Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Attributes for Improved Attributes: A Multi-Task Network Utilizing Implicit and Explicit Relationships for Facial Attribute Classification

Emily M. Hand, Rama Chellappa
{emhand, rama}@umiacs.umd.edu
University of Maryland, College Park
College Park, MD, 20742

Abstract

Attributes, or mid-level semantic features, have gained popularity in the past few years in domains ranging from activity recognition to face verification. Improving the accuracy of attribute classifiers is an important first step in any application which uses these attributes. In most works to date, attributes have been considered independent of each other. However, attributes can be strongly related, such as *heavy makeup* and *wearing lipstick* as well as *male* and *goatee* and many others. We propose a multi-task deep convolutional neural network (MCNN) with an auxiliary network at the top (AUX) which takes advantage of attribute relationships for improved classification. We call our final network MCNN-AUX. MCNN-AUX uses attribute relationships in three ways: by sharing the lowest layers for all attributes, by sharing the higher layers for spatially-related attributes, and by feeding the attribute scores from MCNN into the AUX network to find score-level relationships. Using MCNN-AUX rather than individual attribute classifiers, we are able to reduce the number of parameters in the network from 64 million to fewer than 16 million and reduce the training time by a factor of 16. We demonstrate the effectiveness of our method by producing results on two challenging publicly available datasets achieving state-of-the-art performance on many attributes.

Introduction

Attributes are mid-level representations used for the recognition of activities, objects, and people (Duan et al. 2012) (Zheng et al. 2014) (Zhang et al. 2014a). Attributes provide an abstraction between low-level features and high-level object, or identity labels. They have seen the most success in face recognition and verification (Kumar et al. 2009) (Kumar et al. 2011). In this domain, attributes include *gender*, *race*, *age*, *hair color*, *facial hair*, etc. These semantic features are very intuitive, and they allow for human-understandable descriptions of objects, people, and activities. Reliable estimation of facial attributes is useful for many different tasks. HCI applications may require information about gender in order to properly greet a user (i.e. Mr. or Ms.) and other attributes such as expression in order to determine the mood of the user. Facial attributes can be used for identity verification in low quality imagery, where other verification methods may fail. Persons of interest - suspects, or missing persons - are often described in terms of their physical attributes, and so attributes can be used to automatically search for individuals in surveillance videos. Attributes have also found success in image search and retrieval as they can be used to search a database of images very quickly (Kumar et al. 2009) (Kumar et al. 2011) (Siddiquie, Feris, and Davis 2011).

Improving the accuracy of attribute classifiers is a challenging problem in itself and has been of recent interest due to the release of several large-scale attribute datasets (Liu et al. 2015). Convolutional neural networks (CNNs) have replaced most traditional methods for feature extraction in many computer vision problems (Krizhevsky, Sutskever, and Hinton 2012) (Vinyals et al. 2015). They have proven to be effective in attribute classification as well (Zhang et al. 2014a) (Abdulnabi et al. 2015) (Levi and Hassner 2015). However, with few exceptions, attributes have been treated as independent from each other. From a simple example - a woman wearing lipstick and earrings - we can see that this is not the case. If the subject is wearing lipstick and earrings, the probability that they are women is much higher than if they did not exhibit those attributes, and the reverse is also true. Treating each attribute as independent fails to use the valuable information provided by the other related attributes. Attributes fit nicely into a multi-task learning framework, where multiple problems are solved jointly using shared information (Argyriou, Evgeniou, and Pontil 2007) (Parameswaran and Weinberger 2010) (Caruana 1997).

We propose a multi-task deep CNN (MCNN) with an auxiliary network (AUX) at the top. The MCNN-AUX network utilizes information provided by all attributes in three ways: first, by sharing the lower layers of the MCNN for all attributes; second, by sharing the higher layers for similar attributes; and finally by utilizing all attribute scores from the trained MCNN in AUX in order to capture attribute relationships at the score level. We are able to achieve state-of-the-art performance on most attributes for two large-scale publicly available datasets: CelebA and LFWA (Liu et al. 2015).

The contributions of our work are as follows:

- We develop MCNN, a multi-task deep CNN for attribute classification.
- We develop AUX, an auxiliary network for MCNN which

allows for learning of attribute relationships at the score level.

