Learning to Disentangle Factors of Variation with Manifold Interaction

Scott Reed REEDSCOT@UMICH.EDU
Kihyuk Sohn KIHYUKS@UMICH.EDU
Yuting Zhang YUTINGZH@UMICH.EDU
Honglak Lee HONGLAK@UMICH.EDU
Dept. of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

Many latent factors of variation interact to generate sensory data; for example, pose, morphology and expression in face images. In this work, we propose to learn manifold coordinates for the relevant factors of variation and to model their joint interaction. Many existing feature learning algorithms focus on a single task and extract features that are sensitive to the task-relevant factors and invariant to all others. However, models that just extract a single set of invariant features do not exploit the relationships among the latent factors. To address this, we propose a higher-order Boltzmann machine that incorporates multiplicative interactions among groups of hidden units that each learn to encode a distinct factor of variation. Furthermore, we propose correspondence-based training strategies that allow effective disentangling. Our model achieves state-of-the-art emotion recognition and face verification performance on the Toronto Face Database. We also demonstrate disentangled features learned on the CMU Multi-PIE dataset.

1. Introduction

A key challenge in understanding sensory data (e.g., image and audio) is to tease apart many factors of variation that combine to generate the observations (Bengio, 2009). For example, pose, shape and illumination combine to generate 3D object images; morphology and expression combine to generate face images. Many factors of variation exist for other modalities, but here we focus on modeling images.

Most previous work focused on building (Lowe, 1999) or learning (Kavukcuoglu et al., 2009; Ranzato et al., 2007; Lee et al., 2011; Le et al., 2011; Huang et al., 2012b;a; Sohn & Lee, 2012) invariant features that are unaffected by nuisance information for the task at hand. However, we argue that image understanding can benefit from retaining information about all underlying factors of variation, because in many cases knowledge about one factor can improve our estimates about the others. For example, a good pose estimate may help to accurately infer the face morphology, and vice versa. From a generative perspective, this approach also supports additional queries involving latent factors; e.g. "what is the most likely face image as pose or expression vary given a fixed identity?"

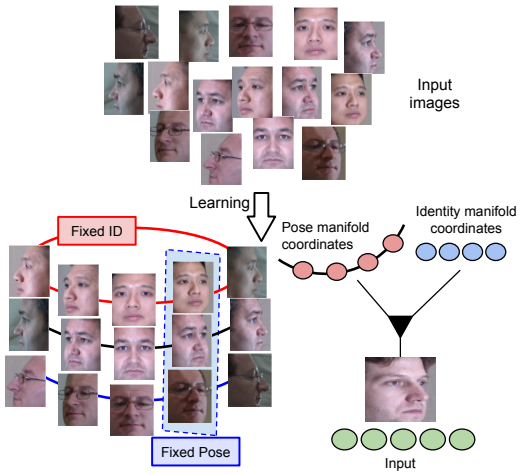

Figure 1. Illustration of our approach for modeling pose and identity variations in face images. When fixing identity, traversing along the corresponding "fiber" (denoted in red ellipse) changes the pose. When fixing pose, traversing across the vertical cross-section (shaded in blue rectangle) changes the identity. Our model captures this via multiplicative interactions between pose and identity coordinates to generate the image.

When the input images are generated from multiple factors of variation, they tend to lie on a complicated manifold, which makes learning useful representations very challenging. We approach this problem by viewing each factor of variation as forming a sub-manifold by itself, and modeling the joint interaction among factors. For example, given face images with different identities and viewpoints, we can envision one sub-manifold for identity and another for view-

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

point. As illustrated in Figure 1, when we consider face images of a single person taken from different azimuth angles (with fixed altitude), the trajectory of images will form a ring-shaped fiber. Similarly, changing the identity while fixing the angle traverses a high-dimensional sub-manifold from one fiber to other.

Concretely, we use a higher-order Boltzmann machine to model the distribution over image features and the latent factors of variation. Further, we propose correspondence-based training strategies that allow our model to effectively *disentangle* the factors of variation. This means that each group of hidden units is sensitive to changes in its corresponding factor of variation, and relatively invariant to changes in the others. We refer to our model variants as *disentangling Boltzmann machines* (disBMs). Our disBM model achieves state-of-the-art emotion recognition and face verification performance on the Toronto Face Database (TFD), as well as strong performance in pose estimation and face verification on CMU Multi-PIE.

2. Preliminaries

In this section, we briefly review the restricted Boltzmann machine (RBM), a bipartite undirected graphical model composed of D binary visible units[1] $\mathbf{v} \in \{0,1\}^D$ and K binary hidden units $\mathbf{h} \in \{0,1\}^K$. The joint distribution and the energy function are defined as follows:

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp(-E(\mathbf{v},\mathbf{h})),$$

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{D}\sum_{k=1}^{K} v_i W_{ik} h_k - \sum_{k=1}^{K} b_k h_k - \sum_{i=1}^{D} c_i v_i,$$

where Z is the partition function, W_{ik} is a weight between i-th visible and k-th hidden units, b_k are hidden biases, and c_i are visible biases. In the RBM, the units in the same layer are conditionally independent given the units in the other layer. The conditional distributions are computed as:

$$P(v_i = 1 \mid \mathbf{h}) = \sigma(\sum_k W_{ik} h_k + c_i),$$

$$P(h_k = 1 \mid \mathbf{v}) = \sigma(\sum_i W_{ik} v_i + b_k),$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is a logistic function. The RBM can be trained to maximize the log-likelihood of data using stochastic gradient descent. Although the gradient is intractable, we can approximate it using contrastive divergence (CD) (Hinton, 2002).

