Transfer Learning vs. Batch Effects: what can we expect from neural networks in computational biology?

Alan M. Moses[1,2], Alex X. Lu[1], Amy X. Lu[1,3] and Marzyeh Ghassemi[1,3]

[1]Department of Computer Science, University of Toronto
[2]Department of Cell & Systems Biology, University of Toronto
[3]Vector Institute for Artificial Intelligence

25 Harbord Street, Toronto, Canada M5S 3G5

Abstract

The diverse applications of deep learning in computational biology include single-cell microscopy image analysis and prediction of transcription factor binding from DNA sequence. Although it is clear that CNNs and their derivatives will revolutionize these fields, it is not yet clear to what extent deep models will be transferred, reused or retrained for each application. For single cell identification/segmentation in microscope images, one study found remarkable generalization capacity of a mask-RCNN: with no parameter tuning, performance across microscopy datasets is competitive with conventional methods that have been highly tuned for each dataset. This type of generalization implies that a single model can be deployed over the web for all users. On the other hand, for protein subcellular localization classification in images, there is evidence for sensitivity to 'batch' or 'out-of-sample' effects, such that performance degrades on test sets taken at different times and on different instruments. We discuss similar issues in deep learning methods applied to transcription factor binding. We conclude that the issue of when models can generalize and when they must be retrained is largely unexplored, but will be critical in shaping how deep learning is applied to computational biology.

Introduction

Given the encouraging results of deep learning applications in many areas of computational biology[1], the widespread adoption of these techniques into mainstream bioinformatics methods seems likely. However, it is currently unclear what the future deep computational biology will look like. Recent perspectives emphasize the importance of model complexity and "big data" [2], [3]. These models require extensive compute resources and expertise to train. Current practice in the IT industry is that only a few large players design and train the state-of-the-art models, and then these are deployed by others in applications through programmatic interfaces[4], [5]. Smaller, niche-specific problems are solved by either transferring directly or "fine-tuning" based on pre-trained model designs and parameters. There are conflicting reports about the efficacy of these strategies, exemplified by results from classification of natural images: on the one hand, models trained on large datasets have shown remarkable generalizability to other problems in image analysis, such that direct use of pre-trained models (so-called "transfer learning") on new problems is a key baseline for any new method[6]. On the other hand, recent research suggests that at a quantitative level, classification results obtained are not generalizable to even to new datasets that have been constructed to be similar to the originals[7]. Given the well-appreciated batch effects and other biases in large biological datasets[8], it is likely that similar issues will arise in computational biology applications as well: deep, non-linear encoders have unprecedented power to capture subtle biological signals, but may also have unprecedented sensitivity to subtle data distribution non-stationarity, batch effects, imbalance, etc. Indeed "batch" effects were given as motivation for the release of a new collection of microscopy images[9] and associated competition[10].

It is currently unclear how to design deep learning methods for biological data so that they can generalize to data from new experiments that were not available during the training process. Key questions surrounding design decisions that may affect generalizability include: training data size and diversity, architecture depth and complexity, supervised vs. unsupervised training, data augmentation procedures and losses. Here we review some recent findings that relate to generalizability vs. sensitivity to out-of-sample effects. We first focus on analysis methods for high-throughput microscopy image datasets, where convolutional neural networks have easily achieved state-of-the-art performance[1]. We then discuss possible generalization issues for deep learning approaches to functional genomics data, and finally a recent attempt to use deep language models to predict protein structure.

Conflicting results from single cell image analysis: remarkable generalizability for cell segmentation, but not for classification

Two key problems in single cell microscopy image analysis are cell segmentation and classification. The first task is simply to identify the cells or nucleii (often based on a nuclear or cell periphery stain or marker). Like many classical problems in image analysis, when the number of cells is larger than a handful and not known *a priori*, and when cells may be clumped together, dividing or touching, solutions are sensitive to changes in signal-to-noise, lighting, magnification, etc. As expected, UNets[11] easily achieved state-of-the-art performance on nucleus identification in fluorescence images [12] but the authors noted large performance drops when testing on new datasets that were collected by different labs with different instruments [12]. These results indicated that generalization appears to be a significant challenge for deep learning methods.

