Human Factors and Simulation, Vol. 83, 2023, 184–191
https://doi.org/10.54941/ahfe1003576

Developing Confidence in Machine Learning Results

Jessica Baweja, Brett Jefferson, and Corey Fallon

Pacific Northwest National Laboratory, Richland, WA 99354, USA

ABSTRACT

As the field of deep learning has emerged in recent years, the amount of knowledge and expertise that data scientists are expected to absorb and maintain has correspondingly increased. One of the challenges experienced by data scientists working with deep learning models is developing confidence in the accuracy of their approach and the resulting findings. In this study, we conducted semi-structured interviews with data scientists at a national laboratory to understand the processes that data scientists use when attempting to develop their models and the ways that they gain confidence that the results they obtained were accurate. These interviews were analysed to provide an overview of the techniques currently used when working with machine learning (ML) models. Opportunities for collaboration with human factors researchers to develop new tools are identified.

Keywords: Data science, Machine learning, Deep learning, Neural networks

INTRODUCTION

Recent years have brought changes to the ways that humans work with and develop technology. The emergence of deep learning, a subfield of machine learning, has been a revolution for data science and for society in general, transforming the ways that we live, work, and play. A major part of what is so revolutionary about deep learning is that it allows data scientists to develop models without the need to first generate features or representation of the dataset (Jordan & Mitchell, 2015; Khanafer & Shirmohammadi, 2020; LeCun et al. 2015). As a result, data scientists can build models on increasingly high-dimensional data without the need to first process the data and build meaningful features for prediction. Instead, the model itself optimizes an output based on some input, allowing the relevant features to be discovered in the dataset. This capability is important to the success of deep learning, allowing data scientists to tackle challenges that were previously untenable due to the size or complexity of the dataset.

However, with the value of deep learning comes new challenges. First, the field of deep learning itself has grown at an incredible rate, from around 19,000 publications on Google Scholar in 2015 for the query *deep learning* to around 277,000 in 2020. With this growth comes an emergence of new techniques, methods, and approaches that data scientists can explore. This rate of change makes it nearly impossible for scientists to remain aware and knowledgeable of the new methods and practices available to them when

conducting research. Second, deep learning methods themselves are opaque due to the reliance on representation learning—that is, the capability of the models to automatically represent the data through features. It is therefore non-trivial to decompose how and why the model provides a recommendation or generates a specific recommendation (Castelvecchi, 2016; Goebel et al., 2018). This makes it challenging for a data scientist to understand why a result is provided and whether the model should be trusted (above and beyond any mathematical metrics indicating accuracy). Finally, creation of a deep learning model relies on the selection of appropriate hyperparameters, such as the number of layers, the learning rate of the model, or the activation function used (Young et al. 2015). These hyperparameters can have a strong impact on the performance of the final model. However, the large number of hyperparameters to be tuned and the complexity of their relationships with the output makes it challenging for data scientists to select the correct hyperparameters and to develop confidence that the results they have obtained from their model are, in fact, optimal.

In this study, we applied human factors methods to understand how data scientists work with deep learning models and the ways they strive to validate their results. Using semi-structured interviews with data scientists working with deep learning models at a national laboratory, we identified the ways that deep learning researchers develop confidence in their findings and revealed opportunities for human factors researchers to contribute to the development of better tools and methods for that process.

METHOD

Eleven self-identified data scientists working at a national laboratory who work with artificial neural networks (ANNs) were interviewed; only 10 of their responses are included because one reported on a project that did not rely on a neural network. The included participants had an average of 3.7 years (SD = 3.6 years) of experience working as a data scientist with expertise in machine learning and deep learning approaches.

Procedure

Participants were contacted via email to participate in a 90-minute interview on Microsoft Teams. Interviews were conducted by one ($n = 6$) or two ($n = 4$) interviewers. Interviews were recorded for later review, but interviewers also took notes throughout the conversation. Participants were asked to think about a study where they worked with machine learning to either improve the performance of the model, to simplify the model, or to improve understanding of the model.

