An Efficient Approach to Informative Feature Extraction from Multimodal Data

Lichen Wang[1*], Jiaxiang Wu[2], Shao-Lun Huang[1], Lizhong Zheng[3], Xiangxiang Xu[4], Lin Zhang[1], Junzhou Huang[5]

[1] Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, [2] Tencent AI Lab
[3] Department of EECS, Massachusetts Institute of Technology
[4] Department of Electronic Engineering, Tsinghua University
[5] Department of CSE, The University of Texas at Arlington
Email: wlc16@mails.tsinghua.edu.cn, jonathanwu@tencent.com, shaolun.huang@sz.tsinghua.edu.cn, lizhong@mit.edu, xuxx14@mails.tsinghua.edu.cn, linzhang@tsinghua.edu.cn, jzhuang@uta.edu

Abstract

One primary focus in multimodal feature extraction is to find the representations of individual modalities that are maximally correlated. As a well-known measure of dependence, the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation becomes an appealing objective because of its operational meaning and desirable properties. However, the strict whitening constraints formalized in the HGR maximal correlation limit its application. To address this problem, this paper proposes Soft-HGR, a novel framework to extract informative features from multiple data modalities. Specifically, our framework prevents the "hard" whitening constraints, while simultaneously preserving the same feature geometry as in the HGR maximal correlation. The objective of Soft-HGR is straightforward, only involving two inner products, which guarantees the efficiency and stability in optimization. We further generalize the framework to handle more than two modalities and missing modalities. When labels are partially available, we enhance the discriminative power of the feature representations by making a semi-supervised adaptation. Empirical evaluation implies that our approach learns more informative feature mappings and is more efficient to optimize.

Introduction

Human perception is typically more accurate when objects are presented in multiple modalities, as information from one sense often augments information from another. The idea has risen recent interests to develop learning machines which can extract correlation across modalities, through the perception of equivalence, dependence or association. However, compared to the ease of human perception, identifying the relationship among multiple sources is much harder for machines. The reason lies in the facts that the varying statistic properties carried by data from each source obscure the correlation among modalities, which could be vital for learning effective feature representations (Baltrušaitis, Ahuja, and Morency 2018; Sohn, Shang, and Lee 2014). Existing methods approaches this problem by Canonical Correlation Analysis (CCA) (Hotelling 1936; Akaho 2006; Andrew et al. 2013), Euclidean distance minimization (Frome et al. 2013), enforcing partial order (Vendrov et al. 2015), etc.

In statistic, the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation (Hirschfeld 1935; Gebelein 1941; Rényi 1959), as a generalization from the Pearson's correlation (Pearson 1895), is well-known for its legitimacy as a measure of dependence. Such notion is appealing to multimodal feature extraction for many reasons. For example, maximizing the HGR maximal correlation enables us to determine the nonlinear transformations of two variables that are maximally correlated (Feizi et al. 2017). In the perspective of the information theory, the HGR transformation carries the maximum amount of information of X about Y, and vice versa (Huang et al. 2017). As for generality, CCA (Hotelling 1936) and its variants (Bach and Jordan 2002; Akaho 2006; Andrew et al. 2013) can be regarded as the realizations of the HGR maximal correlation with different designs of transformation functions.

However, the HGR maximal correlation suffers from two limitations. Firstly, HGR maximal correlation involves whitening constraints which require each feature to be strictly uncorrelated. Most commonly, the orthogonal geometry is preserved by a whitening process (Andrew et al. 2013; Wang et al. 2015b), which relies on the computation of matrix inversion or decomposition. These operations are of high-complexity and may have numerical stability issues for large feature dimensions. Secondly, discriminativeness is not explicitly formulated in the objective of the HGR maximal correlation. In fact, it can lead to desirable performance in downstream supervised tasks only if all the discriminative information "accidentally" lies in the common subspace of different modalities. Such assumption may not hold true when input modalities are weakly correlated and do not possess much common information. In this case, the underlying discriminative information is more likely to be omitted after feature mapping, which leads to performance degradation.

To address these problems, we propose Soft-HGR, a novel framework to learn correlated representation across modalities without hard whitening constraints. The objective of Soft-HGR consists of two inner products, one between the feature mappings and the other between feature covariances.

[*] This work was done when Lichen Wang was an intern at Tencent AI Lab.

While the formulation rules out the whitening constraints, our model is still able to preserve the same feature geometry as in the original HGR formulation. Therefore, no additional decorrelation process is required in optimization, which promises scalability and stability to the algorithm. Besides, the simple formulation of the Soft-HGR provides additional generalizability to the framework. Soft-HGR can be readily extended to manage more than two modalities and missing modalities. In the semi-supervised settings, we adapt the model to extract the information not only about the dependence between different modalities, but has good predictive power to the labels. Empirically, our method reveals superior efficiency, stability and discriminative performance on real data.

In summary, our main contributions are as follows:

- We proposed Soft-HGR, based on the HGR maximal correlation, to extract informative features from multimodal data. The objective is simple and easy to implement;
- We proposed an alternative strategy to learn the HGR transformations without explicit whitening constraints. The optimization is more efficient and reliable;
- We generalize our framework to handle more than two modalities and missing modalities, and to incorporate discriminative information for semi-supervised tasks.

Background: The HGR Maximal Correlation

The HGR maximal correlation (Hirschfeld 1935; Gebelein 1941; Rényi 1959) generalizes the well-known Pearson's correlation (Pearson 1895) as a general measure of dependence. While it was originally defined on one feature, the multi-feature extension is straightforward. For joint distributed random variables X and Y with ranges $\mathcal{X}$ and $\mathcal{Y}$, the HGR maximal correlation with k features is defined by:

$$\rho^{(k)}(X,Y) = \sup_{\substack{\mathbf{f}:\mathcal{X}\to\mathbb{R}^k,\frac{1}{k}\mathbb{E}[\mathbf{f}]=0,\mathrm{Cov}(\mathbf{f})=\mathbf{I}\\ \mathbf{g}:\mathcal{Y}\to\mathbb{R}^k,\mathbb{E}[\mathbf{g}]=0,\mathrm{Cov}(\mathbf{g})=\mathbf{I}}} \mathbb{E}\left[\mathbf{f}^{\mathrm{T}}(X)\mathbf{g}(Y)\right] \tag{1}$$

where $\mathbf{f} = [f_1, f_2, ..., f_k]^{\mathrm{T}}$, $\mathbf{g} = [g_1, g_2, ..., g_k]^{\mathrm{T}}$, and the supremum is taken over all sets of Borel measurable functions with zero-mean and identity covariance. As a legitimate measure of dependence, the HGR maximal correlation satisfies many fundamental properties which are rarely provided. For example, the correlation coefficient is bounded by 0 and 1, corresponding to the case when two random variables are independent, or there exists a deterministic relationship between X and Y (Rényi 1959).

