EXTENDED-ABSTRACT

A Mobile Tool that Helps Nonexperts Make Sense of Pretrained CNN by Interacting with Their Daily Surroundings

CHAO WANG, Honda Research Institute Europe GmbH, Offenbach, Hessen, Germany

PENGCHENG AN, David R. Cheriton School of Computer Science, Waterloo, ON, Canada

Published: 27 September 2021

MobileHCI '21: 23rd International Conference on Mobile Human-Computer Interaction
September 27 - October 1, 2021
Toulouse & Virtual, France

Conference Sponsors:
SIGCHI

https://doi.org/10.1145/3447527.3474873
ISBN: 9781450383295

A Mobile Tool that Helps Nonexperts Make Sense of Pretrained CNN by Interacting with Their Daily Surroundings

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

Pengcheng An
David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada

ABSTRACT

Current research on explainable AI (XAI) is primarily aimed at expert users (data scientists or AI developers). However, there is an increasing emphasis on making AI more understandable to non-experts who are expected to use AI techniques but have limited knowledge about AI. We propose a mobile application to help non-experts understand convolutional neural networks (CNN) in an interactive way; it allows users to taking pictures of surrounding objects and use pre-trained CNN to recognize it. We use the latest XAI (Class Activation Map) technology to visualize the model decision (the most important image area leading to a specific result). This playful learning tool was implemented in college courses and found to help design students gain a vivid understanding of the functions and limitations of pre-trained CNN in the real world. We thereby contribute an online tool that could be used for twofold purposes: first, it could help non-experts interactively learn how a pre-trained CNN works. Second, it can be used by researchers to probe and characterize the non-experts' process of sensemaking, which could contribute insights into explainable AI design beyond expert users.

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. A Mobile Tool that Helps Nonexperts Make Sense of Pretrained CNN by Interacting with Their Daily Surroundings. In *Adjunct Publication of the 23rd International Conference on Mobile Human-Computer Interaction (MobileHCI '21 Adjunct), September 27-October 1, 2021, Toulouse & Virtual, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3447527.3474873

MobileHCI '21 Adjunct, September 27-October 1, 2021, Toulouse & Virtual, France

ACM ISBN 978-1-4503-8329-5/21/09.
https://doi.org/10.1145/3447527.3474873

1 INTRODUCTION

Convolutional Neural Networks (CNN) are wildly used for object classification in various domains e.g., autonomous driving, healthcare and robotics. With emerging open-sourced toolkits (e.g., TensorFlow by Google), non-expert users, such as user experience designers, are expected to more easily access, or adapt a pretrained CNN model for their own pragmatic purposes. However, CNN is a complex type of deep learning models which normally consist of multiple convolutional, pooling and fully connected layers. Non-expert users normally have limited knowledge about AI, which makes it very difficult for them to make sense of how a pretrained CNN works and what its capabilities and limitations are in real-world contexts. This creates obstacles for them to utilize pretrained models as "building blocks" [6] 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 [9]. Nowadays, machine learning experts can use various Explainable Artificial Intelligence (XAI) approaches, such as Class Activation Maps [5] or Deconvnet [10] to interpret the internal state of the CNN and reason the possible problems of the trained model. But such techniques are developed for professional machine learning engineers, whose knowledge level and pragmatic goals differ from non-expert user. Aiming to explore how to help non-experts to practically understand pretrained 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 practically make sense of CNN and understand its capabilities and limitations. We see the application as a technology probe to surface how non-experts' sense-making process could be scaffolded, and gather implications for future design of similar tools. In this demo, we present the design rationale of this playful mobile application in light of the prior research and design regarding explainable CNN techniques and tools. Subsequently, we detail the design elements and interaction flow of this application, with providing a functioning link (and QR code) to access the complete system. An ongoing evaluation of this tool has been conducted to explore how nonexperts' sensemaking processes were supported. The goal is to contribute both design of and insights into interactive, playful XAI systems targeted on nonexperts, for both AI education and democratization.

