Explaining Contrasting Categories

Michael Pazzani, Amir Feghahati, Christian Shelton, Aaron Seitz
University of California, Riverside
Riverside, CA, United States
pazzani@ucr.edu, sfegh001@ucr.edu, cshelton@cs.ucr.edu, aseitz@ucr.edu

ABSTRACT
This paper describes initial progress in deep learning capable not only of fine-grained categorization tasks, such as whether an image of bird is a Western Grebe or a Clark's Grebe, but also explaining contrasts to make them understandable. Knowledge-discovery in databases has been described as the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1]. In spite of this, much of machine learning has focused on "valid" and "useful" with little attention paid to "understandable" [2- 6]. Recent work in deep learning has showed remarkable accuracy on a wide range of tasks [7], but produces models that are more difficult to interpret than most earlier approaches to artificial intelligence and machine learning. Our ultimate goal is to learn to annotate images to explain the difference between contrasting categories as found in bird guides or medical books.

Author Keywords
Explainable Artificial Intelligence, Machine Learning, Categorization, Deep Learning

ACM Classification Keywords
1.2.6 Artificial Intelligence: Learning (K.3.2).

HISTORY
The first author's research on learning explainable models from data started in the mid-1990s after interacting with doctors on models for medical diagnosis [2-5]. Although some have focused on which representation formalism is more "understandable" (e.g., [8]), the research has focused on how to constrain or bias an algorithm within a particular representation to produced results that are acceptable to human experts [6]. In this paper, we investigate how people explain contrasting categories and develop algorithms to create explanations of the category of objects in images (e.g., [9]). We focus not on explaining why an object belongs to a certain category, but rather why it belongs to that category and not a contrasting category. Figures 1 & 2 show examples of such explanations that people use to explain contrasting categories. Medical diagnosis is an area where explanations are of importance, particularly when treatments are risky or painful. For example, deep learning systems [10] have proven accurate at analyzing images to identify melanoma, but not at explaining the diagnosis in a way that gives patients or doctors confidence in following treatments. Figure 1 is from the web site. http://tiphero.com/skin-cancer/ and lists several signs of melanoma and provides examples of what patients and physicians look for and a model for what explainable learning should aspire to produce. Dermatology as well as histology and radiology [11] are examples where visual clues are important to initial differential diagnosis. Figure 2 shows an explanation from a bird web site on how to distinguish two varieties of grebes. We aspire for our deep learning algorithms to create explanations similar to those in figures. In the remainder of this paper, we concentrate on bird identification in the remainder of the paper due to the availability of large existing image datasets and the ease of finding amateur bird watchers. We first show that amateur bird watchers attend to the distinguishing characteristics, known in the bird watching community as "field marks," such as those in Figure 2. Next, we, describe an extension to a deep learning system that automatically identifies field marks. We leave it to future research on how to describe field marks.

BIRD IDENTIFICATION: A PILOT STUDY.
We prepared images of 12 birds divided into 6 sets of contrasting birds (e.g., Spotted Towhee and Eastern Towhee). Images were shown to four experienced bird watchers who were asked a yes-or-no question about the bird identification (e.g., "Is this a Spotted Towhee?"). Using an eye tracking system, we recorded the parts of the image that received attention. Figure 3 shows an example of where one subject focused on the wing of eastern towhee and spotted towhee, two similar pictures distinguished in part due to the spots on the wing. In contrast to distinguish a Clark's Grebe from a Western Grebe that subject concentrated on the area around the eye and the bill.

Although suggestive of how bird watchers learn and attend to field marks to distinguish similar species, the data is preliminary and requires more subjects and statistical tests. Experiments are planned to show two contrasting images simultaneously to experienced bird watchers and to track attentional changes in novices as they learn to identify birds.