There are many reasons why HGR maximal correlation is appealing to multimodal feature extraction. For example, finding the HGR maximal correlation also leads us to the non-linear transformation $\mathbf{f}$ and $\mathbf{g}$. These transformations are the most "informative" ones, in the view of information theory, as $\mathbf{f}(X)$ carries the maximum amount of the information towards Y and vice versa (Huang et al. 2017).

Connections to CCA Based Models

One strand of research on correlation extraction is based on the work of Hotelling on CCA (Hotelling 1936), which is later extended to Kernal CCA (Bach and Jordan 2002; Akaho 2006) and Deep CCA (Andrew et al. 2013). In fact, CCA based models share a very similar objective to the HGR maximal correlation, except their transformation functions are restricted to certain forms. More specifically, CCA and Kernel CCA find optimal feature mappings in linear and reproducing kernel Hilbert space, respectively. Deep CCA takes a different approach, in which the $\mathbf{f}$ and $\mathbf{g}$ are implemented as deep neural networks. Assuming the infinite expressive power of the neural structure, the $\mathbf{f}$ and $\mathbf{g}$ have the capability to approximate the HGR transformations.

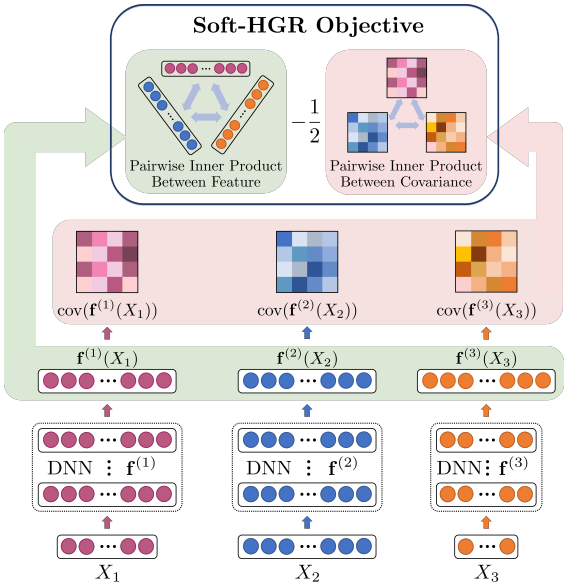

Figure 1: Architecture of Soft-HGR

Limitations

An impediment to HGR maximal correlation is that the whitening constraints bring high computational complexity to the optimization. Existing models introduce a decorrelation step which forces the covariance to be an identity matrix. The decorrelation process is not scalable since it relies on the computation of the matrices inversion and decomposition, whose time complexity is $O(k^3)$. Besides, the optimization in practice often encounters gradients explosion as we choose large k, because the covariance matrices become ill-posed. Some works are proposed to address the problem. Soft-CCA (Chang, Xiang, and Hospedales 2018) introduces a decorrelation regularizer based on the l_1 penalty to replace the hard whitening constraints. Correlational Neural Network (Chandar et al. 2016), inspired by autoencoder, introduces an addition reconstruction loss to replace the whitening constraints. However, both methods break the original feature geometry of the HGR maximal correlation.

Besides, the features extracted from the HGR maximal correlation are not necessarily suitable for downstream discriminative tasks. As a dimension deduction process, there are inevitably some information about data that is discarded during transformation $\mathbf{f} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^k$. This is acceptable if the primary goal is to model the correlation between modalities.

However, if $\mathbf{f}$ is utilized for future discriminative tasks, we may expect some performance loss.

Soft-HGR

In this section, we detail our framework for Soft-HGR. We commence by deriving the optimal solution for the HGR maximal correlation with the whitening constraints. Then we propose an alternative strategy, the low-rank approximation, to approach the HGR problem. we show our proposed objective escapes whitening constraints but still arrives at an equivalent optimum. Finally, we generalize Soft-HGR to handle more than two data modalities and missing modalities, and to incorporate supervised information.

The Optimal Feature Transformations

To simplify the discussions, we assume that X and Y are discrete random variables with range $\mathcal{X} = \{1, 2, ..., |\mathcal{X}|\}$ and $\mathcal{Y} = \{1, 2, ..., |\mathcal{Y}|\}$, respectively. However, the discussion is still valid when X and Y are multivariate and continuous in nature.

We first introduce matrix $\mathbf{B} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ as a function of joint distribution P_{XY} (Huang et al. 2017). The (x, y)-th entry is defined as:

$$B_{x,y} = \frac{P_{XY}(x, y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}} \tag{2}$$

As a summarization of the data, $\mathbf{B}$ has the following property:

Lemma 1. *The largest singular value of* $\mathbf{B}$ *is* 1*, with the corresponding left and right singular vectors given by:*

$$\begin{aligned}\mathbf{u}_0 &= \left[\sqrt{P_X(1)}, \sqrt{P_X(2)}, \ldots, \sqrt{P_X(|\mathcal{X}|)}\right]^{\mathrm{T}} \\ \mathbf{v}_0 &= \left[\sqrt{P_Y(1)}, \sqrt{P_Y(2)}, \ldots, \sqrt{P_Y(|\mathcal{Y}|)}\right]^{\mathrm{T}}\end{aligned} \tag{3}$$

