MOON: A *Mixed Objective Optimization Network* for the Recognition of Facial Attributes

Ethan M. Rudd, Manuel Günther, and Terrance E. Boult

University of Colorado at Colorado Springs

Abstract. Multi-task vision problems can often be decomposed into separate tasks and stages, e.g., separating feature extraction and model building, or training independent models for each task. Joint optimization has been shown to improve performance, but can be difficult to apply to deep convolution neural networks (DCNN), especially with unbalanced data. This paper introduces a novel mixed objective optimization network (MOON), with a loss function which mixes errors from multiple tasks and supports domain adaptation when label frequencies differ between training and operational testing. Experiments demonstrate that not only does MOON advance the state of the art in facial attribute recognition, but it also outperforms independently trained DCNNs using the same data.

Keywords: Facial Attributes, Deep Neural Networks, Multi-Task Learing, Multi-Label Learning

1 Introduction

Given an input image or video, there are often multiple vision tasks to be accomplished, i.e., multiple objectives to be optimized. A few examples of multi-task problems include the *simultaneous* detection *and* localization of multiple objects, detection *and* tracking of objects, and the *simultaneous* computation of multiple labels or attributes associated with an image. Often, such multi-task problems appear to be readily decomposed into N independent optimization problems, each of which can be solved separately. While in some cases separately optimizing each objective is a valid approach, under certain constraints, e.g., when task feeds into each other, or there is a need to share computed features or representations, then the different task objectives must be mixed and often balanced between one another. For these sorts of computer vision problems, multi-task learning has benefited many areas, including multi-label image tagging and retrieval [1,2,3,4], tracking [5], facial landmark estimation [6,7], face verification [8], and face detection and head pose estimation [9,10,11].

This paper examines the multi-task problem of facial attribute recognition. While this problem may seem far removed from previous problems to which multi-objective learning has been successfully applied, we hypothesize that it is well suited to multi-objective optimization because a solution can be approached with similar intuition to that behind other successful multi-objective techniques.

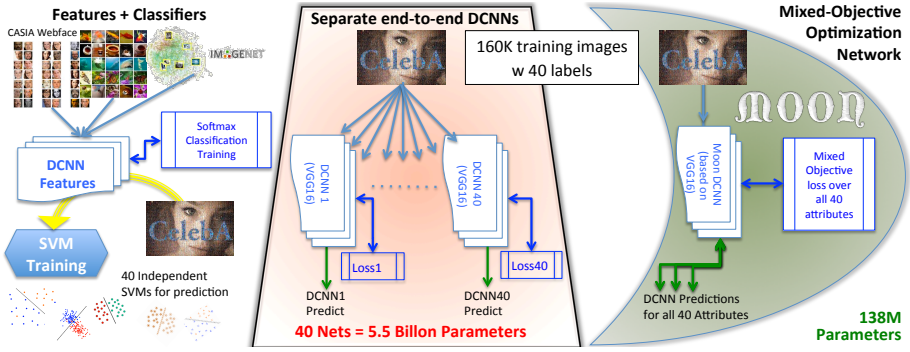


Fig. 1: Three approaches to attribute learning (and other multi-task problems). *On the left is a conceptual model of the current and recent state-of-the-art approaches, with features trained for classification problems then adapted as inputs to independent SVMs for prediction. The middle approach attacks the problem with separately trained Deep Convolution Neural Networks (DCNNs). While we demonstrate that this advances the state of the art in attribute accuracy, the cost is prohibitive for practical use. This papers asks, "How can a single network outperform the separately trained DCNNs?" On the right, is our answer, the mixed objective optimization network (MOON) architecture with a domain adapting multi-task DCNN loss. The MOON approach allows one network to efficiently learn to simultaneously output predictions for all attributes, with reduced training and storage costs, while producing better accuracy than independently trained DCNNs.*

Facial attributes have a shared, albeit latent correlation, which imposes soft constraints on the space of attributes, e.g. $p(\texttt{Male}|\texttt{Mustache}, \texttt{Bushy Eyebrows}) \approxeq 1$.

The most common approach to multi-label problems is to allow each task to independently optimize its choice of features and recognition model, which we call Features+Classifiers in Fig. 1. This was the original approach taken by Kumar et al. [12] to learn facial attribute classifiers: The AdaBoost algorithm was used to optimize feature selection separately for each facial attribute, and independent SVM classifiers were built using those customized features. The current state of the art [13] trains DCNN features and then uses SVMs for classification. Features+Classifiers type models have held the state of the art in facial attribute recognition since the first papers on the topic, and have also been widely used for other attribute work [14,15,16].

We suspect that the Features+Classifiers approach for attributes cannot exploit all correlation because the feature extraction and attribute modeling are separate phases. While correlations could be learnt in each independent classifier, or leveraged by some post-hoc fusion of several classifiers, these approaches *neglect implicit latent structure*. Furthermore, any post-hoc fusion would rely on some ad-hoc manner of selecting correlated attributes, which easily leads to non-regular, inefficient, and cumbersome classifiers as the number of attributes increases.

