Your contrastive learning problem is secretly a distribution alignment problem

Zihao Chen*, Chi-Heng Lin, Ran Liu, Jingyun Xiao, Eva L. Dyer
School of Electrical & Computer Engineering
Georgia Tech, Atlanta, GA

NeurIPS 2024

Outline

- Introduction: Contrastive Learning
  - Background
  - Motivation
- Generalized Contrastive Alignment (GCA)
  - Framework and Algorithm
  - Connections to Different CL Objectives
- Experiments
- Conclusion

Background

- Contrastive learning (CL) is a representation learning method that uses positive and negative pairs to define similarity in the latent space.
- Let $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ denote the dataset. For each sample $\boldsymbol{x}_i$ in a batch of training data with size B, we create two augmented copies $\boldsymbol{x}_i'$ and $\boldsymbol{x}_i''$ independently. The $(\boldsymbol{x}_i', \boldsymbol{x}_i'')$ is a positive pair, while $(\boldsymbol{x}_i', \boldsymbol{x}_j'')$ for $j \neq i$ is a negative pair.
- **InfoNCE (INCE)**: The loss function can be defined as :

$$\mathcal{L}_{\text{INCE}} = -\log\left(\frac{e^{s_{ii}}}{e^{s_{ii}} + \sum_{i\neq j} e^{s_{ij}}}\right) \tag{1}$$

where $s_{ij} = \varepsilon^{-1} f_\theta(\boldsymbol{x}_i')^\top f_\theta(\boldsymbol{x}_j'') / \|f_\theta(\boldsymbol{x}_i')\| \|f_\theta(\boldsymbol{x}_j'')\|$

Background

- **Robust InfoNCE (RINCE)** is a robust variant for InfoNCE with adjustable parameters λ and q to handle noisy:

$$\mathcal{L}_{\text{RINCE}}^{\lambda,q} = \frac{1}{q}\left(-e^{qs_{ii}} + \lambda^q\left(e^{s_{ii}} + \sum_{i\neq j} e^{s_{ij}}\right)^q\right) \tag{2}$$

- RINCE is a robust contrastive loss function characterized by its symmetric properties to noisy. [1]
  - When $q \to 1$, RINCE becomes a contrastive loss that fully satisfies the symmetry property (Ghosh et al. show that symmetric loss functions that is noise-tolerant [2])
  - When $q \to 0$, RINCE becomes asymptotically equivalent to InfoNCE.

[1] Robust contrastive learning against noisy views (CVPR 2022)
[2] Making risk minimization tolerant to label noise (Neurocomputing 2015)

Background

- **Limitations of Traditional CL**
  - CL's theoretical foundations and mechanisms for building representations are not fully understood.
  - Traditional CL methods face challenges in handling real-world data, such as noise in views and the need for domain generalization.

- **Proposed Methods**
  - Introduce the Generalized Contrastive Alignment (GCA) framework, which reinterprets CL as a distributional alignment problem.
  - INCE and RINCE can be considered special cases under the GCA framework.
  - Provide theoretical and experimental evidence of the benefits of GCA for noisy and generalized contrastive alignment.

GCA Framework

- **Defining the Kernel Space**:
  - The augmentation kernel $K_\theta(\boldsymbol{x}_i', \boldsymbol{x}_j'') = \exp\left(-\text{dist}(\tilde{f}_\theta(\boldsymbol{x}_i'), \tilde{f}_\theta(\boldsymbol{x}_j''))/\varepsilon\right)$ is defined, where $\text{dist}(\cdot)$ can be an arbitrary distance metric, and $\tilde{f}_\theta(\boldsymbol{x}_i')$ is the normalized output of f_θ.
- **Main Objective Formalization**:
  - The main objective can be defined as:

$$\min_\theta d_M(\boldsymbol{P}_{\text{tgt}} \| \boldsymbol{P}_\theta), \text{with } \boldsymbol{P}_\theta = \arg\min_{\boldsymbol{P} \in \mathcal{B}} \{h(\boldsymbol{P}) + d_\Gamma(\boldsymbol{P} \| \boldsymbol{K}_\theta)\} \tag{3}$$

  - where $h(x)$ is a convex function (typically an indicator function),
  - $\mathcal{B}$ is a closed convex constraint set (e.g., Birkhoff polytope) 双随机矩阵多面体,
  - d_Γ is a Bregman divergence,
  - and d_M is a convex function (e.g., KL - divergence) that measures the divergence between $\boldsymbol{P}_\theta$ and $\boldsymbol{P}_{\text{tgt}}$.