Proof. For any $\boldsymbol{\psi} = [\sqrt{P_Y(y)}\, g(y),\ y = 1, 2, \ldots, |\mathcal{Y}|]^{\mathrm{T}}$ that satisfies $||\boldsymbol{\psi}||_2 = 1$, we have

$$\begin{aligned}||\mathbf{B}\boldsymbol{\psi}||_2^2 =& \sum_x \left(\sum_y \frac{P_{XY}(x, y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}} \sqrt{P_Y(y)}\, g(y) \right)^2 \\ =& \sum_x P_X(x) \left(\sum_y \frac{P_{XY}(x, y)}{P_X(x)}\, g(y) \right)^2 \\ =& \sum_x P_X(x) \mathbb{E}^2 \left[g(Y) | X = x \right] \\ \leq& \sum_x P_X(x) \mathbb{E} \left[g^2(Y) | X = x \right] \\ =& \mathbb{E} \left[g^2(Y) \right] = ||\boldsymbol{\psi}||_2^2 = 1\end{aligned} \tag{4}$$

Therefore the largest singular value $\sigma_0 = \sup ||\mathbf{B}\boldsymbol{\psi}||_2 \leq 1$. The equality only holds when $g(Y)$ is the constant 1 and $\boldsymbol{\psi} = \mathbf{v}_0$. The derivation is similar for $\mathbf{u}_0$. □

Below, we show that finding the most correlated feature transformations for the maximal HGR correlation is equivalent to solving the SVD for $\tilde{\mathbf{B}} = \mathbf{B} - \mathbf{u}_0\mathbf{v}_0^{\mathrm{T}}$.

Theorem 1. *(Huang et al. 2017) Given the SVD of* $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} = \sum_{i=0}^{K} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}$*, with* $1 = \sigma_0 \geq \sigma_1 \geq ... \geq \sigma_K$*, then optimal feature transformations for the HGR maximal correlation are given by:*

$$\begin{aligned}f_i^*(x) &= U_{x,i}/\sqrt{P_X(x)}, i = 1, ..., k, x \in \mathcal{X} \\ g_i^*(y) &= V_{y,i}/\sqrt{P_Y(y)}, i = 1, ..., k, y \in \mathcal{Y}\end{aligned} \tag{5}$$

Proof.

$$\begin{aligned}&\mathbb{E}\left[\mathbf{f}^{\mathrm{T}}(X)\mathbf{g}(Y)\right] \\ =& \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{XY}(x, y) \mathbf{f}^{\mathrm{T}}(x)\mathbf{g}(y) \\ =& \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \sqrt{P_X(x)}\mathbf{f}^{\mathrm{T}}(x) \frac{P_{X,Y}(x, y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}} \sqrt{P_Y(y)}\mathbf{g}(y) \\ =& \operatorname{tr}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{B}\mathbf{\Psi})\end{aligned} \tag{6}$$

In (6) we introduce new variables $\mathbf{\Phi} \in \mathbb{R}^{|\mathcal{X}|\times k}$ and $\mathbf{\Psi} \in \mathbb{R}^{|\mathcal{Y}|\times k}$, which are connected to $\mathbf{f}$ and $\mathbf{g}$ by:

$$\begin{aligned}\mathbf{\Phi} &= \left[\sqrt{P_X(1)}\mathbf{f}(1), \ldots, \sqrt{P_X(|\mathcal{X}|)}\mathbf{f}(|\mathcal{X}|)\right]^{\mathrm{T}} \\ \mathbf{\Psi} &= \left[\sqrt{P_Y(1)}\mathbf{g}(1), \ldots, \sqrt{P_Y(|\mathcal{Y}|)}\mathbf{g}(|\mathcal{Y}|)\right]^{\mathrm{T}}\end{aligned} \tag{7}$$

Following the variables substitution, the objective of the HGR maximal correlation can be reformulated as follows:

$$\rho_k(X, Y) = \max_{\substack{\mathbf{f}:\mathcal{X}\to\mathbb{R}^k, \mathbb{E}[\mathbf{f}]=\mathbf{0}, \mathrm{Cov}(\mathbf{f})=\mathbf{I} \\ \mathbf{g}:\mathcal{Y}\to\mathbb{R}^k, \mathbb{E}[\mathbf{g}]=\mathbf{0}, \mathrm{Cov}(\mathbf{g})=\mathbf{I}}} \mathbb{E}\left[\mathbf{f}^{\mathrm{T}}(X)\mathbf{g}(Y)\right] \tag{8}$$

$$= \max_{\substack{\mathbf{\Phi}:\mathbf{\Phi}^{\mathrm{T}}\mathbf{u}_0=\mathbf{0}, \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}=\mathbf{I} \\ \mathbf{\Psi}:\mathbf{\Psi}^{\mathrm{T}}\mathbf{v}_0=\mathbf{0}, \mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi}=\mathbf{I}}} \operatorname{tr}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{B}\mathbf{\Psi}) \tag{9}$$

$$= \max_{\substack{\mathbf{\Phi}:\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}=\mathbf{I} \\ \mathbf{\Psi}:\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi}=\mathbf{I}}} \operatorname{tr}(\mathbf{\Phi}^{\mathrm{T}}\tilde{\mathbf{B}}\mathbf{\Psi}) \tag{10}$$

As for the optimization problem in (10), the optimal $\mathbf{\Phi}^*$ and $\mathbf{\Psi}^*$ should align the left and right singular vectors of $\tilde{\mathbf{B}}$ respectively. Substituting $\{\mathbf{\Phi}^*, \mathbf{\Psi}^*\}$ back to $\{\mathbf{f}, \mathbf{g}\}$ leads us to the solution in (5). □

For the maximization problem in (10), the whitening constraints over $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ and $\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi}$ are inevitable as they assure the selected features to be mutually orthogonal in the functional space. In the next subsection, we show an alternative formulation for this problem.

