# Facial Attributes Classification using Multi-Task Representation Learning

Max Ehrlich[*], Timothy J. Shields[*], Timur Almaev[†], and Mohamed R. Amer
SRI International
Princeton, NJ 08540
FirstName.LastName@sri.com

## Abstract

*This paper presents a new approach for facial attribute classification using a multi-task learning approach. Unlike other approaches that uses hand engineered features, our model learns a shared feature representation that is well-suited for multiple attribute classification. Learning a joint feature representation enables interaction between different tasks. For learning this shared feature representation we use a Restricted Boltzmann Machine (RBM) based model, enhanced with a factored multi-task component to become Multi-Task Restricted Boltzmann Machine (MT-RBM). Our approach operates directly on faces and facial landmark points to learn a joint feature representation over all the available attributes. We use an iterative learning approach consisting of a bottom-up/top-down pass to learn the shared representation of our multi-task model and at inference we use a bottom-up pass to predict the different tasks. Our approach is not restricted to any type of attributes, however, for this paper we focus only on facial attributes. We evaluate our approach on three publicly available datasets, the Celebrity Faces (CelebA), the Multi-task Facial Landmarks (MTFL), and the ChaLearn challenge dataset. We show superior classification performance improvement over the state-of-the-art.*

## 1. Introduction

Attribute prediction is an important topic in the computer vision field and is applied in different fields such as entertainment, advertising, and security. It is a challenging problem because faces can vary dramatically from one person to the other and can be viewed under a variety of different poses, occlusions, and lighting conditions. Attributes have been used for object classification [12], part-based recognition [4], comparison [32], scene understanding [39], face identification [35] and verification [21]. The focus of this work is facial attributes prediction.

Recent work has been successful at predicting attributes [25] using Convolution Neural Networks. We propose a new model that learns a shared feature representation using multi-task learning. We use Restricted Boltzmann Machines (RBMs) [15] as our building block. We formulate this problem as hybrid model that enhances the RBM model with a multi-task component based on the work of [22] by extend their formulation to account for multiple tasks resulting in Multi-Task Restricted Boltzmann Machines (MT-RBMs). We use an iterative learning approach consisting of a bottom-up/top-down passes of contrastive divergence [14] to learn the shared representation of our model and at inference we use a bottom-up pass to predict the different tasks. Our approach operates directly on faces and facial landmark points and learns a joint feature representation over all attributes. We use an off the shelf face detector [17] and landmark point detector [43] as inputs to our model. This work leads to a superior classification performance as well as efficient representation shared between the different tasks. Figure 1 shows an block diagram of our approach. Our model is trained jointly on normalized faces and facial landmark points which are treated as multimodal inputs.

We evaluate our approach on three publicly available datasets, the Celebrity Faces (CelebA) [35], the Multi-Task Facial Landmarks (MTFL) [43], and the ChaLearn challenge dataset. We show superior classification performance improvement over the state-of-the-art with reduced number of model parameters.

*Our contributions:*

- New multi-task model for facial attributes detection.
- Evaluations on three multi-task public datasets.

*Paper organization:* In sec. 2 we discuss prior work. In sec. 3 we give a brief background of similar models that motivate our approach, followed by a description of our model. In sec. 4 we describe the inference and learning algorithms. In sec. 5 we show quantitative results of our approach, followed by the conclusion in sec. 6.

[*] Both authors equally contributed to this work.
[†] Author contributed to this work during his internship at SRI International as a part of the final year of his PhD at the University of Nottingham.

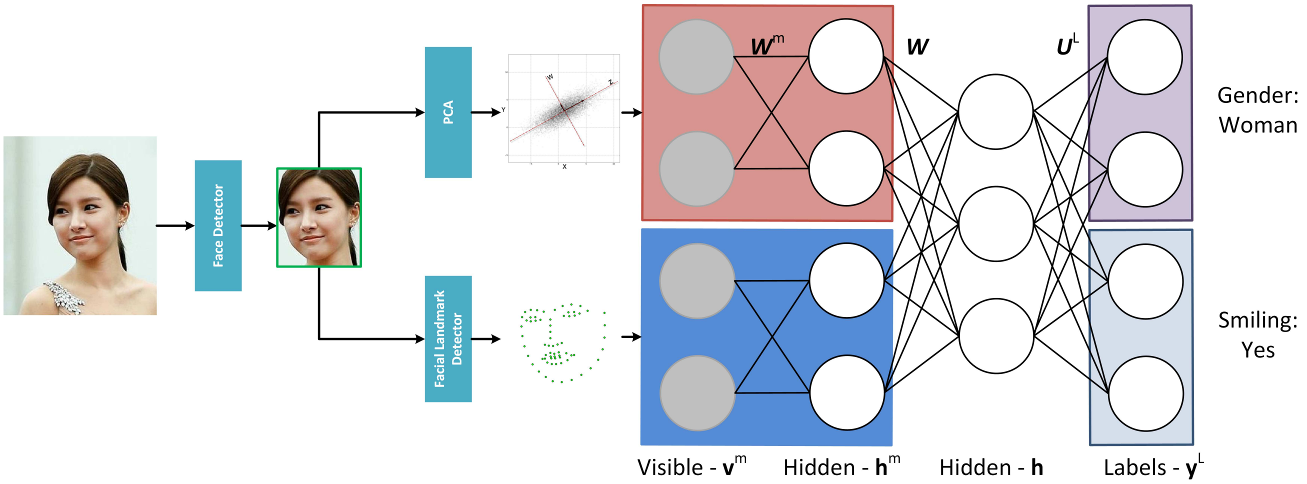


Figure 1: Our approach starts with landmark localization on detected faces and PCA for dimensionality reduction and noise elimination. Obtained landmark points and the PCA components are then fed into our MT-RBM model as two individual modalities of the data.

