HuTCH: Human Teachable Concept Highlighter for Post-hoc Visual Explanations

Erfan Mirhaji[1][0009−0002−9404−0750], Nikhil Krishnaswamy[1][0000−0001−7878−7227], Jill Zarestky[2][0000−0003−1728−1796], Lisa Mason[3][0009−0004−8847−5920], Sarath Sreedharan[1][0000−0002−2299−0178], and Nathaniel Blanchard[1][0000−0002−2653−0873]

[1] Department of Computer Science, Colorado State University, USA
{erfan.mirhaji, nikhil.krishnaswamy, sarath.sreedharan, nathaniel.blanchard}@colostate.edu
[2] School of Education, Colorado State University, USA
jill.zarestky@colostate.edu
[3] Colorado State University Extension, USA
lisa.mason@colostate.edu

Abstract. In recent years, deep learning (DL) models have surpassed human experts in a variety of tasks. However, when it comes to educational contexts, human experts possess something these high-performing models currently lack—the ability to provide clear and approachable explanations tailored to human learners. Efforts to develop explainable methods for these models have typically been designed for DL-experts rather than learners. In this study, we propose a novel approach that combines high-performance models with the teachability of human expert explanations. Specifically, we focus on generating post-hoc, human-understandable explanations for key expert-defined concepts on a novel task: training citizen scientists to identify pollinators from images. Our method, HuTCH, transforms the representational space and highlights relevant segments of an image for learners based on essential concepts — for example, automatically highlighting hair in an image as a key concept for bee identification. We evaluate HuTCH's performance by comparing against traditional saliency maps and expert annotations, and show that HuTCH concepts better align with expert annotations. The proposed framework bridges the gap between models' accuracy and human-teachable features, contributing to the advancement of explainable AI for use in pedagogy.

Keywords: Explainable AI · Concept-Based Teaching · Concept Highlighting · Feature Localization.

1 Introduction

In recent years, deep learning has made rapid advances. We currently have deep learning systems that can match or even surpass human-level performance in

many tasks. Concurrently, there have been various efforts within the Explainable AI (XAI) field to generate explanations for why the system made specific decisions [5, 9, 15]. Unfortunately, these systems still do not provide a source of lessons for teaching non-expert learners about the target task. This is in no small part because most existing explanations are designed to help AI or domain experts debug and verify these systems. Nonetheless, if we are able to capitalize on these models to generate easy-to-understand lessons about tasks these systems excel at, this could open novel opportunities to automatically generate lessons for many learning problems. Towards this end, assuming access to domain experts, we develop an entirely new XAI paradigm that combines expert-specified rules with internal representations used by AI systems to automatically generate task-specific lesson content.

Here, we explicitly focus on a visual learning task, namely identifying different pollinator groups from their pictures. Convolutional Neural Networks (CNNs) can be trained to solve this task with a near-expert level accuracy. We introduce a novel method called HuTCH, which extends existing XAI methods, to generate lessons as to why the current image belongs to a specific group, which could be easily followed by learners with no familiarity with AI concepts or expertise in entomology. Specifically, HuTCH leverages concepts and rules used by domain experts to provide a conceptual explanation as to why the insect in the picture corresponds to a specific pollinator group. HuTCH will make use of internal representations learned by the CNN to determine what rules and concepts are applicable for the given image. We then make it further accessible to non-domain experts by illustrating visually each concept used in the rule to make the determination.

2 Related Work

Feature-attribution methods, for instance, gradient-based methods, have been popular in generating explanations for vision models [9]. Saliency maps are one example of such methods [20]. Other methods like LIME [18] and Shapley Additive Explanations (SHAP) [15] have been introduced to generate explanations. Even though many methods besides LIME and SHAP have been established, they can generate misleading explanations on the true reason for a prediction by a model [2, 3, 7, 26]. These methods are designed for domain experts who can associate meaningful concepts with visual features highlighted in the explanation.

