This work has been published in CHI 2020 Late Breaking Work. This is a resubmission for the ICML HILL workshop.

SCRAM: Simple Checks for Realtime Analysis of Model Training for Non-Expert ML Programmers

Eldon Schoop [1] **Forrest Huang** [1] **Björn Hartmann** [1]

Abstract

Many non-expert Machine Learning users wish to apply deep learning models to their own domains but encounter hurdles in the model training process. We introduce SCRAM, a tool which uses heuristics to detect potential error conditions in model output and suggest best practices to help such users tune their models. Inspired by metaphors from software engineering, SCRAM extends high-level deep learning development tools to check model metrics during training and produce human-readable error messages. We validate SCRAM through three author-created examples with image and text datasets, and by collecting informal feedback from ML researchers with teaching experience. We reflect upon their feedback for the design of future ML debugging tools.

1. Introduction

Many domain experts, hobbyists, and makers wish to adopt Machine Learning (ML) models such as neural networks into their applications, but lack formal training in ML. These users often have some programming expertise and their own novel datasets for a particular domain problem. For example, a farmer may want to classify the types of cucumbers from their farm, or an independent app developer may want to recommend workouts in their fitness app. Several Deep Learning (DL) toolkits, including Keras (Chollet, 2015) and Apple Create ML (Inc, 2019a), make these tasks more approachable by providing high-level APIs to preprocess data, train, and evaluate DL models (neural networks). However, when non-expert ML developers use these APIs to train models on novel datasets, they can produce unexpected output without explicitly throwing errors.

While experts can rely on experience and tools such as TensorBoard (Abadi et al., 2016) and tfdbg (Cai, 2017) to inspect and correct model behavior, non-experts often lack the theoretical and practical knowledge to interpret results from these tools (Hill et al., 2016; Cai & Guo, 2019) and could benefit from guidance through this unstructured process (Amershi et al., 2019; Patel et al., 2008).

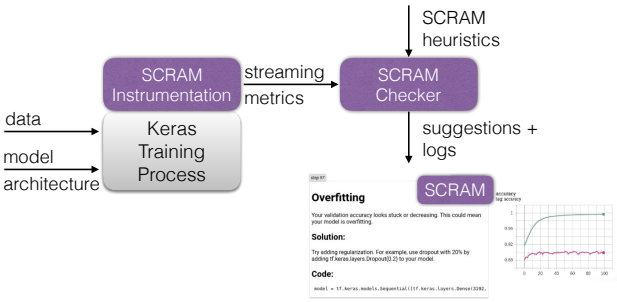

Figure 1. The Keras framework outputs data batches and model metrics to SCRAM (left), and SCRAM outputs error messages and visualizations to Tensorboard (right).

We introduce SCRAM, a prototype system which can interpret potential error conditions in the DL training phase and provide descriptive, actionable warning messages to help users debug and produce well-trained models (see Figure 1). SCRAM draws inspiration from tools in software engineering which inspect code to provide warning messages and suggestions to developers. Our goal is to develop a system that can encode this tacit knowledge of experts into heuristics which check model output over time during training. This system will guide non-expert users to correct errors with human-readable error messages that explain best practices and code recipes to bridge theoretical and practical knowledge gaps. During the tuning phase, users interpret these error messages to make changes to hyperparameters, correcting model behavior.

In this paper, we describe the SCRAM prototype; share three heuristics for detecting common problems during neural network training; and validate error messages produced by SCRAM in three author-created scenarios with experienced ML instructors for future iterations of our system.

[1]EECS, UC Berkeley, Berkeley, CA, USA. Correspondence to: Eldon Schoop <eschoop@berkeley.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

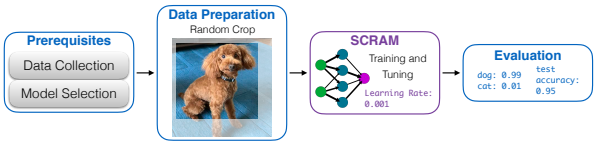

Figure 2. Once a dataset is collected and an ML algorithm is selected, users must (1) preprocess data, (2) train and tune their model, and (3) evaluate their model on test data.

2. Related Work

2.1. Interactive ML Development

HCI research has produced novel interfaces which allow users to interactively train and tune ML models as early as 2003 (Fails & Olsen, 2003). Gestalt is toolkit which adds structure to the ML development process, allowing developers to iteratively modify and analyze their models in an IDE (Patel et al., 2010). Makers can alternatively use ESP to interactively train and deploy gesture recognition models on Arduino hardware (Mellis et al., 2017). While these tools support the feature engineering workflow required for classical ML, SCRAM focuses on training and tuning DL models which enable powerful contemporary applications in domains where manual feature selection is infeasible (e.g., image/speech recognition), but come with the trade-off of an extended and hard-to-interpret training process.