## 2. Prior work

We first review literature on Facial Attribute Classification; second we review Multi-Task Learning; Finally, we review Representation Learning approaches.

**Facial Attribute Classification:** There are two main directions to facial attribute classification, either local or global methods. Local methods focus on extracting features from landmarks to train a classifier for the different attributes [4, 5, 8, 21, 26]. The problem with these methods is that if the landmark detector fails, due to occlusion or lighting noise, then their method would fail. Global methods focused on processing the full face image to extract a feature representation that is not reliant on the landmark points [25, 34–36, 41, 43]. These methods were able to outperform the local methods with a significant margin. Our work follows the work of [25, 43], in addition, we also benefit from the local methods by processing both the image and the landmark points. [25] used Deep Neural Networks to address the multi-task problem and applied it to facial landmark detection and attribute classification.

**Multi-Task Learning:** Multi-task learning is a natural approach for problems that require simultaneous solutions of several related problems [6]. Multi-task learning approaches can be grouped into two main sets. The first set focuses on regularizing the parameter space. The main assumption is that there is an optimal shared parameter space for all tasks. These approaches regularize the parameter space by using a specific loss [11], methods that manually define relationships [10], or more automatic ways that estimate the latent structure of relationships between tasks [9, 19, 27, 28, 44]. The second set focuses on correlating relevant features jointly [3, 18, 30, 40]. Other work focused on the schedule of which tasks should be learned [29]. Multi-task learning achieved good results on vision problems such as: person re-identification [33], multiple attribute recognition [7], and tracking [42]. Deep Neural Networks (DNNs) were used for multi-task learning and were applied successfully to facial landmark detection [43], object localization and segmentation [38], and attribute prediction [2]. Other work used multi-task autoencoders [45] for object recognition in a generalized domain [13], where the tasks are different domains.

**Representation Learning:** Rather than using hand-crafted features suited for a specific problem [4, 5, 12], by using HOG features, or mid-level features, deep learning solved this problem by enabling automatically learned features. It has been successfully applied to attribute classification problems [16]. The main two directions are Convolutional Neural Networks (CNNs) [23] and Restricted Boltzmann Machines (RBMs) [15]. CNNs were applied to applications on facial attributes, landmark detection, verification and identification [8, 36, 41, 43]. RBMs have not been used as extensively. RBMs form the building blocks of energy-based deep networks [15]. RBMs are trained using the Contrastive Divergence (CD) algorithm [14]. CD demonstrated the ability of deep networks to capture feature distributions efficiently and learn complex representations. RBMs can be stacked together to form deeper networks known as Deep Boltzmann Machines (DBMs). Discriminatively trained RBMs are a natural extension of RBMs which have an additional discriminative term for classification [22].

## 3. Model

Rather than immediately defining our Multi-Task Multimodal RBM (MTM-RBM) model, we discuss a sequence of models, gradually increasing in complexity. We start with the basic RBM model (sec. 3.1), move to a discriminative D-RBM (sec. 3.2), which we then extend to the multi-task model MT-RBM (sec. 3.3), and then finally introduce the multimodal MTM-RBM (sec. 3.4).

### 3.1. Restricted Boltzmann Machines

RBMs [31] shown in Figure 2(a), define a probability distribution $p_{\mathrm{R}}$ as a Gibbs distribution (1), where $\mathbf{v}$ is a vector of visible nodes, $\mathbf{h}$ is a vector of hidden nodes, $E_{\mathrm{R}}$ is the energy function, $Z$ is the partition function, $\boldsymbol{\theta}$ are the model parameters. $\mathbf{a}$ and $\mathbf{b}$ are the biases for $\mathbf{v}$ and $\mathbf{h}$ respectively, and $W$ is the weight matrix. The RBM is fully connected between layers, with no lateral connections. This architecture implies that $\mathbf{v}$ and $\mathbf{h}$ are factorial given one of the two vectors. This allows for the exact computation of $p_{\mathrm{R}}(\mathbf{v}|\mathbf{h})$ and $p_{\mathrm{R}}(\mathbf{h}|\mathbf{v})$.

$$\begin{aligned} p_{\mathrm{R}}(\mathbf{h},\mathbf{v}) &= \frac{\exp[-E_{\mathrm{R}}(\mathbf{h},\mathbf{v})]}{Z(\boldsymbol{\theta})}, \\ Z(\boldsymbol{\theta}) &= \textstyle\sum_{\mathbf{h},\mathbf{v}} \exp[-E_{\mathrm{R}}(\mathbf{h},\mathbf{v})], \\ \boldsymbol{\theta} &= \begin{bmatrix} \{\mathbf{a},\mathbf{b}\} & \text{-bias,} \\ \{W\} & \text{-fully connected} \end{bmatrix} \end{aligned} \tag{1}$$

In case of binary valued data $v_i$ is defined as a logistic function. In case of real valued data, $v_i$ is defined as a multivariate Gaussian distribution with a unit covariance. A binary valued hidden layer $h_j$ is defined as a logistic function such that the hidden layer is sparse [37]. The probability distributions over $v$ and $h$ are defined as in (2).