- We combine MCNN and AUX to create MCNN-AUX, a multi-task attribute network which utilizes implicit and explicit attribute relationships for improved classification.
- We demonstrate the effectiveness of our approach by evaluating on two challenging publicly available datasets: LFWA and CelebA.
- We achieve state-of-the-art performance for many attributes, some showing up to a 15% improvement over other methods, without the expensive pre-training, alignment, or part extraction steps.
- We significantly decrease the number of parameters - over four times - and the amount of training time - over 16 times - required for the attribute classifier.

The remainder of the paper is organized as follows: We first discuss the related work in CNNs, multi-task learning, and attribute classification. This is followed by a discussion of the proposed MCNN and MCNN-AUX architectures. In the final sections, we detail our extensive experiments and results, and discuss the impact of our work.

Related Work

There are large bodies of work on CNNs, multi-task learning, and attributes. We draw from all three areas to design the proposed method, MCNN-AUX. The relevant literature is reviewed in the following sections.

CNN

Deep CNNs have been widely used for feature extraction and have shown great improvement over hand-crafted features for many problems including object recognition, automatic caption generation, face detection, face recognition and verification, and activity recognition (Girshick et al. 2014) (Krizhevsky, Sutskever, and Hinton 2012) (Vinyals et al. 2015). CNNs have quickly gained popularity since the introduction of open-source software tools which allow for straight-forward construction, training, and testing of deep CNNs. Caffe, Torch, and TensorFlow are among the most popular packages for implementing CNNs (Jia et al. 2014)(Abadi et al. 2015). The first big success for deep CNNs on a large-scale problem was in the 2012 Imagenet Challenge with a network that significantly outperformed the then existing methods for object recognition (Krizhevsky, Sutskever, and Hinton 2012). Since then, a wide variety of CNN architectures have been proposed for many computer vision problems.

CNNs have dominated the field of face recognition and verification. One of the most notable works in this domain is that of DeepFace, which utilized a large dataset and applied both a deep Siamese CNN and a classification CNN in order to maximize the distance between impostors and minimize the distance between true matches (Taigman et al. 2014). Motivated by the success on the challenging LFW dataset, researchers focused more on CNNs for face recognition and the networks have become deeper and more complex (Sun et al. 2014) (Sun, Wang, and Tang 2014a) (Sun, Wang, and Tang 2014b) (Sun et al. 2015).

In this work, we take advantage of the discriminative power of the CNN to learn semantic attribute classifiers as a mid-level representation for subsequent use in recognition and verification systems.

Multi-Task Learning

Multi-task learning (MTL) is a way of solving several related problems simultaneously, utilizing shared information (Argyriou, Evgeniou, and Pontil 2007) (Parameswaran and Weinberger 2010) (Caruana 1997). MTL has found success in the domains of facial landmark localization, pose estimation, action recognition, face detection, and many more (Zhang et al. 2014b) (Zhou et al. 2013) (Yim et al. 2015) (Zhang and Zhang 2014) (Devries, Biswaranjan, and Taylor 2014) (Ranjan, Patel, and Chellappa 2015).

In (Wang and Forsyth 2009), (Wang and Mori 2010), and (Hwang, Sha, and Grauman 2011) attributes and object classes are learned jointly to improve overall object classification performance. (Wang and Forsyth 2009) use Multiple Instance Learning to detect and recognize objects in images by learning attribute-object pairs. (Wang and Mori 2010) use an undirected graph to model the correlation amongst attributes in order to improve object recognition performance. In (Hwang, Sha, and Grauman 2011), attributes and objects share a low-dimensional representation allowing for regularization of the object classifier. In our work, all attributes share the lower layers in the CNN, so information common to all the attributes can be learned. Applying MTL to attribute prediction is very natural given the strong relationships among the facial attributes.

Attributes

(Kumar et al. 2009) introduced the concept of attributes as image descriptors for face verification. They used a collection of 65 binary attributes to describe each face image. They later extended this work with the addition of eight attributes and applied their method to the problem of image search in addition to face verification (Kumar et al. 2011). Berg et al. created classfiers for each pair of people in a dataset and then used these classifiers to create features for a face verification classifier (Berg and Belhumeur 2012). Here, rather than manually identifying attributes, each person was described by their likeness to other people. This is a way of automatically creating a set of attributes without having to exhaustively hand-label attributes on a large dataset. Prior to this, there were decades of research on gender and age recognition from face images (Fu, Guo, and Huang 2010)(Ng, Tay, and Goi 2012).