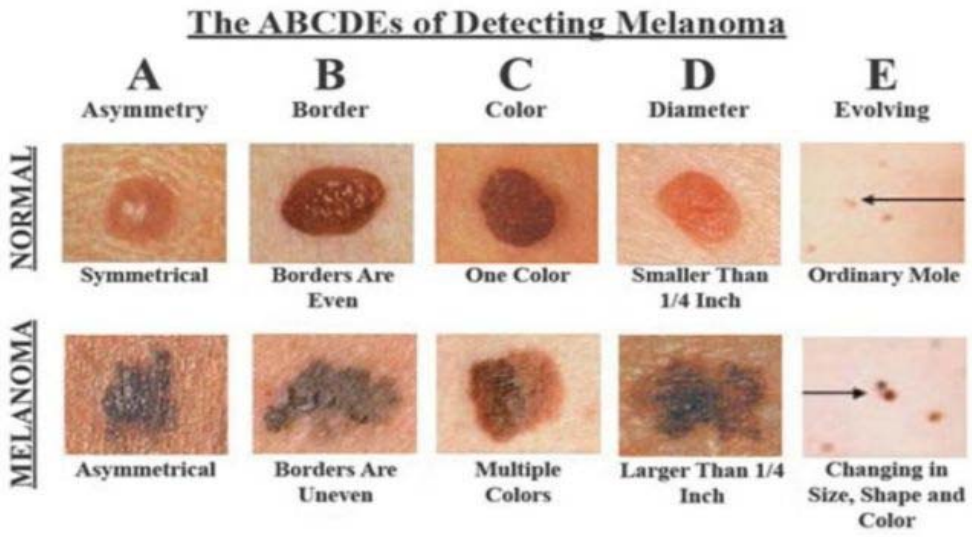

Figure 1. Explanation of how to differentiate moles from melanoma.

Figure 2. Explanation of how to distinguish a Clark's grebe from a Western grebe.

Figure 3. Eye tracking data shows an experienced bird watcher concentrates on the wing to distinguish a spotted towhee (left) from an eastern towhee (right)

Figure 4. Eye tracking data shows an experienced bird watcher concentrates on the eye to distinguish a Western grebe (top) from a Clark's grebe (bottom).

LEARNING DISCRIMINATIVE REGIONS OF CONTASTING CATEGORIES

Deep learning for image classification has shown great results [12, 13], surpassing the previous best computer vision systems. Although classification is an interesting and challenging problem, we want to go one step further and augment the deep network to create a *contrasting visual explanation.* This explanation identifies the regions of an image that discriminate the selected category from the second most likely category. In bird identification, these regions should correspond to the field marks described in bird guides and the areas that bird watchers focus on when identifying bird species.

We demonstrate the approach with a fine-grained classification task of birds. We seek to find out which regions of a bird image are the most important to distinguish it from images of the most similar contrasting class of birds. To do so, we train a known deep network [12] on the bounding boxes of a frequently used birds dataset [15]. The network has been chosen because of its simplicity. Our proposed method is not dependent on any specific network architecture. As contrasting visual explanation, we will highlight those sections of the input image that were most important to the network for distinguishing the two classes: that of the output class and that of the next closest class. The next closest class is chosen since the network has the most difficulty in distinguishing them. As a simple extension, it is possible for the user to ask for an explanation with respect to any other category.

To find these image regions, we backward propagate an output vector consisting of +1 for the most likely class and -1 for the next most likely class. This propagation is similar to the "backprop" neural network algorithm used to train deep networks, but in this case applied for explanation and not for training network weights.

This process identifies the most important pixels, i.e., those for which changes will cause the network to assign the given image to the most-similar (i.e., the next most likely) class instead of the correct class. This raw set of pixels is too sporadically distributed to provide a human-consumable explanation. To get larger coherent regions, we convolve windows with different sizes on the image and record the maximum change in each window. The windows with the maximum changes are the regions that are mostly contributing to misguide the network and the most important regions to explain the features that discriminate the contrasting categories. The process is depicted in Figure 5.

As one example, the network correctly identified a test image as a cerulean warbler and the second most likely classification as a black throated blue warbler Figure 6 (left) highlights the regions of the image that were found to be most important in distinguishing the two classes including the eye and throat. Figure 6 (right) is an image of a black throated blue warbler which shows the difference in throat and neck.

