Implementing a Robust Explanatory Bias in a Person Re-identification Network

Esube Bekele
National Research Council Fellow
Washington, DC
esube.bekele.ctr@nrl.navy.mil

Wallace E. Lawson
Naval Research Laboratory
Washington, DC
ed.lawson@nrl.navy.mil

Zachary Horne
Arizona State University
Phoenix, AZ
zachary.horne@asu.edu

Sangeet Khemlani
Naval Research Laboratory
Washington, DC
sunny.khemlani@nrl.navy.mil

Abstract

Deep learning improved attributes recognition significantly in recent years. However, many of these networks remain "black boxes" and providing a meaningful explanation of their decisions is a major challenge. When these networks misidentify a person, they should be able to explain this mistake. The ability to generate explanations compelling enough to serve as useful accounts of the system's operations at a very high human-level is still in its infancy. In this paper, we utilize person re-identification (re-ID) networks as a platform to generate explanations. We propose and implement a framework that can be used to explain person re-ID using soft-biometric attributes. In particular, the resulting framework embodies a cognitively validated explanatory bias: people prefer and produce explanations that concern inherent properties instead of extrinsic influences. This bias is pervasive in that it affects the fitness of explanations across a broad swath of contexts, particularly those that concern conflicting or anomalous observations. To explain person re-ID, we developed a multi-attribute residual network that treats a subset of its features as either inherent or extrinsic. Using these attributes, the system generates explanations based on inherent properties when the similarity of two input images is low, and it generates explanations based on extrinsic properties when the similarity is high. We argue that such a framework provides a blueprint for how to make the decisions of deep networks comprehensible to human operators. As an intermediate step, we demonstrate state-of-the-art attribute recognition performance on two pedestrian datasets (PETA and PA100K) and a face-based attribute dataset (CelebA). The VIPeR dataset is then used to generate explanations for re-ID with a network trained on PETA attributes.

1. Introduction

Few deep learning systems can generate explanations of their own computations in a manner that is comprehensible to human end users. As a result, researchers have begun to explore techniques for building explainable AI systems, but, as Miller and his colleagues described [14], the designers of such systems seldom consult results from the behavioral sciences on how humans generate and evaluate explanations. Miller et al. [14] argue that progress on building explainable AI systems will be limited until they can recognize and adapt to human-level explanatory biases. Otherwise, these systems will produce descriptions that have limited explanatory value.

Person re-identification (re-ID) involves identifying a person from a full body query image and searching through a gallery of existing images. Re-ID networks have numerous applications: they can help track individuals in real time (i.e., as they enter and leave a particular video frame), integrate data from multiple-camera surveillance setups, and track pedestrians at different angles and viewpoints [11]. Although there is a significant body of research on pedestrian re-ID, most do not have any way of explaining their decision making process in a way that would be easily understood by a human collaborator. Most re-ID networks focus on end-to-end mapping of the input image(s) to the person's identity label without reporting any intermediate human understandable features. Soft biometric attributes describe a person using both inherent (e.g., age, gender, build) and extrinsic (e.g., clothing, objects carried) properties. In this respect, soft-biometrics attributes can provide a way of providing an explanation in terms that are commonly used and readily understood by these collaborators. The attributes serve as an intermediate way of accounting for what went into the decision process of re-identification. Some recent

re-ID networks use soft biometric attributes to generate descriptions of images in the hope of making the network's underlying processes more transparent. However, descriptions of images are difficult for human end-users to interpret: they can be long and uninformative. We argue instead that re-ID networks can serve as an ideal platform on which to generate short concise explanations.

As attribute recognition is an important intermediate step of the explanation generation process, here we describe briefly the state-of-the-art in attribute recognition. Attribute recognition has recently attracted a significant amount of attention, partially due to the the increasing availability of larger labeled attribute datasets [10][3]. The number of attributes also is growing with the size of these datasets. Early work in this area used the shallower AlexNet architecture (8 layers) pre-trained with ImageNet [8] (e.g., ACN [16], DeepMAR [9], MLCNN [19]), and ANet [13]. More recent work utilized the deeper GoogleNet network (22 layers total) pre-trained with ImageNet [18, 15].