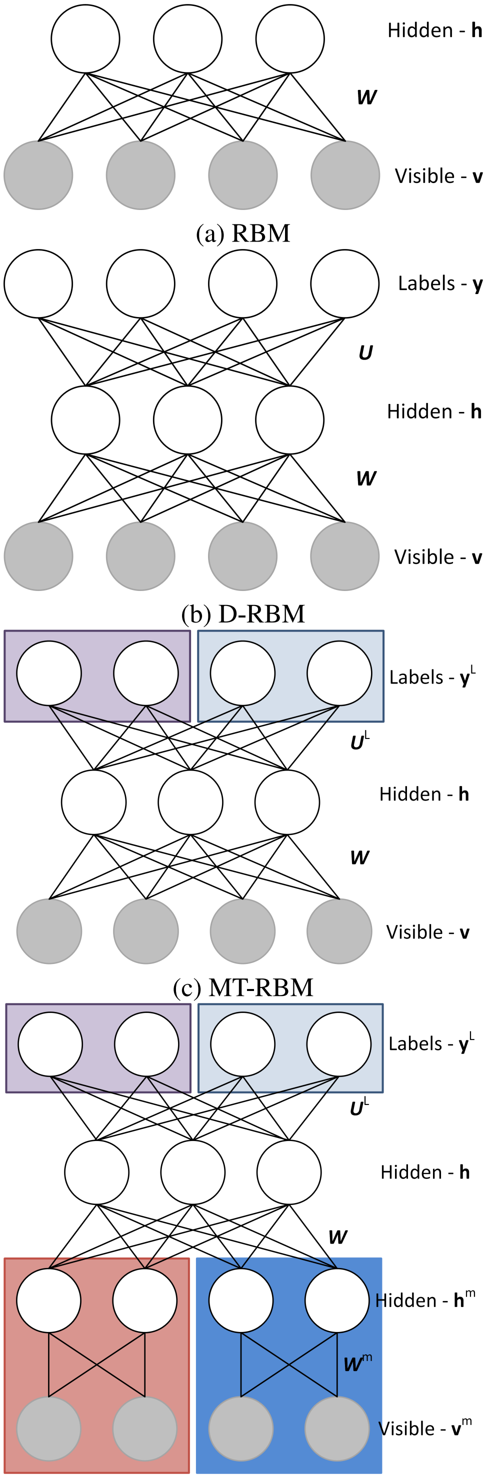
$$\begin{aligned} p_{\mathrm{R}}(v_i = 1|\mathbf{h}) &= \sigma(a_i + \textstyle\sum_j h_j w_{ij}), && \text{Binary,} \\ p_{\mathrm{R}}(v_i|\mathbf{h}) &= \mathcal{N}(a_i + \textstyle\sum_j h_j w_{ij}, 1), && \text{Real,} \\ p_{\mathrm{R}}(h_j = 1|\mathbf{v}) &= \sigma(b_j + \textstyle\sum_i v_i w_{ij}), && \text{Binary.} \end{aligned} \tag{2}$$

The energy $E_{\mathrm{R}}$ for the real valued $\mathbf{v}$ is defined as in (3).

$$E_{\mathrm{R}}(\mathbf{h},\mathbf{v}) = \sum_i \frac{(a_i - v_i)^2}{2} - \sum_j b_j h_j - \sum_{i,j} v_i w_{ij} h_j \tag{3}$$

### 3.2. Discriminative Restricted Boltzmann Machines

DRBMs, shown in Figure 2(b), are a natural extension of RBMs which have an additional discriminative term for classification [22]. The DRBM defines a probability distri-



(a) RBM

(b) D-RBM

(c) MT-RBM

(d) MTM-RBM

Figure 2: The representation learning models described in sections 3.1, 3.2, 3.3, and 3.4: (a) RBM (b) DRBM (c) MT-RBM (d) MTM-RBM.

bution $p_{\text{DR}}$ as a Gibbs distribution (4).

$$\begin{aligned} p_{\text{DR}}(\mathbf{y},\mathbf{v},\mathbf{h}|\mathbf{v}) &= \tfrac{\exp[-E_{\text{DR}}(\mathbf{y},\mathbf{v},\mathbf{h})]}{Z(\boldsymbol{\theta})}, \\ Z(\boldsymbol{\theta}) &= \textstyle\sum_{\mathbf{y},\mathbf{v},\mathbf{h}} \exp[-E_{\text{DR}}(\mathbf{y},\mathbf{v},\mathbf{h})], \\ \boldsymbol{\theta} &= \begin{bmatrix} \{\mathbf{a},\mathbf{b},\mathbf{s}\} & \text{-bias,} \\ \{W,U\} & \text{-fully connected} \end{bmatrix} \end{aligned} \tag{4}$$

The probability distribution over the visible layer will follow the same distributions as in (2). The hidden layer $\mathbf{h}$ is defined as a function of the labels $y$ and the visible nodes $\mathbf{v}$. Also, a new probability distribution for the classifier is defined to relate the label $y$ to the hidden nodes $\mathbf{h}$ as in (5).

$$\begin{aligned} p_{\text{DR}}(v_i|\mathbf{h}) &= \mathcal{N}(a_i + \textstyle\sum_j h_j w_{ij}, 1), \\ p_{\text{DR}}(h_j = 1|y_l,\mathbf{v}) &= \sigma(b_j + u_{jl} + \textstyle\sum_i v_i w_{ij}), \\ p_{\text{DR}}(y_l|\mathbf{h}) &= \tfrac{\exp[s_l + \sum_j u_{jl} h_j]}{\sum_{l*} \exp[s_{l*} + \sum_j u_{jl*} h_j]}. \end{aligned} \tag{5}$$