Assuming multiple correlated vision tasks, intuition might suggest that if the relationships between tasks are latent, the correlation between the tasks would still be maximally exploited by independently trained models, each of which train with all of the data. This suggestion indicates that separate optimization of an end-to-end deep convolution neural network for each task would improve accuracy and maximize the extraction of information in the data. We show that this approach does improve accuracy over the Features+Classifiers approach, but at significant computational costs. While these separate end-to-end DNCCs are more accurate than the state of the art, building separate DCNNs is just engineering and it does not answer if they maximally exploited the data. We address more fundamental questions: "Can a single network outperform the separately trained DCNNs, and if so, how?"

This paper demonstrates that the intuition that separately optimizing over multiple tasks/labels maximally exploits the data does not hold, at least not when using DCNNs with large but limited training data. We explore a joint optimization by scalarizing the multi-objective classification problem into a single *mixed objective optimization network* (MOON). MOON's loss function uses a mix of per-task squared errors that incorporate domain adapted weighting. How can this approach improve the overall performance? *We conjecture that the mixed objective formulation provides a form of regularization which uses implicit latent structure to constrain the space of possible models and we show that a single MOON network outperforms separately trained networks, both in terms of accuracy and speed.* In summary, the contributions of this paper include:

- A novel *mixed objective optimization network* (MOON) architecture, which learns multiple attribute labels simultaneously via a single convolutional neural network and which supports domain adaption for multi-task DCNNs by adjusting for different label frequencies between training and operation,
- A fair evaluation technique which incorporates source and target distributions into the classification metric, leading to the *balanced* CelebA (CelebAB) evaluation protocol,
- Experiments which demonstrate that the MOON architecture significantly advances state-of-the-art attribute recognition on the CelebA dataset, improving both accuracy and efficiency, and
- Evaluation of stability of the MOON architecture to fiducial perturbations and data set imbalance.

Our experiments demonstrate that optimizing over all attributes simultaneously offers a noticeable reduction in classification error compared to optimizing single attributes over the same dataset and network topology.

2 Related Work

Multi-task learning has been applied to several areas of computer vision, which rely on learning fine-grained discriminations or localizations under the constraint of a global correlating structure. In these problems, multiple target labels or objective functions must simultaneously be optimized.

A common application of multi-label learning in the vision community is image, or more generally, multi-modal tagging/retrieval [4,17]. In these problems, representations of the contents of an image across modalities (e.g., textual descriptions, voice descriptions) are jointly inferred in a representation, which is additionally derived from the raw images themselves. The resulting classifiers can then be used to generate descriptions of novel images (tagging) or to query images based on their descriptions (retrieval).

Facial model fitting and landmark estimation [18,19] is another multi-task problem, which requires a fine-grained fit due to tremendous diversity in facial features, poses, lighting conditions, expressions, and many other exogenous factors. Solutions also benefit from global information about the space of face shapes and textures under different conditions. Optimization with respect to local gradients and textures is necessary for a precise fit, while considering the relative locations of all points is important to avoid violating facial topologies.

The reverse of facial model fitting – cross-pose synthesis – is similarly well formulated under a multi-task/multi-objective approach: by simultaneously minimizing extrapolation and reconstruction errors, Yim et al. [11] were able to achieve state-of-the-art results.

Applications of facial attributes include searches based on semantically meaningful descriptions (e.g., "Caucasian female with blond hair") [12,20,21], verification systems which explain in a human-comprehensible form *why* verification succeeded or failed [22], relative relations among attributes [15], social relation/sentiment analysis [23], and demographic profiling. Facial attributes also provide information that is more or less independent of that distilled by conventional recognition algorithms, potentially allowing for the creation of more accurate and robust systems, narrowing down the search space, and increasing efficiency at match time. Finally, facial attributes are interesting due to their ability to convey meaningful identity information about a previously unseen face, e.g., not enrolled in a gallery or used to train a classifier.

The classification of facial attributes was first pioneered by Kumar et al. [22]. Their classifiers depended heavily on face alignment, with respect to a frontal template, each attribute using AdaBoost-learnt combinations of features from hand-picked facial regions (e.g., cheeks, mouth, etc.). The feature spaces were simplistic by today's standards, consisting of various normalizations and aggregations of color spaces and image gradients. Different features were learnt for each attribute, and a single RBF-SVM per attribute was independently trained for classification. Although novel, the approach was cumbersome, due to high dimensional varying length features for each attribute, leading to inefficiencies in feature extraction and classification [24].

In recent years, approaches have been developed to leverage more sophisticated feature spaces. For example, gated CNNs [25] use cross-correlation across an aligned training set to determine which areas of the face are *most relevant* to attributes. The outputs of an ensemble of CNNs, one trained for each of the relevant regions, are then joined together into a global feature vector. Final classification is performed via independent binary linear SVMs. Zhang et al. [23] use

CNNs to learn facial attributes, with the ultimate goal of using these features as part of an intermediate representation for a siamese network to infer social relations between pairs of identities within an image.

Liu et al. [13] use three CNNs – a combination of two *localization networks* (LNets), and an *attribute recognition network* (ANet) to first localize faces and then classify facial attributes in the wild. The localization network proposes locations of face images, while the attribute network is trained on face identities and attributes, and is used to extract features which are fed to independent linear SVMs for final attribute classification. This approach is the current state of the art on the CelebA dataset. In these recent works, the same deep feature space is learnt for all attributes, but is not necessarily attribute derived, and independent binary classifiers are used to perform the attribute classifications.