GCA Framework-Advantages

- **Reframing CL as a Distribution Alignment Problem**:
  - Instead of just bringing positive pairs closer, GCA focuses on controlling the alignment of samples by defining a target transport plan $\boldsymbol{P}_{\text{tg}t}$.
  - For example, setting $\boldsymbol{P}_{\text{tg}t}$ to resemble a diagonal matrix encourages each positive to align with itself or its augmentations, minimizing the deviation from an identity matrix (e.g., using KL - divergence).
- **Objective of GCA**:
  - Learn an encoder f_θ that minimizes the transport cost between positive samples. By defining $\boldsymbol{P}_{\text{tg}t}$ with specific alignment rules, we can influence how samples are organized in the latent space.
  - This flexibility allows for encoding more nuanced forms of similarity and adapting to various learning tasks.

GCA Framework - Proximal Point Algorithm

Algorithm 1 Proximal-Point Algorithm for Generalized Contrastive Alignment (GCA)

1: **Initialization:** Initial encoder parameters θ, target transport plan $\mathbf{P}_{\text{tgt}}$, kernel function $\mathbf{K}_\theta$, the function $h(x)$, divergences d_Γ and d_M (KL or W_1). Initialize transport plan $\mathbf{P}_\theta$ based on θ.

2: **Compute the transport coupling $\mathbf{P}_\theta$:** Update $\mathbf{P}_\theta$ using the proximal operator scaling for fixed θ as described in Eq. (8):

$$\mathbf{P}_\theta = \arg\min_{\mathbf{P}\in\mathcal{B}} \{h(\mathbf{P}) + d_\Gamma(\mathbf{P}||\mathbf{K}_\theta)\}.$$

3: **Calculate the loss:** Calculate deviation between the target and current transport plans

$$\mathcal{L}_{GCA} = d_M(\mathbf{P}_\theta, \mathbf{P}_{\text{tgt}}).$$

Update networks f_θ (encoder) and g_θ (projector) to minimize $\mathcal{L}_{GCA}$.

4: **Repeat until convergence:** Repeat steps 2 and 3 until convergence.

- Solve $d_\Gamma(\boldsymbol{P} \,\|\, \boldsymbol{K}_\theta)$ by Sinkhorn Algorithm based on iterative process;

Sinkhorn Algorithm

- **Concept of Optimal Transport**:
  - Entropy-regularized OT (EOT) objective can be defined as :

$$\min_{\boldsymbol{P}\in\mathcal{B}} \langle \boldsymbol{P}, \boldsymbol{C}\rangle - \varepsilon H(\boldsymbol{P}) \tag{4}$$

    - where $H(\boldsymbol{P}) = -\sum_{i,j} \boldsymbol{P}_{ij} \log(\boldsymbol{P}_{ij})$,
    - Cost matrix: $C(x, y) = 1 - \langle x, y\rangle / \|x\|\|y\|$.
    - Two contrains: $C_1^{(\mu)} := \{\boldsymbol{P} : \boldsymbol{P}\mathbf{1}_B = \mu\}$ and $C_2^{(\nu)} := \{\boldsymbol{P} : \boldsymbol{P}^\top \mathbf{1}_B = \nu\}$.
- **Sinkhorn Algorithm and Bregman Projection**:
  - The first step finds the minimizer $\boldsymbol{P}^{(1)} = \arg\min\{\epsilon KL(\boldsymbol{P}||\boldsymbol{K}) : \boldsymbol{P}\mathbf{1}_B = \mu\}$ by the proximal operator $Prox^{KL}_{C_1^{(\mu)}}(\boldsymbol{K})$, compute its derivatives with respect to $\boldsymbol{P}$, set them to 0, and get $\boldsymbol{P}$, where $\boldsymbol{K}_{i,j} = \exp(-\boldsymbol{C}_{i,j}/\varepsilon)$.
  - Subsequent steps project $\boldsymbol{P}^{(1)}$ onto the column constraint set.
    $\boldsymbol{P}^{(2)} := \text{Prox}^{KL}_{C_2^{(\nu)}}(\boldsymbol{P}^{(1)})$

