Wizard of Oz Prototyping for Machine Learning Experiences

Jacob T. Browne
Twickle
San Diego, CA
jake@twickle.co

ABSTRACT

Machine learning is being adopted in a wide range of products and services. Despite its adoption, design and research processes for machine learning experiences have yet to be cemented in the user experience community. Prototyping machine learning experiences is noted to be particularly challenging. This paper suggests Wizard of Oz prototyping to help designers incorporate human-centered design processes into the development of machine learning experiences. This paper also surfaces a set of topics to consider in evaluating Wizard of Oz machine learning prototypes.

1 INTRODUCTION

Machine learning (ML) allows a system to perform a task based upon examples of how to perform the task [12]. ML has the potential to dramatically alter the experiences of our products and services. Designing and implementing ML remains a contemporary challenge for designers and ML practitioners [33].

Designers aren't currently involved in the early design of ML experiences [34]. This contrasts with the typical role of the designer, where designers are typically responsible for user feedback and insights throughout development. In contemporary ML development processes, the end user's ability to affect the resulting model is limited, manifesting most strongly in Interactive Machine Learning (iML) systems [1]. In iML, the user tunes the system with their activity [1]. The training

CHI'19 Extended Abstracts, May 4-9, 2019, Glasgow, Scotland, UK.

ACM ISBN 978-1-4503-5971-9/19/05.
DOI: https://doi.org/10.1145/3290607.3312877

KEYWORDS

Machine Learning; Wizard of Oz; Prototyping; User Experience; Design; User Interface

of the model is framed as an HCI task and the user iteratively refines the model [1, 11]. This process assumes that the model selected is the best one chosen, the mechanisms in which the model receives input is the best way to receive input, and that users understand how the model works. Waiting on feedback from users once the model is fully developed is expensive, time consuming, and dependent on large data sets [3].

Contemporary UX methodology is perceived to be lacking a prototyping method for ML experiences [9]. Design prototyping literature lacks any mention of how to test ML based experiences [22, 30]. Early prototyping of ML experiences is important given the interface is critical to the success of ML systems, with poor interaction design leading to degraded performance [11].

Low-fidelity prototyping methods (e.g. paper prototypes) offer designers frugal means of generating insights early on in a system's development [30]. A low-fidelity method to prototype ML experiences would bridge the gap in getting designers involved earlier on in the process, helping designers understand the models more sincerely, and generating user feedback on the model design. This paper suggests Wizard of Oz (WOz) prototyping for ML experiences and offers a set of topics to help designers evaluate their prototypes.

2 WIZARD OF OZ PROTOTYPING AND MACHINE LEARNING

2.1 Wizard of Oz Prototyping

WOz prototyping is a prototyping method where participants are made to believe that they are acting with a functional system, but instead the experimenter acts as the "wizard", a proxy for the system behind the scenes [18]. The "wizard" emulates the system's intelligence and interacts with the user through a real or mock computer interface [21]. WOz prototyping allows designers to test ideas at a lower cost than developing a functioning system, eliciting feedback early and throughout the development process [10, 14]. Researchers have used WOz prototyping to test systems in an array of contexts; interface designs, UIs for teaching an unmanned aerial vehicle, natural language dialogue systems, speech recognition systems, recommender systems, and even to garner children's mental models of ML systems [7, 13, 16, 21, 26, 27, 31].

2.2 Wizard of Oz for Machine Learning

WOz surfaces as an ideal prototyping method for testing ML experiences given the experimenter can mimic the model's computations and get feedback early in the development process [10]. Insights surrounding the design, user perception and behavior of the model can be gleaned from users at a low cost, rather than waiting until the model is fully developed. WOz prototyping allows the designer to test out the experience of their model without sacrificing the benefits of low-fidelity testing. This requires the designer to understand the constraints and affordances of the models to be tested while being inventive in garnering the data necessary to simulate the experience.