There has been significant prior work in visual domain adaptation [26], including more recent work for CNNs [27]. While the latter work is related to ours, it addressed the more general problem of adaptation with unlabeled data. We, however, are addressing one of the simpler forms of class imbalance adaptation within our multi-task problem via a frequency reweighting and, hence, our approach is a special case of the recent model unifying multiple domains with multi-task learning [28].

3 Approach

For multi-task problems, the high level goal is to maximize accuracy over all tasks, where each task has its own objective. In our case, the task is attribute prediction, and we seek to simultaneously maximize prediction accuracy over all attributes.

Formally, let $\mathbb{I}$ be the space of allowable images, and let M be the number of attributes. For a given sample $x \in \mathbb{I}$, let $y_i : x \to \{-1, +1\}$ be a function yielding the binary ground truth label for x, where $i \in \{1, \ldots, M\}$ is the attribute index. Let $\mathcal{H}$ be the space of allowable decision functions and $f_i(x; \theta_i) \in \mathcal{H}$ be the decision function, with parameters θ_i, learnt for the ith attribute classifier. Given a set of loss functions $L_i(f_i(x; \theta_i), y_i)$, each of which defines the cost of an error on input x with respect to attribute i, let $\mathbb{E}(f_i(x; \theta_i), y_i(x))$ be the expected value of that loss over the range of inputs $\mathbb{I}$. Then the idealized problem is to minimize the loss for each attribute, i.e.:

$$\forall i \colon f_i^* = \underset{f_i \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}(f_i(x; \theta_i), y_i(x)). \tag{1}$$

For input x and attribute i, the classification result $c_i(x)$ and its corresponding accuracy $c_i(x, y)$ are obtained by thresholding the associated network:

$$c_i(x) = \begin{cases} +1 & \text{if } f_i(x) > 0 \\ -1 & \text{otherwise,} \end{cases} \quad \text{and} \quad c_i(x, y) = \begin{cases} +1 & \text{if } y_i(x) c_i(x) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Intuitively, this appears to lead to M independent optimization problems, for which one should be able to optimize each f_i separately. Accordingly, the most common approach to attribute classification in prior work is to use independent binary classifiers in some characteristic feature space to classify each

attribute [22,13]. The original approach taken by Kumar et al. [22] used separate per-attribute AdaBoost learnt feature space representations. Recent state of the art approaches use convolutional neural networks, trained on face identification and verification datasets to arrive at an underlying feature space representation [13], and extract features by truncating the network prior to the final softmax layer. Both [22] and [13] learn M independent binary classifiers trained with a hinge-loss objective. The hinge-loss objective function is:

$$\underset{\theta_i}{\operatorname{argmin}} L_i(x, \theta_i, y_i) = max(0, 1 - y_i(x) f_i(x; \theta_i)). \tag{3}$$

When the classifier is a dot product, i.e., $f_i(x) = \theta_i^T (1, x^T)^T$, solving this objective function results in a binary *support vector machine* (SVM) – the hyperplane that separates the two binary classes of data (+1 and −1) with maximum soft-margin. Given M attributes, this approach leads to M binary classifiers, each of which outputs a decision score. A positive decision score corresponds to the predicted presence of an attribute, while a negative decision score corresponds to its absence.

In order to learn latent correlations, it is also important to use attribute data directly to derive the feature space. Although Liu et al. [13] claim that latent features of attributes are learnt by their feature space representation while optimizing over a dataset for an identification task, the extent to which this is true for attributes which have little to do with facial identity (e.g., `Smiling`) is questionable. Rather, intuition suggests the opposite – that networks trained for identification of individuals would learn to ignore such attributes. To uncover such correlations, the network used to learn the feature space should be directly trained on attribute data and the distribution of attributes in training should match the operational or testing distribution.

This leads to the problem of how to appropriately balance the dataset used to learn attribute features. A perfectly balanced dataset can be obtained by collecting separate images for each attribute, but this leads to an enormous dataset, with different identities for different attributes, effectively yielding a relatively small number of training images per attribute in proportion to the size of the dataset [22]. This approach also does not account for label correlations. Using a multi-label dataset, e.g., CelebA [13] allows us to leverage multiple labels in a mixed objective, but the distribution is highly imbalanced for many attributes (cf. Sec. 4). Unfortunately, the attribute distribution of a given target population does not always follow the dataset bias.

In a separate per-class training, balancing the number of positive and negative examples that are input to the classifier is easy, e.g., by weighting or sampling. However, balancing is nearly impossible for multi-task training. Furthermore, for many tasks, the training frequencies and the operational/test frequencies need not match. Our solution to both problems is to define a mixed objective function which includes domain adapted weights that incorporate the difference between the source and target distributions. First, we compute the source distribution S_i from the training set for each attribute i by counting the relative number of occurrences of the positive S_i^+ and negative S_i^- samples. Given a

binary target distribution, T_i^+ and T_i^-, for each attribute i we assign a probability for each class:

$$p(i|+1) = \begin{cases} 1 & \text{if } T_i^+ > S_i^+ \\ \frac{S_i^- T_i^+}{S_i^+ T_i^-} & \text{otherwise} \end{cases} \quad \text{and} \quad p(i|-1) = \begin{cases} 1 & \text{if } T_i^- > S_i^- \\ \frac{S_i^+ T_i^-}{S_i^- T_i^+} & \text{otherwise.} \end{cases} \tag{4}$$