3. Model description

The disBM is an undirected graphical model with higher-order interactions between observations and multiple groups of hidden units, as in Figure 2. Each group of hidden units can be viewed as manifold coordinates for a distinct factor of variation. Our proposed model is shown in Figure 2. For simplicity, we assume two groups of hidden units $\mathbf{h}$ and $\mathbf{m}$, although it is straightforward to add more groups. If labels are available, they can be incorporated with the $\mathbf{e}$ units (see Section 4.1).

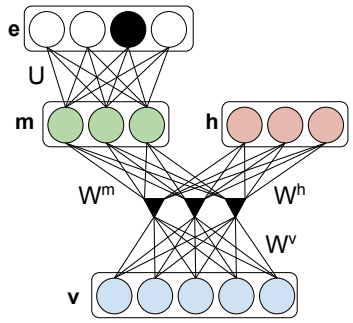

Figure 2. An instance of our proposed model with two groups of hidden units. We can optionally include label units (e.g., label units $\mathbf{e}$ are connected to hidden units $\mathbf{m}$).

3.1. Energy function

As shown in Figure 2, our model assumes 3-way multiplicative interaction between D visible units $\mathbf{v} \in \{0,1\}^D$ and two groups of hidden units $\mathbf{h} \in \{0,1\}^K$ and $\mathbf{m} \in \{0,1\}^L$. We define the energy function as:

$$E(\mathbf{v},\mathbf{m},\mathbf{h}) = -\sum_f (\sum_i W^v_{if} v_i)(\sum_j W^m_{jf} m_j)(\sum_k W^h_{kf} h_k) - \sum_{ij} P^m_{ij} v_i m_j - \sum_{ik} P^h_{ik} v_i h_k \quad (1)$$

We have used factorization of 3D weight tensor $W \in \mathbb{R}^{D\times L\times K}$ into three weight matrices $W^v \in \mathbb{R}^{D\times F}$, $W^m \in \mathbb{R}^{L\times F}$, $W^h \in \mathbb{R}^{K\times F}$ with F factors as

$$W_{ijk} = \sum_{f=1}^{F} W^v_{if} W^m_{jf} W^h_{kf} \quad (2)$$

to reduce the number of model parameters (Memisevic & Hinton, 2010). We also include additive connections with weight matrices $P^m \in \mathbb{R}^{D\times L}$ and $P^h \in \mathbb{R}^{D\times K}$ between visible units and each group of hidden units. We omit the bias terms for clarity of presentation. Although the hidden units are not conditionally independent given the visible units, units in each group are conditionally independent given units in all other groups. The conditional distributions are as follows:[2]

$$P(v_i = 1 \mid \mathbf{h},\mathbf{m}) = \sigma(\sum_{jk} W_{ijk} m_j h_k + \sum_j P^m_{ij} m_j + \sum_k P^h_{ik} h_k) \quad (3)$$

$$P(m_j = 1 \mid \mathbf{v},\mathbf{h}) = \sigma(\sum_{ik} W_{ijk} v_i h_k + \sum_i P^m_{ij} v_i) \quad (4)$$

$$P(h_k = 1 \mid \mathbf{v},\mathbf{m}) = \sigma(\sum_{ij} W_{ijk} v_i m_j + \sum_i P^h_{ik} v_i) \quad (5)$$

[1] The RBM can be extended to model the real-valued visible units (Hinton & Salakhutdinov, 2006).

[2] W_{ijk} denotes factorized weights as in Equation (2).

The conditional independence structure allows efficient 3-way block Gibbs sampling.

3.2. Inference and learning

Inference. The exact posterior distribution is intractable since $\mathbf{h}$ and $\mathbf{m}$ are not conditionally independent given $\mathbf{v}$. Instead, we use variational inference to approximate the true posterior with a fully factorized distribution $Q(\mathbf{m},\mathbf{h}) = \prod_j \prod_k Q(m_j)Q(h_k)$. By minimizing $\mathrm{KL}\left(Q(\mathbf{m},\mathbf{h}) \| P(\mathbf{m},\mathbf{h} \mid \mathbf{v})\right)$, we obtain the following fixed-point equations:

$$\hat{h}_k = \sigma(\sum_{ij} W_{ijk} v_i \hat{m}_j + \sum_i P^h_{ik} v_i) \tag{6}$$

$$\hat{m}_j = \sigma(\sum_{ik} W_{ijk} v_i \hat{h}_k + \sum_i P^m_{ij} v_i) \tag{7}$$

where $\hat{h}_k = Q(h_k = 1)$ and $\hat{m}_j = Q(m_j = 1)$. Initialized with all 0's, the mean-field update proceeds by alternately updating $\hat{\mathbf{h}}$ and $\hat{\mathbf{m}}$ using Equation (6) and (7) until convergence. We found that 10 iterations were enough in our experiments.

