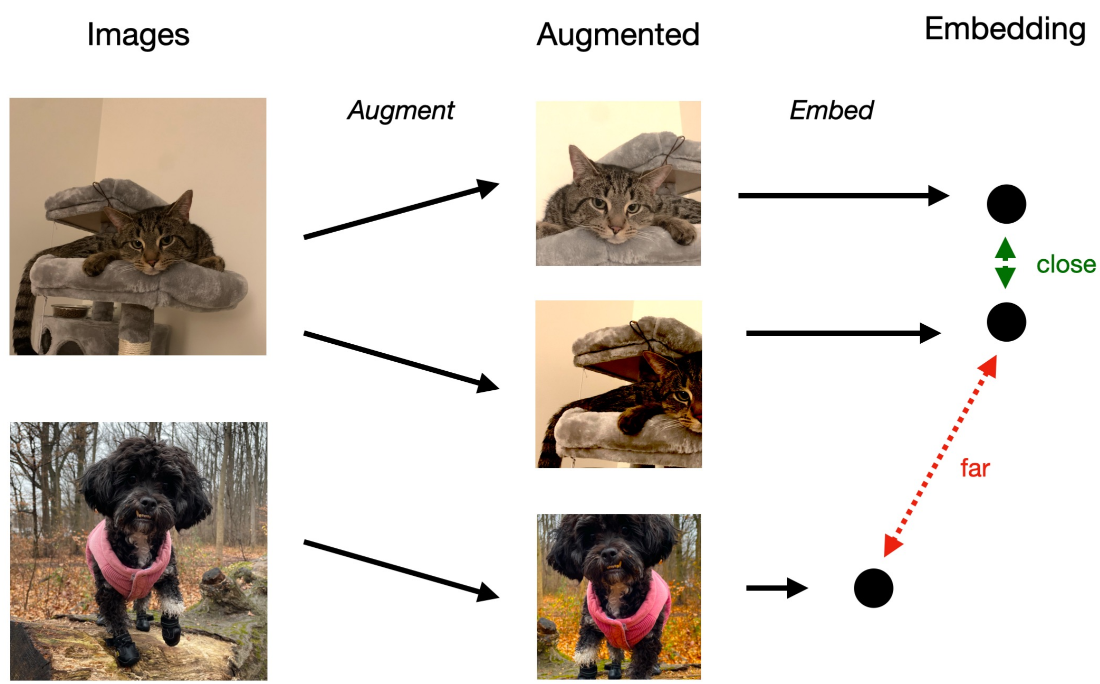
The simple idea works, often zero-shot, often with no knowledge of the downstream task

Contrastive Learning

Example 1:

SimCLR
[Chen et. al., 2020]

Partners from (natural image specific) augmentations like cropping, rotation Etc.