We would like to incorporate this domain adaptation directly into a loss function, but we need a loss function which additionally mixes all attribute predictions and simultaneously infers latent correlations between attribute labels and image data. One approach would be to combine all of the objective functions for each attribute into one joint objective function, e.g.:

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{M} L_i(x, \theta_i, y_i), \tag{5}$$

where $\theta = \{\theta_1, \ldots, \theta_M\}$ are the network weights, which for legibility reasons are left out in the following equations. We can then solve that optimization problem via backpropagation using raw attribute images and labels as a training set. While we could use many potential loss functions, for MOON, we optimize a weighted mixed task squared error:

$$L(x, y) = \sum_{i=1}^{M} p(i|y_i(x)) \; ||f_i(x) - y_i(x)||^2, \tag{6}$$

where $f_i(x)$ is the network output for attribute i, and for which the output dimensionality is the number of attributes M. Across an N element training set X with labels Y this yields:

$$L(X, Y) = \sum_{j=1}^{N} \sum_{i=1}^{M} p(i|Y_{ji}) \; ||f_i(X_j) - Y_{ji}||^2. \tag{7}$$

Replacing the standard loss layer of a deep convolutional neural network with a layer implementing Eq. (7) results in the *mixed objective optimization network* (MOON) architecture. MOON incorporates attribute correlations and can adapt the bias of the training dataset to a target distribution.

4 Experiments

4.1 Dataset

For comparison with other attribute benchmarks, we conducted our experiments on the CelebA dataset [13]. The dataset consists of batches of 20 images from approximately 10K celebrities, resulting in a total of more than 200K images. The first 8K identities (160K images) are used for training, and the remaining 2K identities are used in the validation and test sets; 1K for validation, 1K for test. Each image is annotated with 5 keypoints (both eyes, the mouth corners and the nose tip), as well as binary labels of 40 attributes. These attributes are shown in Fig. 2, which also shows the relative number of images in which the attribute is hand-labeled as present (blue) or absent (tan), respectively. As one can observe, for many of the attributes, there is a strong bias for either of the

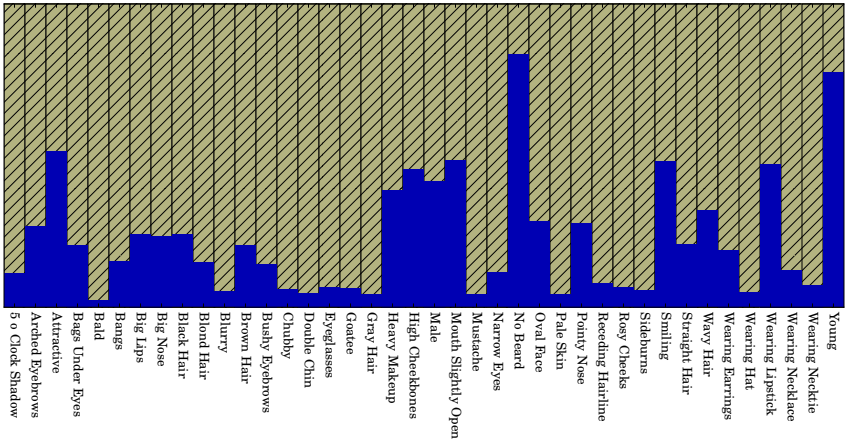


Fig. 2: CelebA Dataset Bias. *This figure shows the distribution of the attribute labels throughout the CelebA dataset: presence (blue) or absence (tan).*

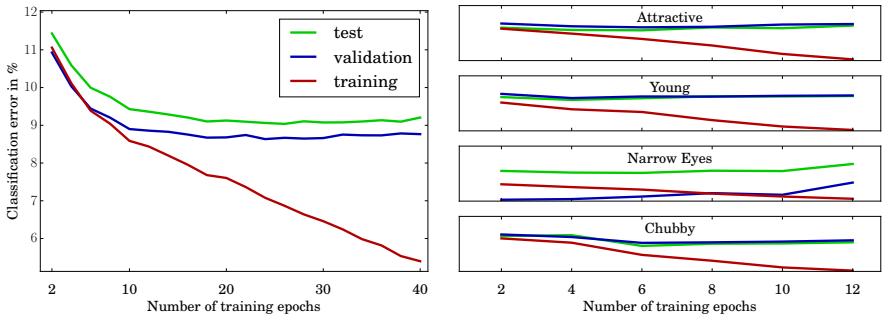
two classes. This is especially the case for certain attributes, e.g., few images are labeled as `Bald` or `Wearing Hat`, while the majority of the facial images are labeled as `Young`.

The CelebA dataset provides a set of pre-cropped face images, which were aligned using hand-labeled keypoints. For our experiments we use these images, but later (cf. Sec. 5.1) we show that the trained classifier can also work with faces which are not perfectly aligned, and we introduce ideas to make our MOON network more robust to mis-alignment.

4.2 Evaluating MOON on CelebA

In order to compare with existing approaches, which do not account for dataset bias, we evaluate MOON on the CelebA dataset, setting the target distribution to the source distribution, i.e., $\forall i\ T_i \equiv S_i$.