2 XAI FOR CNN

It seems that there is not yet a common definition of Explainable Artificial Intelligence (XAI). We would refer to the definition that summarised prior research by Arrieta et al in [2]: "Given an audience, an explainable Artificial Intelligence is one that produces

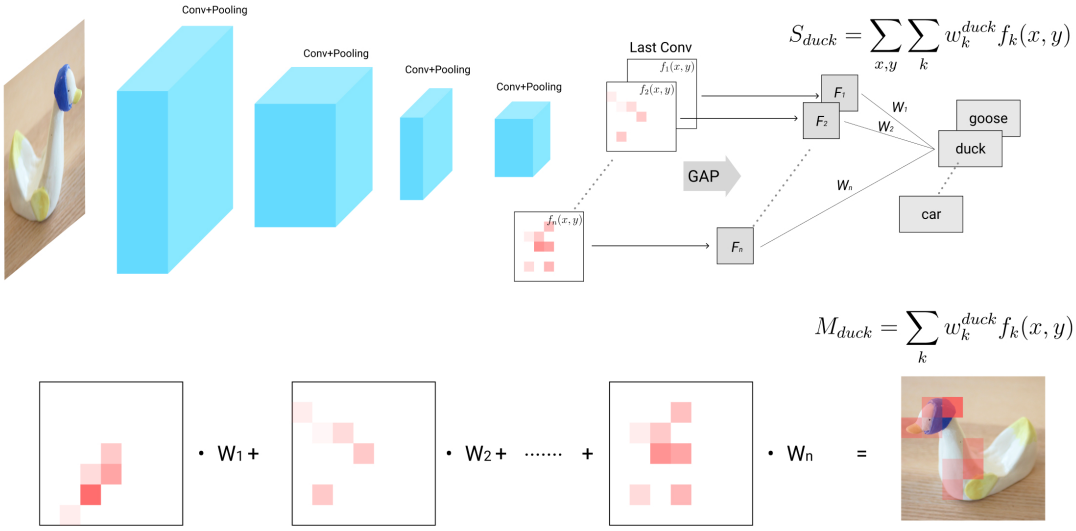

Figure 1: Visualizing discriminative regions with Class Activation Mapping [11]. fk(x, y) represents the activation of unit k in the last convolutional layer at spatial location (x, y). Fk is the result of performing global average pooling for the unit k. Then, for a certain prediction result, w1, w2 . . . wn are the weights for Fk to calculate the softmax input S (e.g., Sduck). As Fk comes from the global averaged pooling, the corresponding weight also indicates the importance of unit k. Thus, the aggregated heatmap (Mduck) reflects 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).

details or reasons to make its functioning clear or easy to understand". From the definition, one crucial consideration of XAI would be the audience, as the purposes of the explanation can vary with different audience. However, most of the existing XAI approaches are developed for expert users like AI specialists or data scientists. Recently, some researchers point out the necessity of supporting non-experts to practically make sense of AI models and proposed some related ideas [9][6]. Our approach focuses on such non-expert users (designers), who are expected to benefit from utilizing pretrained CNN models in their domain-specific tasks, but only have limited knowledge about AI.

Regarding "how" to enhance the explainability by XAI techniques, Guidotti et al [4] and Arrieta et al [2] suggested a clear distinction between transparent models and post-hoc explainability. Post-hoc explainability targets models that are not readily interpretable by design, also called black-box models, such as Support Vector Machines (SVM). One of the most focused topics of black-box XAI is to investigate the approaches for interpretation of Convolutional Neural Network (CNN), which is a widely used model aiming for computer vision problems, such as image classification or object detection. CNN consists of a sequence of convolutional layers and pooling layers to automatically learn features increasingly from low to high levels. The structure of CNN is extremely complex and its internal logic is difficult to explain. Fortunately, the road to explainability for CNN is even easier than other deep-learning models, because researchers found an approach which can match the intrinsic skills of processing visual data by human's brain [2]: explaining the decision process of CNN by propagating the output to the input space to visualize which parts of the input were "important" for the output (see figure 1). This approach has been investigated by researchers extensively. One of the seminal works is proposed by Zeiler et al [10]., who tried to use Deconvnet [10] 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. modified CNN though global average pooling [5] to generate class activation maps (CAM), which indicates the discriminative image regions used by CNN to identify the prediction [11]. Based on their work, Selvaraju et al. [7] further proposed Gradient-weighted Class Activation Mapping (Grad-CAM), which is also able to show important part of the input respect to the prediction, but without any modification in the network architecture. In general, highlighting the most discriminative area upon the input image is an important approach for explain and evaluate a pretrained CNN model. Inspired by these works, our solution utilizes a similar approach to visualize the activation of CNN to more intuitively show non-experts how the classification results are generated.

