# On Deep Multi-View Representation Learning (2015 ICML)

Weiran Wang, Raman Arora, Karen Livescu, Jeff Bilmes

Discussed by: Yizhe Zhang

April 15th, 2016

# Outline

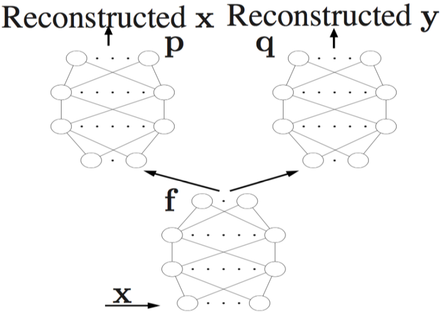
# Multi-view representation learning

- In many applications, we have access to multiple views of data:
  - Audio + video (Kidron et al., 2005; Chaudhuri et al., 2009)
  - Audio + articulation (Arora & Livescu, 2013)
  - Images + text (Hardoonetal., 2004; Socher&Li, 2010; Hodosh et al., 2013)
  - Parallel text in two language (Vinokourov et al., 2003; Haghighi et al., 2008)
- The task is to leverage multiple-view information to learn a better representation (than single view).
- At training time, one attempt to learn the *latent representations* from paired two-view training set.
- At test time, only primary view is available.

# DNN-based multiview feature learning

- Suppose we have access to paired observations from two views, denoted $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_N, \boldsymbol{y}_N)$, where $N$ is the sample size.
- Split autoencoders (Ngiam et al. (2011)) minimize the sum of reconstruction errors for the two views

$$\min_{\mathbf{W}_{\boldsymbol{f}}, \mathbf{W}_{\boldsymbol{p}}, \mathbf{W}_{\boldsymbol{q}}} \frac{1}{N} \sum_{i=1}^{N} (||\boldsymbol{x}_i - \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{x}_i))||^2 + ||\boldsymbol{y}_i - \boldsymbol{q}(\boldsymbol{f}(\boldsymbol{x}_i))||^2) \qquad (1)$$




(a) SplitAE

# Canonical correlation analysis (CCA)

- Given two data matrix $\mathbf{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^T, \mathbf{Y} = [\boldsymbol{y}_1, ..., \boldsymbol{y}_M]^T$, canonical-correlation analysis (CCA) first seeks vectors $\boldsymbol{u}_1 \in \mathbb{R}^N$ and $\boldsymbol{v}_1 \in \mathbb{R}^M$ to maximize the correlation $\rho = \text{corr}(\boldsymbol{u}_1^T\mathbf{X}, \boldsymbol{v}_1^T\mathbf{Y})$.
- For identifiability issue, one may add constraint on $\boldsymbol{u}_1^T\mathbf{X}\mathbf{X}^T\boldsymbol{u}_1$ and $\boldsymbol{v}_1^T\mathbf{Y}\mathbf{Y}^T\boldsymbol{v}_1$
- Then one seeks vectors minimizing the same correlation, subject to the constraint that $(\boldsymbol{u}_2^T\mathbf{X}, \boldsymbol{v}_2^T\mathbf{Y})$ are uncorrelated with the $(\boldsymbol{u}_1^T\mathbf{X}, \boldsymbol{v}_1^T\mathbf{Y})$
- This procedure may be continued up to $\min(N, M)$ times.

# Deep canonical correlation analysis (DCCA)

- Andrew et al. (2013) propose a DNN extension of CCA termed deep CCA.
- In DCCA, the canonical correlation of the extract features for each view is maximized