Using the CelebA training set, we trained a deep convolutional network to predict attributes under a MOON architecture. As the basic network configuration, we adopted the 16 layer VGG network from [29], where we replaced the final loss layer with the loss in Eq. (7). We also changed the dimension of the RGB image input layer from 224×224 pixels to 178×218 pixels, the resolution of the aligned CelebA images. In opposition to [29], we do not incorporate any dataset augmentation or mirroring, but train the network purely on the aligned images. Due to memory limitations, the batch size was set to 64 images per training iteration and, hence, the training requires approximately 2500 iterations to run a full epoch on the training set. The learning rate was chosen to be 0.00001, as higher learning rates would lead the network to learn only the dataset bias. During training we update the convolution kernel weights using the backpropagation algorithm with an *RMSProp* update rule and an inverse learning rate decay policy.



(a) (b)

Fig. 3: NETWORK TRAINING ERRORS. *This figure shows the error trends during network training in terms of average classification errors as given in Eq.* (9).

We ran two types of network training, one training a separate network for each attribute, and one optimizing the combined MOON network. Separately training classifiers is the most common approach taken in the literature. By training one network for each of the attributes individually, the network can concentrate on a single attribute and can therefore ignore all parts of the image that are not required to classify that attribute. During the separate training, we presented each network with all images from the training set, and a single input to the loss layer encoded with labels that denoted the presence (+1) or the absence (−1) of the attribute. Loss was computed according to Eq. (6). As each network required several hours of training time on an NVIDIA Titan-X GPU, we chose to train each network for ≈ 2 epochs (5000 iterations). To check if 2 epochs iterations are sufficient to attain convergence to a maximum validation accuracy, we continued the network training for four attributes. We selected these attributes – `Attractive`, `Chubby`, `Narrow Eyes`, and `Young` – to have varying statistics from the dataset: While `Attractive` is relatively balanced, images with `Chubby` and `Narrow Eyes` are mostly not contained in the dataset, whereas `Young` is over-represented. The error trends (evaluated at every 2 epochs) are shown in Fig. 3(b). Although the error on the training set decreases with additional epochs, the errors on the validation and test sets start to *increase* after approximately 4 - 6 epochs, with only a little improvement over the 2 epochs network. This leads us to believe that improvements in validation accuracy beyond 2 epochs are negligible.

When training our MOON network, we use a single network with 40 outputs to learn all attributes simultaneously. The loss layer of the MOON network is set up such that – following Eq. (7) – it minimizes the average weighted Euclidean distance[1] between the network output and the 40 binary attribute values. We trained the network for 40 epochs since the validation error after 10

[1] The weights in this experiment are all equal to one, following from identical source and target distributions $S_i = T_i$.

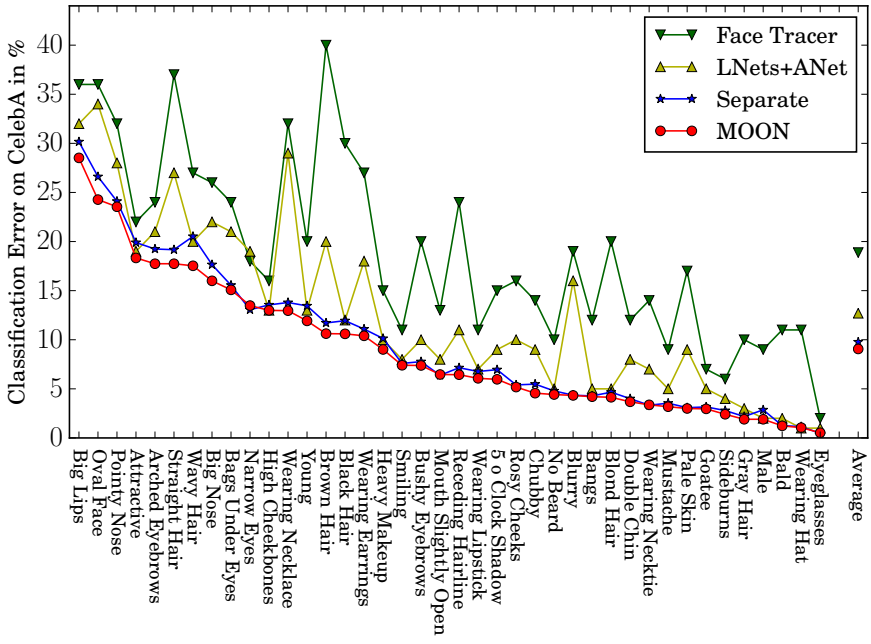


Fig. 4: Error Rates on CelebA. *This figure shows the classification errors on the test set of the CelebA dataset for several algorithms. The results of Face Tracer and LNets+ANet are taken from Liu et al. [13]. The Separate networks were evaluated after 2 epochs, while the MOON network is evaluated after 24 training epochs.*

epochs was still decreasing. The complete error trends can be seen in Fig. 3(a). Based on the minimum validation set error, we choose our final MOON network after 24 epochs. Note that this is not the optimal number of epochs in terms of minimizing test error, but it is the proper way to evaluate our classifier with respect to the CelebA protocol. While individual classifiers seem to take fewer training iterations than MOON to minimize their validation error, the total training time of the MOON network is still lower than the sum of the separate network training times. We suspect that the additional iterations required by the MOON network are needed to learn a more sophisticated latent structure than those learnt by the separate networks.