Sinkhorn Algorithm

- The iterative updates can be succinctly expressed as the Sinkhorn iterations:

$$\boldsymbol{P}^{(2t+1)} = \mathrm{diag}(\boldsymbol{u}^{(t+1)})\boldsymbol{K}\mathrm{diag}(\boldsymbol{v}^{(t)}), \tag{5}$$

$$\boldsymbol{P}^{(2t+2)} = \mathrm{diag}(\boldsymbol{u}^{(t+1)})\boldsymbol{K}\mathrm{diag}(\boldsymbol{v}^{(t+1)}), \tag{6}$$

with the scaling vectors $\boldsymbol{u}^{(t)}$ and $\boldsymbol{v}^{(t)}$ updated according to:

$$\boldsymbol{u}^{(t+1)} \stackrel{\text{def}}{=} \frac{\mu}{\boldsymbol{K}\boldsymbol{v}^{(t)}}, \tag{7}$$

$$\boldsymbol{v}^{(t+1)} \stackrel{\text{def}}{=} \frac{\nu}{\boldsymbol{K}^T\boldsymbol{u}^{(t)}}. \tag{8}$$

GCA Framework - GCA - UOT Method

- **Introduction to GCA - UOT**:
  - GCA - UOT is an extension of the GCA framework that utilizes unbalanced optimal transport (UOT) to relax the marginal constraints.
- **Objective Function and Relaxation**:
  - The objective function is

$$\min_{\theta} d_M(\boldsymbol{P}_{\text{tgt}} \| \boldsymbol{P}_\theta) + \lambda_1 h_F(\boldsymbol{P}_\theta \mathbf{1} \| \mu) + \lambda_2 h_G(\boldsymbol{P}_\theta^\top \mathbf{1} \| \nu) + \varepsilon H(\boldsymbol{P}_\theta),$$

    where h_F and h_G are divergence measures (e.g., KL divergence) that penalize deviations from the desired marginals μ and ν, and λ_1 and λ_2 are regularization parameters.
- **Benefits and Impact**:
  - This relaxation allows the method to better handle data with imbalances or noise. It leads to different types of proximal operators and affects the coupling matrix. The impact of the entropy regularization parameter ε on the coupling matrix is studied, along with the number of iterations and corresponding sensitivity.

GCA Framework - Modifying the Target Transport Plan to Encode Matching Constraints

- **Beyond Traditional Constraints**
  - In standard contrastive learning, the objective is often to minimize the deviations between the transport plan $\boldsymbol{P}_\theta$ and the identity matrix ($\boldsymbol{P}_{\mathrm{tg}t} = I$). However, the GCA framework offers the flexibility to go beyond this simple one - to - one constraint.
  - By modifying $\boldsymbol{P}_{\mathrm{tg}t}$, we can incorporate additional meaningful constraints informed by domain knowledge or specific problem requirements.

GCA Framework - Modifying P_{tgt} to Encode Matching Constraints

- **Domain Generalization Scenario**
  - Consider a domain generalization setting where each batch of data contains samples from multiple domains (e.g., Photo, Cartoon, Sketch, Art).
  - We can structure the target alignment plan $\boldsymbol{P}_{\text{tgt}}$ as follows:

$$\boldsymbol{P}_{\text{tgt}}[i,j] = \boldsymbol{I}[i=j] + \alpha \cdot I(D_i = D_j, i \neq j) + \beta \cdot I(D_i \neq D_j, i \neq j)$$

    Here, $I(\cdot)$ is the indicator function (equals 1 if the condition is true, 0 otherwise), D_i represents the domain of sample i, and $\alpha \geq 0$ and $\beta \geq 0$ are parameters.
  - When $\alpha = 0$ and $\beta > 0$, we prioritize cross-domain alignment. This encourages the model to learn similarities and relationships between samples from different domains.
  - Conversely, when $\alpha > 0$ and $\beta = 0$, the focus is on intra-domain alignment, emphasizing the similarities within each domain.