The new energy $E_{\text{DR}}$ is defined similar to (6),

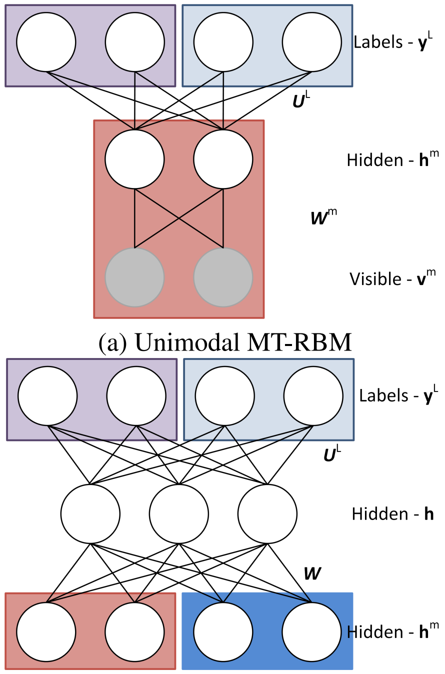
$$E_{\text{DR}}(\mathbf{y},\mathbf{v},\mathbf{h}) = \underbrace{E_{\text{R}}(\mathbf{v},\mathbf{h})}_{\text{Generative}} - \underbrace{\sum_l s_l y_l - \sum_{j,l} h_j u_{jl} y_l}_{\text{Discriminative}} \tag{6}$$

### 3.3. Multi-Task Restricted Boltzmann Machines

In the same way the RBMs can be extended to the DC-RBMs by adding a discriminative term to the model, we can extend the RBMs to be multi-task MT-RBMs shown in Figure 2(c). MT-RBMs define the probability distribution $p_{\text{MT}}$ as a Gibbs distribution (7). The MT-RBMs learn a shared representation layer for all tasks.

$$\begin{aligned} p_{\text{MT}}(\mathbf{y}^L,\mathbf{h},\mathbf{v}) &= \tfrac{\exp[-E_{\text{DC}}(\mathbf{y}^L,\mathbf{h},\mathbf{v})]}{Z(\boldsymbol{\theta})}, \\ Z(\boldsymbol{\theta}) &= \textstyle\sum_{\mathbf{y},\mathbf{h},\mathbf{v}} \exp[-E_{\text{MT}}(\mathbf{y}^L,\mathbf{h},\mathbf{v})], \\ \boldsymbol{\theta} &= \begin{bmatrix} \{\mathbf{a},\mathbf{b},\mathbf{s}^L\} & \text{-bias,} \\ \{W,U^L\} & \text{-fully connected.} \end{bmatrix} \end{aligned} \tag{7}$$

The probability distribution over the visible layer will follow the same distributions as in (5). The hidden layer $\mathbf{h}$ is defined as a function of the multi-task labels $y^L$ and the visible nodes $\mathbf{v}$. A new probability distribution for the multi-task classifier is defined to relate the multi-task labels $y^L$ to the hidden nodes $\mathbf{h}$ as shown in (8).



(a) Unimodal MT-RBM

(b) Fusion MT-RBM

Figure 3: We first classify the unimodal data by activating the corresponding hidden layers $\mathbf{h}^m$ as shown in (a), followed by classifying the multimodal data by activating the fusion layer $\mathbf{h}$ as shown in (b).

$$\begin{aligned} p_{\text{MT}}(v_i|\mathbf{h},) &= \mathcal{N}(c_i + \textstyle\sum_j h_j w_{ij}, 1), \\ p_{\text{MT}}(h_j = 1|y^L,\mathbf{v}) &= \sigma(d_j + \textstyle\sum_{l,k} y_k^l u_{jk}^l + \sum_i v_{i,t} w_{ij}), \\ p_{\text{MT}}(y_k^l|\mathbf{h}) &= \tfrac{\exp[s_k^l + \sum_j u_{jk}^l h_j]}{\sum_{k*} \exp[s_{k*}^l + \sum_j u_{jk*}^l h_j]}. \end{aligned} \tag{8}$$

The energy for the model shown in Figure 2(c), $E_{\text{MT}}$, is defined as in (9).

$$E_{\text{MT}}(\mathbf{y}^L,\mathbf{h},\mathbf{v}) = \underbrace{E_{\text{C}}(\mathbf{v},\mathbf{h})}_{\text{Generative}} - \underbrace{\sum_{k,l} s_k^l y_k^l - \sum_{j,k,l} h_j u_{jk} y_k^l}_{\text{Multi-Task}} \tag{9}$$

### 3.4. Multi-Task Multimodal Restricted Boltzmann Machines

We can naturally extend MT-RBM to MTM-RBM shown in Figure 2(d). A MTM-RBM combines a collection of unimodal MT-RBMs, one for each visible modality. The hidden representations produced by the unimodal MT-RBMs are then treated as the visible vector of a single fusion MT-RBMs. The result is a MTM-RBM model that relates multiple temporal modalities to multi-task classification labels.

MTM-RBMs define the probability distribution $p_{\text{MTM}}$ as a Gibbs distribution (10). The MTM-RBMs learn an extra representation layer for each of the modalities, which learns a modality specific representation as well as the shared layer for all the tasks.

$$
\begin{aligned}
&p_{\text{MTM}}(\mathbf{y}^L, \mathbf{h}, \mathbf{h}^{1:M}, \mathbf{v}^{1:M}) = \\
&\qquad \exp[-E_{\text{MTM}}(\mathbf{y}^L, \mathbf{h}, \mathbf{h}^{1:M}, \mathbf{v}^{1:M}]/Z(\boldsymbol{\theta}), \\
&Z(\boldsymbol{\theta}) = \textstyle\sum_{\mathbf{y},\mathbf{v},\mathbf{h}} \exp[-E_{\text{MTM}}(\mathbf{y}^L, \mathbf{h}, \mathbf{h}^{1:M}, \mathbf{v}^{1:M}), \\
&\boldsymbol{\theta} = \begin{bmatrix} \{\mathbf{a}^{1:M}, \mathbf{b}^{1:M}, \mathbf{e}, \mathbf{s}^L\} & \text{-bias,} \\ \{W^{1:M}, U^{1:M}, W, U^L\} & \text{-fully connected.} \end{bmatrix}
\end{aligned}
\tag{10}
$$