To compare with the results of Liu et al. [13], we measure the success of our training in terms of classification error,[2] i.e., the number of cases, where our classifier f predicted the incorrect label, relative to the total number of test images:

$$e_i(X, Y) = \frac{1}{N_{test}} \sum_{j=1}^{N_{test}} \big(1 - c_i(X_j, Y_j)\big). \tag{8}$$

[2] Liu et al. [13] presented their results in terms of classification success, which is simply $1 - E(X, Y)$.

The `Average` classification error is computed by taking the average of the classification errors over all (M) attributes:

$$E(X,Y) = \frac{1}{M}\sum_{i=1}^{M} e_i(X,Y). \tag{9}$$

Note that this error does not differentiate between positive and negative values. Hence, for very biased attributes, a random classifier which always predicts the dominant class would reach a low classification error, e.g., for `Bald` the random classification error would be as low as 2.24 %!

The classification errors for all the attributes are visually displayed in Fig. 4. There, we also included two results[3] from Liu et al. [13]. The Face Tracer results reflect the best non-neural-network based algorithm that has been evaluated so far on the CelebA dataset. LNets+ANet represent the state-of-the-art results on this dataset obtained by combining three different deep convolutional neural networks with support vector machines.

The average classification errors over all attributes for each classifier are: Face Tracer: 18.88 %, LNets+ANet: 12.70 %, Separate: 9.78 %, and MOON: 9.06 %. Thus, our MOON network achieves a relative reduction of 28.7 % of the error over the state of the art, and a 7.4 % reductuion over the separately trained networks. For almost all attributes, the results of our two approaches outperform the LNets+ANet state-of-the-art results, and the MOON network gives a lower error than the Separate networks trained specifically on a single attribute.

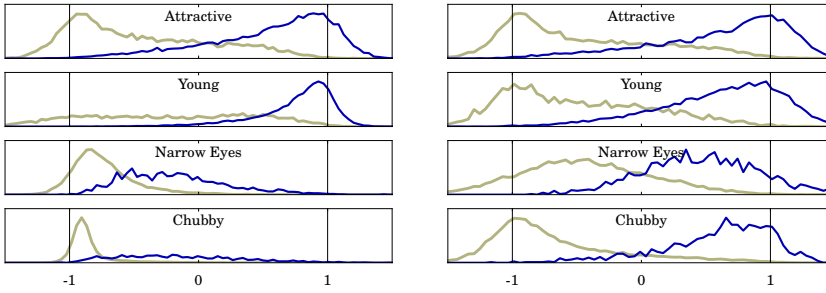
Interestingly, for several attributes that are traditionally not considered to be useful in face recognition, such as hair color (e.g. `Brown Hair`), hair style (e.g. `Straight Hair`), accessories (e.g. `Wearing Necklace`), and non face-related attributes (e.g. `Blurry`), our approach outperforms the LNets+ANet combination by an especially large margin. We suspect that this effect is due in part to the fact that in [13], the ANet network's feature space was derived from training on a *face recognition* benchmark, and later adapted to the attribute classification task, which offers little direction for inferring the hidden representations of non facial identity related attributes.

4.3 CelebAB: A Balancing Act

As demonstrated in Sec. 4.2, MOON obtains state-of-the-art classification accuracies on the CelebA dataset. However, it is unclear how meaningful these results are for target distributions with different attribute frequencies.

Since our objective is to learn the network outputs to be $+1$ or -1 corresponding to presence or absence of attributes, respectively, we plotted the score distributions of the validation set for four of the attributes. From Fig. 2 we observe a strong bias for several attributes in the CelebA dataset, which we can find in the score distribution plots of Fig. 5(a), too. Note that the positive and negative score distributions have been normalized independently, otherwise the positive scores for `Narrow Eyes` and `Chubby` would not be visible. For attributes

[3] For the two algorithms from [13] we have converted classification success into classification error, and averaged the numbers to recompute the final average.



(a) Unbalanced (b) Balanced

Fig. 5: SCORE DISTRIBUTIONS. *This figure shows the distributions of the network outputs for four different attributes, when classifying images with present (blue) and absent (tan) attributes. In (a) network outputs after training with unbalanced data are shown, while in (b) the outputs of the network after training with the balancing loss layer are presented. Positive and negative score distributions are normalized independently.*

with a balanced number of positive and negative examples, such as `Attractive`, the distributions of negative (tan) and positive (blue) scores are also balanced. On the other hand, for unbalanced attributes, such as `Young`, `Narrow Eyes` or `Chubby`, the dominant class is well distributed around its desired value, but the other class has not been learnt well. Interestingly, a comparably small bias in the training set (for `Young` there are 77 % positives and 23 % negatives) can destroy the capability of the network to learn the inferior class.

Intuitively, when having such unbalanced score distributions, one would expect that the threshold of 0 that we use for classification should be adapted. However, given that the validation and test set follow the same bias as the training set, a threshold of 0 works well for the CelebA dataset. Even more astonishingly, a wide range of thresholds around 0 will lead to approximately the same classification error and, hence, the network has learnt to balance between false positives and false negatives – including the dataset bias.