Connections to Different CL Objectives - Connection to INCE

- **INCE as a Single Step in GCA**
  - An interesting connection exists between the GCA framework and the widely used InfoNCE (INCE) objective. We can interpret INCE as a single step within an iterative GCA objective.
  - Under specific conditions (such as using the augmentation kernel $\boldsymbol{K}_\theta$ with cosine similarity, setting d_Γ and d_M to KL - divergence, and using the constraint set $\mathcal{C}_1^{(\mu)}$), the INCE objective can be re - expressed as a GCA problem.
  - Specifically, the INCE objective in Equation (1) can be written as $\min_\theta KL\left(\boldsymbol{I} \| Prox^{\boldsymbol{K}_\theta}_{\mathcal{C}_1^{(\mu)}}(\boldsymbol{K}_\theta)\right)$. This shows that INCE can be seen as solving the matching problem with row normalization constraints $\mathcal{C}_1^{(\mu)}$.

Connections to Different CL Objectives

Table 1: *Comparison of different contrastive alignment objectives.* Here we have C_1^μ and C_2^ν as constraint sets (denoted as $\mathcal{B}$) defined in Equation (4) with their corresponding indicator function. "Iter" refers to iterative methods.

Methods	d_M	d_Γ	$\mathcal{B}$	Iter
INCE	KL	KL	C_1^μ	
GCA-INCE	KL	KL	$C_1^\mu \cap C_2^\nu$	✓
RINCE (q=1)	W1	KL	C_1^μ	
GCA-RINCE (q=1)	W1	KL	$C_1^\mu \cap C_2^\nu$	✓
BYOL	KL	L2	$R^{B\times B}$	

- Modification of the different parts of the main objective $(d_\Gamma, d_M, \mathcal{B})$ can be connected to different contrastive losses.
- GCA-INCE and GCA-RINCE are iterative processes.
- INCE is a single step in GCA-INCE. RINCE is a single step in GCA-RINCE.

Experiments

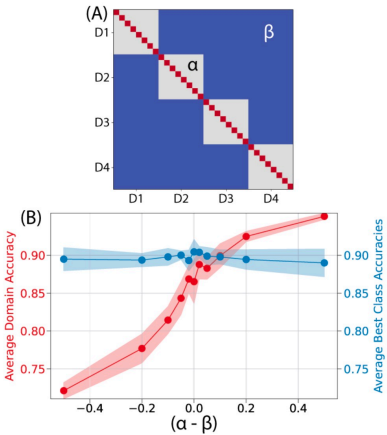
- Methods:
  - InfoNCE(INCE), Robust InfoNCE (IRINCE),
  - the GCA - based alternatives (IGCA-INCE, IGCA-RINCE)
  - ISimCLR, IBYOL, and IOT
  - GCA-UOT
- Datasets:
  - SVHN (Resnet-50)
  - ImageNet-100, CIFAR-10, CIFAR-100(Resnet-18)
  - The extreme angmentations datastes:CIFAR-10(Ex),CIFAR-100(Ex), CIFAR-10C(contains various corruptions), CIFAR-10C(Ex) (Resnet-18)