Learning. We train the model to maximize the data log-likelihood using stochastic gradient descent. The gradient of the log-likelihood for parameters $\Theta = \{W^v, W^m, W^h, P^m, P^h\}$ can be computed as:

$$-\mathbb{E}_{P(\mathbf{m},\mathbf{h}|\mathbf{v})}\left[\frac{\partial E(\mathbf{v},\mathbf{m},\mathbf{h})}{\partial \theta}\right] + \mathbb{E}_{P(\mathbf{v},\mathbf{m},\mathbf{h})}\left[\frac{\partial E(\mathbf{v},\mathbf{m},\mathbf{h})}{\partial \theta}\right]$$

Unlike in the RBM case, both the first (i.e., data-dependent) and the second (i.e., model-dependent) terms are intractable. We can approximate the data-dependent term with variational inference and the model-dependent term with persistent CD (Tieleman, 2008) by running a 3-way sampling using Equation (3),(4),(5). A similar approach has been proposed for training general Boltzmann machines (Salakhutdinov & Hinton, 2009).

3.3. Computing gradients via backpropagation

When the training objective depends on hidden unit activations, such as correspondence (Section 4.2) or sparsity (Lee et al., 2008; Hinton, 2010), the exact gradient can be computed via backpropagation through the recurrent neural network (RNN) induced by mean-field inference (See Figure 3). The forward propagation proceeds as:

$$\hat{h}^{(t+1)}_k = \sigma(\sum_{ij} W_{ijk} v_i \hat{m}^{(t)}_j + \sum_i P^h_{ik} v_i) \tag{8}$$

$$\hat{m}^{(t+1)}_j = \sigma(\sum_{ik} W_{ijk} v_i \hat{h}^{(t)}_k + \sum_i P^m_{ij} v_i) \tag{9}$$

A similar strategy was rigorously developed by Stoyanov et al. (2011) and was used to train deep Boltzmann machines (Goodfellow et al., 2013).

4. Training strategies for disentangling

Generative training of the disBM does not explicitly encourage disentangling, and generally did not yield well-disentangled features in practice. However, we can achieve better disentangling by exploiting correspondences between images (e.g. matching identity, expression or pose), and by using labels.

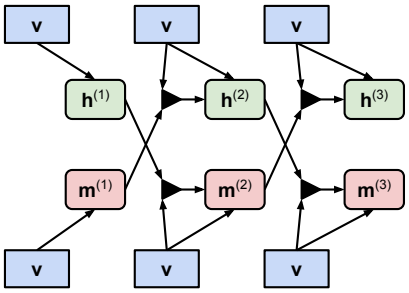

Figure 3. Visualization of the RNN structure of our model. Arrows show the direction of the forward propagation.

4.1. Learning with partial labels

We can use labels to improve disentangling, even when they are only provided for a subset of factors. Figure 2 illustrates how label units $\mathbf{e}$ are connected to the corresponding hidden units $\mathbf{m}$ but not to the other group. In this way, we can make $\mathbf{m}$ sensitive to the variation related to $\mathbf{e}$ while the other group of hidden units focus on other types of variation in the data. To accommodate labels, we augment the energy function as:

$$E_{\text{label}}(\mathbf{v},\mathbf{m},\mathbf{h},\mathbf{e}) = E(\mathbf{v},\mathbf{m},\mathbf{h}) - \sum_{jl} m_j U_{jl} e_l \tag{10}$$

subject to $\sum_l e_l = 1$.[3] The posterior inference is intractable, and we use variational inference resulting in the following fixed-point equations:

$$\hat{h}_k = \sigma(\sum_{ij} W_{ijk} v_i \hat{m}_j + \sum_i P^h_{ik} v_i) \tag{11}$$

$$\hat{m}_j = \sigma(\sum_{ik} W_{ijk} v_i \hat{h}_k + \sum_i P^m_{ij} v_i + \sum_l U_{jl} \hat{e}_l) \tag{12}$$

$$\hat{e}_l = \frac{\exp(\sum_j U_{jl} \hat{m}_j)}{\sum_{l'} \exp(\sum_j U_{jl'} \hat{m}_j)} \tag{13}$$

The model is trained to maximize the hybrid objective $\log P(\mathbf{v},\mathbf{e}) + \eta \log P(\mathbf{e}|\mathbf{v})$ (Larochelle & Bengio, 2008).