On the other hand, at least two highly general methods have been reported. Cell segmentation was the subject of a 2018 Kaggle competition[13], and the top-ranking approaches trained mask-RCNNs, an advanced CNN-based model (developed to segment objects in natural images[14]), on a diverse collection of microscope images of mostly mammalian cells with segmented nucleii. Remarkably, (at least) one of these models had unexpected generalization capacity to identify yeast cells: with no parameter tuning at all, it outperformed conventional segmentation methods (based on 2D-HMMs and watershed refinement[15]) that had been developed and trained specifically for high-throughput yeast fluorescence microscopy image collections (Table 1). In addition, the mask-RCNN (which was named YeastSpotter[16]) obtained competitive accuracy on several benchmark datasets with tools for segmenting yeast cells in brightfield images, all of which had to be tuned to obtain good results for each dataset[17]. Unpublished results[18] suggest that similar generalization capacity may be possible for human cells, such that a single method can identify cells in any type of images (although this method has not yet been evaluated on out-of-sample datasets.) These methods have been made available as webtools, offering the first truly general microscope image segmentation to biologists: users simply upload images and obtain results.

Although it is not currently clear whether the more sophisticated architectures (mask-RCNNs vs. Unet) or differences in the training data diversity or augmentation are responsible for the observed differences in generalization capacity, these recent results suggest that, at least in principle, deep-learning methods can provide unprecedented generality in single cell identification in microscope images.

Table 1. Of-the-shelf transfer learning for a mask-RCNN on fluorescent yeast micrographs (from[16])

Method	Ellipses Matched	Mean	Standard Deviation	Correlation	Run Time
YeastSpotter	97.5%	1.58	0.99	0.969	1172
Engineered	92.3%	1.41	1.21	0.928	13851
CellProfiler	89.0%	2.23	1.80	0.876	231

Percent of manual ellipses with a matched single-cell segmentation within 10 pixels, the mean and standard deviation of distance (in pixels) between the centers of the manual ellipse and segmentation, the correlation between their areas, and the time (in seconds) to process the evaluation image set (68 images). YeastSpotter is a mask-RCNN trained to identify nuclei used “off-the-shelf” on yeast images that do not resemble any images in its training set. Engineered is a method designed for this dataset[15] and CellProfiler[19] is a field standard general microscope image analysis package.

CNN-based methods have also easily achieved state-of-the-art performance in single cell classification[1], another widely studied problem in microscopy image analysis. In this problem, batch-effects are well-known. In one of the first applications of deep learning to single cell image data, an 11-layer CNN (DeepLoc) was trained to classify yeast cells into subcellular localization classes[20]. While some transfer capacity to a different dataset was reported, retraining using labelled data for each class was required for reasonable performance. In human cell phenotype classification, authors reported significant challenges due to batch effects in generalizing their classification results, even within the same dataset[21]. A recent study designed to directly test generalization of classifiers used a large, diverse dataset of mouse cells from 7 localization classes (COOS-7 [22]). In this dataset, out-of-sample (different days, microscopes, etc.) and within-sample (random subset) test image datasets of the same cells were compared directly to test generalization capacity. While, overall, CNN-based methods show the highest classification accuracies, they still suffer performance degradation when the test set is from a different sample than the training data (Table 2). We emphasize that this is not due to simple "overfitting" because even the within-sample test set is unseen during training. Thus, unlike for single cell segmentation, in single cell classification, highly generalizable models have not yet been obtained.

Table 2. Class-Balanced Error (%) of classification models on out-of-sample tests (from[22])

Method	Train	Test1	Test2	Test3	Test4
Supervised (DeepLoc)	1.2	1.2	1.5	7.4	5.4
Transfer (VGG16)	2.8	3.9	3.9	8.0	6.8
Self-Supervised (PCI)	1.0	1.4	1.7	9.2	7.4
Texture features	6.4	6.8	6.5	12.0	12.1

Comparison of best performing classifiers on COOS-7. 'Train' is a diverse training set of >41k images, 'Test1' is a within-sample randomly held out set of 10k images, and Test2, Test3 and Test4 are sets of >15k images each from other imaging wells, days and miscroscope. DeepLoc is an end-to-end 11-layer supervised CNN-based classifier[20]. Transfer learning following [23] using a L1-logistic regression classifier on the second layer of the 4th convolutional block of VGG16 features. For self-supervised, the features from the 3rd convolutional layer of a CNN trained using Paired-Cell-Inpainting (PCI,[24]) were used in an L1-logistic regression classifier. Texture features are classical rotation-invariant features used for microscopy image analysis used as input to an L1-logistic regression classifier. In each case, the models were trained only on the training data set, and for the unsupervised methods, L1 logistic regression was the best performing classification strategy.

Generalization may be a challenge for deep-learning models for transcription factor binding

One area where deep models have achieved state-of-the-art performance in genomics is the prediction of protein-DNA binding based on *in vivo* and *in vitro* binding data [25]. Architectures have been comprehensively tested, and performance on held out data from a different assay was reported[25]. Although the two assays are not directly comparable (there may be *bona fide* biological differences in binding), and only three datasets were compared, the results are consistent with substantial performance decreases when test data are derived from other experiments (Table 3).

