Some Notes on Deep CCA

Karl Stratos

1 CCA

Let $\boldsymbol{X} \in \mathbb{R}^{d \times N}$ and $\boldsymbol{Y} \in \mathbb{R}^{d' \times N}$ denote two views of N samples. Let $\mu_{\boldsymbol{X}}$ and $\mu_{\boldsymbol{X}}$ denote the column means of $\boldsymbol{X}$ and $\boldsymbol{Y}$ and assume the following centered, scaled matrices

$$\overline{\boldsymbol{X}} = \frac{1}{\sqrt{N}} (\boldsymbol{X} - \mu_{\boldsymbol{X}}) \qquad\qquad \bar{\boldsymbol{Y}} = \frac{1}{\sqrt{N}} (\boldsymbol{Y} - \mu_{\boldsymbol{Y}})$$

where matrix-vector subtraction is understood as columnwise. Then the cross-covariance and covariance matrices (omitting regularization) are estimated by

$$\boldsymbol{C}_{\boldsymbol{XY}} = \overline{\boldsymbol{X}} \bar{\boldsymbol{Y}}^\top \qquad\qquad \boldsymbol{C}_{\boldsymbol{X}} = \overline{\boldsymbol{X}} \overline{\boldsymbol{X}}^\top \qquad\qquad \boldsymbol{C}_{\boldsymbol{Y}} = \bar{\boldsymbol{Y}} \bar{\boldsymbol{Y}}^\top$$

Given $m \leq \min(d, d')$, CCA calculates m-dimensional representions of the two views $\underline{\boldsymbol{X}} \in \mathbb{R}^{m \times N}$ and $\underline{\boldsymbol{Y}} \in \mathbb{R}^{m \times N}$ by

$$\underline{\boldsymbol{X}} = \boldsymbol{A}^\top \overline{\boldsymbol{X}} \qquad\qquad \underline{\boldsymbol{Y}} = \boldsymbol{B}^\top \bar{\boldsymbol{Y}}$$

where $\boldsymbol{A} = \boldsymbol{C}_{\boldsymbol{X}}^{-1/2} U_m$ and $\boldsymbol{B} = \boldsymbol{C}_{\boldsymbol{Y}}^{-1/2} V_m$. The columns of $U_m \in \mathbb{R}^{d \times m}$ and $V_m \in \mathbb{R}^{d' \times m}$ are the left/right singular vectors of

$$\boldsymbol{\Omega} = \boldsymbol{C}_{\boldsymbol{X}}^{-1/2} \boldsymbol{C}_{\boldsymbol{XY}} \boldsymbol{C}_{\boldsymbol{Y}}^{-1/2}$$

corresponding to the m largest singular values $1 \geq \sigma_1(\boldsymbol{\Omega}) \geq \cdots \geq \sigma_m(\boldsymbol{\Omega})$: each $\sigma_i(\boldsymbol{\Omega})$ is precisely the i-th canonical correlation defined in the original formulation of CCA (i.e., as an iterative optimization problem). Check that these representations satisfy

$$\underline{\boldsymbol{X}}\, \underline{\boldsymbol{X}}^\top = \underline{\boldsymbol{Y}}\, \underline{\boldsymbol{Y}}^\top = I_m \qquad\qquad \underline{\boldsymbol{X}}\, \underline{\boldsymbol{Y}}^\top = \operatorname{diag}\left(\sigma_1(\boldsymbol{\Omega}) \ldots \sigma_1(\boldsymbol{\Omega})\right)$$

Because of the variational characterization of singular vectors (Theorem A.2), we can equivalently state that U_m, V_m are a solution of

$$\max_{\substack{\widehat{U} \in \mathbb{R}^{d \times m} \colon \widehat{U}^\top \widehat{U} = I_m \\ \widehat{V} \in \mathbb{R}^{d' \times m} \colon \widehat{V}^\top \widehat{V} = I_m}} \operatorname{tr}\left(\widehat{U}^\top \boldsymbol{\Omega} \widehat{V}\right) = \sum_{i=1}^{m} \sigma_i(\boldsymbol{\Omega})$$

such that the columns ordered corresponding to nonincreasing singular values. To see why the maximum is achieved at $\sum_{i=1}^{m} \sigma_i(\boldsymbol{\Omega})$, see the proof of the theorem.

2 Deep CCA

In deep CCA [1], we assume that data matrices are already transformed by differentiable parameters to the target dimension m before considering the CCA objective. To be concrete, assume parameters $W_1 \in \mathbb{R}^{m \times d}$ and $W_2 \in \mathbb{R}^{m \times d'}$ yielding $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m \times N}$ by

$$\boldsymbol{X} = \tanh(W_1 \widetilde{\boldsymbol{X}}) \qquad\qquad \boldsymbol{Y} = \tanh(W_2 \widetilde{\boldsymbol{Y}})$$

where in this case we change our notation and write $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{d \times N}$ and $\widetilde{\boldsymbol{Y}} \in \mathbb{R}^{d' \times N}$ to denote the original data matrices. As before, let $\boldsymbol{\Omega} \in \mathbb{R}^{m \times m}$ denote the correlation matrix resulting from these matrices (the operations involved in calculating $\boldsymbol{\Omega}$ are all fully differentiable). Now we consider the CCA objective and note that the optimum