4.2. Learning with correspondence

CLAMPING HIDDEN UNITS FOR PAIRS

If we know two data points $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ match in some factor of variation, we can "clamp" the corresponding hidden units to be the same for both data points. For example, given two images from the same person, we clamp the $\mathbf{h}$ units so that they focus on modeling the common face morphology while other hidden units explain the differences such as pose or expression. To do clamping, we augment

[3] Although we restrict the label units to be multinomial, it is straightforward to relax the representation into unrestricted binary units when there are structured labels.

the energy function as follows:

$$E_{\text{clamp}}(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{m}^{(1)}, \mathbf{m}^{(2)}, \mathbf{h}) = E(\mathbf{v}^{(1)}, \mathbf{m}^{(1)}, \mathbf{h}) + E(\mathbf{v}^{(2)}, \mathbf{m}^{(2)}, \mathbf{h}) \quad (14)$$

Note that we can incorporate labels via Equation (10) when available. The fixed-point equations are the same as before, except that Equation (6) changes to reflect the contributions from both $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$:

$$\hat{h}_k = \sigma(\sum_{ij} W_{ijk} v_i^{(1)} \hat{m}_j^{(1)} + \sum_i P_{ik}^h v_i^{(1)} + \sum_{ij} W_{ijk} v_i^{(2)} \hat{m}_j^{(2)} + \sum_i P_{ik}^h v_i^{(2)}) \quad (15)$$

The model is trained to maximize the joint log-likelihood of data pairs $\log P(\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$.

MANIFOLD-BASED TRAINING

In the manifold learning perspective, we want each group of hidden units to be a useful embedding with respect to its factor of variation. Specifically, corresponding data pairs should be embedded nearby, while the non-corresponding data pairs should be far apart. Clamping forces corresponding pairs into exactly the same point within a sub-manifold, which may be too strong of an assumption depending on the nature of the correspondence. Furthermore, clamping does not exploit knowledge of non-correspondence. Instead, we propose to learn a representation $\mathbf{h}$ such that

$$||\mathbf{h}^{(1)} - \mathbf{h}^{(2)}||_2^2 \approx 0 \quad , \text{ if } (\mathbf{v}^{(1)}, \mathbf{v}^{(2)}) \in \mathcal{D}_{sim}$$
$$||\mathbf{h}^{(1)} - \mathbf{h}^{(3)}||_2^2 \geq \beta \quad , \text{ if } (\mathbf{v}^{(1)}, \mathbf{v}^{(3)}) \in \mathcal{D}_{dis}$$

where $\mathcal{D}_{sim}$ is a set of corresponding data pairs and $\mathcal{D}_{dis}$ is a set of non-corresponding data pairs. Formally, the manifold objective for $\mathbf{h}$ is written as:

$$||\mathbf{h}^{(1)} - \mathbf{h}^{(2)}||_2^2 + \max(0, \beta - ||\mathbf{h}^{(1)} - \mathbf{h}^{(3)}||_2)^2 \quad (16)$$

This approach does not directly use label units, but labels can be used to construct correspondence sets $\mathcal{D}_{sim}$ and $\mathcal{D}_{dis}$. The formulation is similar to the one proposed by Hadsell et al. (2006). However, our goal is not dimensionality reduction and we consider multiple factors of variation jointly. Furthermore, we can combine the manifold objective together with the generative objective. Since our model uses mean-field inference to compute the hidden units, we compute gradients via RNN backpropagation as discussed in Section 3.3.

5. Related Work

Manifold learning methods (Tenenbaum et al., 2000; Roweis & Saul, 2000; Hadsell et al., 2006) model the data by learning low-dimensional structures or embeddings. Existing manifold learning methods can learn intrinsically low-dimensional structures such as viewpoint manifolds from face images of a single person, but it becomes challenging to model complex high-dimensional manifolds such as the space of face images from millions of people. Deep learning has shown to be effective in learning such high-dimensional data manifolds, as suggested by Rifai et al. (2011). However, it remains a challenge to jointly model multiple factors of variation and their interacting manifolds.

Our work is related to multi-task learning (Caruana, 1997; Argyriou et al., 2007) if one views each factor as a "task" feature to be learned jointly. However, our approach considers joint interaction among the factors, and benefits from a synergy in which knowledge of one factor can help infer about the others. In addition, our model is generative and can answer higher-order queries involving the input and multiple factors.

There are several related works that use higher-order interactions between multiple latent variables. For example, bilinear models (Tenenbaum & Freeman, 2000) were used to separate style and content within face images (pose and identity) and speech signals (vowels and speaker identity). The tensor analyzer (TA) (Tang et al., 2013) extended factor analysis by introducing a factor loading tensor to model the interaction among multiple groups of latent factor units, and was applied to modeling lighting and face morphology. Our approach is complementary to these, and is also capable of exploiting correspondence information.

The higher-order spike and slab RBM (ssRBM) (Desjardins et al., 2012) extends the ssRBM (Courville et al., 2011) with higher-order interactions. Our motivation is similar, but our model formulation is different and we propose novel training strategies that significantly improve the disentangling. Finally, we show state-of-the-art performance on several discriminative tasks on face images.

The factored gated Boltzmann machine (FGBM) (Memisevic & Hinton, 2010; Susskind et al., 2011) models the relation between data pairs (e.g. translation, rotation of images, facial expression changes) via 3-way interactions. Both the FGBM and disBM are variants of higher-order Boltzmann machines, but the FGBM assumes two sets of visible units interacting with one set of hidden units, whereas the disBM assumes multiple sets of hidden units interacting with a single set of visible units.

The point-wise gated Boltzmann machine (Sohn et al., 2013) is an instance of a higher-order Boltzmann machine that jointly learns and selects task-relevant features. Contractive discriminative analysis (Rifai et al., 2012) also learns groups of task-relevant and irrelevant hidden units using a contractive penalty, but only uses additive interactions between the input and each group of hidden units. These models are complementary to ours in that they learn to separate task-relevant from task-irrelevant features.