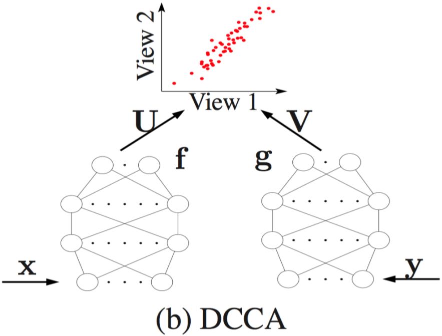



Figure: DCCA

# Deep canonical correlation analysis (DCCA) Cont'd

- Specifically, the optimization function can be written as [1]

$$\min_{\mathbf{W}_{\boldsymbol{f}},\mathbf{W}_{\boldsymbol{g}},\mathbf{U},\mathbf{V}} \frac{1}{N}\text{tr}(\mathbf{U}^T\boldsymbol{f}(\mathbf{X})\boldsymbol{g}(\mathbf{Y})\mathbf{V}^T) \tag{2}$$

$$\text{s.t.}\,\mathbf{U}^T(\frac{1}{N}\boldsymbol{f}(\mathbf{X})\boldsymbol{f}(\mathbf{X})^T + r_x\mathbf{I})\mathbf{U} = \mathbf{I}, \text{(whitening)} \tag{3}$$

$$\mathbf{V}^T(\frac{1}{N}\boldsymbol{g}(\mathbf{Y})\boldsymbol{g}(\mathbf{Y})^T + r_y\mathbf{I})\mathbf{V} = \mathbf{I}, \text{(whitening)} \tag{4}$$

$$\boldsymbol{u}_i^T\boldsymbol{f}(\mathbf{X})\boldsymbol{g}(\mathbf{Y})^T\boldsymbol{v}_j = 0, \text{for}\, i \neq j \text{(orthogonal)} \tag{5}$$

where $\mathbf{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]$, $\mathbf{Y} = [\boldsymbol{y}_1, ..., \boldsymbol{y}_N]$, $\mathbf{U} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_L]$ and $\mathbf{V} = [\boldsymbol{v}_1, ..., \boldsymbol{v}_L]$. $L$ denote the feature dimension. $r_x, r_y$ are regularization parameters (De Bie & De Moor, 2003).

[1] optimization the DCCA requires full data by whitening constraints. The authors claims a sufficiently large minibatch will be enough for estimating the covariances, thus SGD is still valid

# Outline

# Deep canonically correlated autoencoders (DCCAE)

- They propose a model that optimizes the combination of *canonical correlation* and the *reconstruction errors* of the autoencoders.
- **Interpretation:** *"The DCCAE objective offers a trade-off between the information captured in the (input, feature) mapping within each view, and the information in the (feature, feature) relationship"*

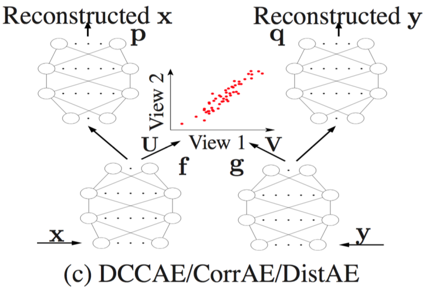



Figure: DCCAE

# Deep canonically correlated autoencoders (DCCAE) Cont'd

- Specifically, the optimization function can be written as

$$\min_{\mathbf{W}_{\boldsymbol{f}},\mathbf{W}_{\boldsymbol{g}},\mathbf{W}_{\boldsymbol{p}},\mathbf{W}_{\boldsymbol{q}},\mathbf{U},\mathbf{V}} -\frac{1}{N}\mathrm{tr}(\mathbf{U}^T\boldsymbol{f}(\mathbf{X})\boldsymbol{g}(\mathbf{Y})^T\mathbf{V}) \tag{6}$$

$$+\frac{\lambda}{N}\sum_{i=1}^{N}(||\boldsymbol{x}_i-\boldsymbol{p}(\boldsymbol{f}(\boldsymbol{x}_i))||^2+||\boldsymbol{y}_i-\boldsymbol{q}(\boldsymbol{g}(\boldsymbol{y}_i))||^2) \tag{7}$$

$$\text{s.t.}\mathbf{U}^T(\frac{1}{N}\boldsymbol{f}(\mathbf{X})\boldsymbol{f}(\mathbf{X})^T+r_x\mathbf{I})\mathbf{U}=\mathbf{I}, \tag{8}$$

$$\mathbf{V}^T(\frac{1}{N}\boldsymbol{g}(\mathbf{Y})\boldsymbol{g}(\mathbf{Y})^T+r_y\mathbf{I})\mathbf{V}=\mathbf{I}, \tag{9}$$

$$\boldsymbol{u}_i^T\boldsymbol{f}(\mathbf{X})\boldsymbol{g}(\mathbf{Y})^T\boldsymbol{v}_j=0, \text{for } i\neq j \tag{10}$$