Experiments

	Standard Setting				Noisy Setting			
Method	CIFAR-10	CIFAR-100	SVHN	ImageNet100	CIFAR-10 (Ex)	CIFAR-100 (Ex)	CIFAR-10C	CIFAR-10C (Ex)
INCE	92.01 ± 0.40	70.07 ± 0.42	90.60 ± 0.17	73.01 ± 0.61	82.03 ± 0.32	54.70 ± 0.43	87.20 ± 0.37	74.84 ± 0.21
GCA-INCE	92.36 ± 0.24	70.11 ± 0.45	90.40 ± 0.16	73.04 ± 0.76	82.18 ± 0.69	54.91 ± 0.56	87.34 ± 0.34	76.00 ± 0.17
Δ	+0.35	+0.04	-0.20	+0.03	+0.15	+0.21	+0.14	+1.16
RINCE	91.05 ± 0.50	69.06 ± 0.64	90.97 ± 0.19	71.91 ± 0.43	82.60 ± 0.63	55.43 ± 0.48	88.62 ± 1.33	77.05 ± 0.82
GCA-RINCE	92.09 ± 0.22	69.72 ± 0.27	91.45 ± 0.41	73.44 ± 0.55	82.76 ± 0.49	55.90 ± 0.41	88.76 ± 0.72	77.23 ± 0.76
Δ	+1.04	+0.66	+0.48	+1.53	+0.16	+0.47	+0.14	+0.18
SimCLR	92.16 ± 0.16	69.95 ± 0.14	90.24 ± 0.24	72.20 ± 0.78	81.87 ± 0.53	54.54 ± 0.79	86.98 ± 1.59	73.79 ± 0.32
BYOL	90.56 ± 0.59	69.75 ± 0.37	89.50 ± 0.46	69.75 ± 0.83	81.55 ± 0.50	54.18 ± 0.46	87.88 ± 1.02	69.40 ± 1.11
IOT [51]	90.99 ± 0.54	67.19 ± 0.21	90.15 ± 0.21	72.27 ± 0.53	80.59 ± 0.64	52.40 ± 0.48	67.36 ± 1.97	58.75 ± 1.96
IOT-uni [51]	90.89 ± 0.57	67.03 ± 0.40	90.54 ± 0.20	72.88 ± 0.71	80.79 ± 0.24	53.04 ± 0.52	69.58 ± 1.25	59.05 ± 1.86
GCA-UOT	**92.61 ± 0.32**	**71.45 ± 0.37**	**91.96 ± 0.15**	**74.09 ± 0.40**	**83.18 ± 0.44**	**56.30 ± 0.51**	**89.61 ± 0.30**	**77.60 ± 0.54**

- The GCA- versions of INCE and RINCE exhibited performance gains in most settings.
- GCA - UOT achieved the top performance across all four datasets tested.
- The GCA-based strategies effectively enhanced the model's generalization ability and adaptability to aggressive data augmentations, especially in handling complex and noisy data.

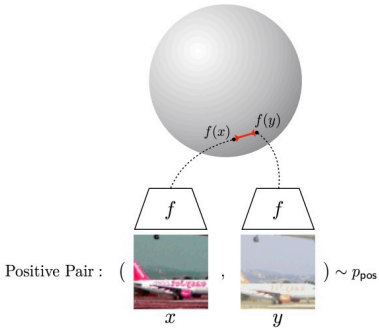
Experiments



- Domain generalization task, samples originate form different domains(e.g. Photo, Cartoon, Sketch, Arts).
- The training was conducted on the PACS dataset using a ResNet18 encoder with the GCA-INCE objective.
- For $\boldsymbol{P}_{tgt}$, we set $\{\alpha = 0, \beta > 0\}$ to prioritize cross-domain alignment and $\{\alpha > 0, \beta = 0\}$ to focus on intra-domain alignment;
- Revealed that increasing the domain alignment weight enhances the accuracy of domain classification (from 72.11% to 95.16%) without diminishing classification performance.

Conclusion

- In this work, we introduced the Generalized Contrastive Alignment (GCA) framework, which reinterprets contrastive learning as a distribution alignment problem using optimal transport.
- We demonstrated the connections between GCA and different contrastive learning objectives, such as INCE, RINCE, and BYOL. Through theoretical analysis and experimental validation, we showed that GCA methods can improve alignment and uniformity in the latent space, leading to more discriminative and resilient representations.
- Theoretical Analysis:
  - Improved alignment with GCA;
  - Improved Uniformity of Representations Through GCA;
  - Impacts of GCA on a downstream classification task;



Alignment: Similar samples have similar features.
(Figure inspired by Tian et al. (2019).)

Uniformity: Preserve maximal information.