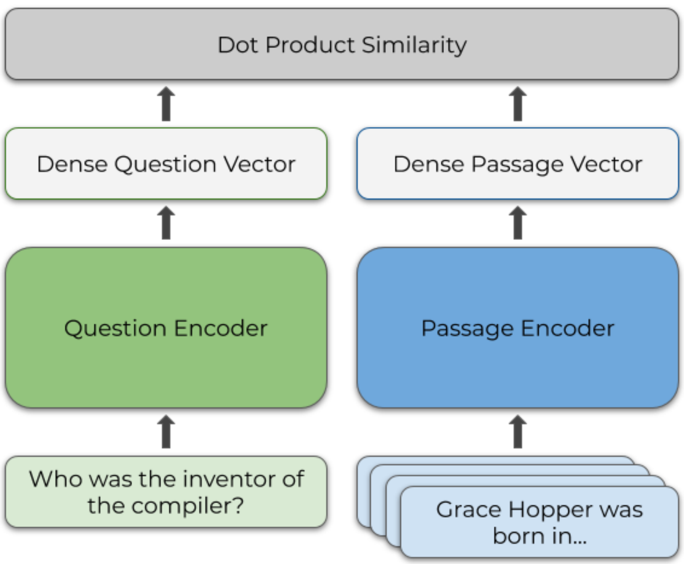


Trunk model trained using SimCLR + last layer fine-tuned on imagenet-1k

beat previous state of the art (and fully supervised training) by 7% in accuracy

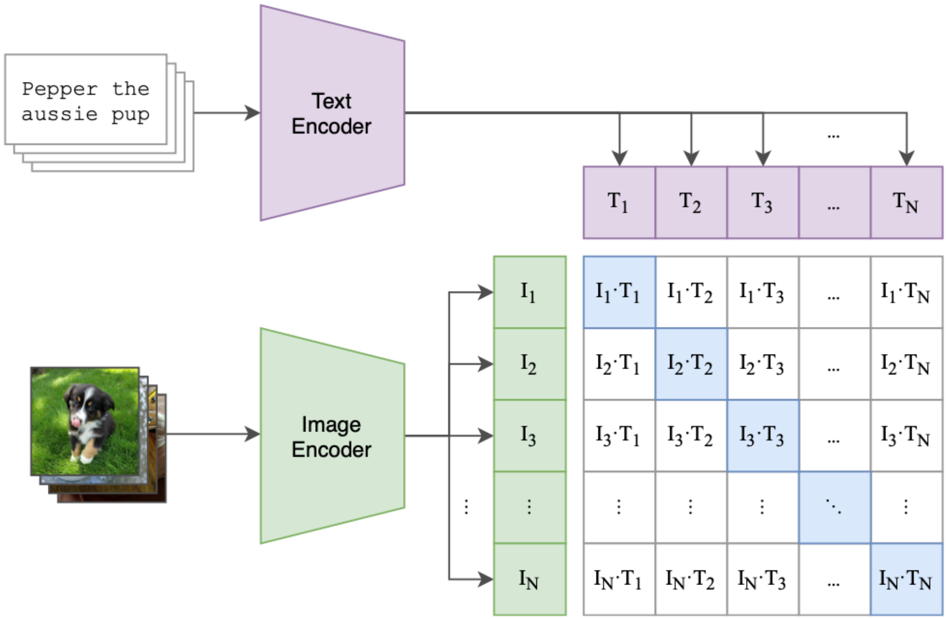
Contrastive Learning

Example 2:
Dense Passage Retrieval (DPR) [Karpukhin et. al]



“multimodal”

Example 3:
CLIP [Radford et. al.]



multimodal

Contrastive Losses: InfoNCE

Data is $(x, \hat{x})$ **pairs**

Point and its partner are close

$$\min_f \; -\mathbb{E}_x \left[\log \left(\frac{\exp(f(x)^T f(\hat{x}))}{E_y[\exp(f(x)^T f(y))]} \right) \right]$$

Point is far from random other point

Linear case: $f(x) = Ax$

Contrastive Losses: CLIP

Data is (x_V, x_T) **pairs**

Point and its (other modality) partner are close

$$\min_{f_V, f_T} \; -\mathbb{E}_x \left[\log \left(\frac{\exp(f_V(x_V)^T f_T(x_T))}{E_y[\exp(f_V(x_V)^T f_T(y_T))]} \right) \right]$$

$$-\mathbb{E}_x \left[\log \left(\frac{\exp(f_V(x_V)^T f_T(x_T))}{E_y[\exp(f_T(x_T)^T f_V(y_V))]} \right) \right]$$

Point is far from random other point (in other modality)

Linear case:

$$f_T(x_T) = A_T x_T \qquad f_V(x_V) = A_V x_V$$

Understanding Contrastive Learning

***<u>Why</u>* does the simple idea of contrastive learning, using pairs of points, work so well in learning representations …. ?**

i.e. what is so special about the "partner points" that makes this idea powerful ?

To get some insight, we study contrastive learning in a simple context:

Linear representation learning for Gaussian Mixtures Models

Part 1: Gaussian Mixture Models (for InfoNCE-style losses)

Part 2: "Multi-modal" Gaussian Mixture Models (for CLIP style losses)

Part 1: Gaussian Mixture Models
(Single modality)

(and InfoNCE loss)

Background: Gaussian Mixture Models (GMMs)

- Gaussian Mixture Model :

$$F = \sum_{k \in [K]} w_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Spherical GMMs

Identity Covariance

$$F = \sum_{k \in [K]} w_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{I})$$

Shared Covariance GMMs

$$F = \sum_{k \in [K]} w_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

Linear Representations for GMMs

Task: Find a projection $A \in \mathbb{R}^{d \times r}$

so that the **projected components are better separated** than they were originally