3 SYSTEM DESIGN

In this demo section, we propose a mobile application, which enables non-experts to take photos of surrounding objects and visualizes the discriminative image regions for CNN to identify the object through Class Activation Map approach. For maximumly clearing the obstacles using the application, it does not require user to install any package on their smartphone. Thus a website for mobile phone is established and a web-based machine learning technique, TensorFlow.js [8] is applied. To generate heatmaps, this

application modifies a pretrained CNN MobileNet [1] model trained to recognize 1000 image classes. Non-expert can use the APP by logging on the website with the browser of their smartphone.

As our prior work, a desktop web interface similar to Demidov's application [3] was created for users to play with a pretrained CNN with online images. Through the web page, user can paste any online image to the slot and check the activation map (Figure 2). However, through informal evaluation with participants, we realized the restriction that users could only input online images to the application could considerably limit users' understandings about how a CNN model would work in real-world environments. For this reason, we decided to develop a mobile tool that would enable users to apply the pretrained CNN with pictures they take on-the-fly, in their own environment, so that they could have richer possibilities for explorations and establish more vivid understandings about CNNs.

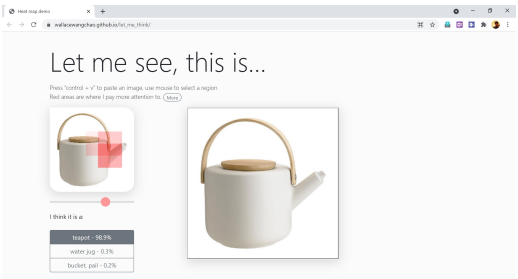

Figure 2: Prior work: a desktop interface for exploring pretrained CNN with online images (https://wallacewangchao.github.io/let_me_think/, current available).

3.1 Interaction Workflow

User can visit the address (https://hri-eu.github.io/guessCNN/) or simply scan the QR code (see Figure 3) to access to the website by their browser. At the beginning, user need to give the permission for the webpage to use the back camera (Figure 4). At the same time, the pretrained CNN model is downloading and the progress is shown. After the model is downloaded, the buttons turn to available status. When clicking "i" button on the bottom right, an overlay page will pop-up to show the introduction of the application and a simple tutorial through a picture; User can aim to an object and take a picture by clicking bottom middle, then the app will jump to the recognition page. By switching the toggle button on the bottom left, user can choose which mode the app will jump into.

3.2 Generating Class Activation Map

The interface will show 3 highest prediction results of the image of object the pre-trained CNN model (VGG-19) identified. The model was modified through Class Activation Map (CAM) approach to generative heatmap. The last fully connected layer is replaced by the average pooling layer. 7x7 red squares indicate the discriminative image regions which contributes to the result of the classification. By clicking the toggle buttons, 3 predictive results can be switched. User can also use slider with the red dot on the bottom to

Figure 3: User simply logs in via the QR code (currently available).