Similar to the MT-RBMs (8), the hidden layer $\mathbf{h}$ is defined as a function of the labels $y^L$ and the visible nodes $\mathbf{v}$. A new probability distribution for the classifier is defined to relate the label $y^L$ to the hidden nodes $\mathbf{h}$ as in (11).

$$
\begin{aligned}
p_{\text{MTM}}(v_i^m|\mathbf{h}^m) &= \mathcal{N}(a_i^m + \textstyle\sum_j h_j^m w_{ij}^m, 1), \\
p_{\text{MTM}}(h_j^m = 1|y^L, \mathbf{v}^m) &= \sigma(b_j^m + \textstyle\sum_{l,k} y_k^l u_{jk}^l + \sum_i v_i^m w_{ij}^m), \\
p_{\text{MTM}}(y_k^l|\mathbf{h}^m) &= \frac{\exp[s_k^l + \sum_j u_{jk}^{m,l} h_j^m]}{\sum_{l*} \exp[s_{k*}^l + \sum_j u_{jk*}^{m,l} h_j^m]}, \\
p_{\text{MTM}}(h_n = 1|y^L, \mathbf{h}^{1:M}) &= \sigma(e_n + \textstyle\sum_{l,k} y_k^l u_{nk}^l + \sum_{m,j} h_j^m w_{jn}^m), \\
p_{\text{MTM}}(y_k^l|\mathbf{h}) &= \frac{\exp[s_k^l + \sum_j u_{nk}^l h_n]}{\sum_{k*} \exp[s_{k*}^l + \sum_n u_{nk*}^l h_n]}.
\end{aligned}
\tag{11}
$$

The new energy function $E_{\text{MTM}}$ is defined in (12) similar to that of the MT-RBMs (7).

$$
\begin{aligned}
E_{\text{MTM}}(\mathbf{y}^L, \mathbf{h}, \mathbf{h}^{1:M}, \mathbf{v}^{1:M}) = \underbrace{\sum_m E_{\text{MT}}(\mathbf{y}^L, \mathbf{h}^m, \mathbf{v}^m)}_{\text{Unimodal}} \\
- \underbrace{\sum_j f_n h_n - \sum_{j,k,m} h_j^m w_{jn} h_n}_{\text{Fusion}} - \underbrace{\sum_{k,l} s_k^l y_k^l - \sum_{n,k,l} h_n u_{nk}^l y_k^l}_{\text{Multi-Task}}
\end{aligned}
\tag{12}
$$

## 4. Inference and Learning

**Inference** is described for the MTM-RBM since it is the most general case. To perform classification at time $t$ given $\mathbf{v}^{1:M}$ we use a bottom-up approach, computing the mean of each node given the activation coming from the nodes below it; that is, we compute the mean of $\mathbf{h}^m$ using $\mathbf{v}^m$ for each modality, then we compute the mean of $\mathbf{h}$, followed by computation of the mean of $\mathbf{y}^L$ for each task using $\mathbf{h}$, obtaining the classification probabilities for each task. Figure 3 illustrates our inference approach. Inference in the MT-RBM is the same as the MTM-RBM, except there is only one modality, and inference in the D-RBM is the same as the MT-RBM, except there is only one task.

**Learning** our model is done using Contrastive Divergence (CD) [14], where $\langle\cdot\rangle_{data}$ is the expectation with respect to the data and $\langle\cdot\rangle_{recon}$ is the expectation with respect to the reconstruction. The learning is done using two steps: a bottom-up pass and a top-down pass using sampling equations from (5) for D-RBM, (8) for MT-RBM, and (11) for MTM-RBM. In the bottom-up pass the reconstruction is generated by first sampling the unimodal layers $p(h_j^m = 1|\mathbf{v}^m, y_l)$ for all the hidden nodes in parallel. This is followed by sampling the fusion layer $p(h_n = 1|y_k^L, \mathbf{h}^{1:M}$. In the top-down pass the unimodal layer is generated using the activated fusion layer $p(h_j^m = 1|\mathbf{h}, y_k^L)$. This is followed by sampling the visible nodes $p(v_i^m|\mathbf{h}^m)$ for all the visible nodes in parallel. The gradient updates are described in (13). A similar learning algorithm can be applied to D-RBM and MT-RBM could be done.

$$
\begin{array}{llll}
\Delta a_i & \propto & \langle v_i^m \rangle_{data} & - \quad \langle v_i^m \rangle_{recon}, \\
\Delta b_j & \propto & \langle h_j^m \rangle_{data} & - \quad \langle h_j^m \rangle_{recon}, \\
\Delta e_n & \propto & \langle h_n \rangle_{data} & - \quad \langle h_n \rangle_{recon}, \\
\Delta s_k^l & \propto & \langle y_k^l \rangle_{data} & - \quad \langle y_k^l \rangle_{recon}, \\
\Delta w_{i,j}^m & \propto & \langle v_i^m h_j^m \rangle_{data} & - \quad \langle v_i^m h_j^m \rangle_{recon}, \\
\Delta w_{j,k} & \propto & \langle h_j^m h_n \rangle_{data} & - \quad \langle h_j^m h_n \rangle_{recon}, \\
\Delta u_{nk}^L & \propto & \langle y_k^l h_n \rangle_{data} & - \quad \langle y_k^l h_n \rangle_{recon}.
\end{array}
\tag{13}
$$

## 5. Experiments

We now describe the datasets in (sec 5.1), specify the implementation details in (sec 5.2), and present our quantitative results in (sec 5.3).