In this paper, we describe a proof-of-concept system that generates its own explanations in a way that implements a cognitively validated bias in explanatory reasoning - the inherence bias. In what follows, we present our approach for generating these explanatory attributes and demonstrate state-of-the-art performance on three attribute datasets as intermediate step towards the explanation generation. We then describe the bias and review its pervasiveness and importance. Finally, we discuss how the bias can be built into a re-identification network for the purpose of classifying and assessing whether two images of pedestrians depict the same individual or a different one. We conclude by discussing the advantages and limitations of this approach.

2. Human-biases in Explanatory Reasoning

Explanations are core features of human rationality that serve as a means to communicate our understanding of the world [7]. When applied to deep learning systems, explanations have the potential to help make a complex system transparent by highlighting its most pertinent components and causal dependencies. Hence, explanations are a form of dimensionality reduction. In this section we briefly review an experimentally validated explanatory bias, i.e. "inherence bias". We apply this known biased explanation generation process to explain re-ID networks in a way that attempts to mimic human explanatory reasoning.

2.1. Biased Explanations in Human Reasoning

A recent psychological proposal argues that the cognitive process of generating explanations operates heuristically to yield biased explanations, i.e., explanations that exhibit certain structural and semantic patterns over others [7]. Humans exhibit a wide variety of explanatory biases: an explanation's simplicity, scope, and completeness affect whether it is considered good or bad [7]. They do not perform an exhaustive search through the space of all possible explanations. In sum, explanations are systematically constrained, and AI and robotic systems that produce explanations for the purpose of helping human end users understand their underlying operations need to simulate human biases.

2.1.1 The Inherence Bias

Consider the following question: Why do lions roar? A compelling explanation might be: because they are ferocious. It cites an inherent property of lions, i.e., ferociousness. A more accurate explanation is that lions roar as means of communication: they often roar as a way of locating one another. The accurate, extrinsic explanation is more complex but more difficult to comprehend, whereas the inherent explanation may be more superficially attractive because it is easier to understand. Recent research suggests that human reasoners perform a shallow search through the contents of their memory to explain a particular observation or regularity [2]. As a result of their shallow search, humans tend to construct explanations based on accessible information about the inherent properties of a particular phenomenon instead of inaccessible information about extrinsic factors. The bias is pervasive: it affects how reasoners generate and evaluate explanations across a broad swath of contexts [2].

We argue that this pervasive semantic bias, i.e., the bias to produce inherent explanations in situations of conflict, can help certain kinds of deep learning systems build better explanations. We show how this bias can help a certain class of convolutional neural networks yield compelling explanations for re-ID. We formulate the explanation generation using human understandable attribute labels that are common in the attribute recognition and person re-ID literature [1]. For instance, in the example network shown in Fig. 1 there are 35 attributes that describe a pedestrian in the PETA dataset. Some of these attributes are inherent such as gender and age and many are extrinsic attributes such as clothing and accessories carried or worn by the pedestrian.

3. Developing an Inherence-biased Network

This section describes how we developed the attribute recognition and re-ID network in generating the inherence biased explanations as a way to communicate the network's operation to a human operator. We present the general process of the explanation generation followed by how we assess the techniques for attribute recognition as well as explanations generation as a way of explaining re-identification.

3.1. Inherence-biased Explanations and Re-ID

Re-ID networks tend to output a similarity metric, a discriminative model, or a description of a particular identified individual [11]. No existing re-ID network provides an explanation of its operations, but such networks provide a feasible platform for which to generate biased explanations that adhere to human expectations, because person re-identification tasks often require networks to learn features that correspond to inherent properties (e.g., whether the person is male or female) or else features that correspond to extrinsic properties (e.g., what the person is carrying). We qualified these hidden features into attributes that are understandable to a human operator. Hence, the process of generating the explanations is a multi-stage recognition process.

