Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Explainability via Interactivity? Supporting Nonexperts' Sensemaking of Pretrained CNN by Interacting with Their Daily Surroundings

Chao Wang, Pengcheng An

2020

Preprint:

Copyright ACM 2020. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in CHI PLAY 2021, https://doi.org/https://doi.org/10.1145/3450337.3483487.

Explainability via Interactivity? Supporting Nonexperts' Sensemaking of pre-trained CNN by Interacting with Their Daily Surroundings

Chao Wang
chao.wang@honda-ri.de
Honda Research Institute Europe
Offenbach, Hessen, Germany

Pengcheng An
pengcheng.an@uwaterloo.ca
David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada

ABSTRACT

Current research on Explainable AI (XAI) heavily targets on expert users (data scientists or AI developers). However, increasing importance has been argued for making AI more understandable to nonexperts, who are expected to leverage AI techniques, but have limited knowledge about AI. We present a mobile application to support nonexperts to interactively make sense of Convolutional Neural Networks (CNN); it allows users to play with a pre-trained CNN by taking pictures of their surrounding objects. We use an up-to-date XAI technique (Class Activation Map) to intuitively visualize the model's decision (the most important image regions that lead to a certain result). Deployed in a university course, this playful learning tool was found to support design students to gain vivid understandings about the capabilities and limitations of pre-trained CNNs in real-world environments. Concrete examples of students' playful explorations are reported to characterize their sensemaking processes reflecting different depths of thought.

CCS CONCEPTS

• **Human-centered computing → User interface programming**.

KEYWORDS

Explainable AI, Class Activation Map, Mobile Application, Convolutional Neural Networks

ACM Reference Format:
Chao Wang and Pengcheng An. 2021. Explainability via Interactivity? Supporting Nonexperts' Sensemaking of pre-trained CNN by Interacting with Their Daily Surroundings. In *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '21), October 18–21, 2021, Virtual Event, Austria.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3450337.3483487

1 INTRODUCTION

With its advantages of solving computer vision problems such as object detection or image classification, Convolutional Neural Networks (CNN) are wildly applied in various domains e.g., autonomous driving, healthcare and robotics. Especially after the AI community open-sourced frameworks such as TensorFlow [1], Caffe [8] or PyTorch[12], non-expert users, such as user experience designers, healthcare practitioners or even AI amateurs, could affordably utilize, or adapt a pre-trained CNN model (e.g., VGG-19 [15]) for their own pragmatic purposes, which is expected to democratize CNN models for real-world usage. However, CNN is a complex type of deep learning models which normally consist of multiple convolutional, pooling and fully connected layers. For non-expert users who have limited knowledge about AI, it is rather difficult for them to make sense of how a pre-trained CNN works and what is its capabilities and limitations in real-world contexts. This creates obstacles for them to utilize pre-trained models as "building blocks" [11] to create their own domain-relevant applications. Therefore, non-experts were often baffled when model performance did not fit their expectations, and hence might abandon their tasks [20]. Nowadays, machine learning experts can use various Explainable Artificial Intelligent (XAI) approaches, such as Class Activation Maps [9] or Deconvnet [22] to interpret the internal state of the CNN and reason the characteristics of the trained model. But such techniques are mainly developed for professional machine learning engineers, whose knowledge level and pragmatic goals differ from non-expert users. Aiming to explore how to help non-experts to intuitively understand pre-trained CNN models, we design a playful tool that allows users to apply a CNN model to interact with their daily surroundings. This application has been deployed in a university course to help 30 design students (who have zero or little experience with machine learning) to interactively make sense of CNN and understand its capabilities and limitations. We built this application as both a novel interface to evaluate and a technology probe to research with, in order to understand how non-experts' process of sensemakingcould be scaffolded, and thereby inform related design and research in the future. Via semi-structured interviews with a few students who particpated in the courase, we gathered their experiences of playing with the tool in their daily surroundings, including vivid examples reflecting how their sensemaking processes were facilitated. Accordingly, three types of sensemaking activities that could be supported by the application were generalized, reflecting different depth of thought in the students: (1) collection of "impressive" or "funny" results, (2) inferences of how environmental factors influenced the results, and (3) experimentation on-the-spot to test assumptions. We thereby contribute a playful tool, as well as empirical implications of helping nonexperts to make sense of pre-trained CNN models, via interactions with their everyday environments.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI PLAY '21, October 18–21, 2021, Virtual Event, Austria
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8356-1/21/10.
https://doi.org/10.1145/3450337.3483487