A hurdle in developing a representative WOz prototype is in understanding the abilities and constraints of the model. The wizard's behavior should be based on an algorithm and maintain a constrained interaction model [21]. This will require clarity on model behavior and limitations. This understanding benefits the designer in informing the design of the explanation of the model in the help and documentation, while increasing their ML domain knowledge. This hurdle will require closer collaborations between data scientists and designers earlier on in the product development process as the UX field adapts ML as a design material.

An additional hurdle is in garnering the least amount of data needed to emulate the experience and "wizard" that information into the prototype. Prototypes lack the data that functional ML systems work with [2]. This demands the designer to be resourceful in getting enough data. For instance, if an experience based on a user's preferences is being tested, the "wizard" could be a survey sent to the user before the test to garner their preferences [20]. The UI could then be sketched to reveal those preferences in a way that reflects the envisioned model as the user progresses through the test. If the model is more dependent on in the moment decisions and activities of the user, the experimenter could sketch the imagined model output in the moment before presenting the next state of the prototype with those decisions reflected. The designer will mimic the model by making on the fly decisions on how the model would behave based on the user's actions and preferences, remaining within the constraints of the model.

These hurdles present limitations to WOz prototyping for ML. ML systems tend to use large data sets, with systems behaving differently depending on the data they were given. Additionally, different model types may be harder to mimic and explain than others (e.g. neural networks). However, the benefits of getting user feedback on the system early in its development far outweigh the cons of having a less than perfect simulation.

3 EVALUATING THE EXPERIENCE

A ML WOz prototype presents specific topics to consider when evaluating; how the user controls, gives feedback to, and understands the model [6].

3.1 User Control

Through WOz prototyping, designers will evaluate how well the user can control the system through the UI. Designing how the user manipulates the model deserves the most attention in designing ML systems [11]. The user ought to be in control of the system, being able to steer it, turn it off, turn it on, undo actions done by the system, and tweak the system when necessary [4, 6, 24]. It's important to evaluate if the user can discover how to manipulate the model through the UI, with such actions represented as explicitly as possible [6]. Uncovering if the user understands the task and how the interactions help the system achieve a goal is a necessity [11]. In evaluating user control in a WOz prototype, the experimenter asks the user how they would do such actions.

3.2 Testing Model Conditions

It's important to understand how users react to the system's output when given true positives, false positives, true negatives, and false negatives [20]. Experimenters can test these conditions to understand how users react to possible suggestions from the model, how users correct and provide feedback on the model, and how the outputs affect their trust in the system.

An important aspect of testing the model's conditions is in learning if the user understands how to contest and correct the model's output. For instance, if the model fails in a certain instance, it should surface why it made that decision and allow the user to give it feedback [6]. This informs the accuracy of the model, increases the quality of the model, increases user comprehension of the model, allows for the surfacing of implicit biases in the model, and increases the feeling of control [6, 15]. For instance, if a recommender system recommends content that is faulty, the user should be able to supply input to the model that it was faulty and correct it [1]. This puts the user more in control of identifying a misspecification of the learner and gives them the tools to potentially debug the model [8]. In a WOz prototype, the user can be given an output with each of these conditions and asked how they might go about reaffirming or rejecting the output.

3.3 Explanations and Status of the Model

Designers will evaluate users' understanding of how the model works. In order to build trust with the system, the user ought to be able to understand at a high-level how the system works and how it is working. This notion follows long standing principles of documentation and system status in interaction design, in the ML field known as Explainable AI (XAI) [19, 23]. Helping the user develop an appropriate mental model of the model is important in making the output understandable, building a sense of the expected model behavior, and in informing the user's later inputs [8, 24, 25]. An inaccurate mental model could have detrimental effects on the system. An explanation should offer cognitive value to the user and clearly communicate the type of explanation relevant for their context [5]. The explanation should encompass model performance on a specific sample instance, model performance in a more general context, and constraints of the model [11]. If the goals of the model aren't clear to the user, users may give incorrect inputs [11].

Successful explanations range from showing the relationship between inputs and outputs, to allowing the user to test inputs to see the effect on the output, to showing confidence or accuracy scores to convey the degree of certainty with each instance [6, 15, 17, 29]. All cases embrace human-readable illustrations and interactions [6]. Legibility of the model is intrinsically coupled to user's trust and willingness to adopt the system, with the model being perceived as more accurate as legibility increases [15]. Successful explanations can be seen as an extension of Nielsen's system status, where the system is confident a certain amount in its' evaluation [23].

