The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Examining CNN Representations with Respect to Dataset Bias

Quanshi Zhang,[†] **Wenguan Wang,**[†‡] **Song-Chun Zhu**[†]
[†]University of California, Los Angeles [‡]Beijing Institute of Technology

Abstract

Given a pre-trained CNN without any testing samples, this paper proposes a simple yet effective method to diagnose feature representations of the CNN. We aim to discover representation flaws caused by potential dataset bias. More specifically, when the CNN is trained to estimate image attributes, we mine latent relationships between representations of different attributes inside the CNN. Then, we compare the mined attribute relationships with ground-truth attribute relationships to discover the CNN's blind spots and failure modes due to dataset bias. In fact, representation flaws caused by dataset bias cannot be examined by conventional evaluation strategies based on testing images, because testing images may also have a similar bias. Experiments have demonstrated the effectiveness of our method.

Introduction

Given a convolutional neural network (CNN) that is pre-trained to estimate image attributes (or labels), how to diagnose black-box knowledge representations inside the CNN and discover potential representation flaws is a crucial issue for deep learning. In fact, there is no theoretical solution to identifying good and problematic representations in the CNN. Instead, people usually just evaluate a CNN based on the accuracy obtained using testing samples.

In this study, we focus on representation flaws caused by potential bias in the collection of training samples (Torralba and Efros 2011). As shown in Fig. 1, if an attribute usually co-appears with certain visual features in training samples, then the CNN may be learned to use the co-appearing features to represent this attribute. When the used co-appearing features are not semantically related to the target attribute, we consider these features as biased representations. This idea is related to the disentanglement of the local, bottom-up, and top-down information components for prediction (Wu, Xia, and Zhu 2007; Yang, Wu, and Zhu 2009; Wu and Zhu 2011). We need to clarify correct and problematic contexts for prediction. CNN representations may be biased even when the CNN achieves a high accuracy on testing samples, because testing samples may have a similar bias.

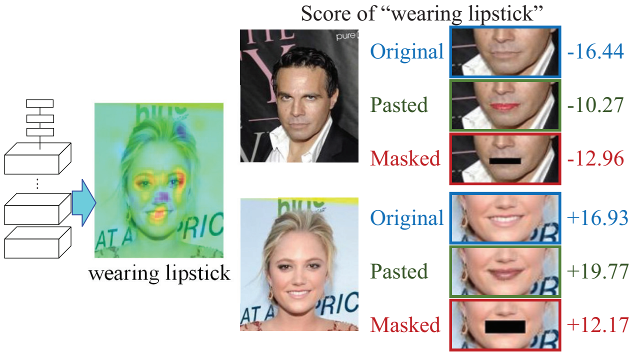

Figure 1: Biased representations in a CNN. Considering potential dataset bias, a high accuracy on testing images cannot always ensure that a CNN learns correct representations. The CNN may use unreliable co-appearing contexts to make predictions. For example, we manually modify mouth appearances of two faces by masking mouth regions or pasting another mouth, but such modifications do not significantly change prediction scores for the *lipstick* attribute. We show heat maps of inference patterns of the *lipstick* attribute, where patterns with red/blue colors are positive/negitive with the attribute score. The CNN mistakenly considers unrelated patterns as contexts to infer the lipstick. We propose a method to automatically discover such biased representations from a CNN without any testing images.

In this paper, we propose a simple yet effective method that automatically diagnoses representations of a pre-trained CNN without given any testing samples. *I.e.,* we only use training samples to determine the attributes whose representations are not well learned. We discover blind spots and failure modes of the representations, which can guide the collection of new training samples.

Intuition, self-compatibility of network representations: Given a pre-trained CNN and an image I, we use the CNN to estimate attribute A for I. We also mine inference patterns[1] of the estimation result, which are hidden in conv-

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] We regard a neural pattern as a group of units in a channel of a conv-layer's feature map, which are activated and play a crucial role in the estimation of the attribute A.