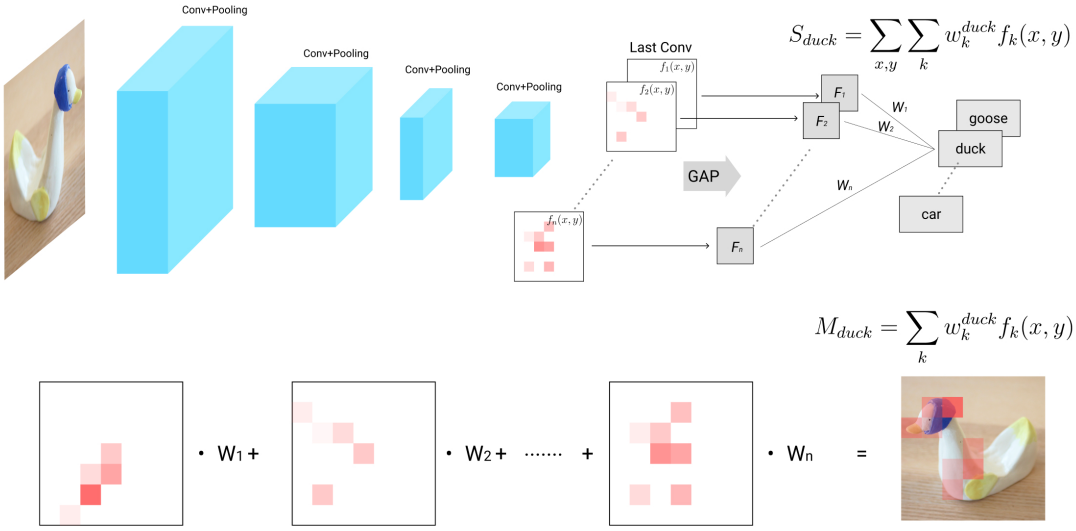

Figure 1: Visualizing heatmap with Class Activation Mapping [23]. $f_k(x, y)$ represents the activation of unit k in the last convolutional layer at spatial location (x, y). After performing global average pooling, unit k becomes F_k. Then, for a certain prediction result, $w_1, w_2 \ldots w_n$ are the weights for F_k to calculate the softmax input S (e.g., S_{duck}). As F_k is the result of the global averaged pooling, the corresponding weight also indicates the importance of unit k. Thus, the aggregated heatmap (M_{duck}) casts the activation of the last convolutional layer, hence indicating the most important regions (red squares) that have made the CNN to output a certain prediction result (e.g., "duck").

2 RELATED WORK

2.1 XAI for non-expert users

There is not a unitary definition of Explainable Artificial Intelligence (XAI). As Arrieta et al. [3] put, "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand". Accroding to this interpretation, who the audience is hence should be the first question to ask when developing an XAI system, and both the purposes and the strategy of the explanation can vary with varying audience. However, most of the existing XAI approaches, by default, target on expert users: e.g., AI specialists or data scientists. Increasingly, researchers point out the necessity of supporting non-experts to practically understand AI models [20][11]. Our approach focuses on such non-experts (namely designers), who are expected to benefit from utilizing pre-trained AI model in their domain-specific tasks [18][19][5], but only have limited knowledge about AI.