ML practitioners can add instrumentation and visualizations to their DL models using toolkits such as TensorWatch (Shah et al., 2019) and Lucid (Olah et al., 2017), but the choice of visualization and its interpretation requires expertise. SCRAM uses heuristics to produce text error messages which can guide novices in the DL debugging process. One recent commercial tool may help non-experts fine-tune pretrained models, but does not interpret model output (AI, 2019). SCRAM integrates with a open source Python framework, Keras (Chollet, 2015), which has a large support community and can provide more advanced functionality as needed.

2.2. Software Engineering Support Tools

SCRAM draws upon established paradigms in software engineering such as linting (Johnson, 1978), unit testing, dynamic analysis (Myers et al., 2011), and explanation-based debugging (Ko & Myers, 2009) to help users interpret the behavior and inspect the points of failure of their ML applications. We draw additional inspiration from software visualization (Stasko et al., 1997) and tutorial systems for complex user interfaces (Grossman et al., 2009) which guide novices through complex tasks.

2.3. Model Visualization and Inspection Tools

Because of the intrinsic relationship between training data and a model, evaluation tools can highlight relevant training data contributing to outliers (PAIR, 2019) and refine the model itself (Amershi et al., 2015). TensorFuzz can assist debugging by adapting coverage-based fuzzing to identify inputs which generate numerical errors (Odena et al., 2019). Other tools track and visualize test results to help select models for large-scale deployments (Li & Wang, 2019; Inc, 2019b; Rolando Garcia, 2019; Murugesan et al., 2019). Evaluating the performance of ML models is a critical step, but depends on having an already trained model. SCRAM assists users in the training step required *before* evaluation.

2.3.1. Explanations and Interpretability

SCRAM is inspired by systems which help practitioners *interpret* the output and behavior of their ML models. Deep neural networks often have too many parameters to easily understand, and explaining their output is an active area of research (Gilpin et al., 2018). Activation Maps highlight the parts of an input image used to make a prediction (Olah et al., 2017). A more recent algorithm, Concept Activation Vectors (CAV), can explain the higher-level concepts used in an output classification (Kim et al., 2017). Training and tuning neural networks similarly produces output which is difficult to interpret (Kaparthy, 2016; 2019), relying on tacit knowledge and expertise to understand (Hill et al., 2016). We believe SCRAM is an early step in providing explanations of neural network output during the training process.

A key component of SCRAM is an automated checking infrastructure that enables running tests over model runtime behavior to flag problems. Other systems in HCI research use this approach to assist debugging electrical circuits and embedded systems (Drew et al., 2016; McGrath et al., 2017). SCRAM adapts this approach to ML debugging.

3. Design Considerations

SCRAM targets scenarios when users have an existing problem formulation for applying ML to their applications. In these cases, a novel dataset has already been collected and a neural network architecture chosen. The remaining steps, shown in Figure 2, are: (1) *data preparation*, in which the data are split into training and tests sets, normalized, and formatted for input; (2) *training and tuning*, where model hyperparameters are tuned during training to help the model fit the data; and (3) *evaluation*, in which model performance is tested and compared. SCRAM focuses on guiding users through the second phase, *training and tuning*. During this phase, hyperparameters, such as the optimizer learning rate or model regularization, tune how the model fits batches

of data, so it can generalize accurate predictions to new input data. However, many non-experts are confused by the model output during this phase, leading many to abandon ML approaches altogether (Cai & Guo, 2019). Experts rely on tacit knowledge to interpret model output, e.g., by visually inspecting the model loss and accuracy curves, or running small scale tests (Kaparthy, 2016; 2019).

We chose to integrate our system with existing, popular DL frameworks, Keras (Chollet, 2015), and Tensorboard (Abadi et al., 2016). Keras is a popular choice for ML novices because it requires little code to construct and train neural networks, but its capabilities can also expand to meet advanced needs such as those of ML researchers. SCRAM outputs plots, suggestions, and error messages to TensorBoard, a visualization framework built for DL instrumentation that also integrates with Keras. Tensorboard supports real-time data loading during training as well as keeping track of runs.