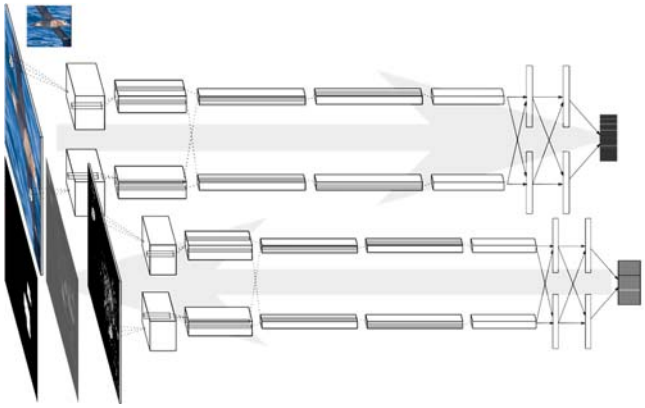

Figure 5. (top) Standard forward propagation. (bottom) Our backward propagation, starting with a vector difference of the best class from the second-best class, passing through the network to produce the derivative of the input plane with respect to this class difference, and then convolution and region finding to identify the most important regions of the image. Note that while not shown, the lower propagation depends on the upper, as the derivatives are at the points defined through the upper propagation.

Figure 6. An image of a cerulean warbler (left) with highlighting indicating the regions that distinguish it from a black-throated blue warbler (right).

To illustrate the impact of having a contrasting category, Figure 7 shows the regions that contribute most to producing the correct category without regard to finding the difference between a contrasting categories. (i.e. not propagating -1 for the next most similar class). Note that neither the throat nor eye are highlighted.

Figure 7. Important regions in categorizing a cerulean warbler without the contrasting categories.

CONCLUSION

We have begun to explore how machine learning may emulate how humans explain contrasting categories. Preliminary data show the features that experienced bird watchers use to differentiate contrasting categories. A network architecture learned similar features. In future work, we will explore how to label differentiating features using techniques similar to [9] with an ultimate goal to automate explanations similar to those found in bird guides and medical books.

ACKNOWLEDGEMENTS

This was developed with funding from the DARPA Explainable AI Program under a contract from NRL. The views, opinions, and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the DoD or the U.S. Government.

REFERENCES

1. U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," Advances in Knowledge Discovery and Data Mining, U.M. Fayyad et al., eds., AAAI/MIT Press, Menlo Park, Calif., 1996, pp. 1–34. 2.
2. Pazzani, M. (2000). Knowledge discovery from data? IEEE Intelligent Systems 15(2): 10-13 (2000)
3. M. J. Pazzani, S. Mani, W. R. Shankle (2001). Acceptance of Rules Generated by Machine Learning among Medical Experts. Methods of Information in Medicine; 40: 380-385.
4. Pazzani, M., Mani, S. & Shankle, W. R. (1997). Beyond concise and colorful: learning intelligible rules. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, CA. AAAI Press, 235-238.
5. Pazzani, M., Mani, S. & Shankle, W. R. (1997). Comprehensive knowledge-discovery in databases. In M. G. Shafto & P. Langley (Ed.), Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, pp. 596-601. Mahwah, NJ:Lawrence Erlbaum.
6. Pazzani, M. J. & Bay, S. D. (1999). The Independent Sign Bias: Gaining Insight from Multiple Linear Regression. In Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society.
7. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
8. B. Gaines, "Transforming Rules and Trees into Comprehensible Knowledge Structures," Advances in Knowledge Discovery and Data Mining, U.M. Fayyad et al., eds., MIT Press, Cambridge, Mass., 1996, pp.205–226.
9. Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating Visual Explanations. n: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9908. Springer, Cham
10. Premaladha, J., and K. S. Ravichandran. "Novel approaches for diagnosing melanoma skin lesions through supervised and deep learning algorithms." Journal of medical systems 40.4 (2016): 96.
11. Geoffrey R. Norman, Donald Rosenthal, Lee R. Brooks, Scott W. Allen, Linda J. Muzzin. The Development of Expertise in Dermatology. Arch Dermatol. 1989;125(8):1063–1068.
12. A. Krizhevsky, I. Sutskever, I., & G. E. Hinton,(2012) ImageNet classification with deep convolutional neural net-works. In NIPS 2012
13. K. He, X. Zhang, S. Ren, & J. Sun (2015). "Deep residual learning for image recognition," arXiv:1512.03385, 2015
14. Wah, C. and Branson, S. and Welinder, P. and Perona, P. & Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.