2.2 Approaches of enhancing explainability of CNN

Regarding the HOW in XAI, Guidotti et al [7] and Arrieta et al [3] suggested a general distinction between transparent models and post-hoc explainability. Post-hoc explainability is pertinent to models that are by design, not readily interpretable, namely the "black-box" models. One of the most focused topics of black-box XAI is to help people understand Convolutional Neural Network (CNN), a widely used model for computer vision problems (e.g., image classification or object detection). CNN contains a series of convolutional layers and pooling layers to automatically learn features incrementally from low to high levels. The functioning and architecture of CNN is extremely complex and difficult to explain. Fortunately, the road to explainability for CNN has become much easier, after researchers found the Activation Map approach, which leverages human brains' intrinsic skills of processing visual data [3]. This approach visualizes the decision process of CNN by propagating the output to the input space to depict which parts of the input were "important" for the output (see Figure 4). A large body of research has investigated this approach. One of the influential works is proposed by Zeiler et al [22], who tried to use Deconvnet [21] to reconstructs the maximum activations. Through a saliency map, human can get an intuitive image about which parts of the image significantly contribute to the activations. However, these methods are not class discriminative. Later, Zhou et al. explored generating class activation maps (CAM) by modifing CNN though global average pooling [9], which visualizes the CNN's prediction by indicating the discriminative image regions [23]. Building further upon their work, Selvaraju et al. [14] presented Gradient-weighted Class Activation Mapping (Grad-CAM), which does not require any modification in the network archetecture to show important part of the input in respect to the prediction. In general, highlighting the most discriminative area upon the input image is an important approach for explain and evaluate a pre-trained CNN model. Based upon these works, we have relied on a similar technique of visualizing the activation of CNN, which could intuitively help non-experts understand how the a CNN generates classification results to .

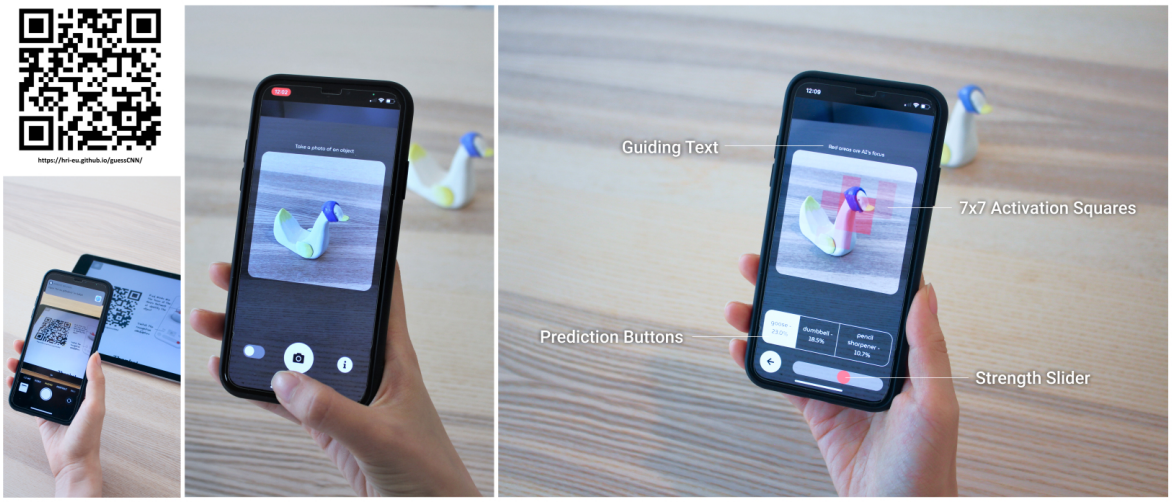

Figure 2: Left: log in via the QR code (currently available). Middle: user takes photo of an object. Right: afterwards, user can see the predictions (three results with highest confidence) and their corresponding Class Activation Maps.