Alternative: The Low-rank Approximation

Instead of solving the SVD, we approach this problem by discovering the low-rank approximation of $\tilde{\mathbf{B}}$, where all the cross-modal interactions lies in. Recall the variable equivalence in (7), we approximate the $\tilde{\mathbf{B}}$ by:

$$\begin{aligned}\min_{\mathbf{f},\mathbf{g}} \quad & \frac{1}{2}\|\tilde{\mathbf{B}} - \mathbf{\Phi}\mathbf{\Psi}^{\mathrm{T}}\|_{\mathrm{F}}^2 \\ s.t. \quad & \mathbb{E}\left[\mathbf{f}(X)\right] = \mathbb{E}\left[\mathbf{g}(Y)\right] = \mathbf{0}.\end{aligned} \tag{11}$$

Note that we do not impose constraints on the $\mathrm{cov}(\mathbf{f}(X))$ or $\mathrm{cov}(\mathbf{g}(Y))$. We will soon argue that this formulation leads to the same feature geometry as the one in (10). In order to solve this problem, we introduce the following theorem:

Theorem 2. ***(Eckart-Young-Mirsky Theorem)*** *(Eckart and Young 1936) Suppose* $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$*, then* $\mathbf{A}_r = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^{\mathrm{T}} = \sum_{i=1}^{r}\sigma_i\mathbf{u}_i\mathbf{v}_i^T$ *is the optimal solution to the following low-rank approximation problem:*

$$\begin{aligned} \min_{\mathbf{A}_r} \quad & \|\mathbf{A} - \mathbf{A}_r\|_{\mathrm{F}}^2 \\ s.t. \quad & \mathrm{rank}(\mathbf{A}_r) \leq r. \end{aligned} \tag{12}$$

Therefore, the optimal $\mathbf{\Phi}^*$ and $\mathbf{\Psi}^*$ should follow:

$$\mathbf{\Phi}^*\mathbf{\Psi}^{*\mathrm{T}} = \sum_{i=1}^{k}\sigma_i\mathbf{u}_i\mathbf{v}_i^T = \mathbf{U}_{1:k}\mathbf{\Sigma}_{1:k}\mathbf{V}_{1:k}^{\mathrm{T}} \tag{13}$$

The $\mathbf{\Phi}$ and $\mathbf{\Psi}$ is not unique. Given any constant decomposition of $\mathbf{\Sigma}_{1:k} = \mathbf{H}_1\mathbf{H}_2^{\mathrm{T}}$, there is an associated solution $\mathbf{\Phi}^* = \mathbf{U}_{1:k}\mathbf{H}_1, \mathbf{\Psi}^* = \mathbf{V}_{1:k}\mathbf{H}_2$. Equivalent expression for $\mathbf{f}$ and $\mathbf{g}$ is:

$$\begin{aligned} f_i^*(x) &= \left[\mathbf{U}_{1:k}\mathbf{H}_1\right]_{x,i} / \sqrt{P_X(x)}, i = 1, ..., k, x \in \mathcal{X} \\ g_i^*(y) &= \left[\mathbf{V}_{1:k}\mathbf{H}_2\right]_{x,i} / \sqrt{P_Y(y)}, i = 1, ..., k, y \in \mathcal{Y} \end{aligned} \tag{14}$$

Since $\mathbf{H_1}$ are $\mathbf{H_2}$ are invertable, one can conclude that the optimal feature transformation for Soft-HGR (14) and for the HGR maximal correlation (5) are linearly transformable from the one to the other. Namely, they span the same feature space, *i.e.* $\mathrm{span}\{f_1, f_2, ..., f_k\} = \mathrm{span}\{f_1^*, f_2^*, ..., f_k^*\}$ (resp. for $\mathbf{g}$) and therefore describe same amount of information. One way to understand this equivalence is to imagine that the HGR features are feed into a linear dense layer, and output Soft-HGR features with same dimensions.

The Soft-HGR Objective

Thus far, we prove that the low-rank approximation of $\tilde{\mathbf{B}}$ also leads to the optimal feature transformation. Based on this idea, now we develop the operational objective for Soft-HGR. By expanding (11), we have:

$$\frac{1}{2}\|\tilde{\mathbf{B}} - \mathbf{\Phi}\mathbf{\Psi}^{\mathrm{T}}\|_{\mathrm{F}}^2 \tag{15}$$

$$= \frac{1}{2}\|\tilde{\mathbf{B}}\|_{\mathrm{F}}^2 - \mathrm{tr}(\mathbf{\Phi}^{\mathrm{T}}\tilde{\mathbf{B}}\mathbf{\Psi}) + \frac{1}{2}\mathrm{tr}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi}) \tag{16}$$

where the norm of $\tilde{\mathbf{B}}$ given the data is a constant. Minimizing the last two terms with respect to $\mathbf{f}$ and $\mathbf{G}$ leads us to the Soft-HGR objective:

$$\begin{aligned} \max_{\mathbf{f},\mathbf{g}} \quad & \mathbb{E}\left[\mathbf{f}^{\mathrm{T}}(X)\mathbf{g}(Y)\right] - \frac{1}{2}\mathrm{tr}\left(\mathrm{cov}(\mathbf{f}(X))\,\mathrm{cov}(\mathbf{g}(Y))\right) \\ s.t. \quad & \mathbb{E}\left[\mathbf{f}(X)\right] = \mathbb{E}\left[\mathbf{g}(Y)\right] = \mathbf{0}. \end{aligned} \tag{17}$$

The proposed Soft-HGR consists of two inner products, one between feature mappings and the other between feature covariance. The first term in (17) is consistent to the objective of the HGR maximal correlation, and the second term is considered as a soft regularizer to replace the whitening constraints.

Follow the practice of Deep CCA, we design transformation functions $\mathbf{f}$ and $\mathbf{g}$ as parametric neural networks. As long as the reachable functional space of the neural structures covers the optimal feature transformation, the Soft-HGR and the HGR maximal correlation will always lead us to the equivalent solution.