First, we train multi-task deep residual networks for multi-attribute predictions as an intermediate step of converting the hidden feature vector (which is not understandable by humans) into a set of easy-to-understand human attributes. The network is trained using either pedestrian or facial attribute datasets to produce the attribute labels. Some of these labels express inherent features such as gender and age while others could be used to describe extrinsic features such as carried objects, clothing, and shoes. Hence, once probabilities were assigned using the attribute recognition intermediate step, we collapse the various attributes into a few inherent and extrinsic categories that are suitable for generating the explanations.

Once the attribute recognition models are trained using a larger attribute recognition and re-ID datasets, we predict on a smaller re-ID dataset to get the attributes for both images as shown in Fig. 2. Using the predicted probabilities to the set of attributes, cosine similarity is computed between the features of these two images. Based on the similarity score, inherent explanations are generated if there is conflict, i.e. the similarity score between the two sets of features is too based on a set threshold (e.g., $< .01$). Likewise, if the similarity score is above a certain set threshold (e.g., $> .99$), extrinsic explanations are generated.

3.2. Assessments and Evaluation Metrics

Assessing the performance of the overall system should be decomposed into stages similar to the overall techniques used to train it as described above. For the attribute recognition stage of the system, we built on the evaluation metrics proposed in [10], i.e. the mean average accuracy (mA) and example-based metrics of example (sample) accuracy (Acc), sample-based recall (Rec), sample-based precision (Prec), sample-based F1 score (F1), and the area under the curve (AUC). These metrics are comprehensive and avoid biased predictions due to intra-class imbalance in the dataset [1].

The evaluation of the explanation generation stage is qualitative at this point. As the explanation generation are based on a well known behavioral human explanation bias [14], a full user study is warranted to assess the comprehensibility of the generated explanations as indicated by preference of human operators. We left that as future work. However, there are qualitative metrics that were used to control the explanations such as length of the generated explanations. Explanations should be short and not a semi-static laundry list of the attributes in some form of long description. Another important aspect is the generation of the explanations should not impact on the speed of the attribute recognition and person re-ID.

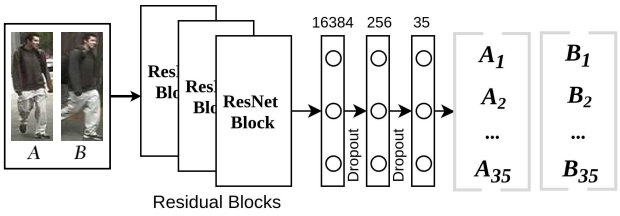

Figure 1. An example deep residual network architecture that was trained using the 35 binary attributes of PETA dataset for person re-ID.

4. Implementing the Network

In this section we present the details of the proposed method including the network architecture, techniques used for training the network efficiently and accurately, and the process of constructing the convincing explanations.

4.1. Network Architecture

Here, we present the network architecture used to predict explanatory attributes. This architecture was selected in order to maximize our prediction accuracy with a limited size and biased dataset. We discuss the related pooling schemes, loss functions, sample weighting and probability calibration needed to produce state-of-the-art performance on a wide range of datasets.

Figure 1 shows an instance of the network architecture. We consider residual networks (ResNet) [5] and dense residual networks (DenseNet) [6] for creating richer feature sets that could achieve comparable performance to the state-of-the-art without complicated pooling and branching strategies [18], and view-based predictions [15]. In addition to the base feature extractor network, the type of feature pooling used in combination with specific classifier layers could have performance and computational implications for joint multi-attribute recognition. Several methods of pooling at the feature stage have been proposed. They range from the simple global pooling at the last stage of the feature extraction pipeline to complex pyramidal pooling schemes [18]. In practice, for deeper networks such as ResNet [5], GoogleNet [17] and DenseNet [6], the last features are pooled in a global manner. Local pooling strate-

gies could benefit shallower networks. More complicated feature pooling schemes are discussed in [18, 15]. We experimented with both logistic and multi-layer dense classifiers.

