Deep Multi-view Representation Learning Based on Adaptive Weighted Similarity

Tetsuya Hada[1,3], **Akifumi Okuno**[2,3], **Hidetoshi Shimodaira**[2,3]
[1]Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, Japan
[2]Graduate School of Informatics, Kyoto University, 36-1 Yoshida Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
[3]RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

Many real-world datasets consist of several types of data such as texts, images, and sounds, and these different kinds of data are referred to as views or domains. One of the best-known approaches for analyzing a multiple-view dataset is canonical correlation analysis (CCA) that linearly transforms data vectors into their low-dimensional representations. Having two-view data matrices $\mathbf{X}^1 := (\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, \ldots, \boldsymbol{x}_n^1)^\top \in \mathbb{R}^{n \times p_1}, \mathbf{X}^2 := (\boldsymbol{x}_1^2, \boldsymbol{x}_2^2, \ldots, \boldsymbol{x}_n^2)^\top \in \mathbb{R}^{n \times p_2}$, CCA considers linear transformations $(\mathbf{A}^1)^\top \boldsymbol{x}_i^1 \in \mathbb{R}^K$, $(\mathbf{A}^2)^\top \boldsymbol{x}_i^2 \in \mathbb{R}^K$ for any fixed $K \leq q := \min\{p_1, p_2\}$. Precisely, CCA finds linear transformation matrices $\mathbf{A}^1 \in \mathbb{R}^{p_1 \times K}, \mathbf{A}^2 \in \mathbb{R}^{p_2 \times K}$, that maximize the total sum of similarities

$$\sum_{i=1}^{n} \langle (\mathbf{A}^1)^\top \boldsymbol{x}_i^1, (\mathbf{A}^2)^\top \boldsymbol{x}_i^2 \rangle, \tag{1}$$

with a quadratic constraint $n^{-1} \sum_{i=1}^{n} [(\mathbf{A}^1)^\top \boldsymbol{x}_i^1 (\boldsymbol{x}_i^1)^\top \mathbf{A}^1 + (\mathbf{A}^2)^\top \boldsymbol{x}_i^2 (\boldsymbol{x}_i^2)^\top \mathbf{A}^2] = \mathbf{I}_K$, that prevent the objective function (1) from diverging. $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{k=1}^{K} a_k b_k$ denotes the inner product. The optimal matrices $\hat{\mathbf{A}}^1, \hat{\mathbf{A}}^2$ are obtained through eigenvalue decomposition of $\hat{\mathbf{S}} := \hat{\boldsymbol{\Sigma}}_{11}^{-1/2} \hat{\boldsymbol{\Sigma}}_{12} \hat{\boldsymbol{\Sigma}}_{22}^{-1/2}$, where $\hat{\boldsymbol{\Sigma}}_{11} := n^{-1} (\mathbf{X}^1)^\top \mathbf{X}^1, \hat{\boldsymbol{\Sigma}}_{22} := n^{-1} (\mathbf{X}^2)^\top \mathbf{X}^2, \hat{\boldsymbol{\Sigma}}_{12} := n^{-1} (\mathbf{X}^1)^\top \mathbf{X}^2$. By substituting the solution to Eq.(1), we obtain the optimal value of the objective function as $\sum_{k=1}^{K} \lambda_k(\hat{\mathbf{S}})$, where $\lambda_k(\cdot)$ denotes the k-th largest eigenvalue.

Although CCA is widely-applicable, CCA sometimes fails to discover a complex structure underlying real-world datasets, because of its linearity. To address the issue, a non-linear extension of CCA, called DCCA (Andrew et al. 2013) has been proposed. DCCA non-linearly translates data matrices with neural networks $f_{\boldsymbol{\theta}}^1 : \mathbb{R}^{p_1} \to \mathbb{R}^{o_1}, f_{\boldsymbol{\theta}}^2 : \mathbb{R}^{p_2} \to \mathbb{R}^{o_2}$, and applies CCA to the translated vectors $\boldsymbol{z}_{\boldsymbol{\theta},i}^1 := f_{\boldsymbol{\theta}}^1(\boldsymbol{x}_i^1), \boldsymbol{z}_{\boldsymbol{\theta},i}^2 := f_{\boldsymbol{\theta}}^2(\boldsymbol{x}_i^2)$. Similar to CCA, we have the objective function of DCCA as

$$\sum_{k=1}^{K} \lambda_k(\hat{\mathbf{S}}_{\boldsymbol{\theta}}), \tag{2}$$

where $\hat{\mathbf{S}}_{\boldsymbol{\theta}}$ is computed with $\{\boldsymbol{z}_{\boldsymbol{\theta},i}^1\}_{i=1}^n$ and $\{\boldsymbol{z}_{\boldsymbol{\theta},i}^2\}_{i=1}^n$. DCCA optimizes Eq.(2) with respect to $\boldsymbol{\theta}$, then we obtain the optimal deep neural networks $f_{\hat{\boldsymbol{\theta}}}^1, f_{\hat{\boldsymbol{\theta}}}^2$. We compute feature vectors by applying CCA to the output of the neural networks.

Meanwhile, DCCA does not consider the importance degree of each feature element, which can be computed as the canonical correlation. By attaching weights to the elements depending on their importance degree, we may improve the result of DCCA. For that reason, we replace the inner product of Eq.(1) with

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle_{\boldsymbol{\nu}} := \sum_{k=1}^{q} \nu_k a_k b_k, \tag{3}$$

where $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_q)$ is a weight vector. The flat weighting

$$\nu_k = \mathbf{1}(k \leq K), \tag{4}$$

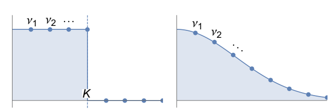
where $\mathbf{1}(\cdot)$ is an indicator function, derives DCCA. We have theoretically showed that the canonical correlations

$$\nu_k = \lambda_k(\hat{\mathbf{S}}_{\boldsymbol{\theta}}), \tag{5}$$

are optimal as the weights, by considering the weighted CCA with a generalized setting. With Eq.(5), the weighted objective function of DCCA becomes

$$\sum_{k=1}^{q} \nu_k \lambda_k(\hat{\mathbf{S}}_{\boldsymbol{\theta}}) = \sum_{k=1}^{q} \lambda_k(\hat{\mathbf{S}}_{\boldsymbol{\theta}})^2, \tag{6}$$

that is the total sum of quadratic eigenvalues. We propose Deep Quadratic CCA (DeQ-CCA) that maximizes the criterion (6). We have verified that our methods outperform existing methods in the experiments on real-world datasets.



(a) DCCA (b) DeQ-CCA

Figure 1: In contrast to DCCA, whose weights are specified as Eq.(4), DeQ-CCA uses smoothly decreasing weights.

References

Andrew, G.; Arora, R.; Bilmes, J. A.; and Livescu, K. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 1247–1255.