Algorithm 1 Evaluate Soft-HGR on a mini-batch

Input:

Paired data samples of two modalities in a mini-batch of size m: $(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \cdots, (\mathbf{x}^{(m)}, \mathbf{y}^{(m)})$

Two branches of parameterized neural networks with k output units: $\mathbf{f}$ and $\mathbf{g}$

Output:

The objective value of Soft-HGR

1: Subtract the mean of features:

$\mathbf{f}(\mathbf{x}^{(i)}) \leftarrow \mathbf{f}(\mathbf{x}^{(i)}) - \frac{1}{m}\sum_{j=1}^{m}\mathbf{f}(\mathbf{x}^{(j)}), i = 1, \cdots, m$

$\mathbf{g}(\mathbf{y}^{(i)}) \leftarrow \mathbf{g}(\mathbf{y}^{(i)}) - \frac{1}{m}\sum_{j=1}^{m}\mathbf{g}(\mathbf{y}^{(j)}), i = 1, \cdots, m$

2: Compute the empirical covariance:

$\mathrm{cov}(\mathbf{f}) \leftarrow \frac{1}{m-1}\sum_{i=1}^{m}\mathbf{f}(\mathbf{x}^{(i)})\mathbf{f}(\mathbf{x}^{(i)})^{\mathrm{T}}$

$\mathrm{cov}(\mathbf{g}) \leftarrow \frac{1}{m-1}\sum_{i=1}^{m}\mathbf{g}(\mathbf{y}^{(i)})\mathbf{g}(\mathbf{y}^{(i)})^{\mathrm{T}}$

3: Compute the empirical Soft-HGR objective:

$\frac{1}{m-1}\sum_{i=1}^{m}f(\mathbf{x}^{(i)})^{\mathrm{T}}g(\mathbf{y}^{(i)}) - \frac{1}{2}\mathrm{tr}(\mathrm{cov}(\mathbf{f})\,\mathrm{cov}(\mathbf{g}))$

Optimization

In practise, we do not usually have access to the joint probability distribution P_{XY}, but rather paired multimodal samples $(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \cdots, (\mathbf{x}^{(m)}, \mathbf{y}^{(m)})$ retrived from this distribution. As common practices, we embrace SGD techniques that operate on mini-batch of data to optimize the Soft-HGR. The prominent concern here is how to the estimate of the sample covariance with only partially seen mini-batches. In fact, we find that simply using the batch covariance as a replacement awards the best performance. This implies the Soft-HGR actually decomposes the empirical $\tilde{\mathbf{B}}$ over every mini-batch. Only in this way the empirical P_{XY} is always consistent with the marginal distribution P_X and P_Y, where the covariance is evaluated on. The detailed procedure to calculate the Soft-HGR objective is summarized in Algorithm 1. The overall complexity of Soft-HGR is $O(mk^2)$, which is significantly less compared to $O(mk^2 + k^3)$ for normal HGR implementation, *i.e.* Deep CCA. It is also worth noting that our method does not impose an upper bound on the feature dimension k. The optimization is consistently stable for very large k.

Extension to More or Missing Modalities

The HGR maximal correlation is originally defined on two random variables. In contrast to reconstruction models (Srivastava and Salakhutdinov 2012; Zhao, Hu, and Wang 2015), the multi-modal extension for correlation based models is not straightforward. New modalities will bring additional whitening constraints, and the computational complexities scales up. However, in Soft-HGR, the "soft" formulation provides

more flexibility. Recall that the core idea behind the Soft-HGR is to find an approximation of the $\tilde{\mathbf{B}}$ matrix defined on two modalities. In order to handle more than two modalities, the multimodal Soft-HGR should be able to learn feature transformations which recover all pairwise $\tilde{\mathbf{B}}$ simultaneously. Landing on this idea, let $X_1, \ldots, X_d$ be d different modalities, and $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(d)}$ be their corresponding transformation functions, the multimodal Soft-HGR is defined as:

$$\begin{aligned} \max_{\mathbf{f}^{(1)},\ldots,\mathbf{f}^{(d)}} \quad & \mathbb{E}\left[\sum_{i\neq j}^{d} \mathbf{f}^{(i)\mathrm{T}}(X_i)\mathbf{f}^{(j)}(X_j)\right] \\ & - \frac{1}{2}\sum_{i\neq j}^{d} \operatorname{tr}\left(\operatorname{cov}(\mathbf{f}^{(i)}(X_i))\operatorname{cov}(\mathbf{f}^{(j)}(X_j))\right) \quad (18) \\ s.t. \quad & \mathbf{f}^{(i)} : \mathcal{X}_i \to \mathbb{R}^k; \mathbb{E}\left[\mathbf{f}^{(i)}(X_i)\right] = \mathbf{0}; \\ & i, j = 1, 2, \ldots, d. \end{aligned}$$

When $d = 3$, Figure 1 provides an illustration for (18) with neural network implementations. The DNN structure for each neural branches may vary, depending on the statistical property of the inputs. The overall model extracts the features from every neural branch, and maximize their pairwise Soft-HGR in an additive manner. From an information theoretical perspective, maximizing (18) is equivalent to extracting the common information from multiple random variables.

Note that this generalization also provides solutions to deal with data with partially missing modalities. To see this, the first term in (18) can be applied only on the presented modalities for each training sample, and the second term is always measurable as it only depends on the marginal distribution of individual modalities.

Incorporating Supervised Information

The primary goal of the above framework is to extract the correlation between modalities. Therefore, any information that is private to the individual modality is eliminated, regardless of its discriminative power. The intuition behind the supervised/semi-supervised adaptation is that feature extraction should be conducted under the guidance of supervised labels, even if they are insufficient.

Assumed that a subset of bi-modal data is associated with discrete labels Z with range $\mathcal{Z} = \{1, 2, \ldots, |\mathcal{Z}|\}$. In order to receive the supervised information from labels, we feed the joint representation, the concatenation of individual feature mappings, into a softmax classifier. The cross entropy loss is added to the overall objective, with a hyper-parameter $\lambda \in [0, 1]$ to trade off the strength of the unsupervised component:

$$\begin{aligned} \mathcal{L} =& (\lambda - 1) \cdot \mathbb{E}\left[\log Q_{Z|XY}\right] - \lambda \mathbb{E}\left[\mathbf{f}^{\mathrm{T}}(X)\mathbf{g}(Y)\right] \\ & + \frac{\lambda}{2} \operatorname{tr}\left(\operatorname{cov}(\mathbf{f}(X))\operatorname{cov}(\mathbf{g}(Y))\right) \end{aligned} \quad (19)$$

where

$$Q_{Z=j|XY} = \frac{\exp\left(\left[\mathbf{f}^{\mathrm{T}}(X), \mathbf{g}^{\mathrm{T}}(Y)\right]\boldsymbol{\theta}_j\right)}{\sum_{i=1}^{|\mathcal{Z}|}\exp\left(\left[\mathbf{f}^{\mathrm{T}}(X), \mathbf{g}^{\mathrm{T}}(Y)\right]\boldsymbol{\theta}_i\right)} \quad (20)$$

Table 1: The linear correlation between features extracted from the Soft-HGR and the HGR maximal correlation.