Figure 4. Samples from flipped MNIST dataset.

Table 1. Test classification errors on flipped MNIST.

MODEL	RBM			DISBM
# HIDDEN UNITS	1,000	2,000	4,000	1,000
RECOGNITION ERROR RATE	5.18	2.68	2.22	**1.84**

6. Experiments

We evaluated the performance of our proposed model on several image databases:

- **Flipped MNIST**. For each digit of the MNIST dataset, we randomly flipped all pixels (0's to 1's and vice versa) with 50% probability. The dataset consists of 50,000 training images, 10,000 validation images, and 10,000 test images.
- **Toronto Face Database (TFD)** (Susskind et al., 2010). Contains $112,234$ face images with $4,178$ emotion labels and $3,874$ identity labels. There are seven possible emotion labels.
- **CMU Multi-PIE** (Gross et al., 2010). Contains $754,200$ high-resolution face images with variations in pose, lighting, and expression. We manually aligned and cropped the face images.[4]

6.1. Flipped MNIST Digits

To understand the role of multiplicative interactions in disentangling, we constructed a variation of the MNIST digits (LeCun & Cortes, 1998) by flipping the binary pixel values. For half of the digit images, we converted 0's to 1's and vice versa. Examples are shown in Figure 4. The factors in the dataset are the flip mode (0 or 1) and the digit shape. We investigate whether it helps to decompose the posterior into a single flip unit and appearance units that interact multiplicatively to generate the image.

We evaluated the digit recognition performance using our disBM compared to the standard RBM. We trained linear SVMs on RBM hidden unit and disBM appearance unit activations for classification.

In Table 1, the disBM achieves significantly lower error rates than RBMs of each size. We hypothesize that the disBM can learn more *compact* representations since it doesn't need to learn separate features for each flip mode.

Predicting the flip mode is easy,[5] and as expected the RBMs achieved 0% error. On the other hand, the disBM appearance units only achieved a random-guessing performance (50.8% accuracy), suggesting that appearance and flip mode were disentangled.

6.2. Reasoning about factors of variation

A good generative model that can disentangle factors of variation should be able to traverse the manifold of one factor while fixing the states of the others. For the case of face images, the model should be able to generate examples with different pose or expression while fixing the identity. It should also be able to interpolate within a sub-manifold (e.g. across pose) and transfer the pose or expression of one person to others. Bengio et al. (2013) showed that linear interpolation across deep representations can traverse closer to the image manifold compared to shallow representations such as pixels or single-layer models. We would like our model to have these properties with respect to *each* factor of variation separately.

To verify that our model has these properties, we constructed a 2-layer deep belief network (DBN), where the first layer is a Gaussian RBM with tiled overlapping receptive fields similar to those used by Ranzato et al. (2011) and the second layer is our proposed disBM. For TFD, our model has identity-related hidden units $\mathbf{h}$ and expression-related hidden units $\mathbf{m}$. For Multi-PIE, our model has identity-related units $\mathbf{h}$ and pose-related units which we will also denote $\mathbf{m}$. For some control experiments we also use label units $\mathbf{e}$, corresponding to one of seven emotion labels in TFD and one of 15 pose labels in Multi-PIE.

We first examined how well the disBM traverses the pose or expression manifolds while fixing identity. Given an input image $\mathbf{v}$ we perform posterior inference to compute $\mathbf{h}$ and $\mathbf{m}$. Then we fixed the pose or emotion label units $\mathbf{e}$ to the target and performed Gibbs sampling between $\mathbf{v}$ and $\mathbf{m}$. Example results are shown in Figure 5(a) and 5(b). Each row shows input image and its generated samples after traversing to the specific target emotion or pose. The identity of the input face image is well preserved across the rows while expressing the correct emotion or pose.

We also performed experiments on pose and expression transfer. The task is to transfer the pose or expression of one image onto the person in a second image. Equivalently, the identity of the second image is transferred to the first image. To do this, we infer $\mathbf{h}$ and $\mathbf{m}$ for both images. Using the pose or expression units $\mathbf{m}$ from the first and identity units $\mathbf{h}$ from the second image, we compute the expect input $\mathbf{v}|\mathbf{h}, \mathbf{m}$. We visualize the samples in Fig-

[4] We annotated two or three fiducial points (e.g., the eyes, nose, and mouth corners) and computed the 2-D similarity transform that best fits them to the predefined anchor locations, which are different for each pose. Then, we warped the image accordingly, and cropped the major facial region with a fixed 4:3 rectangular box. We resized the cropped grayscaled images into 48×48.

[5] One solution is to simply use the ratio between the number of pixels of 0 and 1 in each digit image.

(a) Expression manifold traversal on TFD

(b) Pose manifold traversal on Multi-PIE

Figure 5. Visualization of (a) expression and (b) pose manifold traversal. Each row shows samples of varying expressions or pose with same identity as in input (leftmost).

ure 6(a) and 6(b).