### 5.1. Datasets

Our problem is very particular in that we focus on multi-task learning for facial landmark detection. We found three publicly available datasets that vary in size to evaluate our approach, the Celebrity Faces Attributes dataset (Celeb A) [35], Multi-Task Facial Landmark (MTFL) dataset [43], and the ChaLearn Challenge dataset [1]. In the following subsections we describe each of the datasets.

**Celebrity Faces Attributes (Celeb A) [35]:** We use the CelebFaces dataset which contains 202,599 face images of 10,177 identities (celebrities) collected from the Internet and annotated with 40 attributes and 5 facial landmark points. The dataset is split into 162,770 instances for training and 19,962 instances for testing. For this dataset we compare our results against the state-of-the-art [20,24,25,41]. The CelebA dataset contains annotations of the following 40 binary attributes: {5 o'Clock Shadow, Arched Eyebrows, Attractive, Bags Under Eyes, Bald, Bangs, Big Lips, Big Nose, Black Hair, Blond Hair, Blurry, Brown Hair, Bushy Eyebrows, Chubby, Double Chin, Eyeglasses, Goatee, Gray Hair, Heavy Makeup, High Cheekbones, Male, Mouth Slightly Open, Mustache, Narrow Eyes, No Beard, Oval Face, Pale Skin, Pointy Nose,

| Classifier | Gender (2) | Smile (2) | Glasses (2) | Pose (5) |
|---|---|---|---|---|
| SVM (LMP) | 58.8 | 62.4 | 91.9 | 61.3 |
| SVM (PCA) | 54.5 | 39.1 | 91.9 | 60.9 |
| SVM (LMP, PCA) | 54.5 | 39.2 | 91.9 | 60.9 |
| MT-RBMs(LMP) | 59.5 | 75.2 | 91.9 | 68.4 |
| MT-RBMs(PCA) | 74.2 | 77.2 | 93.3 | 75.1 |
| MT-RBMs(LMP,PCA) | 73.0 | 79.0 | 93.1 | 75.0 |
| MTM-RBMs(LMP,PCA) | 79.0 | 79.0 | 93.1 | 75.7 |

Table 2: Average Classification Accuracy on the MTFL dataset.

| Classifier | Gender (2) | Smile (2) |
|---|---|---|
| SVM (LMP) | 54.6 | 69.5 |
| SVM (PCA) | 55.4 | 63.6 |
| SVM (LMP,PCA) | 55.4 | 63.6 |
| MT-RBM (LMP) | 62.8 | 73.8 |
| MT-RBM (PCA) | 70.3 | 78.0 |
| MT-RBM (LMP,PCA) | 69.9 | 79.0 |
| MTM-RBM (LMP,PCA) | 71.7 | 80.8 |

Table 3: Average Classification Accuracy on the validation partition of the ChaLearn dataset.

Receding Hairline, Rosy Cheeks, Sideburns, Smiling, Straight Hair, Wavy Hair, Wearing Earrings, Wearing Hat, Wearing Lipstick, Wearing Necklace, Wearing Necktie, Young$\}$.

**Multi-Task Facial Landmark (MTFL) [43]:** This dataset was collected for the purpose of facial landmark detection using attributes. However, we decided to use it for attributes classification. The dataset is annotated by 4 different attributes and annotated with 5 facial landmark points. This is a medium sized dataset compared to CelebA. It consists of 10,000 instances for training and 2,995 instances for testing. The MTFL dataset is annotated for Smile (S) $\in$ $\{$Yes, No$\}$, Gender (G) $\in$ $\{$Male, Female$\}$ and Glasses (GL) $\in$ $\{$Yes, No$\}$.

**ChaLearn Smile-Gender [1]:** The goal of this dataset is to classify images from the FotW dataset according to gender and basic expression. The data is challenging even to the human eye in uncontrolled environments, such as those present in the FotW dataset. ChaLearn is a rather small dataset in comparison with CelebA and MTFL. It consists of 6,171 instances for training, 3,087 for validation, and 11,145 for testing. The ChaLearn dataset is only annotated for Smile (S) $\in$ $\{$Yes, No$\}$ and Gender (G) $\in$ $\{$Male, Female$\}$.

## 5.2. Implementation Details

Given an image, we first run an off the shelf face detection algorithm [17]. After that we applied a landmark point (LMP) detector [43] outputting 68 landmark point coordinates in the image plane coordinate space. We use the landmark points to align the faces by averaging the eye points such that there are 6 points for each eye. Using the centroid for each eye and the centroid for the mouth we compute an affine transformation that projects the left and right eyes as well as the mouth to fixed coordinates on the image plane being $(100, 100)$, $(300, 100)$ and $(200, 300)$ respectively. We then use this transformation to warp the original images and crop them to be $400 \times 400$ in size. We then reduce the dimensionality of the obtained images using PCA to 500. There are different tasks defined for each dataset depending on the given annotations.

