Revisiting Deep Generalized Canonical Correlation Analysis

Paris A. Karakasis and Nicholas D. Sidiropoulos
University of Virginia, Charlottesville, USA

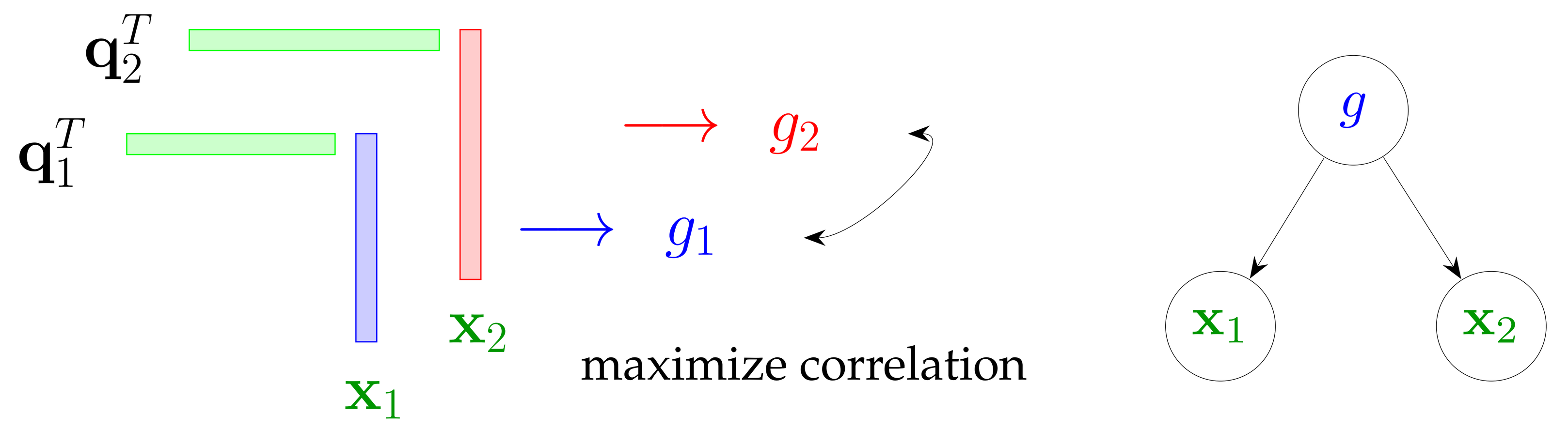
Introduction

- What is Canonical Correlation Analysis (CCA)?

▷ Let $\mathbf{x}_1 \in \mathbb{R}^{M_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{M_2}$ be random vectors (a.k.a. views)

▷ **CCA**: Find vectors $\mathbf{q}_1 \in \mathbb{R}^{M_1}$ and $\mathbf{q}_2 \in \mathbb{R}^{M_2}$ that solve

$$\max_{\mathbf{q}_1,\mathbf{q}_2} \rho\left(\mathbf{q}_1^T\mathbf{x}_1, \mathbf{q}_2^T\mathbf{x}_2\right) := \frac{\text{cov}\left(\mathbf{q}_1^T\mathbf{x}_1, \mathbf{q}_2^T\mathbf{x}_2\right)}{\sqrt{\textbf{Var}\left[\mathbf{q}_1^T\mathbf{x}_1\right]}\sqrt{\textbf{Var}\left[\mathbf{q}_2^T\mathbf{x}_2\right]}}$$



▷ **Goal**: Estimate the common factor g

- Deep Generalized CCA (DGCCA)[1]

  - Generalized → More than 2 views, i.e. $\{\mathbf{x}_k \in \mathbb{R}^{M_k}\}_{k=1}^K$
  - Deep → Nonlinear transformations using Deep Neural Networks

$$\min_{\mathbf{g},\{\boldsymbol{f}_k\in\mathcal{C}\}_{k=1}^K} \sum_{k=1}^K \mathbb{E}\left[\left\|\boldsymbol{f}_k\left(\mathbf{x}^{(k)}\right) - \mathbf{g}\right\|^2\right] \quad \text{s.t.} \quad \mathbb{E}\left[\mathbf{g}\mathbf{g}^T\right] = \mathbf{I}_F,\ \mathbb{E}\left[\mathbf{g}\right] = \mathbf{0}_F$$

- Challenges & Popular Options

⊗ Trivial solutions can occur

▷ **Solution I**: Use $\mathbf{g}$ to reconstruct $\{\mathbf{x}_k \in \mathbb{R}^{M_k}\}_{k=1}^K$

– Deep Canonically Correlated Autoencoders (DCCAE)[2]