	Feature dimensions		
Linear correlation	10	20	40
Upper Bound	10	20	40
$\mathbf{f}_{\text{HGR}}(\mathbf{X})$ and $\mathbf{g}_{\text{HGR}}(\mathbf{Y})$	1.36	2.37	3.40
$\mathbf{f}_{\text{SHGR}}(\mathbf{X})$ and $\mathbf{g}_{\text{SHGR}}(\mathbf{Y})$	1.36	2.37	3.40
$\mathbf{f}_{\text{SHGR}}(\mathbf{X})$ and $\mathbf{f}_{\text{HGR}}(\mathbf{X})$	9.99	20.00	39.99
$\mathbf{g}_{\text{SHGR}}(\mathbf{Y})$ and $\mathbf{g}_{\text{HGR}}(\mathbf{Y})$	10.00	20.00	39.99

In semi-supervised settings, the supervised softmax loss, the first term in (19), is only effective when labels are presented. The last two terms of (19) corresponds to the Soft-HGR loss, which is evaluated independently from labels. The gradients from the label Z are first backpropagated to the individual feature mappings, then affect the feature selection.

Experiments

In this section, we evaluate Soft-HGR in the following aspects:

- To verify the relationship between the HGR features and Soft-HGR feature is linear;
- To compare the efficiency and numerical stability of CCA based models and Soft-HGR;
- To demonstrate the power of semi-supervised Soft-HGR on discriminative tasks with limited labels;
- To show the performance of Soft-HGR on more than two modalities and missing modalities.

Comparing Soft-HGR with HGR

The formulation of the Soft-HGR and the original HGR maximal correlation are equivalent except for the way they control whitening. In this section, we compare two methods in terms of linearity, efficiency and stability.

Linearity Check Based on the theory, the HGR and the Soft-HGR transformations should span the same feature space. To verify this, we randomly generated 100K data samples (x_i, y_i) from a randomly chosen joint distribution P_{XY}, where $X, Y \in \{1, ..., 50\}$ are both discrete random variables. The HGR feature $\{\mathbf{f}_{\text{HGR}}, \mathbf{g}_{\text{HGR}}\}$ is obtained by directly solving the SVD for $\tilde{\mathbf{B}}$, which is calculated from empirical joint distribution $\tilde{P}_{XY}$. In order to retrieve the Soft-HGR features $\{\mathbf{f}_{\text{SHGR}}, \mathbf{g}_{\text{SHGR}}\}$, we first turn the data into one-hot form $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{100K\times 50}$, then feed them into a two-branch one-layer neural network optimized by Soft-HGR objective. Note that when data are one-hot encoded, all possible functions can be captured by linear operations. Finally, we apply all learned functions to data and run linear CCA between every two feature transformations. Recall that the HGR and Soft-HGR features are linearly transformable from the one to the other. Therefore, the linear correlation between $\{\mathbf{f}_{\text{SHGR}}(\mathbf{X}),$

Table 2: Phonetic prediction accuracy obtained by different methods on certain percentages of the labeled data in XRMB.

	Percentages of labels		
Method	10%	50%	100%
Baseline DNN	72.2%	81.2%	86.4%
PCA + DNN	71.5%	80.5%	85.2%
CCA + DNN	70.7%	79.9%	84.4%
Deep CCA + DNN	73.2%	80.1%	84.0%
Soft-HGR + DNN	73.0%	79.9%	83.7%
Soft CCA + DNN	69.4%	76.0%	78.8%
CorrNet	71.2%	79.7%	83.2%
Semi Soft-HGR	**76.3%**	**85.0%**	**88.0%**
Semi Soft CCA	73.6%	82.8%	85.5%

$\mathbf{f}_{\text{HGR}}(\mathbf{X})\}$ and between $\{\mathbf{g}_{\text{SHGR}}(\mathbf{Y}), \mathbf{g}_{\text{HGR}}(\mathbf{Y})\}$, in the ideal case, should reach the upper bound.

Table 1 summarizes the simulation result. The HGR and the Soft-HGR extract exactly the same linear correlation between X and Y on different choice of k. Besides, the correlation between corresponding features from two models is almost identical to the upper bound, which provides an empirical evidence for our theory.

Efficiency and Stability In this subsection, we focus on the efficiency and stability provided by two methods in optimization. In particular, we compare the execution time and maximally reachable feature dimension by applying both models to the MNIST handwritten image dataset (LeCun et al. 1998), which consists of 60K/10K gray-scale digit images of size 28×28 as training/testing sets. We follow the experiment setting in (Andrew et al. 2013), and treat left and right halves of digit images as two modalities X and Y. In order to highlight efficiency difference brought by the objectives, we restrict the all the feature transformation to take the linear form. Therefore, the HGR maximal correlation degrades to linear CCA. Both optimizations are executed on a Nvidia Tesla K80 GPU with mini-batch SGD of 5K batchsize.

Figure 2 compares the execution time on one training epoch with different feature dimensions k. As we expected, Soft-HGR is faster than CCA methods by orders of magnitude. In addition, the execution time of CCA method grows quickly with the feature dimensions. This is undesirable in real-world settings where k could be very large. It is also worth noting that CCA experiences numerical issue when feature dimension exceeds 350. The instability arises in that the empirical covariance matrices over some mini-batches become ill-posed, or even non-invertible.