CNNs have been used for attribute classification recently, demonstrating impressive results. Pose Aligned Networks for Deep Attributes (PANDA) achieved state-of-the-art performance by combining part-based models with deep learning to train pose-normalized CNNs for attribute classification (Zhang et al. 2014a). Focusing on age and gender, (Levi and Hassner 2015) applied deep CNNs to the Adience dataset. Liu et al. used two deep CNNs - one for face localization and the other for attribute recognition - and achieved

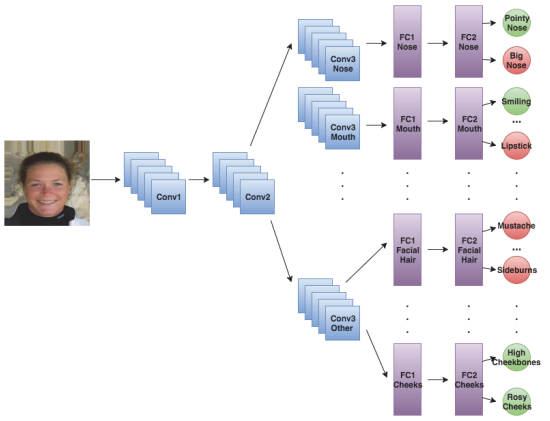

Figure 1: Overview of MCNN. The input image on the left is cropped to 227x227 and the training mean is subtracted. The red attributes indicate the absence of that attribute and the green attributes indicate a positive instance.

Group	Attributes
Gender	Male
Nose	Big Nose, Pointy Nose
Mouth	Big Lips, Lipstick, Mouth Slightly Open, Smiling
Eyes	Arched Eyebrows, Bags Under Eyes, Bushy Eyebrows, Eyeglasses, Narrow Eyes
Face	Attractive, Blurry, Heavy Makeup, Oval Face, Pale Skin, Young
AroundHead	Balding, Bangs, Black Hair, Blond Hair, Brown Hair, Earrings, Gray Hair, Hat, Necklace, Necktie, Receding Hairline, Straight Hair, Wavy Hair
FacialHair	5 o'clock Shadow, Goatee, Mustache, No Beard, Sideburns
Cheeks	High Cheekbones, Rosy Cheeks
Fat	Chubby, Double Chin

Table 1: Attributes and their corresponding groupings.

impressive results on the new CelebA and LFWA datasets, outperforming PANDA on many attributes (Liu et al. 2015). Unlike these methods, our MCNN-AUX requires no pre-training, alignment or part extraction.

Past work has generally considered attributes to be independent, with (Kumar et al. 2009), (Zhang et al. 2014a), and (Liu et al. 2015) training a separate classifier for each attribute. There are a few exceptions, however. (Siddiquie, Feris, and Davis 2011) use the correlation amongst attributes to improve image ranking and retrieval, learning pairwise correlations based on the outputs of independently trained attribute classifiers. Our method goes above and beyond this by training a single attribute network which classifies 40 attributes, sharing information throughout the network, and by learning the relationships among all 40 attributes, not just attribute pairs. (Abdulnabi et al. 2015) use a multi-task network to learn attributes for animals and clothing, rather than faces. They utilize groupings as in (Jayaraman, Sha, and Grauman 2014), but impose constraints at the feature level according to the groups. We incorporate groupings directly into the network by sharing layers amongst attributes in a single grouping. Using a deep CNN, unlike the RBM-based method of (Ehrlich et al. 2016), we achieve higher prediction accuracies.

Multi-Task CNN

The proposed MCNN takes an image as input and outputs 40 separate attribute scores, which are then thresholded to obtain binary outputs. We describe the details of the architecture below.

Architecture

Figure 1 shows the MCNN architecture. Conv1 consists of 75 7x7 convolution filters, and it is followed by a ReLU, 3x3 Max Pooling, and 5x5 Normalization. Conv2 has 200 5x5 filters and it is also followed by a ReLU, 3x3 Max Pooling, and 5x5 Normalization. Conv1 and Conv2 are shared for all attributes. This allows for learning of implicit relationships amongst attributes at a lower level. After Conv2, groupings are used to separate the layers. We use nine groups in our work: *Gender*, *Nose*, *Mouth*, *Eyes*, *Face*, *AroundHead*, *FacialHair*, *Cheeks*, and *Fat*. The attributes in each group are listed in table 1. There are six Conv3s: one each for *Gender*, *Nose*, *Mouth*, *Eyes*, and *Face*, and one for the remaining groups - Conv3Other. Each Conv3 has 300 3x3 filters and is followed by a ReLU, 5x5 Max Pooling and 5x5 Normalization. The Conv3s are followed by fully connected layers, FC1. There are 9 FC1s - one for each group. Each FC1 is fully connected to the corresponding previous layer, with Conv3Other connected to the FC1s for *AroundHead*, *FacialHair*, *Cheeks*, and *Fat*. Every FC1 has 512 units and is followed by a ReLU and a 50% dropout to avoid overfitting. Each FC1 is fully connected to a corresponding FC2, also with 512 units. The FC2s are followed by a ReLU and a 50% dropout. Each FC2 is fully connected to one output node for each of the attributes in that group. For example, FC2*Nose* is connected to output nodes for *Big Nose* and *Pointy Nose*. The grouping of attributes in the Conv3, FC1, and FC2 layers allows for the learning of explicit relationships among attributes from similar locations in the face image.