Concept-based explanations are another type of post-hoc explanations. In a concept-based explanation, explanations go beyond features of each image and identify higher-level human-understandable concepts that are true for the entire dataset [8]. Concept Activation Vectors (CAVs) are another method used in concept-based explanations to provide clarification about a neural net's internal state in terms of human-friendly concepts [13]. Because of the human-explainability focus of these works, they are easier to understand by humans. TCAV and its explainability implications have been widely used in medical applications [6, 16, 22]. Some other methods suggest configuring the architec-

ture of the model to achieve Post-hoc Concept Bottleneck models (PCBMs) [25]. These explanability methods also assume the familiarity of the user with the defined concepts. If a layperson is presented with these explanations, they might not understand them, as a familiarity with said concepts is implicitly expected by concept-based explanations. Our Human Teachable Concept Highlighter (HuTCH) method produces explanations based on human-understandable concepts with the aid of a field expert, and further highlights the region containing the concept. We separate the decision-making process of the model and the explanations that are generated by our model, which are highly teachable and interpretable to learners.

3 Methodology

In this paper, we highlight a region of an input image containing a teachable concept to learners based on information from a human expert using the HuTCH framework. HuTCH's highlights are more similar to that of human expert's, indicating its applicability as a tutor for teaching visual tasks.

3.1 Concepts, Datasets and Model

A visual identification task pertinent to teaching citizen scientists is the classification of bees versus wasps. Entomologists look for specific characteristics when encountering an insect that is either a bee or a wasp; for example, hair on the body, or the narrowness of the abdomen and thorax connection, commonly known as a "pinch-waist". These two are some of the most important concepts differentiating bees and wasps. These concepts are suitable for citizen scientists to aid them in this specific visual identification task. As a result, we tuned our explainability method to these two features.

The data used in this study is gathered from the iNaturalist website [1]. We chose bee species that display hair on their body, and wasp species that have the pinch-waist feature. In total we selected 23 bee species and 17 wasp species, and 226,892 images of bees and 229,921 images of wasps were collected. To highlight regions containing a certain concept using our HuTCH method, we defined a concept dataset with 200 bee and 200 wasp images. A human expert selected one region in each image that contains the corresponding concept, and one region where the concept is absent, resulting in two datasets: one for each concept, containing 200 positive and 200 negative examples. Next, we augmented our dataset, resulting in 9,600 positive and 9,600 negative samples.

We used 60 bee images and 60 wasp images for our comparisons. First, the original images are resized to 224 by 224 and then segmented using the Mask R-CNN ResNet-50 object instance segmentation method [11] with zero threshold to find the largest segment. The results are then given to the expert, the saliency highlighter, and our two HuTCH highlighter methods, each outputting their highlights. For the CNN model, we chose the ResNet-152 architecture [12], and the model was re-trained as a whole, accounting for mean and standard deviation

values of the training dataset. The trained ResNet-152 CNN model was tested on 34,022 bee and 34,481 wasp images, achieving a combined F_1 score of 96%.

3.2 Concept Activation Vectors

CNNs transform images to higher representational spaces and extract many complex features. In the case of ResNet-152, we use the output of the second convolutional layer of the third bottleneck layer of the fourth residual stage, which contains 512 features each with a spatial dimension of 7×7, our layer of interest. We then pass each of our concept datasets to the model and train a linear classifier on the flattened activations of our layer of interest. The vector containing weights of this linear classifier is the Concept Activation Vector. Later on, using the dot product of the flattened activations of a new image and the CAV of a concept, we can calculate the alignment of that image with that concept.

3.3 Comparison with Human Expert and CNN Highlights

Using the selected 120 test images, we ask the expert, the saliency map method, and our HuTCH method to highlight what each considers the most important region. The human expert highlights the relevant concept region in each image; in the case of a bee, the hair region, and in the case of a wasp, the pinch-waist region. The expert highlights only one continuous region. The CNN model, using the normalized gradients with a threshold of 0.05, calculates the saliency map of

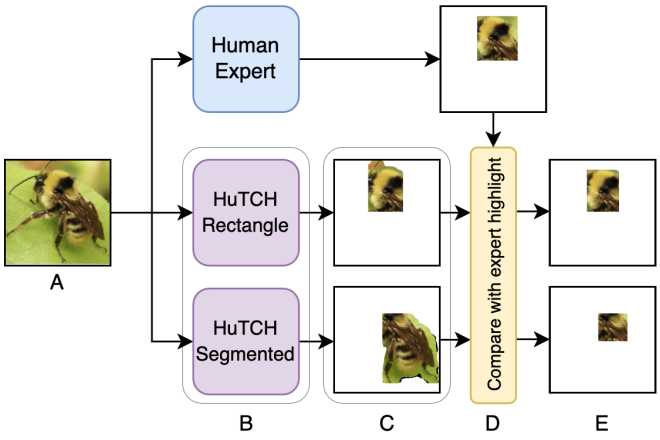

Fig. 1. The workflow of highlighting images by each method. **A** Each input image is filtered to the biggest object segment. **B** The two HuTCH methods partition the image into sub-images via rectangles and masks respectively, and calculate the sub-images with the highest CAV scores. **C** The combined region of top K sub-images is highlighted by each method. **D** The highlights are compared to the expert's highlighted region via IoU and Dice metrics. **E** The overlapping region is shown.