To obtain balanced score distributions, we chose to have a balanced target distribution, i.e., $T_i^+ = T_i^- = \frac{1}{2}$ for each attribute i. In our custom implementation of the loss layer in Eq. (6), we obtain the weights $p(i|y_i(x))$ via sampling. For each attribute i with target value $y_i(x) \in \{-1, +1\}$ we only backpropagate the error with the probability $p(i|y_i(x))$; otherwise we set the gradient for attribute i to 0. The more the source and target distributions differ, the more elements in the gradient get reset.

The resulting validation set score distribution for the same four attributes generated by the rebalanced MOON network after 34 training epochs can be seen in Fig. 5(b). Apparently, the score distributions are much more balanced, and the threshold 0 seems to make more sense now. Thus, one would expect that the classification error would be lower, too. However, due to the high dataset bias, which is also present in the validation and test sets, the total average

classification error of the balanced network on the (unbalanced) CelebA test set is 13.67 %.

Although this classfication error is larger than that obtained by the unbalanced MOON network, *this is an artifact of the significant imbalance in the original test set*; the error measure in Eq. (8) has not been adapted to the target domain. A fair comparison would measure the balanced classification error e_i^B that weights the positive and negative classes according to the target distribution:

$$e_i^B(X,Y) = \sum_{j=1}^{N_{test}} \begin{cases} \frac{c_i(X,Y)T_i^+}{N_i^+} & \text{if } Y_i = +1 \\ \frac{c_i(X,Y)T_i^-}{N_i^-} & \text{if } Y_i = -1, \end{cases} \tag{10}$$

where N_i^+ and N_i^- are the respective numbers of positive and negative examples of attribute i in the test set. When computing classification error of the rebalanced MOON network example with $T_i^+ = T_i^- = \frac{1}{2}$, we obtain an average e_i^B error of 12.98 %.

Note that the unbalanced MOON network, which is not trained to follow the target distribution, obtains an e_i^B error of 21.41 %. This is precisely what we would expect of a domain adaptation system: A classifier adapted to the target distribution does better than a classifier that is not.

5 Discussion

5.1 Handling Mis-aligned Images

In our experiments in Sec. 4, we used aligned images to train and test the networks. To show that MOON is in principle able to deal with badly aligned images, we conducted an additional experiment in which we used perturbed test images. To perturb the images, we applied a random rotation within $\pm 10°$, a random scaling with a scale factor between 0.9 and 1.1, and a random translation of up to 10 pixels in either direction to the pre-aligned faces in the CelebA dataset. We selected these parameters to be well outside of the error range of a reasonable (frontal face) eye detector. Alignment errors of these magnitudes have been shown to *highly* influence the performance of many traditional face recognition algorithms [30].

When running this perturbed test set through our (unbalanced) MOON network, which was trained purely on aligned faces, we obtain a classification error of 11.62 %, which is higher than the 9.06 % obtained with aligned test images, but still better than the current state of the art in [13]. We assume that we can improve the network stability against mis-alignment by incorporating augmented (e.g., misaligned perturbations) training data into the training process, since this has shown to improve the performance of deep convolutional neural networks [31].

Some preliminary experiments seem to verify this claim: When training with mis-aligned and horizontally mirrored images (in total 10 copies for each training image), we were able to decrease the classification error on the mis-aligned test images to 9.50 %. Unfortunately, this also caused a slight performance degra-

dation when evaluating on purely aligned images, causing classification error to increase from the 9.06 % to 9.23 %. Hence, in principle, the MOON architecture is able to work with aligned and mis-aligned images, as long as the conditions during training and testing are similar. These tests further highlight the need to select data augmentation methods appropriate to the respective quality of the actual alignment algorithms used in real end-to-end systems.

5.2 Hinge Loss

In order to test whether we have fully optimized our objective function, we performed an additional SVM training on top of the 40-dimensional attribute vector from the network. We took the finally selected (unbalanced) MOON network after 24 epochs, and extracted attribute vectors for the training, validation and test sets of CelebA. We trained 40 linear SVMs [32] on the training set, and used the validation set to optimize the C parameter for each attribute independently. Then, we classified all extracted test set attributes. The final result was a classification error of 9.11 %, which is very close to, but still above the 9.06 % that we obtained using the classification as given in Eq. (2). Hence, it seems that the MOON network has learnt a representation that is able to perform a multi-objective classification similar to the hinge-loss from Eq. (3).

6 Conclusion

The MOON architecture achieves an accurate, computationally efficient, and compact representation, whose attribute classification performance advances the state of the art on the CelebA dataset. Further, our experiments did not rely on any datasets external to CelebA to train our network, unlike competing approaches. We investigated the dataset bias in CelebA and proposed domain adaptation methods which allow us to define a different target distribution without changing the training set. We incorporate these methods directly into our mixed objective function and perform a demonstration on a novel re-balanced version of CelebA, the *CelebAB* dataset, for which we propose a different evaluation error measure.