2.3 Platforms of XAI for CNN

Although the topic of XAI for CNN is gaining increasing attention, most the of the tools are built for machine learning experts. Few platforms help non-expert users to develop practical and intuitive understandings about CNN. Yet one example is from the PAIR team of Google, called Embedding Projector [17], which can visualize the features of CNN model in a 3D space after reducing dimensions by T-SNE. Input images of various layers are laid out in a 3D array in its GUI, to illustrate how the input is processes layer by layer. Although granting a global view of CNN, it does not mean to explain how a prediction is generated. Mishra et al [11] proposed an interactive transfer learning tool for non-experts. User can drag and drop a grey square on any part of an image and observe the influence of the occlusion on the prediction result. Another similar study is ClickMe.ai, a large-scale online experiment conducted by Linsley et al [10]. Authors made an interface for human participants to select important image parts with their mouse. Meanwhile, these parts are shown to CNN incrementally until the network recognize the object. Later, Demidov [6] created a website, which allows user to input an online image to see a corresponding activation heat map through CAM. Although these works inspired us to explore XAI through interactivity, they were all developed for desktop environments and used online images as input source. By contrast, we present a mobile tool that helps users understand CNN's capabilities and limitations by freely interacting with their real-world surroundings.

3 SYSTEM DESIGN

Our mobile application enables non-expert users to take photos of surrounding objects and visualizes the CNN's decision about a classification result using a Class Activation Map approach. Namely, it highlights the discriminative image regions (i.e., the red blocks as seen in Figure 4), to intuitively show users which regions of the input image have made the model to decide that the image is about a certain object class. To guarantee the ease of access, we built the tool using a web-based machine learning toolkit, TensorFlow.js [16], so that it does not require any installation on users' mobile device. A pre-trained CNN model, MobileNet [2], is integrated in the tool, which is trained upon the ImageNet database (with 10 million labelled images) to recognize over 1000 object categories. This enables users to open-endedly apply the CNN with different kinds of daily objects, and have meaningful recognition results for them to make sense of the CNN's functioning in real-world settings.

3.1 System interaction

Users can visit the URL address (https://hri-eu.github.io/guessCNN/) or simply scan the QR code (see Figure 2, currently available) to access our tool. At the beginning, users need to give the permission for the application to use the back camera. At the same time, the pre-trained CNN model is downloaded with the progress shown. After the model is downloaded, the buttons are turned to available status. When the "i" button on the bottom right is clicked, the introduction of the application and a simple pictorial tutorial will be shown on a pop-up overlay page; User can aim to an object and take a picture by clicking bottom middle, then the app will start applying the CNN to recognize the taken picture (see Figure 2 Right). Via the toggle button on the bottom left, user can go back to the camera mode to take new pictures.

3.2 Generating Class Activation Map

The interface will show top 3 prediction results (with highest confidence) from the overall output by the pre-trained model (Figure 2 Right). The model was modified through Class Activation Map (CAM) approach to generate heatmap (Figure 4). Namely, the last fully connected layer of the CNN is replaced by the average pooling

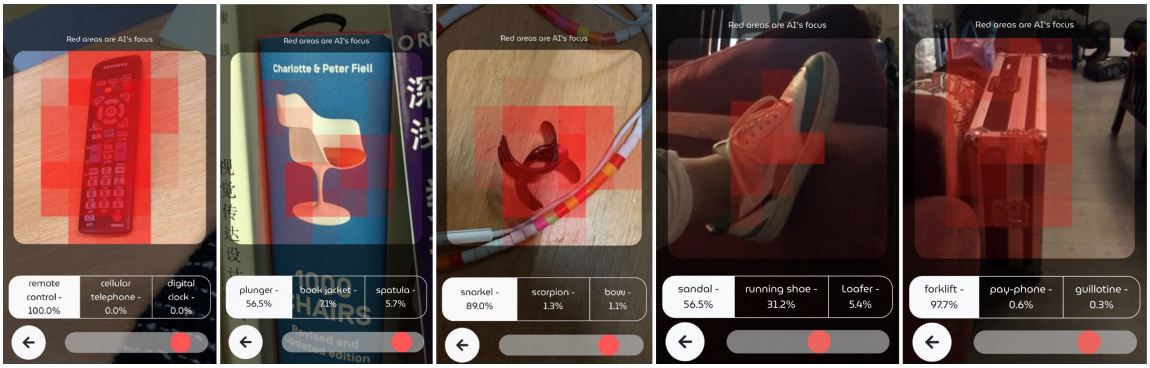

Figure 3: A few examples collected by the students when playing with their daily surrounding, to understand the functioning of the CNN model, from left to right: remote TV control (Jill), book (Sam), hair clip (Jill), sport shoe (Ada), suitcase (Ada).