Note that in our MT-RBM model, the tasks are assumed conditionally independent given the hidden representation. Thus the number of parameters needed for the hidden-label edges is $H \cdot \sum_{k=1}^{L} Y_k$, where $H$ is the dimensionality of the hidden layer and $Y_k$ is the number of classes for task $k$. Contrast this to the number of parameters needed if instead the tasks are flattened as a Cartesian product, $H \cdot \prod_{k=1}^{L} Y_k$. Our factored representation of the multiple tasks uses only linearly many parameters instead of the exponentially many parameters needed for the flattened representation.

| Approach | [20] | [41] | [41] | [24] | [25] | [25] | **MT-RBM** |
|---|---|---|---|---|---|---|---|
| Attribute | | W | L | ANet | LNet | Full | **PCA** |
| 5 O.C. Shadow | 85 | 82 | 88 | 86 | 88 | **91** | 90 |
| Arched Eyebrow | 76 | 73 | 78 | 75 | 74 | **79** | 77 |
| Attractive | 78 | 77 | **81** | 79 | 77 | **81** | 76 |
| Bags Under Eye | 76 | 71 | 79 | 77 | 73 | 79 | **81** |
| Bald | 89 | 92 | 96 | 92 | 95 | **98** | **98** |
| Bangs | 88 | 89 | 92 | 94 | 92 | **95** | 88 |
| Big Lips | 64 | 61 | 67 | 63 | 66 | 68 | **69** |
| Big Nose | 74 | 70 | 75 | 74 | 75 | 78 | **81** |
| Black Hair | 70 | 74 | 85 | 77 | 84 | **88** | 76 |
| Blond Hair | 80 | 81 | 93 | 86 | 91 | **95** | 91 |
| Blurry | 81 | 77 | 86 | 83 | 80 | 84 | **95** |
| Brown Hair | 60 | 69 | 77 | 74 | 78 | 80 | **83** |
| Bushy Eyebrow | 80 | 76 | 86 | 80 | 85 | **90** | 88 |
| Chubby | 86 | 82 | 86 | 86 | 86 | 91 | **95** |
| Double Chin | 88 | 85 | 88 | 90 | 88 | 92 | **96** |
| Eyeglasses | 98 | 94 | 98 | 96 | 96 | **99** | 96 |
| Goatee | 93 | 86 | 93 | 92 | 92 | 95 | **96** |
| Gray Hair | 90 | 88 | 94 | 93 | 93 | **97** | **97** |
| Heavy Makeup | 85 | 84 | **90** | 87 | 85 | **90** | 85 |
| High Cheekbone | 84 | 80 | 86 | 85 | 84 | **87** | 83 |
| Male | 91 | 93 | 97 | 95 | 94 | **98** | 90 |
| Mouth Open | 87 | 82 | 93 | 85 | 86 | **92** | 82 |
| Mustache | 91 | 83 | 93 | 87 | 91 | 95 | **97** |
| Narrow Eyes | 82 | 79 | 84 | 83 | 77 | 81 | **86** |
| No Beard | 90 | 87 | 93 | 91 | 92 | **95** | 90 |
| Oval Face | 64 | 62 | 65 | 65 | 63 | 66 | **73** |
| Pale Skin | 83 | 84 | 91 | 89 | 87 | 91 | **96** |
| Pointy Nose | 68 | 65 | 71 | 67 | 70 | 72 | **73** |
| Recede Hair | 76 | 82 | 85 | 84 | 85 | 89 | **92** |
| Rosy Cheeks | 84 | 81 | 87 | 85 | 87 | 90 | **94** |
| Sideburns | 94 | 90 | 93 | 94 | 91 | **96** | **96** |
| Smiling | 89 | 89 | **92** | **92** | 88 | **92** | 88 |
| Straight Hair | 63 | 67 | 69 | 70 | 69 | 73 | **80** |
| Wavy Hair | 73 | 76 | 77 | 79 | 75 | **80** | 72 |
| Earring | 73 | 72 | 78 | 77 | 78 | **82** | 81 |
| Hat | 89 | 91 | 96 | 93 | 96 | **99** | 97 |
| Lipstick | 89 | 88 | 93 | 91 | 90 | **93** | 89 |
| Necklace | 68 | 67 | 67 | 70 | 68 | 71 | **87** |
| Necktie | 86 | 88 | 91 | 90 | 86 | 93 | **94** |
| Young | 80 | 77 | 84 | 81 | 83 | **87** | 81 |
| Average | 81 | 79 | 85 | 83 | 83 | **87** | **87** |

Table 1: Average Classification Accuracy on the CelebA dataset.

### 5.3. Quantitative Results

We first define baselines and variants followed by the classification accuracy results on the three datasets.

**Baselines and Variants:** We use SVMs as our baseline method and define the following variants of our approach: *MT-RBM*, which is our multi-task model presented in Section 3.3, and *MTM-RBM*, which is the multi-modal multi-task model presented in Section 3.4. We define MT-RBM (LMP) and MT-RBM (PCA) as models that use either facial landmarks or PCA components extracted from face images as input, respectively. MT-RBM (LMP,PCA) performs early fusion of the features and feeds them to the model. MTM-RBM (LMP,PCA) performs multimodal fusion by treating each feature type as a modality.

In order to tune the network parameters we performed a grid search, varying the number of hidden nodes per layer in the range of $\{10, 30, 50, 70, 100, 200, 500, 1000\}$. The best results were obtained using 500 hidden units.

**Attribute Classification:** For the CelebA dataset, Table 1 shows the results of other methods as well as our MT-RBM (PCA). We follow the same evaluation protocol provided in [4]. We compare against the state of the art methods FaceTracer [20], PANDA-W [41], which obtains the face parts by applying the state-of-the-art face detection [20] and alignment [34], PANDA-L [41], which is the same approach as PANDA-W, except that it operates on the ground truth bounding boxes. We also compared against LocalizationNet and AttributeNet [25], which use a set of CNNs bootstrapped together for face detection and attribute classification. Our approach ties with the best performing method [25] on the averaged results, outperforming it in 18 attributes and tying in 3 attributes. Note that our MT-RBM (PCA) does not use any convolutions and is relatively simple compared to that of [25]).