$$\max_{\substack{\widehat{U} \in \mathbb{R}^{d \times m} \colon \widehat{U}^\top \widehat{U} = I_m \\ \widehat{V} \in \mathbb{R}^{d' \times m} \colon \widehat{V}^\top \widehat{V} = I_m}} \operatorname{tr}\left(\widehat{U}^\top \boldsymbol{\Omega} \widehat{V}\right) = \sum_{i=1}^{m} \sigma_i(\boldsymbol{\Omega}) = ||\boldsymbol{\Omega}||_1$$

coincides with the nuclear norm of $\boldsymbol{\Omega}$ because this is the sum of *all* singular values of $\boldsymbol{\Omega}$ without truncation. A main reason to prefer the nuclear norm $||\boldsymbol{\Omega}||_1 := \operatorname{tr}\left((\boldsymbol{\Omega}^\top \boldsymbol{\Omega})^{1/2}\right)$ for differentiability is that it is given by the trace operator which has many nice differentiable properties (e.g., see here). The matrix gradients

$$\frac{\partial \, ||\boldsymbol{\Omega}||_1}{\partial \overline{\boldsymbol{X}}} \in \mathbb{R}^{m \times N} \qquad\qquad \frac{\partial \, ||\boldsymbol{\Omega}||_1}{\partial \overline{\boldsymbol{Y}}} \in \mathbb{R}^{m \times N}$$

are given by an SVD of $\boldsymbol{\Omega}$. So this can be viewed as a final node in the computation graph with $\overline{\boldsymbol{X}}, \overline{\boldsymbol{Y}}$ as parents whose forward and backward passes are calculated by an SVD.

So all we're doing is updating parameters W_1 and W_2 by taking gradient steps on $||\boldsymbol{\Omega}||_1$, which is a function of these parameters. There is no explicit CCA calculation during training! Rather, they are trained to maximize the objective of an implicit trivial CCA (i.e., involving no dimensionality reduction) on top of the representations they induce. Thus to calculate correlations, we need to explicitly perform a trivial CCA after training to obtain actual CCA projection matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times m}$. We would then calculate the sum of canonical correlations on the training data as follows:

1. Apply the trained W_1 and W_2 to obtain m-dimensional data $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m \times N}$.
2. Apply the CCA projection matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times m}$ and return $\operatorname{tr}\left(\boldsymbol{A}^\top \overline{\boldsymbol{X}} \overline{\boldsymbol{Y}}^\top \boldsymbol{B}\right)$.

One final note: the deep CCA objective does not decompose over samples and invalidates the correctness of stochastic gradient descent. This issue is examined more closely in follow-up works [4, 3].

A Matrix Facts

A.1 Trace

Theorem A.1 (Von Neumann, 7.4.11 in [2])**.** Let $A, B \in \mathbb{R}^{m\times n}$ where $m \leq n$. Let $\sigma_1(A) \geq \cdots \geq \sigma_m(A)$ denote the m largest singular values of A in nonincreasing order (likewise for B). Then

$$\operatorname{tr}\left(AB^\top\right) \leq \sum_{i=1}^{m} \sigma_i(A)\sigma_i(B)$$

A.2 SVD

Theorem A.2. Let $A \in \mathbb{R}^{m\times n}$ where $m \leq n$. For any $k \leq m$,

$$(U_k, V_k) \in \underset{\substack{\widehat{U}\in\mathbb{R}^{m\times k}:\, \widehat{U}^\top\widehat{U}=I_k \\ \widehat{V}\in\mathbb{R}^{n\times k}:\, \widehat{V}^\top\widehat{V}=I_k}}{\arg\max} \operatorname{tr}\left(\widehat{U}^\top A\widehat{V}\right)$$

where the columns of $U_k \in \mathbb{R}^{m\times k}$ and $V_k \in \mathbb{R}^{n\times k}$ are the left and right singular vectors of A corresponding to the largest k singular values.

Proof. For any orthonormal $\widehat{U} \in \mathbb{R}^{m\times k}$ and $\widehat{V} \in \mathbb{R}^{n\times k}$,

$$\operatorname{tr}\left(\widehat{U}^\top A\widehat{V}\right) = \operatorname{tr}\left(A\widehat{V}\widehat{U}^\top\right) \leq \sum_{i=1}^{m} \sigma_i(A)\sigma_i(\widehat{V}\widehat{U}^\top) = \sum_{i=1}^{k} \sigma_i(A)$$

To see the last equality, note that $\sigma_i(\widehat{V}\widehat{U}^\top) = \lambda_i(\widehat{U}\widehat{U}^\top)^{1/2}$ where $\widehat{U}\widehat{U}^\top$ is the projection operator onto an k-dimensional subspace of $\mathbb{R}^m$. This upper bound is reached by taking $\widehat{U} = U_k$ and $\widehat{V} = V_k$. □

References

[1] Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255.

[2] Horn, R. A. and Johnson, C. R. (2013). Matrix analysis.

[3] Wang, W., Arora, R., Livescu, K., and Bilmes, J. (2015a). On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092.

[4] Wang, W., Arora, R., Livescu, K., and Srebro, N. (2015b). Stochastic optimization for deep cca via nonlinear orthogonal iterations. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 688–695. IEEE.