– No identifiability conditions

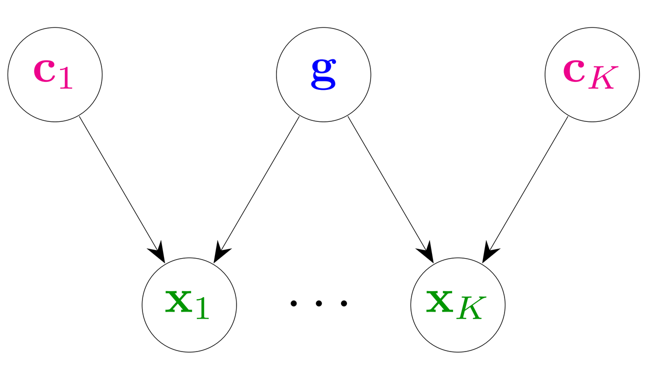
▷ **Solution II**: Introduce also private factors[3]

– Identifiability conditions

– $\{\mathbf{c}_k\}_{k=1}^K$ have to be estimated

– Leakages can take place in practice

Block Independence



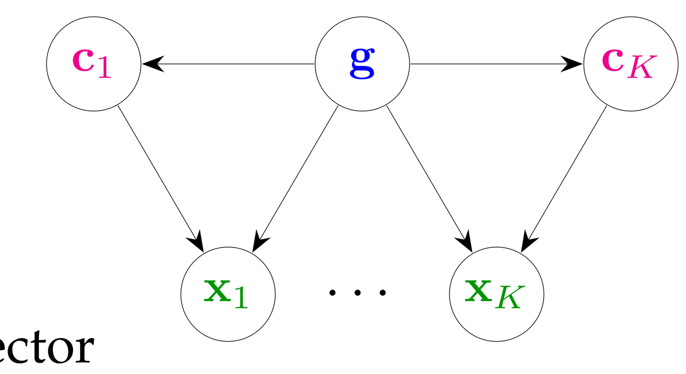
Proposed Approach

- Generative Model

▷ For each view, assume that

$$\mathbf{x}_k = \boldsymbol{v}_k\left(\begin{bmatrix}\mathbf{g}\\ \mathbf{c}_k\end{bmatrix}\right), \text{ for } k = 1, .., K$$

Conditional Independence



- $\mathbf{g} \in \mathbb{R}^F \sim \mathcal{D}_\mathbf{g}$ is the common random vector
- $\mathbf{c}^{(k)} \in \mathbb{R}^{L_k} \sim \mathcal{D}_{\mathbf{c}^{(k)}}$ are the private random vectors

$$p\left(\mathbf{g}, \mathbf{c}^{(1)}, \dots, \mathbf{c}^{(K)}\right) = p\left(\mathbf{g}\right)\prod_{k=1}^K p\left(\mathbf{c}^{(k)}|\mathbf{g}\right)$$

- Problem Formulation

▷ **Ideal formulation**

$$\min_{\substack{\{\boldsymbol{f}_k\in\mathcal{C}_f\}_{k=1}^K,\\ \hat{\mathbf{g}},\\ \{\boldsymbol{w}_k\in\mathcal{C}_w\}_{k=1}^K}} \overbrace{\sum_{k=1}^K \mathbb{E}\left[\left\|\boldsymbol{w}_k\left(\hat{\mathbf{g}}\right) - \mathbf{x}_k\right\|^2\right]}^{\text{reconstruction error}} \quad \text{s.t.} \quad \mathbb{E}\left[\hat{\mathbf{g}}\hat{\mathbf{g}}^T\right] = \mathbf{I}_F,\ \mathbb{E}\left[\hat{\mathbf{g}}\right] = \mathbf{0}_F, \quad \underbrace{\mathbb{E}\left[\left\|\boldsymbol{f}_k\left(\mathbf{x}_k\right) - \hat{\mathbf{g}}\right\|^2\right] = 0}_{\text{consensus}},\ \forall\, k$$

▷ **Practical & Leakage-free formulation**

$$\min_{\substack{\{\boldsymbol{f}_k\in\mathcal{C}_f\}_{k=1}^K,\\ \hat{\mathbf{g}},\\ \{\boldsymbol{w}_k\in\mathcal{C}_w\}_{k=1}^K}} \lambda \frac{\overbrace{\sum_{j\neq k}\mathbb{E}\left[\left\|\boldsymbol{w}_k\left(\boldsymbol{f}_j\left(\mathbf{x}^{(j)}\right)\right) - \mathbf{x}^{(k)}\right\|^2\right]}^{\text{reconstruction error}}}{(K-1)} + (1-\lambda)\overbrace{\sum_{k=1}^K \mathbb{E}\left[\left\|\boldsymbol{f}_k\left(\mathbf{x}_k\right) - \mathbf{g}\right\|^2\right]}^{\text{consensus}}$$
$$\text{s.t.} \quad \mathbb{E}\left[\hat{\mathbf{g}}\hat{\mathbf{g}}^T\right] = \mathbf{I}_F,\ \mathbb{E}\left[\hat{\mathbf{g}}\right] = \mathbf{0}_F$$