The nine groups were manually chosen according to attribute location. Some groupings were separated from others and some were absorbed into others through experimentation on the validation portion of the CelebA dataset giving the groupings in table 1. *Male* was kept separate from all other attributes as we found that *male* classification was improved by sharing layers with other attributes, but the classification of the other attributes suffered. We found the best compromise was to include *male* in the shared Conv1 and Conv2 layers and then to have separate Conv3, FC1, and FC2 layers.

We use the Caffe software for our implementation, training, and testing of MCNN and MCNN-AUX (Jia et al. 2014). We use a sigmoid cross-entropy loss for all attribute scores to facilitate training. As preprocessing steps, the training mean is subtracted from the images and they are cropped randomly with a size of 227x227. This helps the network to be robust to shifts in the input. Unlike other at-

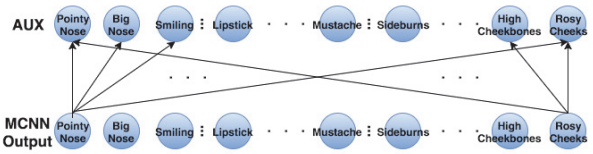

Figure 2: AUX network architecture. The output of the MCNN is fully connected to the final layer creating the 2-layer AUX network.

tribute classification methods, we do not perform any alignment or part extraction in the preprocessing stage. Both alignment and part extraction are expensive and error-prone processes, and so we save time and avoid problems associated with poor and misalignment by skipping these steps. Our method is also more applicable to real-world imagery for which alignment may be challenging.

If we were to use an independent CNN for each attribute, following the architecture of one path in the MCNN - 3 convolutional layers and 3 fully connected layers - each CNN would have over 1.6 million parameters. So, for all 40 attributes, there would be over 64 million parameters. Using MCNN, we reduce the number of parameters to fewer than 16 million, over four times fewer.

MCNN-AUX

After training the MCNN, we add one fully connected layer after the output of the trained MCNN. This layer creates the two-layer AUX network. Figure 2 shows the connection between MCNN and AUX. The input to AUX is the attribute scores from the trained MCNN, and the output is the final attribute scores. Starting with the weights from the trained MCNN, we learn the weights for the AUX portion of the network, keeping the weights from the MCNN constant. The AUX network allows for learning of score-level attribute relationships. The AUX network only adds 16000 parameters to the fewer than 16 million from MCNN.

Experiments

Data

In our experiments, we used two challenging, publicly available datasets: CelebA and LFWA. LFW was originally collected for verification; binary labels were recently added for 40 different attributes making it LFWA (Huang et al. 2007). CelebA was collected for attribute classification and was labeled with the same 40 attributes from LFWA (Zhang et al. 2012). Both datasets are extremely challenging, with large variations in subject pose, illumination and image quality. The CelebA dataset consists of 200,000 images: 160,000 for training and 20,000 each for validation and testing. The LFWA dataset contains 13,143 images with 6,263 for training and 6,880 for testing. Since the CelebA dataset is so large, we did not need to augment it in any way. If we did not augment the LFWA dataset, the network would severely overfit to the training data due to the large number of parameters. We augmented the LFWA dataset by jittering the original images by increments of 10 pixels. After jittering, we had over 75,000 images for training.