In a WOz prototyping test, this can be gleaned from separate questions for the user throughout the test; asking the user how the system works, how they know it works, how they might find out it works, and if they trust it. These can be evaluated instance by instance and more generally concerning the system as a whole.

4 CONCLUSIONS

Wizard of Oz prototyping is an ideal methodology for prototyping machine learning experiences. Designers ought to evaluate their prototypes upon how well they keep the user in control, how users respond to different model conditions, and how clear the explanation of the model is. Future work will present case studies evaluating WOz'd ML experiences, surface methodologies for designers to understand when to apply ML after generating user insights, and drive towards processes for incorporating human-centered practices in the development of ML.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. AI Magazine 35, 4 (2014), 105--120. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2513

[2] Saleema Amershi. 2012. Designing for Effective End-User Interaction with Machine Learning. Ph.D. Dissertation. University of Washington, Seattle, WA.

[3] Tone Bratteteig and Guri Verne. 2018. Does AI make PD obsolete?: exploring challenges from artificial intelligence to participatory design. In Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2 (PDC '18), Liesbeth Huybrechts, Maurizio Teli, Ann Light, Yanki Lee, Carl Di Salvo, Erik Grönvall, Anne Marie Kanstrup, and Keld Bødker (Eds.), Vol. 2. ACM, New York, NY, USA, Article 8, 5 pages. DOI: https://doi.org/10.1145/3210604.3210646

[4] Ajay Chander, Ramya Srinivasan, Suhas Chelian, Jun Wang, and Kanji Uchino. Working with beliefs: AI transparency in the enterprise. In Joint Proceedings of the ACM IUI 2018 Workshops co-located with the 23rd ACM Conference on Intelligent User Interfaces (ACM IUI 2018), 2018.

[5] Ajay Chander and Ramya Srinivasan. 2018. Evaluating Explanations by Cognitive Value. In: Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2018. Lecture Notes in Computer Science, vol 11015. Springer, Cham

[6] Eric Corbett, Nathaniel Saul, and Meg Pirrung. Interactive Machine Learning Heuristics. (October 2018). Retrieved December 1, 2018 from https://learningfromusersworkshop.github.io/papers/IMLH.pdf

[7] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In Proceedings of the 1st international conference on Intelligent user interfaces (IUI '93), Wayne D. Gray, William E. Hefley, and Dianne Murray (Eds.). ACM, New York, NY, USA, 193-200. DOI: https://doi.org/10.1145/169891.169968

[8] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, Philipp Ebel. 2019. The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. 2019. - Hawaii International Conference on System Sciences (HICSS). - Maui, Hawaii, USA.

[9] 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 (CHI '17). ACM, New York, NY, USA, 278-288. DOI: https://doi.org/10.1145/3025453.3025739

[10] Steven Dow, Blair MacIntyre, Jaemin Lee, Christopher Oezbek, Jay David Bolter, and Maribeth Gandy. 2005. Wizard of Oz Support throughout an Iterative Design Process. IEEE Pervasive Computing 4, 4 (October 2005), 18-26. DOI=http://dx.doi.org/10.1109/MPRV.2005.93

[11] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. ACM Trans. Interact. Intell. Syst. 1, 1, Article 1 (March 2018), 37 pages. https://doi.org/10.1145/3185517

[12] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, Nicolas d'Alessandro, Joëlle Tilmanne, Todd Kulesza, and Baptiste Caramiaux. 2016. Human-Centred Machine Learning. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16). ACM, New York, NY, USA, 3558-3565. DOI: https://doi.org/10.1145/2851581.2856492