… so that subsequent tasks (e.g. k-means clustering, nearest neighbors, classification etc.) become easier …

A rudimentary form of "representation learning"

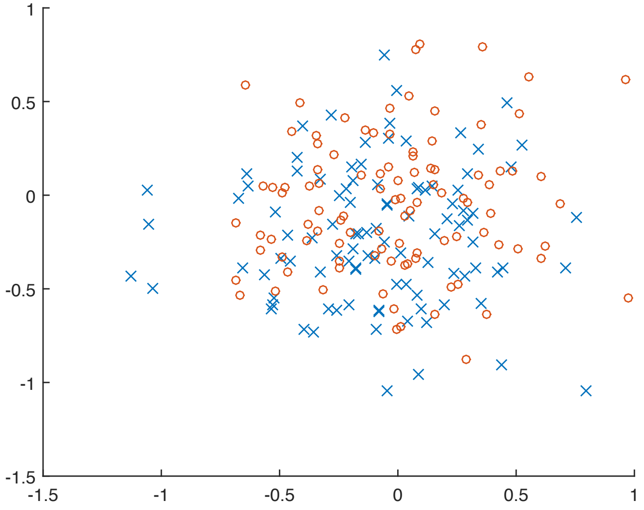
Classical approach: find the top-r SVD-subspace of data matrix

$$\mathbb{E}_{\boldsymbol{x} \sim F}[\boldsymbol{x}\boldsymbol{x}^T]$$

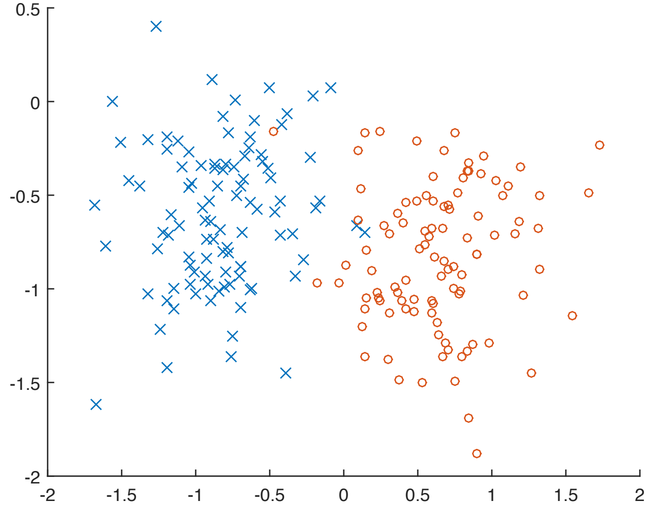
"Spectral methods"

Spectral Clustering

Example: two isotropic clusters in $\mathbb{R}^{50}$, projected down to $\mathbb{R}^2$



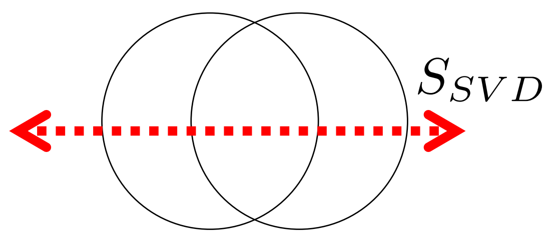
Random 2-d subspace



First two singular vectors

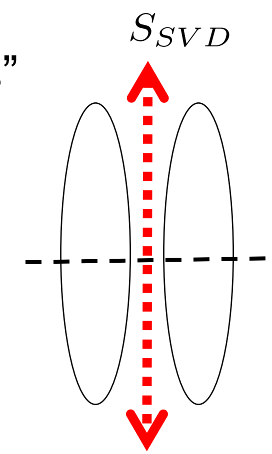
Linear Representations for GMMs

Spherical



Works really well

"Parallel Pancakes"



Works really poorly

Reason: SVD-based approach is equivalent to:

$$\max_{A} \ \mathrm{Tr}\left[A^T \left(\sum_k w_k(\Sigma_k + \mu_k \mu_k^T)\right) A\right]$$

Background: Fisher Discriminant

- Intuitive characterization of a “good” projection subspace
  - Low intra-component variance
  - High inter-component variance