CelebA dataset				
		Accuracy in ordinary images	Accuracy in failure modes	Decrease of accuracy
top-5	Entropy-based	74.10	60.37	13.73
	Our method	73.81	40.22	**33.59**
top-10	Entropy-based	69.22	58.85	10.37
	Our method	72.29	46.43	**25.85**
top-15	Entropy-based	67.49	56.44	11.05
	Our method	68.05	47.95	**20.10**
top-20	Entropy-based	68.06	57.32	10.73
	Our method	66.94	46.57	**20.37**
top-25	Entropy-based	68.24	59.79	8.45
	Our method	67.06	49.23	**17.83**
SUN Attribute database				
top-40	Entropy-based	63.36	35.89	27.47
	Our method	68.65	38.98	**29.68**
top-50	Entropy-based	59.29	35.62	23.67
	Our method	65.73	38.86	**26.87**

Table 1: Average accuracy decrease caused by top-N failure modes, which were mined from the CNN for the CelebA dataset ($N = 5, 10, 15, 20, 25$) and the CNN for the SUN Attribute database ($N = 40, 50$). We compare the entropy-based method with our method.

and regards negative samples with $Y_2^* < 0$ as random samples without sharing common features. In this case, the conditional distribution $P(Y_1^*|Y_2^* > 0)$ will probably control the relationships between A_1 and A_2. Whereas, if the CNN mainly extracts features from negative samples with $Y_2^* < 0$ to represent A_2, then the attribute relationship will not be sensitive to the conditional distribution $P(Y_1^*|Y_2^* > 0)$.

Therefore, as shown in Fig. 8 and Table 1, our method is more effective in the discovery of failure modes than the method based on the entropy of annotation distributions.

Summary and discussion

In this paper, we have designed a method to explore inner conflicts inside representations of a pre-trained CNN without given any additional testing samples. This study focuses on an essential yet commonly ignored issue in artificial intelligence, *i.e.* how can we ensure the CNN learns what we expect it to learn. When there is a dataset bias, the CNN may use unreliable contexts to represent an attribute. Our method mines failure modes of a CNN, which can potentially guide the collection of new training samples. Experiments have demonstrated the high correlations between the mined KL divergences and dataset bias and shown the effectiveness in the discovery of failure modes.

In this paper, we used Gaussian distributions to approximate ground-truth distributions of attribute relationships to simplify the story. However, our method can be extended and use more complex distributions according to each specific application. In addition, it is difficult to say all discovered representation biases are "definitely" incorrect representations. For example, the CNN may use *rosy cheeks* to identify the *wearing lipstick* attribute, but these two attributes are "indirectly" related to each other. It is problematic to annotate the two attributes are either positively related or not related to each other. The *wearing necktie* attribute is directly related to the *male* attribute, but is indirectly related to the *mustache* attribute, because the necktie and the mustache describe different parts of the face. If we label *wearing necktie* is not related to *mustache*, then our method will examine whether the CNN uses mustache as contexts to describe the necktie. Similarly, if we consider such an indirect relationship as reliable contexts, we can simply annotate a positive relationship between *necktie* and *mustache*. Moreover, if neither the "not-related" relationship nor the positive relationship between the two attributes is trustworthy, we can simply ignore such relationships to avoid the risk of incorrect ground truth. In the future work, we would encode ground-truth attribute relationships as a prior into the end-to-end learning of CNNs, in order to achieve more reasonable representations.

Acknowledgement

This work is supported by ONR MURI project N00014-16-1-2007 and DARPA XAI Award N66001-17-2-4029, and NSF IIS 1423305.

References

Adler, P.; Falk, C.; Friedler, S. A.; Rybeck, G.; Scheidegger, C.; Smith, B.; and Venkatasubramanian, S. 2016. Auditing black-box models for indirect influence. *In ICDM*.

Aubry, M., and Russell, B. C. 2015. Understanding deep features with computer-generated imagery. *In ICCV*.

Bansal, A.; Farhadi, A.; and Parikh, D. 2014. Towards transparent systems: Semantic characterization of failure modes. *In ECCV*.

Deng, J.; Ding, N.; Jia, Y.; Frome, A.; Murphy, K.; Bengio, S.; Li, Y.; Neven, H.; and Adam, H. 2014. Large-scale object classification using label relation graphs. *In ECCV*.

Dosovitskiy, A., and Brox, T. 2016. Inverting visual representations with convolutional networks. *In CVPR*.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. *In CVPR*.

Fong, R. C., and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. *In arXiv:1704.03296v1*.

Goyal, Y.; Mohapatra, A.; Parikh, D.; and Batra, D. 2016. Towards transparent ai systems: Interpreting visual question answering models. *In arXiv:1608.08974v2*.

Hu, Z.; Ma, X.; Liu, Z.; Hovy, E.; and Xing, E. P. 2016. Harnessing deep neural networks with logic rules. *In arXiv:1603.06318v2*.

Koh, P., and Liang, P. 2017. Understanding black-box predictions via influence functions. *In arXiv preprint, arXiv:1703.04730*.

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

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Horvitz, E. 2017. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. *In AAAI*.