Attribute	Baseline	Liu et al.	Independent	MCNN	MCNN-AUX
5 o'clock Shadow	90.01	91	93.94	94.41	**94.51**
Arched Eyebrows	71.55	79	83.16	**83.55**	83.42
Attractive	50.41	81	82.22	82.94	**83.06**
Bags Under Eyes	79.73	79	84.83	84.89	**84.92**
Bald	97.88	98	98.85	98.87	**98.90**
Bangs	84.42	95	95.99	96.04	**96.05**
Big Lips	67.29	68	70.80	71.20	**71.47**
Big Nose	78.79	78	84.47	84.50	**84.53**
Black Hair	72.83	88	89.41	**89.87**	89.78
Blond Hair	86.67	95	95.88	95.97	**96.01**
Blurry	94.94	84	96.07	96.08	**96.17**
Brown Hair	82.03	80	88.75	88.99	**89.15**
Bushy Eyebrows	87.04	90	**92.87**	92.80	92.84
Chubby	94.69	91	95.55	95.66	**95.67**
Double Chin	95.42	92	**96.43**	96.41	96.32
Earrings	79.33	82	90.35	90.32	**90.43**
Eyeglasses	93.54	99	**99.67**	99.63	99.63
Goatee	95.41	95	97.13	**97.30**	97.24
Gray Hair	96.81	97	98.07	**98.20**	**98.20**
Hat	95.79	99	98.97	99.04	**99.05**
Heavy Makeup	59.50	90	90.95	91.37	**91.55**
High Cheekbones	51.81	**88**	87.34	87.55	87.58
Lipstick	52.18	93	93.80	93.95	**94.11**
Male	61.34	98	98.02	98.16	**98.17**
Mouth Slightly Open	50.49	92	**93.99**	93.74	93.74
Mustache	96.13	95	96.67	**96.93**	96.88
Narrow Eyes	85.13	81	87.22	87.16	**87.23**
Necklace	86.20	71	86.41	**86.82**	86.63
Necktie	92.99	93	**96.71**	96.53	96.51
No Beard	85.36	95	95.93	**96.11**	96.05
Oval Face	70.43	66	74.70	75.81	**75.84**
Pale Skin	95.79	91	**97.07**	97.01	97.05
Pointy Nose	71.42	72	**77.47**	**77.47**	**77.47**
Receding Hairline	91.51	89	93.41	**93.81**	**93.81**
Rosy Cheeks	92.82	90	95.02	95.13	**95.16**
Sideburns	95.36	96	97.77	97.82	**97.85**
Smiling	50.03	92	92.65	92.66	**92.73**
Straight Hair	79.01	73	82.62	83.39	**83.58**
Wavy Hair	63.59	80	83.24	**83.92**	83.91
Young	75.71	87	87.98	88.30	**88.48**

Table 2: Results for CelebA. The highest accuracy for each attribute is in bold.

Independent CNNs

We train independent CNNs for all the 40 attributes for both datasets in order to compare these results with those from MCNN and MCNN-AUX. We use one portion of our MCNN network for this. Each independent CNN has 3 convolutional layers, and 3 fully connected layers with the parameters specified in previous sections. We train these networks for 22 epochs for both datasets and use a batch size of 100. The independent CNNs each take about an hour to train for the CelebA dataset and about 30 minutes for the

LFWA dataset. For all 40 attributes, training independent CNNs takes over 40 hours for CelebA and over 20 hours for LFWA.

MCNN

To train MCNN, we use batches of size 100, and we train for 22 epochs for both datasets. Training takes about 2.5 hours for the CelebA dataset and about 1 hour for the LFWA dataset. We see a significant reduction in training time from 40 hours to 2.5 hours for CelebA and 20 hours to 1 hour for LFWA using MCNN over independent CNNs.

MCNN-AUX

Taking the trained MCNN, we fix the weights for that portion of the MCNN-AUX network and only train the AUX network. This takes about twenty minutes to train for CelebA and about 10 minutes for LFWA.

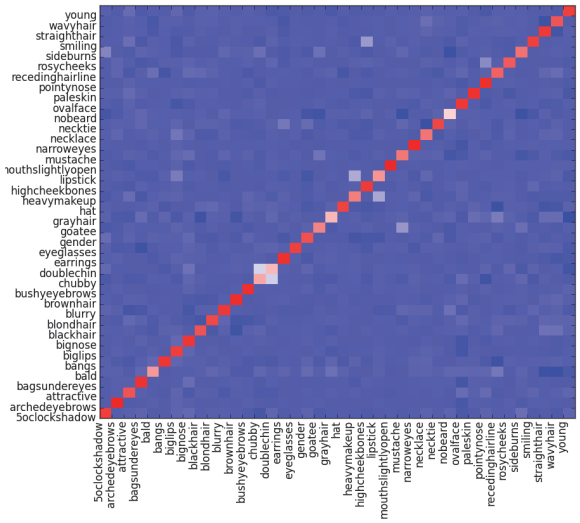

Figure 3: Heatmap of AUX network weights on CelebA. Along the x-axis, we have the MCNN output units and on the y-axis, the AUX units. Red indicates a strong relationship, and blue indicates a strong inverse relationship. Best viewed in color.

