Ten Simple Rules for Deep Learning in Biology

This manuscript (permalink) was automatically generated from Benjamin-Lee/deep-rules@cd442a1 on December 15, 2018.

Authors

- **Benjamin D. Lee**
  0000-0002-7133-8397 · Benjamin-Lee
  School of Engineering and Applied Sciences, Harvard University; Department of Genetics, Harvard Medical School; Lab41, In-Q-Tel

Introduction

Deep learning (DL), a subfield of machine learning (ML) that focuses on deep artificial neural networks, is exploding in popularity and is increasingly being used for biological data analysis [1]. However, DL itself remains an active area of research, and the complexity of this field poses a large barrier of entry to those who would like to correctly utilize state-of-the-art DL technology in biological research applications. While much has been written about the impressive results of DL in biology, there is a comparative dearth of literature articulating best practices for its use, with most instructional literature focusing on ML, rather than DL [2].

To fix this problem, we solicited input from a diverse community of researchers, who wrote this manuscript collaboratively using the GitHub version control platform [3] and Manubot [4]. In the course of our discussions, several themes became clear: the importance of understanding and applying ML fundamentals as a baseline for utilizing DL, the necessity for extensive model comparisons and careful evaluation, and the need for critical thought in interpreting results generated by means of DL, among others. Ultmately, the rules we established range from high-level guidance to the implementation of best practices, and it is our hope that they will provide actionable, DL-specific advice for both new and experienced DL practitioners alike who would like to employ DL in biological research. By increasing the accessibility of DL techniques to biology, we aim to improve the overall quality and reproducibility of DL in the literature, enabling these powerful methods to be properly utilized to generate new scientific insights.

Rule 1: Concepts that apply to machine learning also apply to deep learning

Deep learning is a distinct subfield of machine learning, but it is still a subfield. Deep learning has proven to be an extremely powerful paradigm capable of outperforming “traditional” machine learning approaches, but it is not immune to the many limitations inherent to machine learning. Many best practices for machine learning apply to deep learning as well. For instance, deep supervised learning models should be trained, tuned, and tested on non-overlapping datasets. Those developing deep learning models should select data that are relevant to the problem at hand; non-salient data can hamper performance or lead to spurious conclusions. Furthermore, investigators should begin by thoroughly inspecting their data. When coupled with imprudence, data that is biased, skewed, or of low quality will produce models of dubious performance and limited generalizability. Biases in testing data can also unduly influence measures of model performance. For example, many conventional metrics for classification (e.g. area under the receiver operating characteristic curve or AUROC) have limited utility in cases of extreme class imbalance. As such, model performance should be evaluated with a carefully-picked panel of relevant metrics that make minimal assumptions about the composition of the testing data [5]. Extreme cases warrant testing the robustness of the model and metrics on simulated data for which the ground truth is known. Said simulations can be used to verify the correctness of the model’s implementation as well. Like all computational methods, deep learning should be leveraged in a systematic manner that is reproducible and rigorously tested.

Rule 2: Use traditional methods to establish performance baselines

Rule 3: Understand the complexities of training deep neural networks

Rule 4: Know your data and your question

Rule 5: Choose an appropriate neural network architecture and data representation

Rule 6: Tune your hyperparameters extensively and systematically

Rule 7: Address deep neural networks' increased tendency to overfit the dataset

Rule 8: Do not necessarily consider a DL model as a black box

Rule 9: Interpret predictions in the correct manner

Rule 10: Don't share models trained on sensitive data

One of the greatest opportunities for deep learning in biology is the ability for deep learning techniques to incorporate representation learning to extract information that can not readily be captured by traditional methods [6]. The abundance of features for each training example means that the representation learning of the deep learning models can capture information-rich abstractions of data during the training process. Therefore with both deep learning and traditional machine learning models (*e.g. k*-nearest neighbors models, which learn by memorizing the full

training data), it is imperative not to share models trained on sensitive data. Applying deep learning to images of cats from the internet does not pose significant ethical, legal, or privacy problems; this is not the case when dealing with classified, confidential, trade secret, or other types of biological data that cannot be shared. For example, adversarial training techniques such as model inversion attacks can be used to exploit model predictions to recover recognizable images of people's faces used for training [7]. These risks are even more significant in deep learning compared to traditional machine learning techniques due to the greater representational capacity of the models. This is achieved by the large number of model weights, even in a relatively small project, that allow deep learning to model high-dimensional non-linear relationships among data. It is this enhanced modeling capacity that allows the model to learn more robust and nuanced features of specific data, leading to the danger of revealing the underlying sensitive data. When training deep learning models on sensitive data, be sure not to share the model weights directly, and use privacy preserving techniques [8] such as differential privacy [10,9] and homomorphic encryption [11,12] to protect sensitive data.

Conclusion

References

1. **Opportunities and obstacles for deep learning in biology and medicine**
Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, … Casey S. Greene
Journal of The Royal Society Interface (2018-04) https://doi.org/gddkhn
DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

2. **Ten quick tips for machine learning in computational biology**
Davide Chicco
BioData Mining (2017-12) https://doi.org/gdb9wr
DOI: 10.1186/s13040-017-0155-3 · PMID: 29234465 · PMCID: PMC5721660

3. https://github.com/Benjamin-Lee/deep-review

4. https://greenelab.github.io/meta-review/

5. **Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets**
Alexandru Korotcov, Valery Tkachenko, Daniel P. Russo, Sean Ekins
Molecular Pharmaceutics (2017-11-13) https://doi.org/gcj4p2
DOI: 10.1021/acs.molpharmaceut.7b00578 · PMID: 29096442 · PMCID: PMC5741413

6. **Convolutional Networks on Graphs for Learning Molecular Fingerprints**
David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, Ryan P. Adams
arXiv (2015-09-30) https://arxiv.org/abs/1509.09292v2

7. **Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures**
Matt Fredrikson, Somesh Jha, Thomas Ristenpart
Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15 (2015) https://doi.org/cwdm
DOI: 10.1145/2810103.2813677

8. **A generic framework for privacy preserving deep learning**
Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, Jonathan Passerat-Palmbach
arXiv (2018-11-09) https://arxiv.org/abs/1811.04017v2

9. **Deep Learning with Differential Privacy**
Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, Li

Zhang
Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16 (2016) https://doi.org/gcrnp3
DOI: 10.1145/2976749.2978318

10. **Privacy-Preserving Distributed Deep Learning for Clinical Data**
Brett K. Beaulieu-Jones, William Yuan, Samuel G. Finlayson, Zhiwei Steven Wu
arXiv (2018-12-04) https://arxiv.org/abs/1812.01484v1

11. **SIG-DB: Leveraging homomorphic encryption to securely interrogate privately held genomic databases**
Alexander J. Titus, Audrey Flower, Patrick Hagerty, Paul Gamble, Charlie Lewis, Todd Stavish, Kevin P. O'Connell, Greg Shipley, Stephanie M. Rogers
PLOS Computational Biology (2018-09-04) https://doi.org/gd6xd5
DOI: 10.1371/journal.pcbi.1006454 · PMID: 30180163 · PMCID: PMC6138421

12. **The AlexNet Moment for Homomorphic Encryption: HCNN, the First Homomorphic CNN on Encrypted Data with GPUs**
Ahmad Al Badawi, Jin Chao, Jie Lin, Chan Fook Mun, Sim Jun Jie, Benjamin Hong Meng Tan, Xiao Nan, Khin Mi Mi Aung, Vijay Ramaseshan Chandrasekhar
arXiv (2018-11-02) https://arxiv.org/abs/1811.00778v1