[13] John D. Gould, John Conti, and Todd Hovanyecz. 1983. Composing letters with a simulated listening typewriter. Commun. ACM 26, 4 (April 1983), 295-308. DOI: https://doi.org/10.1145/2163.358100
[14] Bruce Hanington and Bella Martin. 2012. Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions. Rockport Publishers, Beverly, MA.
[15] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. In Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17). ACM, New York, NY, USA, 95-99. DOI: https://doi.org/10.1145/3064663.3064703
[16] Tom Hitron, Iddo Wald, Hadas Erel, and Oren Zuckerman. 2018. Introducing children to machine learning concepts through hands-on experience. In Proceedings of the 17th ACM Conference on Interaction Design and Children (IDC '18). ACM, New York, NY, USA, 563-568. DOI: https://doi.org/10.1145/3202185.3210776
[17] Kristina Höök. 2000. Steps to take before intelligent user interfaces become real. Interacting with Computers 12, 4 (2000), 409–426. DOI: http://dx.doi.org/10.1016/S0953-5438(99)00006-5
[18] J. F. Kelley. 1983. An empirical methodology for writing user-friendly natural language computer applications. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '83), Ann Janda (Ed.). ACM, New York, NY, USA, 193-196. DOI=http://dx.doi.org/10.1145/800045.801609
[19] Michael van Lent, William Fisher, and Michael Mancuso. 2004. An Explainable Artificial Intelligence System for Small-unit Tactical Behavior.
[20] Josh Lovejoy and Jess Holbrook. 2017. Human-Centered Machine Learning. (July 2017). Retrieved December 3, 2018 from https://medium.com/google-design/human-centered-machine-learning-a770d10562cd
[21] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an intelligent agent through Wizard of Oz. In Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93). ACM, New York, NY, USA, 277-284. DOI: https://doi.org/10.1145/169059.169215
[22] Kathryn McElroy. 2017. Prototyping for Designers. O'Reilly Media, Sebastopol, CA.
[23] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90), Jane Carrasco Chew and John Whiteside (Eds.). ACM, New York, NY, USA, 249-256. DOI=http://dx.doi.org/10.1145/97243.97281
[24] Christopher Noessel. 2017. Designing Agentive Technology. Rosenfield Media, Brooklyn, NY.
[25] Don Norman. 2009. The Design of Future Things. Basic Books, New York, NY.
[26] Masayuki Okamoto , Yeonsoo Yang and Toru Ishida. 2001. Wizard of Oz Method for Learning Dialog Agents, Proceedings of the 5th International Workshop on Cooperative Information Agents V, p.20-25, September 06-08, 2001
[27] Raquel Torres Peralta, Tasneem Kaochar, Ian R. Fasel, Clayton T. Morrison, Thomas J. Walsh, and Paul R. Cohen. 2011. Challenges to decoding the intention behind natural instruction. 2011 Ro-Man (2011). DOI:http://dx.doi.org/10.1109/roman.2011.6005273
[28] Edward Tiong, Olivia Seow, Bradley A Camburn, Kenneth Teo, Arlindo Silva, Kristin Lee Wood, Dan Jensen, and Maria Yeng. 2018. The Economies and Dimensionality of Design Prototyping: Value, Time, Cost and Fidelity (DETC2018-85747). ASME. J. Mech. Des. 2018;():. doi:10.1115/1.4042337.
[29] Jasper van der Waa, Jurriaan van Diggelen, and Mark Neerincx. 2018. The design and validation of an intuitive confidence measure. In Workshop On Explainable Smart Systems (EXSS), 2018.
[30] Todd Zaki Warfel. 2009. Protoyping: A Practitioner's Guide. Rosenfeld Media, Brooklyn, NY.
[31] Jason Williams and Steve Young. 2003. Using wizard-of-oz simulations to bootstrap reinforcement-learning-based dialog management systems. In Proc. 4th SIGdial workshop.
[32] 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 (DIS '18). ACM, New York, NY, USA, 573-584. DOI: https://doi.org/10.1145/3196709.3196729
[33] Qian Yang, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. 2016. Planning Adaptive Mobile Experiences When Wireframing. In Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16). ACM, New York, NY, USA, 565-576. DOI: https://doi.org/10.1145/2901790.2901858
[34] Qian Yang. 2018. Machine learning as a UX design material: How can we imagine beyond automation, recommenders, and reminders? In AAAI Spring Symposium Series