layer. 7x7 red squares indicate the discriminative image regions which contribute most to the result of the classification. Users could toggle among the 3 results to see corresponding CAM. A slider on the bottom can modify the thresholds of the activation map display, to adjust the relative intensity and visibility of the squares.

4 STUDY SETUP

The interface was implemented as part of a university course module named "prototyping with AI", involving 30 students (9 master students, 21 bachelors) in an industrial design department. Before this module, these students did not take courses encompassing the topic of artificial intelligence, and they had no prior experience with developing or implementing CNN models. This interface was meant as a playful tool for the design students to get familiarized with the pre-trained CNN model, so that later they were prepared to apply a transfer learning technique to modify a CNN model to perform their own customized tasks (using the Teachable Machine [4] platform developed by Google). Our mobile application were introduced to the students in the beginning of the module via an online video lecture (because of the covid-19 pandemic). This video included examples of how to use this interface. After the introduction, an assignment was given in the video, which asked the students to use this application to interact with their daily surroudnings and try to practically make sense of the functioning of the CNN model in their own environment. The students were encouraged to record important observations via screenshots, and share interesting examples with each other. To gather empirical implications, three students participated in in-depth interviews to talk about their experience with the application, especially concrete examples of how they used the application to interact with their daily surroundings and make sense of the classification results. These interviews were subjected (along with the related screenshots sent by the students) to an affinity diagram analysis.

5 FINDINGS AND IMPLICATIONS

In general, the students had rather positive experience with the application, which they considered to be both fun to play with and useful for understanding how the CNN model works. We now further address their experience and sense-making process using concrete examples identified by themselves, which have been categorized into three main clusters. These examples illustrate how students' sensemaking of the CNN model could be supported in different depths. We refer to the interviewed students as Jill, Ada, and Sam.

Collection of "impressive" or "funny" classification results in the surrounding. The students all mentioned that they enjoyed the playful process of exploring the mobile application with various daily objects, and collecting interesting cases in which the classification results seemed extra impressive or funny to them. They felt that interacting with surrounding objects provided them with more vivid understandings of the model's capabilities and limitations in real-world situations. For instance, one of the examples collected by Jill is a remote (TV) control which was correctly classified by the model with 100%-symbol confidence (see Figure 3). In the introductory lecture, it was mentioned that the web interface could classify certain internet images with 100% confidence, because the image might be part of the training dataset of the model. Therefore, it was noticeable to her that a randomly taken picture of her remote control could also reach 100% confidence. Apart from such impressive classification results, the students also enjoyed collecting the funny results. For example, Sam shared a case that the model classified an image of a chair on his book cover as a plunger (56.5% confidence). This feels quite funny to him, but also rather understandable, due to the visual similarity of the two. As explicated by Jill, such collection of impressive and funny examples helped her to better understand the capability and limitations of such CNN models, and think further about what they can do in the real-world contexts.

Inferences of how the classification results were influenced by environmental factors. Beyond getting a feel for the model's capabilities and limitations, examples shared by the students also revealed the inferences they made about how certain classification results were influenced by environmental factors (e.g., background, or adjacent objects). For example, Jill collected a case showing that how another object next to the target would influence the classification results: her hair clip was classified as a snorkel when put next to some cables. As she reasoned, the is because the cable looked similar to a snorkel tube and thus made the combination identified as snorkel. In another example, Ada found that her shoe was