# Corrlated autoencoders (CorrAE)

- They further relax the uncorrelated feature constraint of DCCAE (i.e. $\boldsymbol{u}_i^T \boldsymbol{f}(\mathbf{X})$, $\boldsymbol{v}_j^T \boldsymbol{g}(\mathbf{Y})$ etc. can be correlated), which they called correlated autoencoders (CorrAE)

$$\min_{\mathbf{W}_{\boldsymbol{f}},\mathbf{W}_{\boldsymbol{g}},\mathbf{W}_{\boldsymbol{p}},\mathbf{W}_{\boldsymbol{q}},\mathbf{U},\mathbf{V}} -\frac{1}{N}\mathrm{tr}(\mathbf{U}^T \boldsymbol{f}(\mathbf{X})\boldsymbol{g}(\mathbf{Y})^T\mathbf{V}) \tag{11}$$

$$+\frac{\lambda}{N}\sum_{i=1}^{N}(||\boldsymbol{x}_i - \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{x}_i))||^2 + ||\boldsymbol{y}_i - \boldsymbol{q}(\boldsymbol{g}(\boldsymbol{y}_i))||^2) \tag{12}$$

$$\text{s.t.}\boldsymbol{u}_i^T \boldsymbol{f}(\mathbf{X})\boldsymbol{f}(\mathbf{X})^T\boldsymbol{u}_i = \boldsymbol{v}_i^T \boldsymbol{g}(\mathbf{Y})\boldsymbol{g}(\mathbf{Y})^T\boldsymbol{v}_i = N,\ 1 \leq i \leq L. \tag{13}$$

- They demonstrate in experiments that this relaxation results in a large performance gap. Therefore, the constraint 10 is necessary.

# Minimum-distance autoencoders (DistAE)

- The whitening constraints complicate the optimization of CCA-based objectives. Thus they also consider another discrepancy objective to substitute CCA objective.

$$\min_{\mathbf{W}_{\boldsymbol{f}},\mathbf{W}_{\boldsymbol{g}},\mathbf{W}_{\boldsymbol{p}},\mathbf{W}_{\boldsymbol{q}}} \frac{||\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{y}_i)||^2}{||\boldsymbol{f}(\boldsymbol{x}_i)||^2 + ||\boldsymbol{g}(\boldsymbol{y}_i)||^2} \tag{14}$$

$$+ \frac{\lambda}{N}\sum_{i=1}^{N}(||\boldsymbol{x}_i - \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{x}_i))||^2 + ||\boldsymbol{y}_i - \boldsymbol{q}(\boldsymbol{g}(\boldsymbol{y}_i))||^2) \tag{15}$$

- This objective is unconstrained and can be factorized by each training sample, so normal SGD applies using small minibatches.

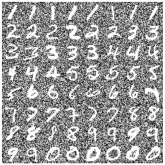
# Outline

# Experiments

- They compare the following methods in the multi-view learning setting on MNIST classification, speech recognition, and word pair semantic similarity.
  - **Baseline**: Original data without any transformation.
  - **DNN-based models**: SplitAE, CorrAE, DCCA, DCCAE, and DistAE.
  - **Linear CCA (CCA)**
  - **Kernel CCA approximations**: Two approximation methods: FKCCA (random Fourier features), NKCCA (Nyström approximation)

# Noisy MNIST digits

- MNIST dataset. Rescale to $[0, 1]$, with 60K/10K images for training/testing.
  - **View 1**: Add independent random uniform noise.
  - **View 2**: Rotate the images at angles uniformly drawn from $[\pi/4, \pi/4]$.
- *"A good multi-view learning algorithm should be able to extract features that disregard the noise"*.
- Criteria:
  - **ACC**: (Clustering accuracy) How well the spectral clustering of projected view 1 matches with Ground truth.
  - **NMI**: (Clustering accuracy) Normalized mutual information.
  - **Error**: Classification error of a linear SVM on the projections.