Results

We present results for our independent CNNs, MCNN, and MCNN-AUX. For comparison, we also provide the state-of-the-art by Liu et al., and a baseline of always choosing the most common label for each attribute.

We can see from Table 2 that our independent CNNs outperform Liu et al. on most attributes for CelebA. The independent CNNs improve on Liu et al. by 15% for *necklace*, 12% for *blurry*, 9% for *straight hair*, and 8% for *big nose*. MCNN makes even further improvements, and finally MCNN-AUX gives the highest accuracy for most attributes.

Attribute	Baseline	Liu et al.	Independent	MCNN	MCNN-AUX
5 o'clock Shadow	58.64	**84**	77.39	77.70	77.06
Arched Eyebrows	74.88	82	81.4	**82.36**	81.78
Attractive	62.87	**83**	80.20	80.42	80.31
Bags Under Eyes	58.29	83	83.24	**83.51**	83.48
Bald	89.37	88	91.51	**91.99**	91.94
Bangs	83.59	88	**90.47**	89.99	90.08
Big Lips	62.86	75	79.06	79.21	**79.24**
Big Nose	68.59	81	84.43	84.76	**84.98**
Black Hair	87.63	90	91.84	92.35	**92.63**
Blond Hair	95.74	97	97.23	**97.45**	97.41
Blurry	84.02	74	**86.71**	85.30	85.23
Brown Hair	64.56	77	80.84	**80.94**	80.85
Bushy Eyebrows	53.70	82	84.79	**85.11**	84.97
Chubby	63.92	73	75.85	**76.90**	76.86
Double Chin	62.44	78	**82.00**	81.17	81.52
Earrings	86.86	94	94.73	94.91	**94.95**
Eyeglasses	81.99	**95**	92.15	91.22	91.30
Goatee	74.68	78	**83.34**	82.52	82.97
Gray Hair	84.25	84	88.98	**89.04**	88.93
Hat	85.52	88	89.79	**90.20**	90.07
Heavy Makeup	89.20	95	95.63	95.84	**95.85**
High Cheekbones	67.74	88	88.02	88.25	**88.38**
Lipstick	85.53	95	94.68	94.89	**95.04**
Male	78.77	94	93.27	93.66	**94.02**
Mouth Slightly Open	58.70	82	82.41	83.47	**83.51**
Mustache	86.62	92	**93.69**	93.53	93.43
Narrow Eyes	65.50	81	82.48	82.73	**82.86**
Necklace	80.49	88	**89.98**	89.66	89.94
Necktie	64.09	79	80.34	80.50	**80.66**
No Beard	70.05	79	81.45	82.13	**82.15**
Oval Face	51.49	74	77.06.	77.38	**77.39**
Pale Skin	52.09	84	94.31	**93.41**	93.32
Pointy Nose	71.10	80	**84.41**	84.18	84.14
Receding Hairline	59.84	85	86.00	**86.26**	86.25
Rosy Cheeks	79.65	78	**89.46**	87.52	87.92
Sideburns	68.72	77	81.70	82.73	**83.13**
Smiling	60.50	91	**92.22**	91.75	91.83
Straight Hair	64.44	76	**81.54**	78.72	78.53
Wavy Hair	55.49	76	81.58	**81.96**	81.61
Young	79.60	**86**	85.11	85.37	85.84

Table 3: Results for LFWA. The highest accuracy for each attribute is in bold.

We see that the largest increase in performance is from the method of Liu et al. to the independent CNNs, with smaller improvements being made with MCNN and MCNN-AUX. From this, we determine that the value in MCNN and MCNN-AUX is in the increased training speed and the decreased parameters, which reduces the chances of overfitting. We do not expect to see an increase in performance with MCNN-AUX for every attribute, as many attributes do not have strong relationships with others. Determining which relationships to use can be done using a set of validation data, however, in this work we chose not to remove any relationships in our testing. All three of our methods outperform the baseline for every attribute in CelebA.