- Fisher Discriminant formalizes this notion for projection matrix A

$$J(\boldsymbol{A}) = \mathrm{Tr}\left(\left[\boldsymbol{A}^T(\sum_k w_k \boldsymbol{\Sigma}_k)\boldsymbol{A}\right]^{-1}\left[\boldsymbol{A}^T(\sum_k w_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T)\boldsymbol{A}\right]\right)$$

Inter-cluster covariance

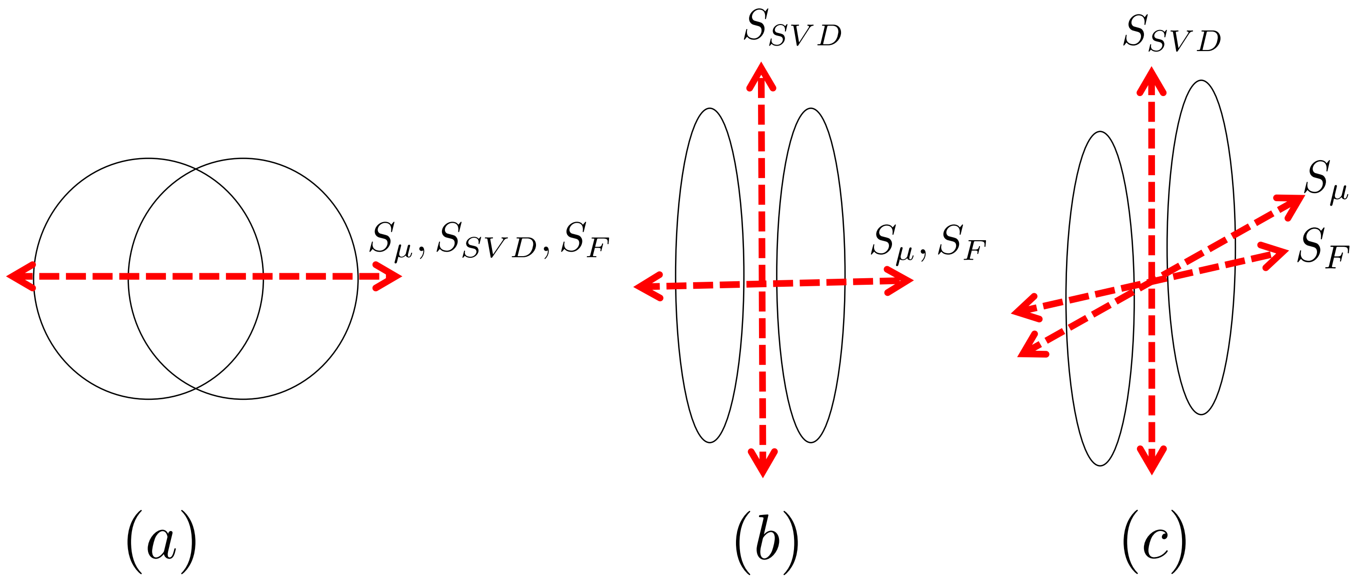
Across cluster covariance

Background: Fisher Subspace

- Minimal Optimal Subspace w.r.t. Fisher Discriminant

$$S_F = \mathrm{Span}\{\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_1, \bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_2, \ldots, \bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_K\}$$

$$\text{where } \bar{\boldsymbol{\Sigma}} = \sum w_k \boldsymbol{\Sigma}_k$$



Our Work

We study the use of contrastive learning for finding projections for GMMs

- Needs a new notion of "augmentations" in GMMs

Shows optimality of contrastive learning

- Using contrastive loss, we can learn the fisher subspace for class of shared covariance matrices, i.e., example (b),(c) in previous slide