4. Using SCRAM

Sam, a molecular biologist, wants to count the number of Gram positive bacteria in samples taken from an experiment on microscope slides. They already have access to thousands of annotated photos from previous experiments, and wish to repurpose a pretrained object detection neural network to count the bacteria in new samples. Formatting the dataset is easy, but once training begins, the model loss seems to increase, then reach NaN. A quick internet search turns up a Twitter thread [1] suggesting a lower learning rate. With the learning rate corrected, the model begins training, but the validation accuracy is much lower than expected and doesn't seem to be improving. Sam tweaks multiple parameters of the network and optimizer, but nothing seems to work. An ML engineer friend takes a quick pass over the code, but doesn't see anything obviously wrong and suggests using SCRAM. SCRAM detects that some input data points are reaching values as high as 255, and produces an error message stating the training data isn't being normalized properly. The message also suggests a code snippet to show how to normalize the training data to the model's expected input distribution, between -1 and 1. After Sam implements the suggested snippet, the model's accuracy increases rapidly. Sam verifies the model's correctness on test data. The model is integrated into Sam's lab workflow, saving hours of cell-counting time.

5. Implementation

SCRAM hooks into the built-in callback mechanism of Keras, which can invoke actions during model training. During training runtime, data batches and model metrics (loss and accuracies) are fed into SCRAM, where they are logged and checked against a list of heuristics to produce error messages. Checks are loaded individually, and can be swapped and customized as needed within SCRAM. Error messages and metrics are emitted directly to Tensorboard via the Tensorflow Summary API. (See Figure 1)

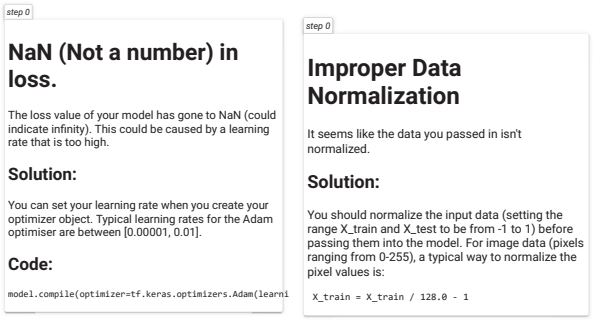

Figure 3. Error messages produced by SCRAM explain high-level concepts as well as suggest code snippets.

5.0.1. MODEL CHECKING HEURISTICS AND ERROR MESSAGES

SCRAM currently includes three heuristics which can identify common problems novices face when training their neural networks (see Table 1). We collect these heuristics from research literature, course notes, and tutorials from ML experts (Kaparthy, 2019; 2016; Shewchuk, 2019; Goodfellow et al., 2016). While these heuristics cover several common scenarios novices may encounter during training, there are many others which could be implemented in the future as well. Each heuristic has an associated checking function that tests if collected metrics violate the heuristic and an associated error message which is authored to give general theoretical advice as well as practical code snippets that can be used.

6. Inital User Experiences

To explore the utility of SCRAM, we constructed 3 example scenarios of errors with 3 different datasets: CIFAR-10 (Krizhevsky, 2009), Fashion-MNIST (Xiao et al., 2017), and Large Movie Reviews (Maas et al., 2011). Each scenario is inoculated with a potentially faulty model setup to generate errors from SCRAM: we use a large fully-connected network without any regularization to overfit on Fashion-MNIST; we set the model learning rate to 1e10 for Large Movie Reviews; and we use unnormalized pixels of the images in CIFAR-10 directly for model training.

To understand how SCRAM may help novices, we solicited feedback from 2 ML researchers with experience teaching introductory ML courses. We showed them our example scenario notebooks and allowed them to interact with the

[1] https://twitter.com/karpathy/status/1013244313327681536

Table 1. SCRAM heuristics and associated detection methods.

Heuristic	Description	Detection Method
Overfitting	When a model too closely fits its training data, it loses its ability to generalize to new data.	Check if validation accuracy decreases over two epochs while training accuracy increases, which likely indicates overfitting (Kaparthy, 2016).
Improper Data Normalization	In transfer learning, new data should be normalized to a similar range as the original data the model was trained on.	Check if the values of input features of current batch lie within the conventional range of $[-1, 1]$ (Shewchuk, 2019).
Unconventional Hyperparameters	Some hyperparameters for training deep models significantly affect model performance (Goodfellow et al., 2016). For instance, using too high of a learning rate will cause the model to produce NaN loss.	Check if the loss value reaches NaN, which indicates a possible incorrect range of hyperparameters (Kaparthy, 2019).

training code and error messages produced by SCRAM. Sessions lasted under half an hour each.