6.3. Discriminative performance

To measure the usefulness of our features and the degree of disentangling, we apply our model to emotion recognition, pose estimation and face verification on TFD and Multi-PIE. For experiments on TFD, we built a 2-layer model whose first layer is constructed with convolutional features extracted using the filters trained with OMP-1 followed by 4×4 max pooling (Coates & Ng, 2011). We used the same model in Section 6.2 for the tasks on Multi-PIE.

We carried out control experiments of our proposed training strategies and provide summary results in Table 2 and 3. We report the performance of pose estimation and face verification for Multi-PIE, and emotion recognition and face verification for TFD. For pose estimation and emotion recognition, we trained a linear SVM and reported the percent accuracy. For face verification, we used the cosine similarity as a score for the image pair and report the AUROC. Both numbers are averaged over 5 folds.

We observed that the naive training without any regularization gets mediocre performance on both datasets. By adding pose or emotion labels, we see improvement in pose estimation and emotion recognition as expected, but also

(a) Expr. transfer. (b) Pose transfer.

Figure 6. Identity units from left column are transferred to (a) expression units and (b) pose units from middle column. Reconstructions shown in right columns.

Table 4. Performance comparison of discriminative tasks on Multi-PIE. RBM stands for the second layer RBM features trained on the first layer RBM features.

MODEL	POSE ESTIMATION	FACE VERIFICATION
RBM	93.06 ± 0.33	0.615 ± 0.002
DISBM	$\mathbf{98.20} \pm 0.12$	$\mathbf{0.975} \pm 0.002$

slightly better verification performance on both datasets. In addition, we observed a modest degree of disentangling (e.g., ID units performed poorly on pose estimation). The clamping method for ID units between corresponding image pairs showed substantially improved face verification results on both datasets. Combined with labels connected to the pose or expression units, the pose estimation and emotion recognition performance were improved. Finally, the best performance is achieved using manifold-based regularization, showing not only better absolute performance but also better disentangling. For example, while the expression units showed the best results for emotion recognition, the ID units were least informative for emotion recognition and vice versa. This suggests that good disentangling is not only useful from a generative perspective but also helpful for learning discriminative features.

We provide a performance comparison to the baseline and other existing models. Table 4 shows a comparison to a standard (second layer) RBM baseline using the same first layer features as our disBM on Multi-PIE. We note that the face verification on Multi-PIE is challenging due to the extreme pose variations. However, our disentangled ID features surpass this baseline by a wide margin. In Table 5, we compare the performance of our model to other existing works on TFD. The disBM features trained with manifold objectives achieved state-of-the-art performance in emotion recognition and face verification on TFD.

To highlight the benefit of higher-order interactions, we

Table 2. Control experiments of our method on Multi-PIE, with naive generative training, clamping identity-related units (ID), using labels for pose-related units (Pose) and using the manifold-based regularization on both groups of units.

MODEL	POSE UNITS FOR POSE EST.	POSE UNITS FOR VERIFICATION	ID UNITS FOR POSE EST.	ID UNITS FOR VERIFICATION
NAIVE	96.60 ± 0.23	0.583 ± 0.004	95.79 ± 0.37	0.640 ± 0.005
LABELS (POSE)	$\mathbf{98.07} \pm 0.12$	0.485 ± 0.005	86.55 ± 0.23	0.656 ± 0.004
CLAMP (ID)	97.18 ± 0.15	0.509 ± 0.005	57.37 ± 0.45	0.922 ± 0.003
LABELS (POSE) + CLAMP (ID)	97.68 ± 0.17	0.504 ± 0.006	49.08 ± 0.50	0.934 ± 0.002
MANIFOLD (BOTH)	$\mathbf{98.20} \pm 0.12$	0.469 ± 0.005	8.68 ± 0.38	$\mathbf{0.975} \pm 0.002$

Table 3. Control experiments of our method on TFD, with naive generative training, clamping identity-related units (ID), using labels for expression-related units (Expr) and using the manifold-based regularization on both groups of units.

MODEL	EXPR. UNITS FOR EMOTION REC.	EXPR. UNITS FOR VERIFICATION	ID UNITS FOR EMOTION REC.	ID UNITS FOR VERIFICATION
NAIVE	79.50 ± 2.17	0.835 ± 0.018	79.81 ± 1.94	0.878 ± 0.012
LABELS (EXPR)	83.55 ± 1.63	0.829 ± 0.021	78.26 ± 2.58	0.917 ± 0.006
CLAMP (ID)	81.30 ± 1.47	0.803 ± 0.013	59.47 ± 2.17	$\mathbf{0.978} \pm 0.025$
LABELS (EXPR) + CLAMP (ID)	82.97 ± 1.85	0.799 ± 0.013	59.55 ± 3.04	$\mathbf{0.978} \pm 0.024$
MANIFOLD (BOTH)	$\mathbf{85.43} \pm 2.54$	0.513 ± 0.011	43.27 ± 7.45	0.951 ± 0.025

Table 5. Performance comparison of discriminative tasks on TFD. RBM stands for the second layer RBM features trained on the first layer OMP features.