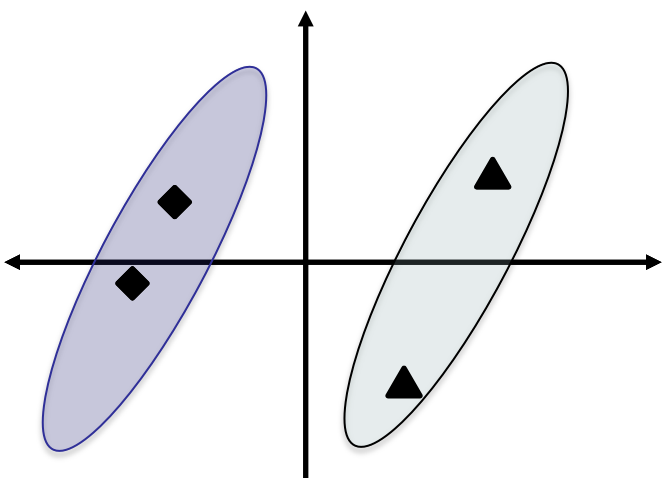
Augmentations in GMMs

Define a new distribution for **pairs of points**

Both points **from same component** with prob δ

$$\hat{F} = \delta \sum_{k \in [K]} w_k \left(\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \times \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$
$$+ (1-\delta) \left(\sum_{k \in [K]} w_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \times \sum_{k' \in [K]} w_{k'} \mathcal{N}(\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'}) \right)$$

Both points **unrelated** with prob $1 - \delta$



▲ ◆ denote augmentation pairs

Result for Single Modality GMMs

Minimizing the InfoNCE loss leads to embeddings of points lying EXACTLY in the fisher subspace for shared covariance gaussians

Theorem 0.1 *Suppose F parameterized by $\{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\}_{k\in[K]}$ be a shared co-variance gaussian mixture model and $\hat{F}$ be its augmentation-distribution with bias δ. Let S_F be the fisher subspace of F and $\boldsymbol{A}^*$ be the optimal solution of the InfoNCE loss :*

$$\boldsymbol{A}^* = \underset{\boldsymbol{A}\in\mathbb{R}^{d\times r}}{\operatorname{argmin}}\ \mathcal{L}(\boldsymbol{A})$$

Then given $r \geq K$, $\mathrm{Col}(\boldsymbol{A}^*) \subseteq S_F$. Moreover if $\delta = 1$, then $\mathrm{Col}(\boldsymbol{A}^*) = S_F$.

(we do not actually have a counter-example of it NOT working if $\delta < 1$)

Part 2: Multimodal Gaussian Mixture Models

(and CLIP loss)

Paired Gaussian Mixture Model

$x_V \in \mathbb{R}^{d_V}$ $x_T \in \mathbb{R}^{d_T}$

pepper the aussie pup

$$\hat{F} = \sum_k w_k \left[\mathcal{N}_{d_V}(\mu_{V,k}\Sigma_{V,k}) \times \mathcal{N}_{d_T}(\mu_{T,k}\Sigma_{T,k}) \right]$$

Each “component” == one Gaussian in each modality

Draw pairs from corresponding components

(here no augmentations needed)

Result for Multi Modality GMMs

Theorem 5.2. *Suppose* $\{w_k, \boldsymbol{\mu}_{V,k}, \boldsymbol{\mu}_{T,k}, \boldsymbol{\Sigma}_V, \boldsymbol{\Sigma}_T\}_{k\in[K]}$ *be a CLIP gaussian mixture model (Def 5.1). Let the fisher subspace of* F_V *be* $S_{V,F}$ *and* F_T *be* $S_{T,F}$ *(Eqn 1). Let* $\boldsymbol{A}_V^*, \boldsymbol{A}_T^*$ *be the optimal solution of the CLIP InfoNCE loss (Eqn 4) :*

$$\boldsymbol{A}_V^*, \boldsymbol{A}_T^* = \underset{\substack{\boldsymbol{A}_V\in\mathbb{R}^{d_1\times r}\\ \boldsymbol{A}_T\in\mathbb{R}^{d_2\times r}}}{\operatorname{argmin}} \mathcal{L}_{clip}(\boldsymbol{A}_V, \boldsymbol{A}_T)$$

Then given $r \geq K$, $\mathrm{Col}(\boldsymbol{A}_V^*) \subseteq S_{V,F}$ *and* $\mathrm{Col}(\boldsymbol{A}_T^*) \subseteq S_{T,F}$

Cross-modality correspondence all you need to find a good within-modality representations

Summary + Discussion

Developed a simple setting in which to understand why contrastive losses work:

Linear representation learning for Gaussian mixtures, in single and multiple modalities

… in which we relied on a new notion of what it means, statistically, to be a "partner point"

Possible next steps:

- Use this understanding to develop a **better contrastive loss**

- Extend this analysis to 1-hidden-layer non-linear networks and **non-shared** covariance GMMs