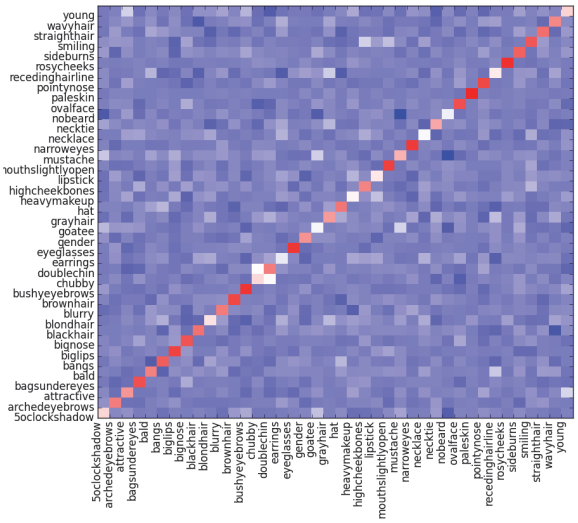

Figure 4: Heatmap of AUX network weights on LFWA. Along the x-axis, we have the MCNN output units and on the y-axis, the AUX units. Red indicates a strong relationship, and blue indicates a strong inverse relationship. Best viewed in color.

Figure 3 shows a heatmap of the weights for the AUX network on the CelebA dataset. From figure 3 we can see that each attribute contributes the most to its final classifier score. This is expected as MCNN already produces strong attribute classification accuracies. Some intuitive relationships can be seen in the heatmap. We see that *bald* is strongly related to *receding hairline* and has an inverse relationship with *straight hair* and *wavy hair* and that *no beard* has an inverse relationship with *5 o'clock shadow*, *mustache*, and *sideburns*. There are many more just like these. We do see some unexpected relationships as well, like *high cheekbones* and *smiling* having a strong connection. This would seem to indicate that people are not very good at determining when someone has *high cheekbones* and therefore the labels for this attribute are somewhat noisy.

Table 3 shows the results for the LFWA dataset. We can see that the accuracies are lower for this dataset than for the CelebA dataset. This is likely due to overfitting because LFWA is much smaller than CelebA. The independent CNNs outperform Liu et al. on most attributes with an improvement of 11% for *blurry*, 11% for *rosy cheeks*, 10% improvement for *pale skin*, and 5% improvements for both *straight hair* and *wavy hair*. MCNN improved the classification accuracy of many attributes, but we see that a few, such as *blurry* and *eyeglasses*, did not improve with MCNN. For *blurry* and *eyeglasses* this makes sense, as both attributes are relatively unrelated to the other attributes, and therefore do not gain anything from shared information. We note that though MCNN-AUX does not improve the results for some attributes, we do not pre-train the networks using a larger dataset, as in Liu et al., which used a much larger dataset to initialize the weights of their networks. Pre-training on external data would likely improve the results, however that is not the focus of this work.

Figure 4 shows a heatmap of the weights for the AUX network on LFWA. There is much more white in this heatmap than in that of figure 3 indicating that there are fewer strong relationships in LFWA than in CelebA. This makes sense, as the classification accuracies for MCNN on LFWA were not as high as on CelebA. Again, we believe that this is due to the small size of the dataset. Though jittering LFWA helps, it does not compare to having a large amount of unique data as in CelebA. As with CelebA, we see that each attribute contributes most to its overall classification accuracy, though not quite as strongly. We again see promising relationships, with *bald* and *receding hairline* being strongly related as well as *heavy makeup* and *lipstick* and several others. We see that there are some noisy labels as in CelebA with *smiling* and *highcheekbones* being strongly related.

Conclusion

In this paper, we have shown that though facial attributes have been treated as independent problems in the past, there is a lot to be gained from shared information amongst attributes. Framing the attribute prediction problem as a multi-task learning problem is very natural and allows for a large reduction in training time and in the number of parameters required for the classifier. The proposed MCNN-AUX reduced the number of parameters from 64 million to fewer than 16 million, and reduced the training time by 16 times. We demonstrated the effectiveness of our independent CNN, MCNN, and MCNN-AUX classifiers on the challenging CelebA and LFWA datasets, achieving state-of-the-art performance for most attributes. Attribute relationships can be exploited in many ways and we presented three ways in this paper: by sharing lower layers of MCNN, by grouping similar attributes in higher layers of MCNN, and by introducing an auxiliary network (AUX), which learns attribute relationships at the score level. Attribute relationships are learned implicitly at the lower levels, and explicitly in the higher grouped layers. Even without pre-training, we were able to outperform the method of Liu et al. for many attributes. Pre-training on external data may improve the results, however that is not the focus of this work. We demonstrated through experiments that a multi-task framework for attribute prediction outperforms independent classifiers. Taking advantage of implicit and explicit relationships among attributes allows for improved attribute prediction which will lead to improved facial recognition.