Table 3. Average AUC over 3 transcription factor binding prediction models (from[25])

Method	Same assay	Different assay
DeepBind	93.5	88.7
ECBLSTM	95.9	91.3

DeepBind[26] was among the first neural network methods for prediction of binding and ECBLSTM is a top-performing method based on an advanced language-model from a recent systematic comparison. ‘Same assay’ is the results of training and testing models on ChIP-seq data, while ‘Different assay’ trains models on SELEX data and tests them on ChIP-seq data.

A closely related problem is prediction of cell-type specific transcription factor binding. The ENCODE-DREAM challenge asked participants to predict transcription factor binding in diverse cell types using genome sequence, DNAase-seq and RNA-seq[27]. Among the best performing methods in the ENCODE-DREAM challenge were discriminative approaches related to regression or ensemble discriminative classification approaches[28]. These results are surprising because the subtle syntax of enhancers (e.g., [29]) should give deep, non-linear encoders a clear advantage over conventional approaches. As described above, CNN-based methods have shown state-of-the-art performance in closely related problems for similar datasets, although they were often evaluated in their ability to discriminate bound sequences from randomized sequences [25], thus balancing the classification problem and improving the signal to noise relative to genomic sequences. Taken together, although the question of generalization has not been explored directly, these reports suggest that generalization to new, diverse genomic datasets may be a challenge for deep learning methods in functional genomics.

Discussion and outlook

Large-scale biological data contain experimental and biological nuisance variation that, when inconsistent across experiments, impacts generalization performance. These so-called covariate shifts, out-of-sample or batch effects have been addressed to some extent in classical statistical approaches [8]. Unlike conventional statistical or model-based machine learning approaches, however, optimizing the hyperparameters of supervised deep learning models requires extensive data and computational resources.

Here we have focused on some initial results relating to these issues in deep-learning applications in microscopy image analysis and transcription factor binding prediction. Even less is known in other applications of deep learning in computational biology. However, a recent study reported end-to-end protein structure prediction based on amino acid sequences[30]. This study included evaluations on predictions of several CASP competitions, but trained hyperparameters on only one. This leaves the results of the other competitions as tests of generalization. Interestingly, this study suggests lower performance on the other competitions, where hyperparameters were not optimized[30]. Once again, we emphasize that this is not overfitting: the test data are always withheld from the model at training time. This report highlights the difficulty of the issue: although the forward pass through the trained deep learning method is millions of times faster than the conventional competitors, the hyperparameters of the model were not optimally tuned for each dataset due to computational resource constraints[30]. The difficulty in optimizing supervised deep learning models suggests that out-of-sample generalization should ideally be incorporated in model design. Unfortunately, seemingly related machine-learning research on meta-learning and domain adaptation has not yet revealed clear direction about which design decisions affect generalizability.

In at least one case, a highly generalizable single cell identification model has been reported[16], suggesting that one-model-fits-all is an attainable goal for some classic problems in computational biology. The extent to which this will be possible for other problems is an open question. Ultimately the generalization capacity of models may shape whether computational biology converges to a small number of big, universal models, or a large number of small models, each trained for new datasets.

References

[1] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, “Deep learning for computational biology,” *Mol. Syst. Biol.*, vol. 12, no. 7, Jul. 2016, doi: 10.15252/msb.20156651.

[2] "Deep learning in bioinformatics: Introduction, application, and perspective in the big data era," *Methods*, vol. 166, pp. 4–21, Aug. 2019, doi: 10.1016/j.ymeth.2019.04.008.

[3] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *J. R. Soc. Interface*, vol. 15, no. 141, 2018, doi: 10.1098/rsif.2017.0387.

[4] "Face API - Facial Recognition Software | Microsoft Azure." [Online]. Available: https://azure.microsoft.com/en-ca/services/cognitive-services/face/. [Accessed: 04-Oct-2019].

[5] "Dialogflow API | Dialogflow," *Google Cloud*. [Online]. Available: https://cloud.google.com/dialogflow/docs/reference/rest/v2-overview. [Accessed: 04-Oct-2019].

[6] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Washington, DC, USA, 2014, pp. 512–519, doi: 10.1109/CVPRW.2014.131.

[7] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet Classifiers Generalize to ImageNet?," in *International Conference on Machine Learning*, 2019, pp. 5389–5400.

[8] J. T. Leek *et al.*, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nat. Rev. Genet.*, vol. 11, no. 10, Oct. 2010, doi: 10.1038/nrg2825.