each image. The acquired saliency map of each image is then further partitioned into rectangles. Each map is partitioned into rectangles of widths and heights of 56, 74 and 112, and are saved separately to create new images referred to as sub-images. The average gradient value of all sub-images is then measured and sorted. The sub-images corresponding to the top K highest average gradient values are overlapped and combined, which is the highlighted region of the saliency map.

3.4 HuTCH Highlights

Our HuTCH method highlights regions in which a certain concept is most present. First, each image is passed to the CNN model to make a prediction of whether it is a bee or a wasp. Then, the concept to look for is set to hair or pinch-waist respectively; this is done to combine the accuracy of CNNs with the teachability of expert-defined concepts. We then further segment the image with a 0.05 threshold to filter out the background from the insect. Next, using two different methods, HuTCH further partitions the image; once using rectangles, called HuTCH Rectangle, and another time using all the segments generated by the Mask R-CNN segmentation method [11], called HuTCH Segmented.

In the HuTCH Rectangle, similar to the saliency partitioning, the image is partitioned into all possible rectangles with sides of 56, 74, and 112 pixels, referred to as sub-images. This dissects every single image into two datasets, one dataset of sub-images of HuTCH Rectangle, one dataset of sub-images of HuTCH Segmented. Finally, each sub-image of each HuTCH dataset is passed through the model to acquire the activation from the layer of interest. The dot products of the flattened activations and the CAV of the respective concept are then measured and sorted [13]. The sub-images corresponding to the K highest dot products are merged together, producing the highlighted region of each HuTCH method.

Table 1. The average and standard deviation of IoU scores between highlighted regions of each method and expert highlights.

Method	**Top 1**	**Top 2**	**Top 3**
Saliency Rectangle Highlight	0.155 ± 0.166	0.189 ± 0.164	0.209 ± 0.164
HuTCH Segmented Highlight	0.237 ± 0.165	0.237 ± 0.124	0.238 ± 0.112
HuTCH Rectangle Highlight	0.275 ± 0.211	0.309 ± 0.190	0.307 ± 0.161

Table 2. The values for Dice coefficient between highlighted regions of each method and expert highlights.

Method	**Top 1**	**Top 2**	**Top 3**
Saliency Rectangle Highlight	0.237 ± 0.288	0.288 ± 0.220	0.317 ± 0.216
HuTCH Segmented Highlight	0.356 ± 0.212	0.368 ± 0.159	0.372 ± 0.140
HuTCH Rectangle Highlight	0.388 ± 0.263	0.439 ± 0.230	0.446 ± 0.198

4 Results and Discussion

In order to compare the highlighted region from each method (Saliency Rectangle, HuTCH Rectangle and HuTCH Segmented) with the highlighted region by the expert, we used two metrics; Intersection over Union (IoU) and Dice coefficient. The HuTCH Rectangle and HuTCH Segmented methods are compared to the region that the expert has marked as relevant for learners to correctly identify the insect. The overall workflow can be seen in Figure 1. The results of the average and standard deviations of comparisons can be viewed in Tables 1 and 2 on the aggregated top K regions respectively.

The concept that the expert highlights is something that is understandable by humans and is used to teach about different taxa in entomology. Compared with salience maps, HuTCH methods more accurately align with expert highlights of key visual concepts for learners. By combining the teachability of our methods and the high accuracy of CNNs, we can teach learners with high certainty and high explainability. In this work, the HuTCH framework bridges the gap between AI models' accuracy and human-teachable features, contributing to the incorporation of explainable AI in teaching. We have demonstrated that concept-based highlighting achieves more alignment with expert highlighting and thus improves the teachability of the explanations. Our method demonstrates how AI models' high performance can be coupled with human-understandable interpretability. This explainability approach can be used as an enhancement for educational settings, complementing educators or fully taking the expert's place for visual teaching tasks.