Acknowledgment

This research is based upon work supported 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

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; and Corrado, G. S. 2015. Tensorflow: Large-scale machine learning on heterogeneous systems.

Abdulnabi, A. H.; Wang, G.; Lu, J.; and Jia, K. 2015. Multi-task cnn model for attribute prediction. *arXiv preprint*.

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

Berg, T., and Belhumeur, P. N. 2012. Tom-vs-pete classifiers and identity-preserving alignment for face verification. *BMVC*.

Caruana, R. 1997. Multitask learning. *Machine Learning*.

Devries, T.; Biswaranjan, K.; and Taylor, G. W. 2014. Multi-task learning of facial landmarks and expression. *CRV*.

Duan, K.; Parikh, D.; Crandall, D.; and Grauman, K. 2012. Discovering localized attributes for fine-trained recognition. *CVPR*.

Ehrlich, M.; Shields, T. J.; Almaev, T.; and Amer, M. R. 2016. Facial attributes classification using multi-task representation learning. *CVPR*.

Fu, Y.; Guo, G.; and Huang, T. S. 2010. Age synthesis and estimation via faces: A survey. *PAMI*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Technical Report, University of Massachusetts, Amherst*.

Hwang, S. J.; Sha, F.; and Grauman, K. 2011. Sharing features between objects and their attributes. *CVPR*.

Jayaraman, D.; Sha, F.; and Grauman, K. 2014. Decorrelating semantic visual attributes by resisting the urge to share. *CVPR*.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *NIPS*.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and simile classifiers for face verification. *ICCV*.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2011. Describable visual attributes for face verification and image search. *PAMI*.

Levi, G., and Hassner, T. 2015. Age and gender classification using convolutional neural networks. *CVPR*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. *ICCV*.

Ng, C. B.; Tay, Y. H.; and Goi, B. M. 2012. Vision-based human gender recognition: A survey. *arXiv preprint*.

Parameswaran, S., and Weinberger, K. 2010. Large margin multi-task metric learning. *NIPS*.

Ranjan, R.; Patel, V. M.; and Chellappa, R. 2015. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint*.

Siddiquie, B.; Feris, R. S.; and Davis, L. S. 2011. Image ranking and retrieval based on multi-attribute queries. *CVPR*.

Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. *NIPS*.

Sun, Y.; Ding, L.; Wang, X.; and Tang, X. 2015. Face recognition with very deep neural networks. *CoRR*.

Sun, Y.; Wang, X.; and Tang, X. 2014a. Deep learning face representation from predicting 10,000 classes. *CVPR*.

Sun, Y.; Wang, X.; and Tang, X. 2014b. Deeply learned face representations are sparse, selective, and robust. *CoRR*.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. *CVPR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. *CVPR*.

Wang, G., and Forsyth, D. 2009. Joint learning of visual attributes, object classes and visual saliency. *CVPR*.

Wang, Y., and Mori, G. 2010. A discriminative latent model of object classes and attributes. *ECCV*.

Yim, J.; Jung, H.; Yoo, B.; Choi, C.; Park, D.; and Kim, J. 2015. Rotating your face using multi-task deep neural network. *CVPR*.

Zhang, C., and Zhang, Z. 2014. Improving multiview face detection with multi-task deep convolutional neural networks. *WACV*.

Zhang, X.; Zhang, L.; Wang, X. J.; and Shum, H. Y. 2012. Finding celebrities in billions of web images. *IEEE Transactions on Multimedia*.

Zhang, N.; Paluri, M.; Ranzato, M. A.; Darrell, T.; and Bourdev, L. 2014a. Panda: Pose aligned networks for deep attribute modeling. *CVPR*.

Zhang, Z.; Luo, P.; Loy, C.; and Tang, X. 2014b. Facial landmark detection by deep multi-task learning. *ECCV*.

Zheng, J.; Jiang, Z.; Chellappa, R.; and Phillips, J. P. 2014. Submodular attribute selection for action recognition in video. *NIPS*.

Zhou, Q.; Wang, G.; Jia, K.; and Zhao, Q. 2013. Learning to share latent tasks for action recognition. *ICCV*.