Soft-HGR for Semi-supervised Learning

In this section we demonstrate how Soft-HGR are applied to improve the performance of discriminative tasks. we evaluate our model on the University of Wisconsin X-ray Microbeam Database (XRMB) (Westbury 1994) for phonetic classification. XRMB is a bi-modal Dataset consisting of articulatory

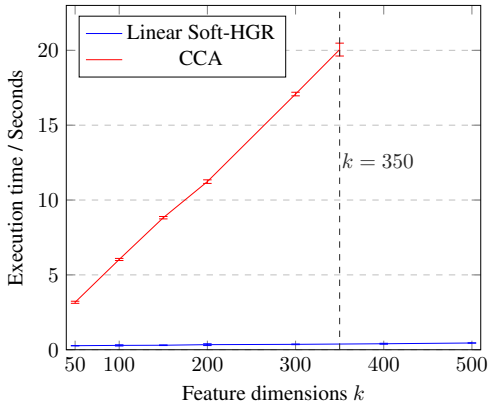

Figure 2: Execution time of SGD on CCA and linear Soft-HGR for one training epoch on MNIST data. When k is larger than 350, CCA experiences numerical issues.

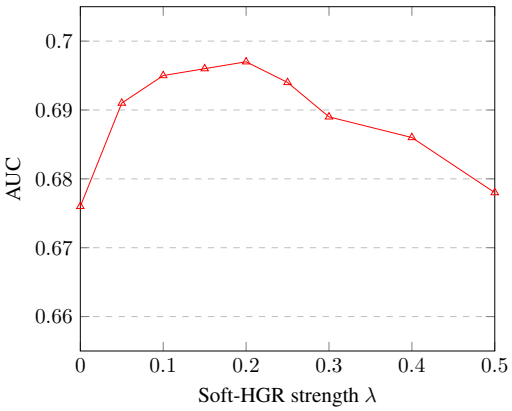

Figure 3: The effect of hyper-parameter λ on AUC

and acoustic data. Followed the same preprocessing and reconstruction procedures as described in (Arora and Livescu 2013; Wang et al. 2015a), we obtain the total number of 160K entries of acoustic and vectors articulatory vectors $X \in \mathbb{R}^{273}$ and $Y \in \mathbb{R}^{112}$, corresponding to 41 classes of labels Z.

Experiment Settings While both modalities X and Y are available in the training phase, Y is not provided at the test time. Namely, the model is evaluated by the classification accuracy with only X observed. We expect using Y during training to improve the classification performance, even if they are absent in the test phase. In addition, we partially mask out some portions of labels Z associated with the training data. These two restrictions are consistent with the real world multimodal settings where facial movement data is usually not obtainable, and labels are limited.

Comparing Models (1) **Supervised DNN**, which has four hidden layers [1K, 1K, 1K, 1K]. It only takes raw feature $\mathbf{X}$ as inputs and makes predictions on Z; Supervised DNN can only deal with labeled data. (2) **DNN on CCA features**: the DNN structure is the same as in (1) but it accepts transformed $\mathbf{f}(\mathbf{X})$ as input. $\mathbf{f}(\mathbf{X})$ is obtained from PCA, CCA (Hotelling

Table 3: AUC obtained by different methods using part of the labels or part of the modalities

		Missing labels			Missing modalities	
Method	No Missing	20%	50%	90%	20%	50%
LR	0.6625	0.6623	0.6618	0.6534	0.6588	0.6286
FM	0.6780	0.6728	0.6723	0.6543	0.6696	0.6449
Deep FM	0.6803	0.6765	0.6756	0.6613	0.6714	0.6450
Neural FM	0.6760	0.6768	0.6746	0.6570	0.6661	0.6574
Semi Soft-HGR	**0.6972**	**0.6935**	**0.6906**	**0.6728**	**0.6823**	**0.6682**

1936), Deep CCA (Andrew et al. 2013), Soft CCA (Chang, Xiang, and Hospedales 2018), Correlational Neural Network (CorrNet) (Chandar et al. 2016) and our model. Except for PCA which extracts feature only from $\mathbf{X}$, all other methods are trying to find the most correlated $\mathbf{f}(\mathbf{X})$ to $\mathbf{g}(\mathbf{Y})$. For Deep CCA, Soft CCA and our model, the selected $\mathbf{f}$ is a DNN with two hidden layers: [1K, 1K] and $\mathbf{g}$ is linear. The output feature dimensions k is chosen to be 80 for all methods, as higher value leads to unstable gradients in Deep CCA. (3) **Semi-supervised Model**: we construct semi-supervised Soft-HGR as described in subsection Incorporating Supervised Information, except that the top layer softmax function only takes $\mathbf{f}(X)$ as input. In particular, we use DNN with four hidden layers [1K, 1K, 1K, 1K] for $\mathbf{f}$ and linear function for $\mathbf{g}$. To see the equivalence, when $\lambda = 0$, the network becomes the supervised DNN. For a fair comparison, we also adapt semi-supervised Soft CCA in the same manner. However, we found Deep CCA fails the adaptation because the training is very unstable which prevents us to get a reliable result. In all DNNs, batch normalization (Ioffe and Szegedy 2015) is applied before ReLU activation function to ensure better convergence. The hyper-parameters for each model are determined by their best average performance on validation set on 5-fold cross validation. Table 2 reports the average phonetic prediction accuracy.

Observations (1) Semi-supervised Soft-HGR achieves the highest accuracy among all models, and the difference becomes more apparent when labels are insufficient. (2) The discriminative performance of Deep CCA and Soft-HGR are similar as they learns equivalent features. (3) $\mathbf{f}(\mathbf{X})$ trained by various unsupervised models is not necessarily more discriminative than raw feature X_1. In fact, they only improve classification when labels are extremely limited. In other cases, their performances are inferior to the end-to-end DNN because valuable information may be lost as $\mathbf{f}$ projects the data into lower dimensions.

Soft-HGR for More or Missing Modalities

In this section we apply our method to recommender system. In such problems, users X_u, items X_i, and context X_c are three natural modalities. Extensive success achieved by collaborative filtering techniques (Breese, Heckerman, and Kadie 1998) demonstrates that the correlations between these modalities are useful to infer user behaviors.