4.2. Training Techniques

To train the attribute feature extractor network described above, we followed the common optimization objective in CNN-based attribute recognition, i.e. weighted binary cross-entropy (BCE) loss [19]. Due to the nature of the presence of multiple attributes in the same example image, a multi-label BCE loss enables learning not only individual attributes but also their joint interrelationship. Equation 1 shows BCE loss.

$$TotalLoss = \sum_{m=1}^{M} \gamma_m \, l_m \tag{1}$$

where l_m is the BCE contribution of m^{th} attribute to the total loss. The γ_m parameter could be used to control the learning to be focused towards a particular attribute. This could be useful if a particular attribute is especially important for a particular recognition scenario. For instance, inherent attributes such as age and gender could be more important than clothing attributes for the inherence biased explanation generation process.

The individual attribute losses, l_m, are computed using the sigmoid binary cross-entropy loss for binary classification of each attribute as shown in Equation 2.

$$l_m = -\frac{1}{N} \sum_{i=1}^{N} \omega_m^i (y_m^i log(\hat{y}_m^i) + (1 - y_m^i) log(1 - \hat{y}_m^i)) \tag{2}$$

where y_m^i is the true and $\hat{y}_m^i$ is the predicted attribute label for m^{th} attribute and i^{th} example. The predicted probabilities are computed via the sigmoid function and is given by Equation 3.

$$\hat{y}_m^i = p(x_m^i) = \frac{1}{1 + exp(-x_m^i)} \tag{3}$$

and w_m is a sample weighting factor introduced to account for the bias that could occur due to inherent class imbalance within an attribute. Li et al.[9] adopted a more controllable weighting scheme given by Equation 4. In this paper we have experimentally tested various weighting schemes and found that Equation 4 is the best and hence used it for the results section unless stated otherwise. This weighted multi-label optimization objective was minimized using stochastic gradient decent (SGD) with Nestrov momentum.

$$\omega_m^i = \begin{cases} exp((1 - p_m)/\sigma^2) & if \; y_m^i = 1 \\ exp(p_m/\sigma^2) & else \end{cases} \tag{4}$$

where p_m is the number of positive examples in the m^{th} attribute. σ is a control parameter. This parameter can be used to control the number of true positives and hence control the recall in effect. However, in this experiment, appropriate attributes were pre-selected by a criteria of meeting a certain threshold of number of positive examples [3], we set this parameter to 1 for all the attributes. Setting this parameter lower improves recall at the expense of mean accuracy. Other techniques employed include probability calibration using set false positive rate (FPR) and calibration method that balances both precision and recall, it favors a strong example-based accuracy and F1 score, we called it *F1 calibration*. Data augmentation as preprocessing step is also another crucial strategy to train deeper networks with limited dataset. In this paper, we employed the most commonly used data augmentation techniques in the pedestrian attribute recognition literature. These include random flip and crop, random image resizing with and without constant aspect ratio, mean subtraction, random RGB color jitter and random rotation. All the data augmentation in this paper was performed on-line, i.e. each image was randomly augmented by the one or more of the listed methods in each iteration separately.

4.3. Datasets

We demonstrate state-of-the-art performance on two pedestrian datasets, i.e. the PEdesTrian Attribute (PETA) recognition dataset [3] and PA-100K dataset [12], and a face-based celebrity faces attributes dataset (CelebA) [13]. PETA contains 19,000 images captured by surveillance cameras in indoor and outdoor scenarios. Originally the images were labeled with 61 binary attributes, which are related to age, clothing, carrying items, wearing accessories etc. There are also 4 multi-class attributes related to color. Deng et al. [3] suggested to use 35 attributes due to severe class imbalance issues in the remaining. Therefore, we adopted these 35 attributes PETA images exhibit wide range of variations in illumination, pose, blurriness, resolution, background and occlusion. We evaluate with the suggested [3] random train/validation/test split of 9500/1900/7600 images for equivalent comparison. The PA-100K contains 100,000 images captured from 598 scenes. Each image is labeled using 26 binary full body attributes. CelebA contains 202,600 images labeled with 40 face-based attributes. We used the provided splits for both PA-100K and CelebA datasets. For the cosine similarity and explanation generation, we trained the network on PETA and predicted on VIPeR dataset [4].