MODEL	EMOTION REC.	FACE VERIFICATION
RBM	81.84 ± 0.86	0.889 ± 0.012
DISBM	$\mathbf{85.43} \pm 2.54$	$\mathbf{0.951} \pm 0.025$
RIFAI ET AL. (2012)	$\mathbf{85.00} \pm 0.47$	–
RANZATO ET AL. (2007)	82.4	–
SUSSKIND ET AL. (2011)	–	**0.951**

Table 6. Comparison of face verification AUC (top) and pose estimation % accuracy (bottom) between 2-way and (2+3)-way disBM with increasingly many factors of variation (e.g., pose, jittering, illumination) on Multi-PIE.

MODEL	2-WAY	(2+3)-WAY
POSE	0.971 ± 0.002	$\mathbf{0.975} \pm 0.002$
POSE + JITTER	0.871 ± 0.005	$\mathbf{0.903} \pm 0.006$
POSE + JITTER + ILLUMINATION	0.773 ± 0.004	$\mathbf{0.822} \pm 0.003$
POSE	97.73 ± 0.20	$\mathbf{98.20} \pm 0.12$
POSE + JITTER	82.58 ± 0.53	$\mathbf{83.68} \pm 0.69$
POSE + JITTER + ILLUMINATION	76.42 ± 1.09	$\mathbf{80.34} \pm 1.29$

performed additional control experiments on Multi-PIE with more factors of variation, including pose, illumination and jittering. We evaluated the performance of the disBM and its 2-way counterpart by setting the higher-order weights to 0, where both are trained using the manifold objective. The summary results in face verification and pose estimation are given in Table 6. When the data have few modes of variation, we found that the 2-way model still shows good pose estimation and face verification performance. However, the higher-order interactions provide increasing benefit with the growth in modes of variation, i.e., joint configurations of pose, lighting or other factors. Such a benefit can be verified in the pose transfer task as well. In Figure 8, we visualize the pose transfer results of 2-way and (2+3)-way disBM models. The (2+3)-way model (fourth column) predicts the pose with given identity well, whereas the 2-way model (third column) produces significantly worse qualitative results, showing overlapping face artifacts and ambiguous identity.

6.4. Invariance and sensitivity analysis

We computed a similarity matrix by randomly selecting 10 identities (that had at least 7 distinct expressions) at a time, computing the cosine similarity for all pairs across all IDs and expressions. Then we averaged this feature similarity matrix over 100 trials. In Figure 7, we show average cosine similarity of several features across expression and identity variation. In ID-major order, the similarity matrix consists of 7×7-sized blocks; for each pair of IDs we compute similarity for all pairs among 7 different emotions. In Expr-major order, the similarity matrix consists of 10×10-sized blocks; for each pair of emotions we compute similarity for all pairs among 10 different IDs.

The ID features show a clear block-diagonal structure in ID-major order, indicating that they maintain similarity across changes in emotion but not across identity. In Expr-major order, our Expr features show similar structure, although there are apparent off-diagonal similarities for (anger, disgust) and (afraid, surprised) emotion labels. This makes sense because those emotions often have similar facial expressions. For the RBM features we see only a faint block diagonal and a strong single band diagonal

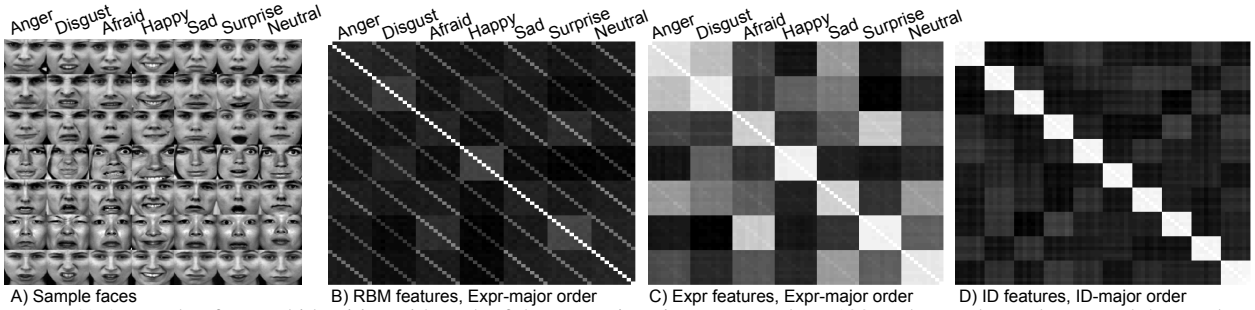

Figure 7. A) A sample of several identities with each of the 7 emotions in TFD. We drew 100 such samples and averaged the results. B) Similarity matrix using RBM features. C) Using our expression-related features (Expr). D) Using our identity-related features (ID).

Figure 8. Comparison of pose transfer results between 2-way and (2+3)-way disBM models on Multi-PIE. The task is pose transfer from faces in the second column onto the face in the first column.