Questions were asked in a series of three sweeps: first, participants described their overall process; next, they discussed each step of the process in greater detail; and finally, they answered questions regarding the challenges experienced in each step and the areas where additional support, tools, or training might be helpful. We report greater detail on the questions used in each sweep in a separate manuscript (see Baweja et al., under review); the results here focus specifically on questions related to model accuracy and

confidence in the results, and the strategies data scientists used to develop them. Questions included:

- How did you decide how confident you should be in the results of your analysis?
- Is there information you wish you had about the model but didn't?
- How did you manage the large number of layers/nodes? Did you have a strategy (e.g., analysis, visualization)?
- Were there critical cues you were paying attention to in the model (e.g., particular nodes, levels)?
- How did you know when something is amiss?
- Did you use a strategy or analysis that helped you notice?

Depending upon the interview, not all questions were asked, and questions may have been elaborated or altered as needed to fit the context of the convesation; the goal was to understand the ways that data scientists were reviewing their results and attempting to understand their outcomes. Following completion of the interviews, the notes from all of them were compiled, and responses were organized by the topic of the question. Responses were coded inductively, with the first author reviewing and extracting themes from the results.

RESULTS

In general, participants discussed projects where the goal was to optimize (i.e., improve) the performance of an ANN for a specific application. When describing their process, although there was some variability, many participants described similar steps of working with machine learning models:

1. Literature Search
2. Data Generation or Collection
3. Data Preprocessing
4. Model Selection
5. Model Construction
6. Model Optimization
7. Model Evaluation
8. Model Explanation.

Notably, this process was not necessarily linear; for instance, participants might select a model and then later explore an alternative based on poor performance. Certainly, the model optimization and model evaluation steps are highly iterative. The next sections describe each of the steps as they pertain to how data scientists develop confidence in their models.

Literature Search

Some data scientists explicitly describe beginning their project with a review of the literature. As it relates to developing confidence in their models, this process generally involved trying to identify how others have approached similar problems in the past. This was especially true for data scientists who had less experience in the domain application where they were working;

the participants sought insight into how others had approached comparable datasets to understand how best to model theirs.

Data Generation Or Collection

Participants described either generating the data for their models themselves or obtaining them from others. In many cases, this was described as their first step. However, there was little discussion here of the ways that participants developed confidence in their dataset outside of idiosyncrasies in the way that experimental data were collected.

Data Preprocessing

Regardless of how participants collected data, they all discussed some aspect of data preprocessing, which refers to the data cleaning and reformatting necessary to prepare data for use within a machine learning model. This was the step most consistently described as tedious and frustrating. Several participants also discussed issues that occurred in the data preprocessing step and how they assessed the accuracy of the final dataset. For example, one participant talked about finding errors in the model results, and not realizing for quite some time that the errors were occurring in the data preprocessing step. Another participant described a situation where the data were preprocessed prior to them receiving it, and they did not realize until later that there were issues in that preprocessing step that translated into errors in the model results. However, participants generally described the process of developing confidence that preprocessing was done correctly as monotonous, tedious, and time-consuming. They attempted to identify issues and verify the data through exploratory analyses, descriptive statistics, and experimental models (i.e., running a model to identify anomalies). Overall, this is an area where data scientists certainly expressed desire for additional support, especially when it comes to verifying that data preprocessing is done accurately or identifying ways to do it more efficiently.

Model Selection

In many ways, although chronologically later in the process, this step overlaps with some of the considerations discussed in the literature search. That is, participants often discussed going to the literature to help guide them as they selected a model for their project. In addition, the model selection process was also dependent on the nature of the dataset. One participant discussed looking at metrics on a test set to assess whether the model was appropriate—for example, exploring if they were underfit or overfit. This participant also discussed developing heuristics based on experience about the appropriate model for a specific task. However, they acknowledged that this judgment is ultimately highly qualitative: comparing the test results to the literature to assess whether it is similar enough to published results to indicate that the participant has selected the correct modelling approach. Overall, participants generally described relying on past research to evaluate whether they selected the correct model for the problem at hand.