4.4. Performance Evaluations

We first present the attribute recognition performance as it directly affects the content and process of how the explanations are generated, results of our residual network based

Table 1. Comparison of multi-attribute recognition performance of proposed model and the state-of-the-art on PETA. The data augmentation employed are mean subtraction, random color jitter, and random flip.

Networks/Method	mA	Acc	Prec	Rec	F1	AUC (micro)
ACN[16]	81.15	73.66	84.06	81.26	82.64	-
DeepMAR [9]	82.89	75.07	83.68	83.14	83.41	-
WPAL-GMP [18]	**85.50**	76.98	84.07	85.78	84.90	-
VeSPA [15]	83.45	77.73	86.18	84.81	85.49	-
Proposed ResNet50 No data augmentation	84.18	78.03	85.78	85.69	85.73	90.55
Proposed ResNet50 With data augmentation	84.68	**78.89**	**86.38**	**86.41**	**86.39**	**90.96**

architecture of 50 layers (ResNet50) is presented in comparison with other state-of-the-art methods. We then present samples of generated explanations and discuss these explanations in terms of length and inherence bias and conclude with Generation speed implications.

4.4.1 Attribute Recognition Performance on PETA

Table 1 compares the performance of our proposed approach against other state-of-the-art attribute recognition approaches on the PETA dataset. ResNet50 with data augmentation techniques such as random flip and RGB color jitter outperforms the state-of-the-art in all the metrics presented with the exception of mean accuracy. However, it is clear from other metrics such as precision, recall and F1 score that ResNet50 is a well balanced recognition approach performing well across a larger number of attributes. Without any data augmentation, ResNet50 has a performance that is comparable to or better than other approaches.

4.4.2 Attribute Recognition Performance on PA-100K

Table 2 compares the performance of our proposed approach against other state-of-the-art attribute recognition approaches on the PA-100K dataset. Our proposed model outperforms the state-of-the-art in this dataset by about 3%

Table 2. Comparison of multi-attribute recognition performance of the proposed model and the state-of-the-art on PA-100K. The data augmentation employed are mean subtraction, random color jitter, and random flip.

Networks/Method	mA	Acc	Prec	Rec	F1	AUC
HydraPlusNet [12]	72.70	70.39	82.24	80.42	81.32	-
DeepMAR [12]	74.21	72.19	82.97	82.09	82.53	-
Proposed ResNet34 F1 calibration	**78.77**	**75.05**	**85.01**	**84.56**	**84.78**	**89.95**
Proposed ResNet50 Sample weighting	78.07	73.51	83.74	83.45	83.59	89.19
Proposed ResNet50 F1 calibration	78.12	74.11	84.42	84.09	84.25	89.54

Table 3. Comparison of multi-attribute recognition performance of the proposed model and the state-of-the-art on CelebA. The data augmentation employed are mean subtraction, random color jitter, and random flip.

Networks/Method	mA	Acc	Prec	Rec	F1	AUC
LNets+ANet [13]	**87**	-	-	-	-	-
FaceTracker [13]	81	-	-	-	-	-
PANDA-l [13]	85	-	-	-	-	-
Proposed ResNet34 FPR@10% calibration	86.55	65.18	**86.55**	72.34	78.81	**88.79**
Proposed ResNet34 F1 calibration	83.16	65.20	77.80	**80.19**	78.98	86.42
Proposed ResNet34 FPR@8% calibration	85.72	**66.67**	82.25	77.79	**79.96**	87.98

on each metric. From the metrics such as precision, recall and F1 score we can see that our model is a well balanced recognition approach performing well across a larger number of attributes.

4.4.3 Attribute Recognition Performance on CelebA

Table 3 compares the performance of our proposed model to that of the state-of-the-art attribute recognition approaches on the CelebA dataset. Our approach performed competitively to the state-of-the-art approaches on the mean accuracy metric. The other approaches only reported the mean accuracy. Hence, it is not clear how balanced their recognition on the other metrics were. To illustrate this point, we run the same network with different probability calibration techniques discussed in the methods section and showed that it is possible to get a mean accuracy competitive to the state-of-the-art at the expense of precision (i.e. proposed system with FPR@10% calibration). However, we would like to point out that the more balanced recognition performance is the F1 score calibration although the mean accuracy is lowered.