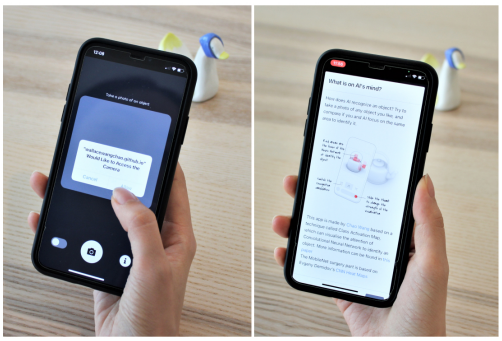

Figure 4: Left: User gives authority to the app to access the camera. Right: instruction of the app.

increase/decrease the thresholds of displaying the activation map, to adjust the intensity of the squares.

3.3 The scribbling Mode

The scribbling mode allows user to compare the most important feature area they assume with the discriminative regions of the CNN (Figure 6). This way, they could compare their own thoughts with the CNN's perception, regarding which areas contribute most to a prediction result. In this mode, the interface only shows the highest predictive result and hides the CAM. Users need use finger to scribble the important part of the image which contributes to the result based on their thought. The user interface adds a 13% transparent blue square on each scribbled area. After user click "compare" button, the app will calculate the total amount of the scribbled "coating" and adjust the threshold of the CAM to display same amount of red square "coating" on the image. By comparing the overlapping area, the guessing score and an "emoji" will be shown to user. The higher the score user gets, the more nervous emoji will be shown, to provide an anthropomorphic feedback of CNN.

4 CONTRIBUTION AND FUTURE WORK

To help people make sense of the decision-making process of CNN, a web-based application that allows users to identify their surrounding objects were created based on the Class Activation Map (CAM)

Figure 5: Left: user can take photo of any object. Right: The predictions and corresponding CAMs is generated according to the image.

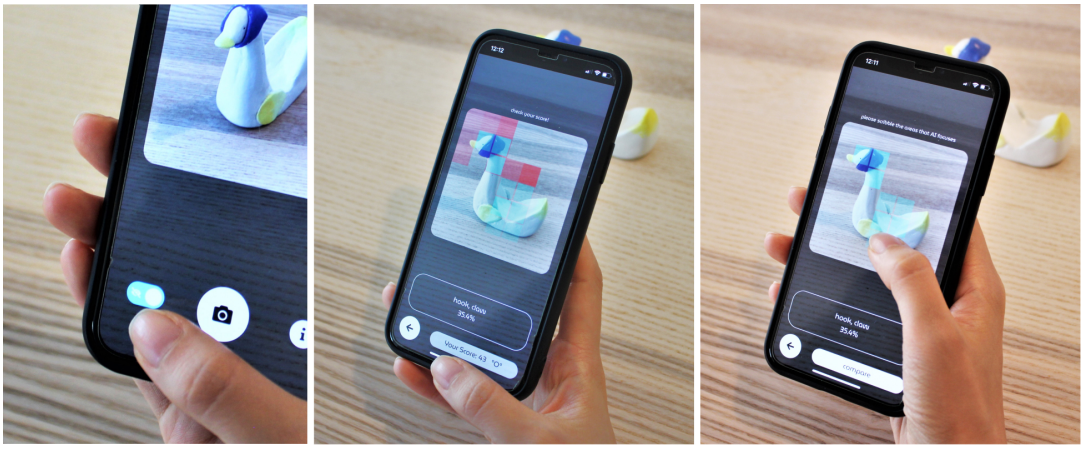

Figure 6: The scribbling Mode

technique. Comparing with other XAI solutions, this tool has the following advantages: Firstly, current XAI solutions focus on the expert users, how to benefit a broader range of professions and how to support nonexperts' learning and exploration is an urgent challenge. This tool could help us probe the implication of this mission. Secondly, current XAI techniques focus on the visibility and explainability of the model, but fewer work has been done on understanding the sense-making process of humans. With the tool, we hope to gather empirical insights into this question. Lastly, this tool could also be used to gather user inputs via real-world cases for improving AI systems.

ACKNOWLEDGMENTS

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

REFERENCES

[1] TensorFlow authors. 2021. MobileNet. https://github.com/tensorflow/tfjs--models/tree/mas

[2] 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

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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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).

[9] 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.

[10] 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

[11] 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