5 Limitation and Future Work

In our segmentation method, we used the mask R-CNN ResNet-50 model, which is trained on Microsoft's COCO dataset; sometimes unrelated background patches found their way into our sub-images, which resulted in erroneous CAV scores. More fine-tuned and specialized segmentation methods that can better segment bees and wasps into different sections would help us work with finer concepts used by human experts. Another limitation was the quality of the dataset as they were captured by amateurs and enthusiasts, sometimes lacking focus or visibility of insects. A cleaner dataset would improve the accuracy of both classification and concept highlights. Currently, we have only used two concepts, body hair and pinch-waist, for training the CAVs and subsequently the linear classifier. We also only trained the CNN on a few bee and wasp species exhibiting these features. For future work, we can expand the concepts and the training dataset to include more features and more species showing those features.

Acknowledgments.

This study was funded by NSF grant 2303019.

References

1. inaturalist. Available from https://www.inaturalist.org, accessed: January 2025
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
3. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6541–6549 (2017)
4. Blanchard, E.G., Mohammed, P.: On cultural intelligence in llm-based chatbots: Implications for artificial intelligence in education. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 439–453. Springer Nature Switzerland, Cham (2024)
5. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8) (2019). https://doi.org/10.3390/electronics8080832, https://www.mdpi.com/2079-9292/8/8/832
6. Clough, J.R., Oksuz, I., Puyol-Antón, E., Ruijsink, B., King, A.P., Schnabel, J.A.: Global and local interpretability for cardiac mri classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 656–664. Springer (2019)
7. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 3681–3688 (Jul 2019). https://doi.org/10.1609/aaai.v33i01.33013681, https://ojs.aaai.org/index.php/AAAI/article/view/4252
8. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
9. Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., Giannotti, F.: A survey of methods for explaining black box models. ACM Computing Surveys **51** (02 2018). https://doi.org/10.1145/3236009
10. Gütl, C., García-Barrios, V.M.: The application of concepts for learning and teaching. In: Proceedings of 8th International Conference on Interactive Computer Aided Learning (ICL 2005). Citeseer (2005)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Kim, B., Wattenberg, M., Gilmer, J., Cai, C.J., Wexler, J., Viégas, F.B., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International Conference on Machine Learning (2017), https://api.semanticscholar.org/CorpusID:51737170
14. Linson, A., Xu, Y., English, A.R., Fisher, R.B.: Identifying student struggle by analyzing facial movement during asynchronous video lecture viewing: Towards an automated tool to support instructors. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 53–65. Springer International Publishing, Cham (2022)

15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
16. M., G., V., A., S., M.M., H., M.: Concept attribution: Explaining cnn decisions to physicians. Computers in Biology and Medicine **123**, 103865 (2020). https://doi.org/https://doi.org/10.1016/j.compbiomed.2020.103865, https://www.sciencedirect.com/science/article/pii/S0010482520302225
17. Ma, Q., Shen, H., Koedinger, K., Wu, S.T.: How to teach programming in the ai era? using llms as a teachable agent for debugging. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 265–279. Springer Nature Switzerland, Cham (2024)
18. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778
19. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization (2016), https://api.semanticscholar.org/CorpusID:217472699
20. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Workshop at International Conference on Learning Representations (2014)
21. Sonkar, S., Liu, N., Mallick, D.B., Baraniuk, R.G.: Marking: Visual grading with highlighting errors and annotating missing bits. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 309–323. Springer Nature Switzerland, Cham (2024)
22. Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A.: Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis **79**, 102470 (2022)
23. Wang, Y., Zhang, T., Guo, X., Shen, Z.: Gradient based feature attribution in explainable ai: A technical review (2024), https://arxiv.org/abs/2403.10415
24. Yao, Y.: Concept formation and learning: a cognitive informatics perspective. In: Proceedings of the Third IEEE International Conference on Cognitive Informatics, 2004. pp. 42–51 (2004). https://doi.org/10.1109/COGINF.2004.1327458
25. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480 (2022)
26. Zhou, Y., Booth, S., Ribeiro, M.T., Shah, J.: Do feature attribution methods correctly attribute features? Proceedings of the AAAI Conference on Artificial Intelligence **36**(9), 9623–9633 (Jun 2022). https://doi.org/10.1609/aaai.v36i9.21196, https://ojs.aaai.org/index.php/AAAI/article/view/21196