[9] "RXRX." [Online]. Available: https://www.rxrx.ai/. [Accessed: 30-Dec-2019].

[10] "Recursion Cellular Image Classification." [Online]. Available: https://kaggle.com/c/recursion-cellular-image-classification. [Accessed: 30-Dec-2019].

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015.

[12] J. C. Caicedo *et al.*, "Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images," *bioRxiv*, p. 335216, Feb. 2019, doi: 10.1101/335216.

[13] "2018 Data Science Bowl." [Online]. Available: https://kaggle.com/c/data-science-bowl-2018. [Accessed: 04-Oct-2019].

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," Mar. 2017.

[15] L.-F. Handfield, Y. T. Chong, J. Simmons, B. J. Andrews, and A. M. Moses, "Unsupervised clustering of subcellular protein expression patterns in high-throughput microscopy images reveals protein complexes and functional relationships between proteins," *PLoS Comput. Biol.*, vol. 9, no. 6, p. e1003085, 2013, doi: 10.1371/journal.pcbi.1003085.

[16] A. X. Lu, T. Zarin, I. S. Hsu, and A. M. Moses, "YeastSpotter: Accurate and parameter-free web segmentation for microscopy images of yeast cells," *Bioinforma. Oxf. Engl.*, May 2019, doi: 10.1093/bioinformatics/btz402.

[17] C. Versari *et al.*, "Long-term tracking of budding yeast cells in brightfield microscopy: CellStar and the Evaluation Platform," *J. R. Soc. Interface*, vol. 14, no. 127, p. 20160705, Feb. 2017, doi: 10.1098/rsif.2016.0705.

[18] R. Hollandi *et al.*, "A deep learning framework for nucleus segmentation using image style transfer," *bioRxiv*, p. 580605, Mar. 2019, doi: 10.1101/580605.

[19] L. Kamentsky *et al.*, "Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software," *Bioinforma. Oxf. Engl.*, vol. 27, no. 8, pp. 1179–1180, Apr. 2011, doi: 10.1093/bioinformatics/btr095.

[20] O. Z. Kraus *et al.*, "Automated analysis of high-content microscopy data with deep learning," *Mol. Syst. Biol.*, vol. 13, no. 4, p. 924, Apr. 2017, doi: 10.15252/msb.20177551.

[21] D. M. Ando, C. Y. McLean, and M. Berndl, "Improving Phenotypic Measurements in High-Content Imaging Screens," *bioRxiv*, p. 161422, Jul. 2017, doi: 10.1101/161422.

[22] A. X. Lu, A. X. Lu, W. Schormann, D. W. Andrews, and A. M. Moses, "The Cells Out of Sample (COOS) dataset and benchmarks for measuring out-of-sample generalization of image classifiers," *arXiv.org*, Jun. 2019.

[23] N. Pawlowski, J. C. Caicedo, S. Singh, A. E. Carpenter, and A. Storkey, "Automating Morphological Profiling with Generic Deep Convolutional Networks," *bioRxiv*, p. 085118, Nov. 2016, doi: 10.1101/085118.

[24] A. X. Lu, O. Z. Kraus, S. Cooper, and A. M. Moses, “Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting,” *PLoS Comput. Biol.*, vol. 15, no. 9, p. e1007348, Sep. 2019, doi: 10.1371/journal.pcbi.1007348.

[25] A. Trabelsi, M. Chaabane, and A. Ben-Hur, “Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities,” *Bioinformatics*, vol. 35, no. 14, pp. i269–i277, Jul. 2019, doi: 10.1093/bioinformatics/btz339.

[26] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning,” *Nat. Biotechnol.*, vol. 33, no. 8, pp. 831–838, Aug. 2015, doi: 10.1038/nbt.3300.

[27] “ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge - syn6131484.” [Online]. Available: https://www.synapse.org/#!Synapse:syn6131484/wiki/415139. [Accessed: 03-Oct-2019].

[28] J. Keilwagen, S. Posch, and J. Grau, “Accurate prediction of cell type-specific transcription factor binding,” *Genome Biol.*, vol. 20, no. 1, p. 9, Jan. 2019, doi: 10.1186/s13059-018-1614-y.

[29] E. K. Farley, K. M. Olson, W. Zhang, D. S. Rokhsar, and M. S. Levine, “Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 23, pp. 6508–6513, 07 2016, doi: 10.1073/pnas.1605085113.

[30] M. AlQuraishi, “End-to-End Differentiable Learning of Protein Structure,” *Cell Syst.*, vol. 8, no. 4, pp. 292–301.e3, Apr. 2019, doi: 10.1016/j.cels.2019.03.006.