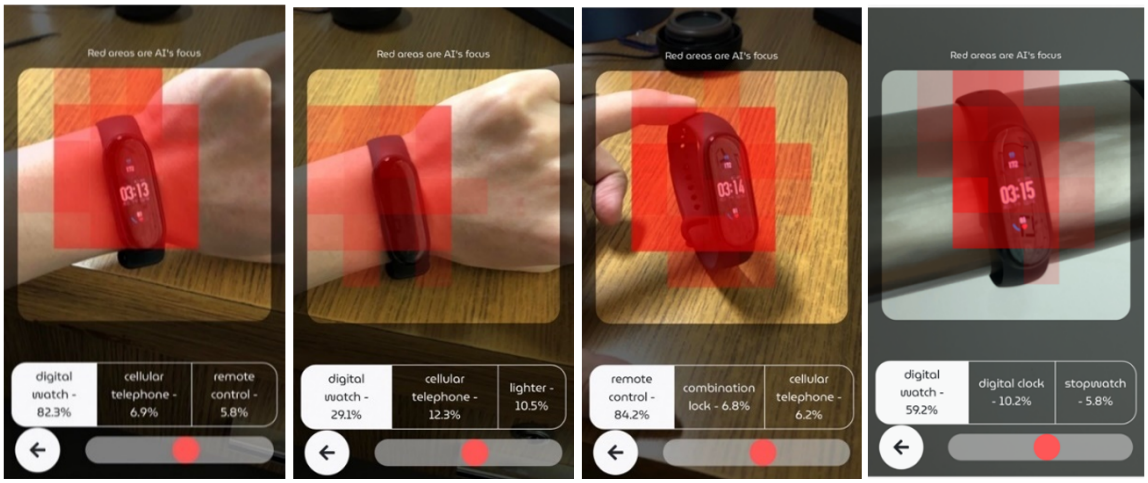

Figure 4: On-the-spot experimentation conducted by Sam to test his hypotheses about what factors may increase the confidence of the pre-trained CNN model recognizing his smartwatch, from left to right: with screen on and worn on the wrist; screen off worn on the wrist, screen on and taken off the wrist, screen on and put on a wrist-like object.

classified as more likely to be a sandal (56.5%) than a running shoe (31.2%). And as she inferred, this is due to that the color division of the shoe pattern (see Figure 3) made the model think that its upper part and lower part were two different objects. And the poor lighting of the environment also contributed to the confidence of classifying the shoe as a sandal. As shown above, the examples in this category suggest that the students' sense-making went beyond simply accumulating case-by-case knowledge of what worked and what did not with the CNN model. In many cases, they also tried to make inferences about why the CNN model generated certain (unexpected) results in given situations. Such inferences echoed what D.A. Schöne referred to as "theory-in-practice" [13], which manifested the practical understandings established by the students to interpret how the model worked in their environment.

Experimentation on the spot with surrounding objects to test assumptions. In addition to making inferences, there are also examples in which the students actively experimented with the CNN model to test their assumptions, e.g., by taking pictures of a target from different perspectives, or in combination with different objects. For instance, by experimenting with the mobile app, Ada found that the CNN model "doesn't seem to understand perspective". This was learnt by taking pictures of her suitcase from multiple perspectives: while the pictures that captured the front view (the largest surface) of the suitcase were correctly classified, the side view of her suitcase was mistakenly classified as a forklift. Another telling example is from Sam. He performed a series of experiments with his smartwatch, to test his assumptions about what features make the CNN model more confident that a target is a digital watch. His first hypothesis was that the appearance of a clock face screen would add to the confidence. To verify this, as Figure 4 shows, he took two pictures from the same perspective, which the only difference being showing/dimming the clock face screen. As a result, the one with the screen dimmed led to lower confidence (29.1%) than the other with the clock face screen (82.3%), Moreover, Sam hypothesized that being attached to an arm also contributes to the confidence of a target being a digital the screen dimmed led to lower confidence (29.1%) than the other with the clock face screen (82.3%), Moreover, Sam hypothesized that being attached to an arm also contributes to the confidence of a target being a digital watch. Hence, he did the next experiment by taking another picture of the smartwatch taken off his arm, from the similar perspective as the last pictures. The result supported his hypothesis: it was classified as a remote control instead of a digital watch. He further hypothesized that putting the watch on a cylinder object might also work like putting it on a real arm, so he attached the smartwatch to his bottle (see Figure 4), which indeed verified his assumption. As recognized by the students, such mentioned experimentations were another advantage of the mobile interface, since in the web interface, it would take more efforts to perform such controlled comparisons, Examples from this category therefore have shown that the mobile interface did not only supported the students' exploration and interpretation, but also enabled them to conduct on-the-spot experimentations to assess their practical assumptions about how the CNN model works.