- First Identifiability Result on DGCCA

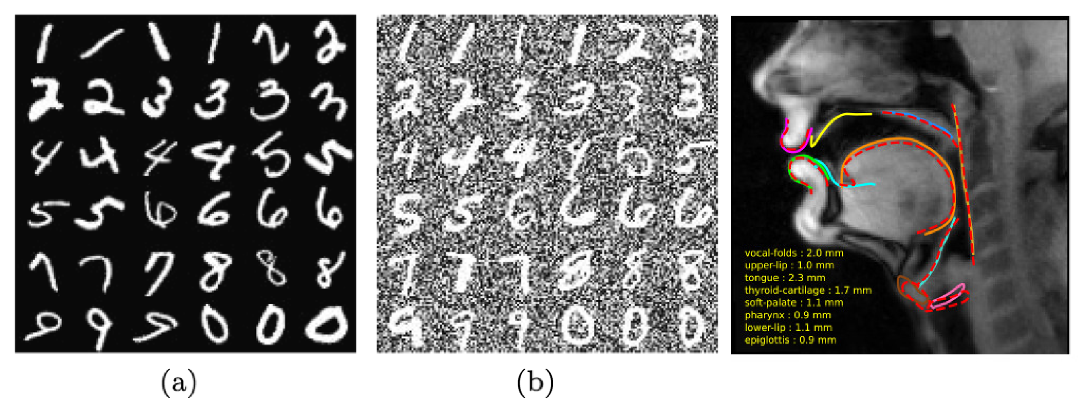
Theorem. *If: (i) functions $\boldsymbol{v}_k$ are partially invertible w.r.t. $\mathbf{g}$,*
(ii) $\exists\, k' \in [K]$ s.t. $\mathbb{E}[\mathbf{x}_{k'} \mid \mathbf{g}] \neq \mathbb{E}[\mathbf{x}_{k'}]$ (mean dependence),
then the learned $\boldsymbol{f}_k(\mathbf{x}_k)$ correspond to $\boldsymbol{\gamma}(\mathbf{g})$, with $\boldsymbol{\gamma}: \mathbb{R}^F \to \mathbb{R}^F$. Moreover, if
(iii) $\mathbb{E}\left[\left[\mathbf{x}_1{}^T, \dots, \mathbf{x}_K{}^T\right]^T \middle| \mathbf{g}\right]$ is an invertible function of $\mathbf{g}$,
then $\mathbf{g}$ is identifiable up to invertible nonlinearities.

- Scalable Algorithm: AO-based with complexity $\mathcal{O}\left(MF^2 + |\mathcal{B}|\sum_{k=1}^K d_k\right)$

Experimental Evaluation

- Dataset

▷ Multiview Noisy MNIST digits
▷ Speech recognition from acoustic-articulatory data (XRMB)



- Baselines Linear CCA, Deep Generalized CCA (DGCCA)[1], Deep Canonically Correlated Autoencoders (DCCAE)[2], Method of Lyu et al[3]

	$\lambda=0.1$	$\lambda=0.3$	$\lambda=0.5$	$\lambda=0.7$	$\lambda=0.9$
CCA	-	-	0.89±0	-	-
DGCCA	-	-	0.97±0	-	-
DCCAE	0.97±0	0.97±0	0.97±0	0.94±0.01	0.49±0.15
Lyu et al. $\beta=0$	0.97±0	0.97 ±0	0.94±0.05	0.94±0.01	0.62±0.11
Lyu et al. $\beta=10^{-3}$	0.21±0.02	0.22±0.02	0.22±0.02	0.21±0.01	0.21±0.02
Proposed	**0.98±0**	**0.98±0**	0.97±0	0.97±0	0.87±0.21

Table 1: Clustering accuracy for the Multiview MNIST for different λs.

	$\lambda=0.1$	$\lambda=0.3$	$\lambda=0.5$	$\lambda=0.7$	$\lambda=0.9$
CCA	-	-	0.35±0	-	-
DGCCA	-	-	0.46±0	-	-
DCCAE	0.47±0.01	0.50±0.01	0.52±0.01	0.54±0.01	0.57±0
Lyu et al. $\beta=10^{-6}$	0.60±0.01	0.60±0.01	0.60±0	0.60±0	0.60±0
Proposed	0.51±0.01	0.54±0.01	0.57±0.01	0.58±0.01	**0.63±0.01**

Table 2: Classification scores on phoneme classification for different λs.

[1] Benton et al. "Deep Generalized Canonical Correlation Analysis." (2017)
[2] Wang et al. "On deep multi-view representation learning." ICML. PMLR, 2015
[3] Lyu et al. "Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective." ICLR, (2021)