Our work raises a philosophical question about the mathematics of attribute recognition: How *should* the attribute recognition problem be treated? Contrary to previous work, in which attribute labels are independently learnt, our approach implicitly leverages attribute correlations and explicitly forces hidden layers in the network to incorporate information from multiple labels simultaneously. As discussed in [33], many attributes can be recognized along a continuous range, some (e.g., `Big Nose`, `Oval Face`, `Young`) more than others (e.g., `Male`, `Eyeglasses`). Unlike other classifiers trained on purely discrete labels, MOON's weighted Euclidean loss allows it to simultaneously learn labels along a continuous range, although continuous labels were not provided in the dataset that we used. Whether our MOON network's output scores resulting from training on discrete labels reflect a reasonable ranking continuity in terms of the degree of expression of the underlying attribute has yet to be investigated.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

1. Zhou, Z.H., Zhang, M.L.: Multi-instance multi-label learning with application to scene classification. In: Advances in Neural Information Processing Systems. (2007) 1609–1616
2. Quattoni, A., Collins, M., Darrell, T.: Transfer learning for image classification with sparse prototype representations. In: Conference on Computer Vision and Pattern Recognition, IEEE (2008) 1–8
3. Huang, Y., Wang, W., Wang, L., Tan, T.: Multi-task deep neural network for multi-label learning. In: International Conference on Image Processing, IEEE (2013) 2897–2900
4. Wu, F., Wang, Z., Zhang, Z., Yang, Y., Luo, J., Zhu, W., Zhuang, Y.: Weakly semi-supervised deep learning for multi-label image annotation. IEEE Transactions on Big Data **1**(3) (2015) 109–122
5. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Robust visual tracking via multi-task sparse learning. In: Conference on Computer Vision and Pattern Recognition, IEEE (2012) 2042–2049
6. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision. Springer (2014) 94–108
7. Zhang, C., Zhang, Z.: Improving multiview face detection with multi-task deep convolutional neural networks. In: Winter Conference on Applications of Computer Vision, IEEE (2014) 1036–1041
8. Wang, X., Zhang, C., Zhang, Z.: Boosted multi-task learning for face verification with applications to web image and video search. In: Conference on Computer Vision and Pattern Recognition, IEEE (2009) 142–149
9. Yan, Y., Ricci, E., Subramanian, R., Lanz, O., Sebe, N.: No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In: International Conference on Computer Vision, IEEE (2013) 1177–1184
10. Ouyang, W., Chu, X., Wang, X.: Multi-source deep learning for human pose estimation. In: Conference on Computer Vision and Pattern Recognition, IEEE (2014) 2329–2336
11. Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., Kim, J.: Rotating your face using multi-task deep neural network. In: Conference on Computer Vision and Pattern Recognition, IEEE (2015) 676–684
12. Kumar, N., Belhumeur, P., Nayar, S.: Facetracer: A search engine for large collections of images with faces. In: European Conference on Computer Vision. Springer (2008) 340–353

13. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: International Conference on Computer Vision, IEEE (2015) 3730–3738
14. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: Conference on Computer Vision and Pattern Recognition, IEEE (2009) 951–958
15. Parikh, D., Grauman, K.: Interactively building a discriminative vocabulary of nameable attributes. In: Conference on Computer Vision and Pattern Recognition, IEEE (2011) 1681–1688
16. Yu, F., Cao, L., Feris, R., Smith, J., Chang, S.F.: Designing category-level attributes for discriminative visual recognition. In: Conference on Computer Vision and Pattern Recognition, IEEE (2013) 771–778
17. Huang, Y., Wang, W., Wang, L.: Unconstrained multimodal multi-label learning. IEEE Transactions on Multimedia **17**(11) (2015) 1923–1935
18. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence (6) (2001) 681–685
19. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(9) (2003) 1063–1074
20. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Describable visual attributes for face verification and image search. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(10) (2011) 1962–1977
21. Scheirer, W.J., Kumar, N., Belhumeur, P.N., Boult, T.E.: Multi-attribute spaces: Calibration for attribute fusion and similarity search. In: Conference on Computer Vision and Pattern Recognition, IEEE (2012) 2933–2940
22. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: International Conference on Computer Vision, IEEE (2009) 365–372
23. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning social relation traits from face images. In: International Conference on Computer Vision, IEEE (2015) 3631–3639
24. Wilber, M.J., Rudd, E., Heflin, B., Lui, Y.M., Boult, T.E.: Exemplar codes for facial attributes and tattoo recognition. In: Winter Conference on Applications of Computer Vision, IEEE (2014) 205–212
25. Kang, S., Lee, D., Yoo, C.D.: Face attribute classification using attribute-aware correlation map and gated convolutional neural networks. In: International Conference on Image Processing, IEEE (2015) 4922–4926
26. Patel, V.M., Gopalan, R., Li, R., Chellappa, R.: Visual domain adaptation: A survey of recent advances. IEEE Signal Processing Magazine **32**(3) (2015) 53–69
27. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: International Conference on Computer Vision, IEEE (2015) 4068–4076
28. Yang, Y., Hospedales, T.M.: A unified perspective on multi-domain and multi-task learning. In: International Conference on Learning Representation. (2015)
29. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. British Machine Vision Conference **1**(3) (2015) 6
30. Dutta, A., Günther, M., El Shafey, L., Marcel, S., Veldhuis, R., Spreeuwers, L.: Impact of eye detection error on face recognition performance. IET Biometrics **4** (2015) 137–150
31. Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: International Conference on Document Analysis and Recognition, IEEE (2003) 958–963

32. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874
33. Parikh, D., Grauman, K.: Relative attributes. In: International Conference on Computer Vision, IEEE (2011) 503–510