Model Construction

Model construction or model engineering is the process of implementing the model in code and creating the necessary computing environment to execute it. Although there were challenges in execution (e.g., difficulty setting up environments), this is another step where there was very little discussed with relation to confidence in the results. Participants who were relying on past research to construct their model, however, did note that past work is often poorly documented, making it challenging to either obtain the same results or to be confident that they used the same approach as the authors of similar work.

Model Optimization

The process of model optimization involves altering values for model variables, known as hyperparameters, that govern process variables, as well as optimizing the model for computing efficiency. Hyperparameter tuning (i.e., the altering of the model variables) was described by all participants as one of the most ambiguous tasks. One participant noted a lack of confidence that they were applying best practices, or even that they knew what the best practices were. Another described this as a trial-and-error process, noting that it would be helpful to have assistance but that the process required human judgment to do it well. One participant expressed that there is no way to know whether the models and the hyperparameters are correct, and that each data scientist has their own metric for success. Finally, another participant noted that all data scientists use different methods to explore hyperparameters, and many of the decisions are manual—and they can have a significant impact on model performance.

To gain confidence in their results overall, participants used a review of comparable results as a baseline for comparison and they also described a process of trial-and-error. However, they also noted that documentation of hyperparameters (particularly those tested but not ultimately selected) is frequently poor, and they expressed significant challenges in determining whether the results were, in fact, a demonstration of the strongest possible performance.

Model Evaluation

As already mentioned, model evaluation often occurred iteratively with model optimization; participants described running a specific model and then heuristically evaluating the results to assess whether additional changes were needed. In this step, participants relied heavily on metrics to determine whether the model was accurate, including examination of loss curves, recall, or precision, amongst others. They discussed looking at confusion matrices and other similar methods to explore when the model was and was not accurate. In addition, one participant discussed testing the model against a variety of datasets to understand when it would and would not perform well, attempting to develop their own heuristics about the robustness of the model. However, they still expressed some scepticism about the accuracy of their models; one data scientist expressed a desire for another person who was

familiar with neural networks to look at the results just to assess whether they looked the way they should. Again, data scientists relied heavily on review of past literature and heuristics developed with experience, in combination with objective metrics, to judge whether the model performed sufficiently well.

Model Explanation

Not all participants included model explanation as part of their process of working with machine learning models. However, some did discuss the desire to understand and explore the reasons why a model generated a specific result. In addition, others discussed attempting to understand models in a more exploratory way by examining when the model was accurate and when it was not to develop intuitive explanations regarding the reasons for model performance. One participant discussed using model explanation as one way of building confidence in the model results: if the model appeared to be producing results for the "right" reasons, they were more confident in the outcome. Overall, given some of the complexities already described, model explanation was in some cases used to improve confidence that the decisions made throughout the model development process resulted in a model that was accurate and interpretable.

DISCUSSION

The results of this research validated previous work that suggested that at least some aspects of machine learning are intuitive, manual, and based on trial-and-error (Young et al. 2015). This was especially true for hyperparameter tuning, where participants expressed significant challenges in determining which values to test, when they had tested sufficiently, and which values were, in fact, the most likely to result in improved performance. To a lesser extent, participants also expressed similar concerns regarding the selection of the model, relying heavily on past literature to determine which approach to use in a specific problem. Overall, these results suggest that a portion of data scientists' work relies on the development of heuristics rather than any objective rules.

In addition, participants expressed frustration with data preprocessing and the identification of errors. This is not surprising, as past work has suggested that the work of cleaning and preparing data is generally disliked and undervalued despite its importance for the final results of the model (Sambasivan et al., 2021). Preprocessing was generally idiosyncratic and manual, and several participants discussed errors introduced into the project due to unidentified issues in the data preprocessing. Especially as datasets increase in size, it is difficult for data scientists to assess when data are formatted correctly and when there might be errors.