4.4.4 Assessing the Network's Explanations

For this task, we determine the similarity between two images by measuring the similarity of predicted attributes. However, we note that this scheme is sufficiently generic to be usable for any approach for person re-identification. We measure similarity using the approach shown in Fig. 1. This shows how the network operates: it takes two separate images as input and uses three residual blocks to learn 35 separate attributes jointly. The attributes pertain to the age and gender of the person in the input image, as well as what the person is wearing, what the person is carrying, and other various descriptors (e.g., whether the person has long hair or not). Some of the 35 attributes concern inherent properties (e.g., age, gender, hair) and some concern extrinsic properties (e.g., whether the person is wearing a hat or carrying a backpack).

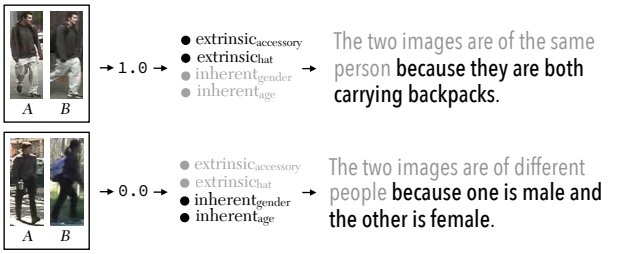

Figure 2. Explanations are constructed using cosine similarity on the predicted attribute probabilities. The model builds an explanation relative to its assessment of re-ID.

The network was trained on the PETA dataset and tested on the VIPeR dataset. After computing the attribute prediction, cosine similarity of the resulting attribute probability vectors given two input images from the dataset were computed. Critically, the overall model reports a judgment paired with an explanation (see Fig. 2). The explanations concern inherent properties when the model detects dissimilarity and extrinsic properties when it detects similarity. Fig. 3 illustrates the inherence biased explanation with further examples. Inherent attributes are preferred over extrinsic attributes in describing difference. We argue that specific similarity in extrinsic attributes such as the person carrying similar backpacks in both images is indicative of the same person. Since the two images in the first example look similar enough in the major inherent attributes, similarity in peculiar extrinsic attributes confirms that it is indeed the same person. In the second and the third examples, the two images are different in a more fundamental way and inherent attributes such as gender and age could be used to distinguish the two people easily. These observations are consistent with the inherence biased explanation hypothesis.

The generated explanations were designed to be short in length in contrast with descriptions that contain the list of attributes in semi-static sentences. Here we make the useful distinction between short meaningful inherence biased explanations and long incomprehensible descriptions. We also note that the generation of the explanations does not add significant overhead to the overall re-ID process. It is important that the system performs with little to no overhead with the addition of the explanation generation.

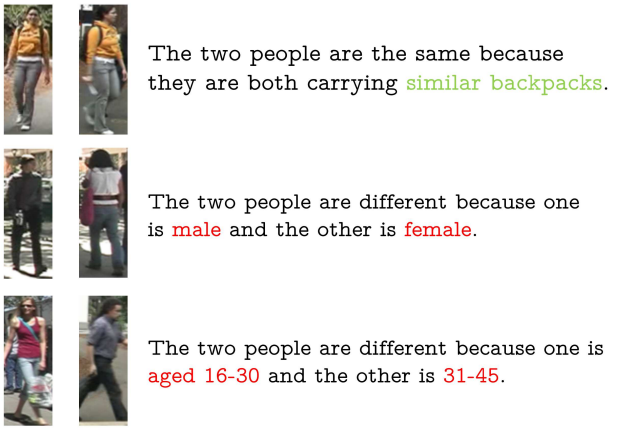

Figure 3. Examples of the generated inherence biased explanations. Inherent attributes were used to highlight differences while extrinsic attributes were employed to explain similarity.