Both participants stated the notifications would be useful to novices, and that the heuristics capture common problems encountered by non-expert ML developers. One participant expressed its potential use to experienced ML developers—since training with large datasets may take days or weeks, notifications produced by SCRAM could direct attention to model training when needed. One participant remarked that SCRAM can catch errors that might not even be detected by a novice at all, such as normalization. Both participants expressed interest in adding other heuristics, such as predicting when a batch size may be too large to fit in memory (usually resulting in an error and program interruption). Other suggestions were for tightening integration between SCRAM and the code itself, by differentiating warnings and errors for conditions that can break execution during runtime, or by identifying the particular lines of code that generated the error (e.g., which part of the model generated a NaN output). Finally, one participant remarked that debugging strategies aren't often taught in ML courses, and SCRAM could serve as an instructional aid.

7. Future Work and Conclusion

SCRAM represents a first step in making the neural network training and tuning process more manageable, thus making applying ML more approachable to non-experts. Beyond adding additional heuristics, we are excited to continue work on SCRAM in the following areas:

Dynamic Error Messages: Error messages produced by SCRAM are written to apply to general cases, and provide explanations to help users narrow down the root cause and implement fixes. Dynamically generated error messages such as those produced by software tutorial systems (Head et al., 2015) could steer users closer to identifying the root causes of error conditions. Future iterations of SCRAM could even learn from examples to dynamically generate messages.

Code-Aware Tutorial Content: Making the error messages from SCRAM interactive could significantly improve its use as a tutorial system. For example, SCRAM could highlight specific lines of user code or Tensorboard visualizations. Another potential approach could be gleaned from the Java Whyline, which allows users to ask questions about program output during runtime to identify bugs (Ko & Myers, 2009).

Integrating Active Tests with SCRAM: Further engineering work could enable SCRAM to run *static* checks of the ML program, enabling many more heuristics (e.g., checking initialization). SCRAM could also be extended to execute operations with the model, such as overfitting on small batches of data or running user-defined unit tests.

Controlled User Evaluation of SCRAM: Our exploratory validation of SCRAM had a limited number of participants and was conducted with experienced instructors, not target users directly. A controlled user evaluation of the next iteration of SCRAM would determine the effectiveness of its heuristics and error messages. One possible experiment design could be that of Gestalt (Patel et al., 2010), in which novices were asked to debug ML models inoculated with errors in randomized conditions.

Communicating Uncertainty of Heuristics: The heuristics used by SCRAM are designed to detect and explain common errors, but these explanations are assumptions of model behavior and may not always be applicable. To mitigate this, the language of the messages are adapted to convey this intrinsic uncertainty, guiding the user to consider multiple possible underlying root causes and offering different solutions to mitigate them.

Recent advances in ML research have impacted numerous aspects of daily living, from transportation to healthcare to entertainment. We believe that artists, makers, domain experts, software engineers, and scientists can benefit from these advances by introducing tools to help understand and adapt their domain-specific data.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., and et al. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pp. 265–283, USA, 2016. USENIX Association. ISBN 9781931971331.

AI, R. Runway ml, 2019. URL https://runwayml.com/.

Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., and Suh, J. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 337–346, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702509. URL https://doi.org/10.1145/2702123.2702509.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and Zimmermann, T. Software engineering for machine learning: A case study. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*, ICSE-SEIP '19, pp. 291–300. IEEE Press, 2019. doi: 10.1109/ICSE-SEIP.2019.00042. URL https://doi.org/10.1109/ICSE-SEIP.2019.00042.

Cai, C. J. and Guo, P. J. Software developers learning machine learning: Motivations, hurdles, and desires. In *2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 25–34, Oct 2019. doi: 10.1109/VLHCC.2019.8818751.

Cai, S. Debug tensorflow models with tfdbg, Feb 2017. URL https://developers.googleblog.com/2017/02/debug-tensorflow-models-with-tfdbg.html.

Chollet, F. keras. https://github.com/fchollet/keras, 2015.

Drew, D., Newcomb, J. L., McGrath, W., Maksimovic, F., Mellis, D., and Hartmann, B. The toastboard: Ubiquitous instrumentation and automated checking of breadboarded circuits. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pp. 677–686, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341899. doi: 10.1145/2984511.2984566. URL https://doi.org/10.1145/2984511.2984566.