At each step in the process of developing machine learning models, data scientists described a process of sensemaking, where they apply knowledge, judgment, and expertise to make decisions about how to select, construct, optimize, or evaluate a model. The decisions that data scientists make during this process can have substantial impacts on the way that the model performs (e.g., Young et al. 2015). This view of machine learning as a process that

relies on human judgment, and to some extent, heuristics and intuition, has also been suggested in past work (e.g., Muller et al., 2019; Young et al., 2015). This is not to say that machine learning or deep learning are deficient; instead, it simply underscores that, despite its reliance on mathematics and computing, it is in some ways highly subjective and dependent upon good human decision making.

Limitations and Future Directions

The results presented here are from a qualitative study of a small group of data scientists working at a national laboratory in the United States. Nonetheless, they correspond to other research in the field suggesting that machine learning scientists face challenges in hyperparameter tuning (e.g., Cooper et al. 2021; Yang & Shami, 2020; Young et al., 2015). In addition, other authors have argued that the field is experiencing a reproducibility crisis, at least partially due to sometimes poor documentation of details necessary to reproduce the work (Hutson, 2018; Kapoor & Narayanan, 2022). Thus, although the results presented here are limited in scope, they nonetheless validate some of the challenges that have already been identified in the field.

These findings highlight several areas where collaboration between human factors researchers and data scientists might be especially fruitful. The first area, is developing additional tools for hyperparameter optimization. Although efforts are underway to create new technologies for the hyperparameter tuning process, human factors researchers are uniquely poised to provide insight into the heuristics that data scientists develop, which might be useful when creating new tools to support human work. Similarly, some of the issues identified in data preprocessing could benefit from human factors insight to help automate those tasks most often perceived as tedious or monotonous. Finally, continued work in model explainability might help to address some of the lack of confidence in the accuracy of the results expressed by the participants in this study. Although metrics can provide evidence for objective performance, data scientists nonetheless sought interpretable evidence that the model was not only right, but that it was right for the right reasons. Human factors professionals can certainly provide valuable insight as to the nature of a good explanation for that specific audience and purpose.

CONCLUSION

Overall, the results of this research suggest that despite the ostensibly objective nature of data science, much of the work still relies heavily on human judgment and expertise. As the field matures, additional work is needed to ensure that results are reproducible. In addition, some of the activities described here could certainly benefit from development of additional tools, relying on machines that can better complement human judgment by automating tasks that are time-consuming and tedious. Developing these tools can help to contribute to a human-machine team that leverages the strengths of both data scientists' expertise as well as the automation and computing power granted to us by machine learning methods.

REFERENCES

Castelvecchi, D., 2016. Can we open the black box of AI?. *Nature News*, *538* (7623), p. 20.

Cooper, A. F., Lu, Y., Forde, J. and De Sa, C. M., 2021. Hyperparameter optimization is deceiving us, and how to stop it. *Advances in Neural Information Processing Systems*, *34*, pp. 3081–3095.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P. and Holzinger, A., 2018, August. Explainable AI: the new 42?. In *International cross-domain conference for machine learning and knowledge extraction* (pp. 295–303). Springer, Cham.

Hutson, M. Artificial intelligence faces reproducibility crisis, 2018. *Science, 359*, pp. 725–726. https://doi.org/10.1126/science.359.6377.725

Jordan, M. I. and Mitchell, T. M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), pp. 255–260.

Kapoor, S. and Narayanan, A., 2022. Leakage and the Reproducibility Crisis in ML-based Science. *arXiv preprint arXiv:2207.07048*.

Khanafer, M. and Shirmohammadi, S., 2020. Applied AI in instrumentation and measurement: The deep learning revolution. *IEEE Instrumentation & Measurement Magazine*, *23*(6), pp. 10–17.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, *521*(7553), pp. 436–444.

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C. and Erickson, T., 2019, May. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems,* pp. 1–15.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L. M., 2021, May. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1-15.

Yang, L. and Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, pp. 295–316.

Young, S. R., Rose, D. C., Karnowski, T. P., Lim, S. H., & Patton, R. M. (2015, November). Optimizing deep learning hyper-parameters through an evolutionary algorithm. In Proceedings of the workshop on machine learning in high-performance computing environments (pp. 1-5).