Specifically, we experiment with KKBox's Music Recommendation Dataset (Chen et al. 2018). The goal is to predict the chances of a user listening to a song repetitively after the first listening event within a month. The binary labels $Y = 1$ represents the user listens to the song again, and $Y = 0$ means the opposite. The user features X_u and item (song) features X_i are explicitly given, and we treat `source_system_tab`, `source_screen_name` and `source_type` as context features X_c. The categorical features are one-/multi-hot encoded, and continuous ones are normalized. The features corresponding to one modality are concatenated into a single vector, resulting in X_u, X_i, and X_c as 34656, 623691, and 45 dimensional feature vectors, respectively. The test labels are not disclosed, therefore we use the last 20% of 7M training data as test set[1]. We test the model under two settings, where labels are insufficient or one modality is missing. In the first setting, we conceal 20%/50%/90% of the labels in training data. In the second scenarios, we randomly mask one of the three modalities as missing in 20%/50% of both training and test data, the status of whether data is missing is constructed as a binary flag in the feature vector. The performances are evaluated by the Area under the ROC curve (AUC).

Comparing Models We compare our model against to the state-of-art predictive model for sparse data. These include **Shallow models**: Logistic regression (LR) and Factorization machines (FM) (Rendle 2010), and **Deep models**: Deep FM (Guo et al. 2017), Neural FM (He and Chua 2017). For models besides LR, the dimension of feature embedding is set to 16. The DNN component for Deep FM, Neural FM and Semi Soft-HGR has consistent structure [100, 100]. **Semi-supervised Soft-HGR**: The architecture for the unsupervised part is designed mainly according to Figure 1. However, since fully connected layers is not effective for sparse features, a Bi-Interaction layer, proposed in (He and Chua 2017), is inserted between the input and DNN structure. The output features from three neural network branches are forwarded to an average pooling layer. The output joint representation is feed to a softmax function for prediction. The hyper-parameter λ controls the participation of the Soft-HGR loss. The comparing result is reported in Table 3. In order to highlight the role of Soft-HGR loss, we plot the AUC versus λ when Semi Soft-HGR is trained with all labels in Figure 3.

[1]The split is suggested by the 1st place solution. The last part of the data is used for test set because the data are speculated to be chronologically ordered.

Observations (1) Semi-supervised Soft-HGR achieves significantly better performance than all the baselines. (2) From Figure 3 we can see the performance decreases as it is eliminated from the objective (*i.e.* $\lambda = 0$). Arguably, the performance gain comes from the introduction of unsupervised Soft-HGR objective.

Conclusion

In this paper, we propose a multimodal feature extraction framework based on the HGR maximal correlation. Further, we replace the intrinsic whitening constraints with a "soft" regularizer which guarantees the efficiency and stability in optimization. Our model is able to cope with more than two modalities, missing modalities, and can be readily generalized to the semi-supervised setting. Extensive experiments show that our proposed model outperforms state-of-the-art multimodal feature selection methods in different scenarios.

Acknowledgement

The research of Shao-Lun Huang was funded by the Natural Science Foundation of China 61807021, and Shenzhen Municipal Scientific Program JCYJ20170818094022586.

References

Akaho, S. 2006. A kernel method for canonical correlation analysis. *CoRR* abs/cs/0609071.

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

Arora, R., and Livescu, K. 2013. Multi-view cca-based acoustic features for phonetic recognition across speakers and domains. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7135–7139.

Bach, F. R., and Jordan, M. I. 2002. Kernel independent component analysis. *Journal of Machine Learning Research* 3(Jul):1–48.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1.

Breese, J. S.; Heckerman, D.; and Kadie, C. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 43–52.

Chandar, S.; Khapra, M. M.; Larochelle, H.; and Ravindran, B. 2016. Correlational neural networks. *Neural Computation* 28(2):257–285.

Chang, X.; Xiang, T.; and Hospedales, T. M. 2018. Scalable and effective deep cca via soft decorrelation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Y.; Xie, X.; Lin, S.-D.; and Chiu, A. 2018. Wsdm cup 2018: Music recommendation and churn prediction. In *ACM International Conference on Web Search and Data Mining (WSDM)*, 8–9.

Eckart, C., and Young, G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* 1(3):211–218.

Feizi, S.; Makhdoumi, A.; Duffy, K.; Kellis, M.; and Medard, M. 2017. Network maximal correlation. *IEEE Transactions on Network Science and Engineering* 4(4):229–247.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NIPS)*, 2121–2129.

Gebelein, H. 1941. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 21(6):364–379.

Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. DeepFM: A factorization-machine based neural network for ctr prediction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1725–1731.

He, X., and Chua, T.-S. 2017. Neural factorization machines for sparse predictive analytics. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 355–364.

Hirschfeld, H. O. 1935. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society* 31(4):520524.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3/4):321–377.

Huang, S.-L.; Makur, A.; Zheng, L.; and Wornell, G. W. 2017. An information-theoretic approach to universal feature selection in high-dimensional inference. In *International Symposium on Information Theory (ISIT)*, 1336–1340.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 448–456.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Pearson, K. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58:240–242.

Rendle, S. 2010. Factorization machines. In *IEEE International Conference on Data Mining (ICDM)*, 995–1000.

Rényi, A. 1959. On measures of dependence. *Acta Mathematica Hungarica* 10(3-4):441–451.

Sohn, K.; Shang, W.; and Lee, H. 2014. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc. 2141–2149.

Srivastava, N., and Salakhutdinov, R. R. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, 2222–2230.

Vendrov, I.; Kiros, R.; Fidler, S.; and Urtasun, R. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. A. 2015a. Unsupervised learning of acoustic features via deep canonical correlation analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4590–4594.

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*, 688–695. IEEE.

Westbury, J. 1994. X-ray microbeam speech production database user's handbook. *Waisman Center on Mental Retardation & Human Development University of Wisconsin Madison*.

Zhao, L.; Hu, Q.; and Wang, W. 2015. Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso. *IEEE Transactions on Multimedia* 17(11):1936–1948.