Fails, J. A. and Olsen, D. R. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, pp. 39–45, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581135866. doi: 10.1145/604045.604056. URL https://doi.org/10.1145/604045.604056.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. Explaining explanations: An overview of interpretability of machine learning, 2018.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Grossman, T., Fitzmaurice, G., and Attar, R. A survey of software learnability: metrics, methodologies and guidelines. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 649–658. ACM, 2009.

Head, A., Appachu, C., Hearst, M. A., and Hartmann, B. Tutorons: Generating context-relevant, on-demand explanations and demonstrations of online code. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 3–12, Oct 2015. doi: 10.1109/VLHCC.2015.7356972.

Hill, C., Bellamy, R., Erickson, T., and Burnett, M. Trials and tribulations of developers of intelligent systems: A field study. In *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 162–170, Sep. 2016. doi: 10.1109/VLHCC.2016.7739680.

Inc, A. Apple create ml, 2019a. URL https://developer.apple.com/machine-learning/create-ml/.

Inc, D. Mlflow, 2019b. URL https://mlflow.org/.

Johnson, S. C. Lint, a c program checker. In *Technical Report*, pp. 78–1273. Bell Telephone Laboratories, 1978.

Kaparthy, A. Training neural networks, part 1. *Convolutional Neural Networks for Visual Recognition. Lecture Slides*, January 2016. URL http://cs231n.stanford.edu/2016/syllabus.html.

Kaparthy, A. A recipe for training neural networks, Apr 2019. URL https://karpathy.github.io/2019/04/25/recipe/.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2017.

Ko, A. J. and Myers, B. A. Finding causes of program output with the java whyline. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1569–1578, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605582467. doi: 10.1145/1518701.1518942. URL https://doi.org/10.1145/1518701.1518942.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Li, L. and Wang, Y. Manifold: A model-agnostic visual debugging tool for machine learning at uber, Aug 2019. URL https://eng.uber.com/manifold/.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

McGrath, W., Drew, D., Warner, J., Kazemitabaar, M., Karchemsky, M., Mellis, D., and Hartmann, B. Bifröst: Visualizing and checking behavior of embedded systems across hardware and software. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, pp. 299–310, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349819. doi: 10.1145/3126594.3126658. URL https://doi.org/10.1145/3126594.3126658.

Mellis, D. A., Zhang, B., Leung, A., and Hartmann, B. Machine learning for makers: Interactive sensor data classification based on augmented code examples. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, pp. 1213–1225, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349222. doi: 10.1145/3064663.3064735. URL https://doi.org/10.1145/3064663.3064735.

Murugesan, S., Malik, S., Du, F., Koh, E., and Lai, T. Deepcompare: Visual and interactive comparison of deep learning model performance. *IEEE Computer Graphics and Applications*, PP:1–1, 05 2019. doi: 10.1109/MCG.2019.2919033.

Myers, G. J., Sandler, C., and Badgett, T. *The Art of Software Testing*. Wiley Publishing, 3rd edition, 2011. ISBN 1118031962.

Odena, A., Olsson, C., Andersen, D., and Goodfellow, I. TensorFuzz: Debugging neural networks with coverage-guided fuzzing. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4901–4911, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/odena19a.html.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

PAIR, G. What-if tool, 2019. URL https://pair-code.github.io/what-if-tool/.

Patel, K., Fogarty, J., Landay, J. A., and Harrison, B. Investigating statistical machine learning as a tool for software development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 667–676, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580111. doi: 10.1145/1357054.1357160. URL https://doi.org/10.1145/1357054.1357160.

Patel, K., Bancroft, N., Drucker, S. M., Fogarty, J., Ko, A. J., and Landay, J. Gestalt: Integrated support for implementation and analysis in machine learning. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pp. 37–46, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450302715. doi: 10.1145/1866029.1866038. URL https://doi.org/10.1145/1866029.1866038.

Rolando Garcia, Vikram Sreekanti, D. C. N. Y. J. G. flor, 2019. URL https://github.com/ucbrise/flor.

Shah, S., Fernandez, R., and Drucker, S. M. A system for real-time interactive analysis of deep learning training. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems, EICS 2019, Valencia, Spain, June 18-21, 2019*, pp. 16:1–16:6, 2019. doi: 10.1145/3319499.3328231. URL https://arxiv.org/abs/2001.01215.

Shewchuk, J. R. Concise machine learning, May 2019. URL https://people.eecs.berkeley.edu/~jrs/papers/machlearn.pdf.

Stasko, J. T., Brown, M. H., and Price, B. A. *Software Visualization*. MIT press, 1997.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.