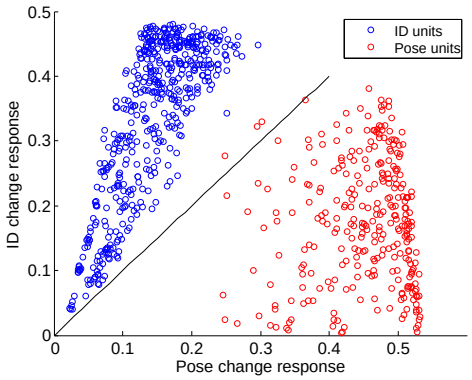

Figure 9. A scatter plot of average sensitivity of ID units (blue) and pose units (red) on Multi-PIE. The black line through the origin has slope 1, and approximately separates ID unit responses from pose unit responses.

corresponding to same-ID, same-expression pairs.

To see whether our disBM features can be both invariant and sensitive to changes in different factors of variation, we generated test set image pairs (1) with the same identity, but different pose, and (2) with different identity, but the same pose. Then we measured the average absolute difference in activation within pose units and within ID units. For every unit k and image pair $(\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$, we compute the average $|h_k^{(1)} - h_k^{(2)}|$. Figure 9 shows that ID units are more sensitive to change in ID than to pose, and pose units are likewise more sensitive to pose change than ID change.

7. Conclusion

We introduced a new method of learning deep representations via disentangling factors of variation. We evaluated several strategies for training higher-order Boltzmann machines to model interacting manifolds such as pose, expression and identity in face images. We demonstrated that our model learns disentangled representations, achieving strong performance in generative and discriminative tasks.

Acknowledgments

This work was supported in part by ONR N00014-13-1-0762, NSF GRFP under Grant No. DGE 1256260, and the Google Faculty Research Award.

References

Argyriou, A., Evgeniou, T., and Pontil, M. Multi-task feature learning. In *NIPS*, 2007.

Bengio, Y. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. Better mixing via deep representations. In *ICML*, 2013.

Caruana, R. Multitask learning. *Machine Learning*, 28(1): 41–75, 1997.

Coates, A. and Ng, A. Y. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, 2011.

Courville, A., Bergstra, J., and Bengio, Y. A spike and slab restricted Boltzmann machine. In *AISTATS*, 2011.

Desjardins, G., Courville, A., and Bengio, Y. Disentangling factors of variation via generative entangling. *arXiv:1210.5474*, 2012.

Goodfellow, I., Mirza, M., Courville, A., and Bengio, Y. Multi-prediction deep Boltzmann machines. In *NIPS*, 2013.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. Multi-PIE. *Image and Vision Computing*, 28(5), 2010.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8): 1771–1800, 2002.

Hinton, G. E. A practical guide to training restricted boltzmann machines. Technical report, 2010.

Hinton, G. E. and Salakhutdinov, R. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, 2006.

Huang, G. B., Lee, H., and Learned-Miller, E. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, 2012a.

Huang, G. B., Mattar, M., Lee, H., and Learned-Miller, E. Learning to align from scratch. In *NIPS*. 2012b.

Kavukcuoglu, K., Ranzato, M., Fergus, R., and LeCun, Y. Learning invariant features through topographic filter maps. In *CVPR*, 2009.

Larochelle, H. and Bengio, Y. Classification using discriminative restricted Boltzmann machines. In *ICML*, 2008.

Le, Q. V., Zou, W. Y., Yeung, S. Y., and Ng, A. Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011.

LeCun, Y. and Cortes, C. The MNIST database of handwritten digits, 1998.

Lee, H., Ekanadham, C., and Ng, A. Y. Sparse deep belief net model for visual area V2. In *NIPS*. 2008.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, 54(10):95–103, 2011.

Lowe, D. G. Object recognition from local scale-invariant features. In *CVPR*, 1999.

Memisevic, R. and Hinton, G. E. Learning to represent spatial transformations with factored higher-order Boltzmann machines. *Neural Computation*, 22(6):1473–1492, 2010.

Ranzato, M., Huang, F. J., Boureau, Y. L., and LeCun, Y. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR*, 2007.

Ranzato, M., Susskind, J., Mnih, V., and Hinton, G. E. On deep generative models with applications to recognition. In *CVPR*, 2011.

Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML*, 2011.

Rifai, S., Bengio, Y., Courville, A., Vincent, P., and Mirza, M. Disentangling factors of variation for facial expression recognition. In *ECCV*, 2012.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000.

Salakhutdinov, R. and Hinton, G. E. Deep Boltzmann machines. In *AISTATS*, 2009.

Sohn, K. and Lee, H. Learning invariant representations with local transformations. In *ICML*, 2012.

Sohn, K., Zhou, G., Lee, C., and Lee, H. Learning and selecting features jointly with point-wise gated Boltzmann machines. In *ICML*, 2013.

Stoyanov, V., Ropson, A., and Eisner, J. Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *AISTATS*, 2011.

Susskind, J., Anderson, A., and Hinton, G. E. The Toronto Face Database. Technical report, University of Toronto, 2010.

Susskind, J., Memisevic, R., Hinton, G. E., and Pollefeys, M. Modeling the joint density of two images under a variety of transformations. In *CVPR*, 2011.

Tang, Y., Salakhutdinov, R., and Hinton, G. E. Tensor analyzers. In *ICML*, 2013.

Tenenbaum, J. B. and Freeman, W. T. Separating style and content with bilinear models. *Neural Computation*, 12 (6):1247–1283, 2000.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Tieleman, T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML*, 2008.