| Method | ACC (%) | NMI (%) | Error (%) |
|---|---|---|---|
| Baseline | 47.0 | 50.6 | 13.1 |
| CCA ($L = 10$) | 72.9 | 56.0 | 19.6 |
| SplitAE ($L = 10$) | 64.0 | 69.0 | 11.9 |
| CorrAE ($L = 10$) | 65.5 | 67.2 | 12.9 |
| DistAE ($L = 20$) | 53.5 | 60.2 | 16.0 |
| FKCCA($L = 10$) | 94.7 | 87.3 | 5.1 |
| NKCCA($L = 10$) | 95.1 | 88.3 | 4.5 |
| DCCA ($L = 10$) | **97.0** | **92.0** | **2.9** |
| DCCAE ($L = 10$) | **97.5** | **93.4** | **2.2** |

Figure: Performance comparison on the test set of noisy MNIST digits (view 1)

# Noisy MNIST digits

- They measure the class separation visually by t-SNE embedding.
- The DCCA and DCCAE manage to map digits of the same identity to similar locations while suppressing the rotational variation.
- Overall, DCCAE gives the cleanest embedding.
- The learned mixing weight hyperparameter $\lambda$ is very small $(10^{-3})$.

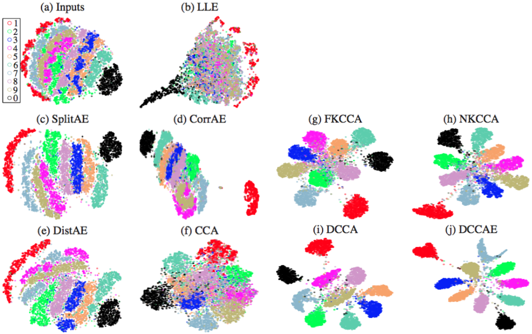



Figure: t-SNE embedding of the projected test set of noisy MNIST (view 1)

# Acoustic-articulatory data for speech recognition

- **Wisconsin X-Ray MicroBeam (XRMB) corpus.** Two views: *speech* and *articulatory measurements*.
- Split the XRMB data to 35/8/2/2 speakers for feature learning/recognizer training/tuning/testing.
- **Criteria:** Phone error rates (PERs) by using an HMM recognizer.

| Method | Mean (std) PER (%) |
|---|---|
| Baseline | 34.8 (4.5) |
| CCA | 26.7 (5.0) |
| SplitAE | 29.0 (4.7) |
| CorrAE | 30.6 (4.8) |
| DistAE | 33.2 (4.7) |
| FKCCA | 26.0 (4.4) |
| NKCCA | 26.6 (4.2) |
| DCCA | **24.8 (4.4)** |
| DCCAE | **24.5 (3.9)** |

Figure: Mean and standard deviations of PERs over 6 folds obtained by each

# Multilingual data for word embeddings

- Learn a vectorial embedding representation of bigram (AN: adjective-noun, VN verb-noun). Two views: English and German.
- Tuning and test splits (of size 649/1,972) for each subset.
- **Criteria:** Spearmans correlation of the bigram representation between two languages.

| Method | AN | VN | Avg. |
|---|---|---|---|
| Baseline | 45.0 | 39.1 | 42.1 |
| CCA | 46.6 | 37.7 | 42.2 |
| SplitAE | 47.0 | **45.0** | 46.0 |
| CorrAE | 43.0 | 42.0 | 42.5 |
| DistAE | 43.6 | 39.4 | 41.5 |
| FKCCA | 46.4 | 42.9 | 44.7 |
| NKCCA | 44.3 | 39.5 | 41.9 |
| DCCA | 48.5 | 42.5 | 45.5 |
| DCCAE | **49.1** | 43.2 | **46.2** |

Figure: Spearmans correlation ($\rho$) for bigram similarities.