Tables 2 and 3 show our average classification accuracy for the MTFL and ChaLearn datasets respectively using different features and baseline combinations as well as the results from our models, since there are no other results available. We can see that the *MTM-RBM* outperforms all the other models in both cases, thereby demonstrating its effectiveness on predicting multi-task labels correctly. The test partition of the ChaLearn was not made publicly available, hence the methods presented above were evaluated on the validation data.

## 6. Conclusion and Future Work

We proposed a new Multi-Task Restricted Boltzmann Machine (MT-RBM), that models the distributions of multiple attributes and classifies them. An extensive experimental evaluation of these models on three different datasets of varying size demonstrates the superiority of our approach over the state-of-the-art for attribute classification. We were able to model global features (PCA) as well as local features (landmark points) and fuse them resulting in a new powerful representation for attribute classification. Our approach did not use any convolutions or processing of the data and was able to tie with the state-of-the-art method that uses CNNs. This improvement in classification performance is done with an efficient use of model parameters via factorization across tasks. The factorization of tasks used in our approach means the number of parameters grows only linearly with the number of tasks and classes. This is seen to be significant when contrasted with a single-task model that uses a flattened Cartesian product of tasks, where the number of parameters grows exponentially with the number of tasks. Our factorized approach makes adding additional tasks a trivial matter. For future work we plan to explore the relationships between different attributes and structured models.

## References

[1] Chalearn lap and fotw challenge and workshop at cvpr2016. In *CVPR*. 5, 6

[2] A. H. Abdulnabi, G. Wang, and J. Lu. Multi-task cnn model for attribute prediction. In *arXiv*, 2016. 2

[3] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. In *Machine Learning*, 2008. 2

[4] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013. 1, 2, 7

[5] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 2011. 2

[6] R. Caruana. Multitask learning. In *Machine Learning*, 1997. 2

[7] L. Chen, Q. Zhang, and B. Li. Predicting multiple attributes via relative multi-task learning. In *CVPR*, 2014. 2

[8] J. Chung, D. Lee, Y. . Seo, and C. D. Yoo. Deep attribute networks. In *NIPS Workshop*, 2012. 2

[9] C. Ciliberto, L. Rosasco, and S. Villa. Learning multiple visual tasks while discovering their structure. In *CVPR*, 2015. 2

[10] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. In *JMLR*, 2005. 2

[11] T. Evgeniou and M. Pontil. Regularized multi?task learning. In *KDD*, 2004. 2

[12] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2

[13] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2015. 2

[14] G. E. Hinton. Training products of experts by minimizing contrastive divergence. In *NC*, 2002. 1, 2, 5

[15] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. In *NC*, 2006. 1, 2

[16] Y. B. Ian Goodfellow and A. Courville. Deep learning. Book in preparation for MIT Press, 2016. 2

[17] B. M. J. Orozco and M. Pantic. Empirical analysis of cascade deformable models for multi-view face detection. *Image and Vision Computing*, 42:47–61, 2015. 1, 6

[18] Z. Kang and K. Grauman. Learning with whom to share in multi-task feature learning. In *ICML*, 2011. 2

[19] A. Kumar and H. D. III. Learning task grouping and overlap in multi-task learning. In *ICML*, 2012. 2

[20] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 2008. 5, 7

[21] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1, 2

[22] H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In *ICML*, 2008. 1, 2, 3

[23] Y. LeCun and Y. Bengio. word-level training of a handwritten word recognizer based on convolutional neural networks. In *ICPR*, 1994. 2

[24] J. Li and Y. . Zhang. Learning surf cascade for fast and accurate object detection. In *CVPR*, 2013. 5, 7

[25] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1, 2, 5, 7

[26] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *ICCV*, 2013. 2

[27] A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. In *ICML*, 2013. 2

[28] A. Maurer, M. Pontil, and B. Romera-Paredes. The benefit of multitask representation learning. In *ArXiv*, 2015. 2

[29] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum learning of multiple tasks. In *CVPR*, 2015. 2

[30] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil. Multilinear multitask learning. In *ICML*, 2013. 2

[31] R. Salakhutdinov and G. E. Hinton. Reducing the dimensionality of data with neural networks. In *Science*, 2006. 3

[32] F. Song, X. Tan, and S. Chen. Exploiting relationship between attributes for improved face verification. In *CVIU*, 2014. 1

[33] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV*, 2015. 2

[34] Y. . Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013. 2, 7

[35] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identication and verification. In *NIPS*, 2014. 1, 2, 5

[36] Y. Sun, X. Wang, and X. Tang. Deep learning face representationfrom predicting 10,000 classes. In *CVPR*, 2014. 2

[37] G. W. Taylor, G. E. Hinton, and S. T. Roweis. Two distributed-state models for generating high-dimensional time series. In *Journal of Machine Learning Research*, 2011. 3

[38] X. Wang, L. Zhang, L. Lin, Z. Liang, and W. Zuo. Deep joint task learning for generic object extraction. In *NIPS*, 2014. 2

[39] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1

[40] Y. Yang and T. M. Hospedales. A unified perspective on multi-domain and multi-task learning. In *ICLR*, 2015. 2

[41] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, 2014. 2, 5, 7

[42] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via structured multi-task sparse learning. In *IJCV*, 2012. 2

[43] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 1, 2, 5, 6

[44] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *NIPS*, 2011. 2

[45] F. Zhuang, D. Luo, X. Jin, H. Xiong, P. Luo, and Q. He. Representation learning via semi-supervised autoencoder for multi-task learning. In *ICML*, 2015. 2