Human-level explanatory biases for person re-identification

Extended Abstract[†]

Esube Bekele*
Naval Research Laboratory
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

Despite the remarkable progress in deep learning in recent years, a major challenge for present systems is to generate explanations compelling enough to serve as useful accounts of the system's operations [1]. We argue that compelling explanations are those that exhibit human-like biases. For instance, humans prefer explanations that concern inherent properties instead of extrinsic influences. The bias is pervasive in that it affects the fitness of explanations across a broad swath of contexts [2], particularly those that concern conflicting or anomalous observations. We show how person re-identification (re-ID) networks can exhibit an inherence bias. Re-ID networks operate by computing similarity metrics between pairs of images to infer whether the images display the same individual. State-of-the-art re-ID networks tend to output a description of a particular individual, a similarity metric, or a discriminative model [3], but no existing re-ID network provides an explanation of its operations. To address the deficit, we developed a multi-attribute residual network that treats a subset of its features as either inherent or extrinsic, and we trained the network against the ViPER dataset [4]. Unlike previous systems, the network reports a judgment paired with an explanation of that judgment in the form of a description. The descriptions concern inherent properties when the network detects dissimilarity and extrinsic properties when it detects similarity. We argue that such a system provides a blueprint for how to make the operations of deep learning techniques comprehensible to human operators.

KEYWORDS

Inherence bias, re-identification, deep learning, explainable AI

1 INTRODUCTION

Many deep convolutional networks approximate human-level pattern recognition behavior [5]. A significant limitation of state-of-the-art deep learning systems is that few are able to 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 observe [1], the designers of such systems seldom consult results from the behavioral sciences on how humans generate and evaluate explanations. Miller et al. argue that progress on building explainable AI systems will be limited until they can recognize and adapt to those biases. Otherwise, they will produce descriptions that have limited explanatory value. To address the deficit, we developed a deep learning system that exploits a robust bias in explanatory reasoning. In what follows, we describe the bias and review its pervasiveness and importance. We then describe 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 the approach.

2 CONSTRUCTING BIASED EXPLANATIONS

Explanations help make a complex system transparent by highlighting its most pertinent components and causal dependencies. Hence, explanations are a form of dimensionality reduction, and a recent psychological proposal argues that the cognitive process of generating them operates heuristically to yield biased explanations, i.e., explanations that exhibit certain structural and semantic patterns over others. Humans exhibit a wide variety of explanatory biases: an explanation's simplicity, scope, and completeness affect whether it is considered good or bad [6]. 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. We argue that a 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.

2.1 The inherence bias

Recent research suggests that human reasoners perform a shallow search through the contents of their memory to explain a particular observation or regularity [2, 7]. 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. For instance, 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. The bias is pervasive: it affects how reasoners generate and evaluate explanations across a broad swath of contexts [2]. We show how the bias can help a certain class of convolutional neural network yield compelling explanations.

2.2 Re-identification and inherence

Person re-identification (re-ID) networks operate by extracting features between two input images and computing similarity metrics between those features to assess whether the images display the same individual. 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 [3, 7]. Re-ID networks tend to output a similarity metric, a discriminative model, or a description of a particular identified individual [3]. 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).

3 BUILDING AN EXPLANATORY RE-ID NETWORK

3.1 MAResNet

We developed a residual network, called MAResNet, that can recognize multiple attributes simultaneously in order to perform re-identification tasks [4]. Fig. 1 shows how the network operates: it takes two separate images as input and uses three residual blocks to learn 35 separate attributes jointly.

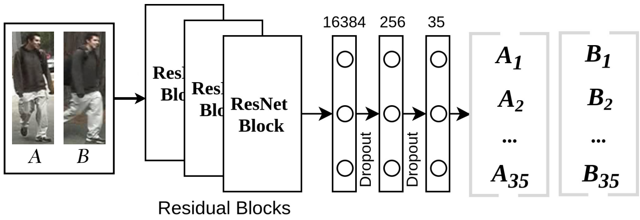

Figure 1: The MAResNet residual network applies residual block layers to extract 35 attributes from images taken from the VIPeR dataset [4, 8].

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).

3.2 Training and application

The network was trained against the VIPeR dataset [8], and its primary application consists in computing the cosine similarity of the resulting attribute vectors given two input images from the dataset. Critically, the network reports a judgment paired with an explanation (see Fig. 2). The explanations concern inherent properties when the network detects dissimilarity and extrinsic properties when it detects similarity.

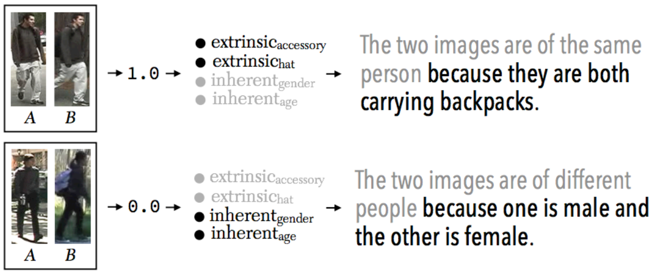
A
B
→ 1.0 →
● extrinsic_accessory
● extrinsic_hat
● inherent_gender
● inherent_age
→
The two images are of the same person because they are both carrying backpacks.
A
B
→ 0.0 →
● extrinsic_accessory
● extrinsic_hat
● inherent_gender
● inherent_age
→
The two images are of different people because one is male and the other is female.

Figure 2: MAResNet analyzes attributes from two images, which results in a cosine similarity score; our model builds an explanation relative to the similarity assessment.

4 CONCLUSIONS

In summary, 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, MAResNet, as well as an analytic pipeline that provides a blueprint for how to make the operations of deep learning techniques comprehensible to human operators.

ACKNOWLEDGMENTS

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

REFERENCES

[1] T. Miller, P. Howe, L. Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum. In *IJCAI-17 Workshop on Explainable AI*, 36.

[2] A. Cimpian, E. Salomon. 2014. 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.

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

[4] E. Bekele, C. Narber, W. Lawson. 2017. Multi-attribute residual network (MAResNet) for soft-biometrics recognition in surveillance scenarios. In *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition*, 386—393.

[5] Y. LeCun, Y. Bengio, G. Hinton. 2015. Deep learning. *Nature* 521, 7553, 436–444.

[6] S. Khemlani. 2018. Reasoning. In S. Thompson-Schill (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience.*

[7] E. Ahmed, M. Jones, T.K. Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 3908—3916.

[8] D. Gray, H. Tao. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision,* 262—275.