5. Conclusion

In this paper, we propose an approach to explaining person re-identification using multiple soft-biometric attributes. Recognition of attributes in the wild can be a very challenging task while providing a possibility for generating compelling explanations. Environmental conditions such as weather, lighting and shadow can impact results. People also can be partially occluded by other things in the environment.

We proposed deep residual networks that learn attributes as a way of generating compelling explanations with human-level biases. Our approach demonstrated state-of-the-art on PETA, PA-100K, and CelebA attribute datasets. The approach generalizes well, ranging from pedestrian attributes to face-based attribute recognition. Our experiments and the ablation studies (see Appendix), indicate that simpler networks with various training strategies result in superior performance than complicated networks.

The attributes were utilized to generate the biased explanations based on preference towards inherence to reflect the well documented inherence bias in explanations generated by humans [14, 7]. We argue that deep learning systems capable of mimicking human explanatory biases can provide meaningful and interpretable explanations of their own internal operations. We developed a system as well as an analytic pipeline that provides a blueprint for how to make the operations of deep learning techniques comprehensible to human operators.

The explanation generation and re-ID pipelines are a work in progress and future versions of this approach would include an end-to-end attribute recognition, re-ID, and explanation generation model by fusing low-level features from the attribute recognition and re-ID branches into a recurrent explanation generator. A large scale human-user study is also underway to assess the pervasiveness and comprehensibility of inherence bias in the explanations generated by the system and to subsequently use the results of the study to improve explanation generation by the deep model.

Acknowledgements

This research was supported by National Research Council Postdoctoral Fellowships to EB and ZH, as well as a grant from the Office of Naval Research to WL and SK.

Appendix

Implementation Details

The networks were trained with similar parameters for fair comparison. The input batch size were 32 for all the experiments. The images were mean subtracted with PETA mean pixel values. The cost function defined in Equation 1 was optimized jointly for all the attributes using SGD optimization with learning rate schedule that starts out at 0.1 and dropped by a factor of 10 when learning flattens. The layers were highly regularized with an l_2 regularizers of value 0.0001 to combat over-fitting.

Ablation of Effects of Strategies

Here, we summarize the effect of various strategies that affect the attribute recognition performance. We show comparisons of different strategies based on ResNet34 as that allows fair comparisons with other types of networks such as inception (GoogleNet with 22 layers) that are used in [18] and [15].

Architectural Comparisons

We compare three deeper architectures: Inception version 3 (a variant of the 22-layer GoogleNet-based architecture), DenseNet34 (with 34 layers), DenseNet121 (121 layers), and ResNet34. Table 4 shows the benefits of residual blocks. ResNet34 performs slightly better than GoogleNet (Inception v3) and both outperforms the two variants of DenseNet (34 and 121 layers). It's important to note here that Inception will quickly overfit with deeper architectures, and with increased data augmentation it trains at least twice slower than ResNet34. As we go deeper without in these networks without additional strategies, the networks overfit too quickly as shown on the performance of DenseNet34 vs. DenseNet121.

Effects of Pooling and Classifiers Types

In practice, simple global average pooling (GAP) works well than complicated pooling strategies. As shown on Table 5, the simple logistic classifier outperforms the dense classifier for ResNet34. Generally, for shallower networks (ResNet34 and below), logistic classifier performs better than dense classifiers as the number of features pooled globally is limited (512 in the case of ResNet34). However, for deeper networks such as ResNet50 and above, the dense classifier works well as the number of globally pooled features is larger (2048 in the case of ResNet50).

Table 4. Architectural comparisons of multi-label attribute recognition on PETA.

Networks/Method	mA	Acc	Rec	Prec	F1	AUC (micro)
Inception v3	**84.77**	72.69	81.36	81.43	81.39	87.69
DenseNet34	81.91	67.76	78.17	78.18	78.17	85.56
DesneNet121	83.19	69.57	79.35	79.40	79.38	86.35
ResNet34	84.54	**72.94**	**81.57**	**81.73**	**81.65**	**87.81**

Table 5. Effect of classifier type on ResNet34 recognition performance on PETA.