6 CONCLUDING REMARKS

In this work, we present a mobile tool that enables nonexperts (design students as an example) to interactively make sense of a pre-trained CNN model. Using a Class Activation Map approach, the tool intuitively visualizes for users how the CNN model makes certain predictions about a picture they take from their daily surroundings. The empirical findings confirmed its usefulness in terms of scaffolding nonexperts' practical understandings encompassing the functioning of the CNN model. Rich examples from the students' playful explorations surface three types of sensemaking processes that differ in the depths of thought. First, students used the tool to collect impressive or funny prediction results that helped them understand the capabilities and limitations of the model in real-world scenarios. Second, some (peculiar) results triggered the students' further inferences about how real-world environmental factors (e.g.,

lighting, perspective, adjacent object) could influence the prediction results. Lastly, the tool enabled the students to conduct on-the-spot experimentations to test their established understandings by creating variations in multiple pictures and comparing the results.

ACKNOWLEDGMENTS

We would like to appreciate Evgeny Demidov for making the code of MobileNet surgery open-source.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and Michael Isard. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 265–283.

[2] TensorFlow authors. 2021. MobileNet. , https://github.com/tensorflow/tfjs–models/tree/mas pages.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 arXiv:1910.10045

[4] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable machine: Approachable Web-based tool for exploring machine learning classification. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[5] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 278–288.

[6] Evgeny Demidov. 2019. Interactive Heat map demo. , https://www.ibiblio.org/e–notes/ml/heatmap.htm pages. https://www.ibiblio.org/e-notes/ml/heatmap.htm

[7] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A survey of methods for explaining black box models. *arXiv* 51, 5 (2018), 1–42. arXiv:1802.01933

[8] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM '14)*. Association for Computing Machinery, New York, NY, USA, 675–678. https://doi.org/10.1145/2647868.2654889

[9] Min Lin, Qiang Chen, and Shuicheng Yan. 2014. Network in network. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* (2014). arXiv:1312.4400

[10] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. 2019. Learning what and where to attend. *7th International Conference on Learning Representations, ICLR 2019* (2019), 1–21. arXiv:1805.08819

[11] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Designing Interactive Transfer Learning Tools for ML Non-Experts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445096

[12] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[13] Donald A Schon. 1984. *The reflective practitioner: How professionals think in action*. Vol. 5126. Basic books.

[14] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (2020), 336–359. https://doi.org/10.1007/s11263-019-01228-7 arXiv:1610.02391

[15] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015). arXiv:1409.1556

[16] Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Ann Yuan, Nick Kreeger, Ping Yu, Kangyi Zhang, Shanqing Cai, Eric Nielsen, and David Soergel. 2019. Tensorflow. js: Machine learning for the web and beyond. *arXiv preprint arXiv:1901.05350* (2019).

[17] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg. 2016. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469* (2016).

[18] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 585–596.

[19] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[20] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 573–584.

[21] Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, and Rob Fergus. 2010. Deconvolutional networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2528–2535. https://doi.org/10.1109/CVPR.2010.5539957

[22] Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus. 2011. Adaptive deconvolutional networks for mid and high level feature learning. *Proceedings of the IEEE International Conference on Computer Vision* (2011), 2018–2025. https://doi.org/10.1109/ICCV.2011.6126474

[23] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), 2921–2929. https://doi.org/10.1109/CVPR.2016.319 arXiv:1512.04150