Networks/Method	mA	Acc	Rec	Prec	F1	AUC (micro)
GAP + Logistic classifier	**84.59**	**72.87**	**81.39**	**81.54**	**81.46**	**87.77**
GAP + Dense classifier	84.35	71.83	80.71	80.79	80.78	87.29

Sample Weights vs. Probability Calibration

From Table 6, it is clear that the naive probability threshold with no sample weighting performed the worst. Next, the sample weighting together with F1 probability calibration performs better. The F1 probability calibration with no sample weighting achieved the best and well balanced performance among these comparisons.

Effects of Image Resizing and Data Augmentation

Other factors such as the way images are resized before being fed to the networks, data augmentation strategies and depth of the networks could affect performance. It is a combination of these factors together with the strategies discussed above that resulted in the state-of-the-art performance on PETA as shown in Table 7 and Table 1. Applying a random flip and RGB color jitter while preserving the aspect ratio were experimented with. Some network inputs are rigid and do not keep the aspect ratio of the input [9] and [15], requiring an image with an input size of 256x256 with crop size of 224x224. But, a dramatic performance increase can was obtained by preserving the aspect ratios of input images with random sampling of the largest dimension from a set. This results in an increase in example-based accuracy (almost 6%) and F1 score (about 4%).

Table 6. Effects of sample weighting and probability calibration strategies on ResNet34 performance on PETA

Networks/Method	mA	Acc	Rec	Prec	F1	AUC (micro)
DeepMar weights + F1 calibration	84.59	72.87	81.39	81.54	81.46	87.77
DeepMar weights + recall @20% FPR	**92.24**	65.73	**93.19**	67.59	78.35	86.60
No weighting/ naive proba thresh	83.07	72.25	79.47	82.95	81.17	87.19
No weighting/ F1 calibration	85.33	**73.54**	82.09	**82.22**	**82.16**	**88.23**

Table 7. Effects of image resizing (Size and Crop) and data augmentation (DA) on performance.

Networks/Method	mA	Acc	Rec	Prec	F1	AUC (micro)
Size: 256x256 Crop: 224x224 DA: flip, crop	84.59	72.87	81.39	81.54	81.46	87.77
Size: 300x300 Crop: 299x299 DA: flip, crop	84.54	72.94	81.57	81.73	81.65	87.81
Size: random AR: preserved DA: None	84.18	78.03	85.69	85.78	85.73	90.55
Size: random AR: preserved DA: flip, jitter, rot	83.19	77.16	84.96	85.12	85.04	90.03
Size: random AR: preserved DA: flip, jitter	**84.68**	**78.89**	**86.41**	**86.38**	**86.39**	**90.96**

References

[1] E. Bekele, C. Narber, and W. Lawson. Multi-attribute residual network (maresnet) for soft-biometrics recognition in surveillance scenarios. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 386–393. IEEE, 2017. 2, 3

[2] A. Cimpian and E. Salomon. The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, 37(5):461–480, 2014. 2

[3] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the ACM International Conference on Multimedia*, pages 789–792. ACM, 2014. 2, 4

[4] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3. Citeseer, 2007. 4

[5] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*, 2016. 3

[6] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016. 3

[7] S. S. Khemlani. Reasoning. In *Stevens Handbook of Experimental Psychology and Cognitive Neuroscience*, pages 385–429. S. Thompson-Schill (Ed.), 2018. 2, 6

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[9] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. 2015. 2, 4, 5, 7

[10] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016. 2, 3

[11] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 1, 3

[12] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. *arXiv preprint arXiv:1709.09930*, 2017. 4, 5

[13] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 2, 4, 5

[14] T. Miller, P. Howe, and L. Sonenberg. Explainable ai: Beware of inmates running the asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 36, 2017. 1, 3, 6

[15] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089*, 2017. 2, 3, 4, 5, 7

[16] P. Sudowe, H. Spitzer, and B. Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 87–95, 2015. 2, 5

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3

[18] K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv preprint arXiv:1611.05603*, 2016. 2, 3, 4, 5, 7

[19] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *Biometrics (ICB), 2015 International Conference on*, pages 535–540. IEEE, 2015. 2, 4