Editorial Commentary

Leveraging well-annotated databases for deep learning in biomedical research

Nam Nhut Phan[1,2,3], Amrita Chattopadhyay[3], Tzu-Pin Lu[3,4], Mong-Hsun Tsai[3,5,6]

[1]Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei; [2]Graduate Institute of Biomedical Electronics and Bioinformatics, Department of Electrical Engineering, National Taiwan University, Taipei; [3]Bioinformatics and Biostatistics Core, Centre of Genomic and Precision Medicine, National Taiwan University, Taipei; [4]Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei; [5]Institute of Biotechnology, National Taiwan University, Taipei; [6]Center of Biotechnology, National Taiwan University, Taipei

Correspondence to: Prof. Mong-Hsun Tsai. 4F, No. 81, Chang-Xing st., Taipei. Email: motiont@gmail.com.

Submitted Nov 02, 2020. Accepted for publication Nov 19, 2020.
doi: 10.21037/tcr-20-3163
View this article at: http://dx.doi.org/10.21037/tcr-20-3163

Buzzwords indicate popular trends in research fields. These terms might last for decades or perish in just a few years (1). Over the last ten years, we have witnessed the rise of a big buzzword-deep learning (DL) (2-7). In brief, DL is a sub-domain of artificial intelligence (AI), a type of representation learning, which automatically finds features in a data and transforms them into a higher abstract data based on matrix operations (3). There are various types of DL algorithms such as convolutional neural networks (8), recurrent neural networks (9), long short-term memory networks (10), convolutional deep belief networks (11), generative adversarial networks (12), and deep residual networks (13), just to name a few. Depending on the specific task/problem, one could use these networks individually or combine them into a pipeline. The biggest advantage of DL algorithms is that they can be trained without pre-defined features/variables, which is especially convenient for complicated data types, such as biomedical images or sequencing data, that are time-consuming and computationally expensive and require a high level of human expertise for feature selection (3). Moreover, high-end facilities such as graphical processing units, central processing units, and random-access memory are needed for processing and training such data within a reasonable amount of time.

A growing body of research related to neural network applications for solving problems in the biomedical field includes diverse research topics that commonly leverage big data. This includes biomedical images and multi-omics datasets either from public domain or in-house data from different populations (6,14-16). Biomedical images can be in 2-dimensional (2D) format such as pathological images, or 3-dimensional (3D) such as with mammography images, computed tomography scans, and magnetic resonance imaging (17-22). A single scanned image could be split into hundreds to several thousands of smaller images, which easily complies with the data demands of neural network training. The data formats for multi-omics data is even more complicated and are highly dependent on the manufacturing platforms. The omics data, such as genomics (sequencing data) (23), transcriptomics (sequencing and expression data) (24,25), proteomics (mass spectrometry data) (26), and metabolomics (metabolite compounds) (27), can be used for DL models as long as the number of samples and features is suitable for training and can achieve acceptable accuracy. From only a single run, these high-throughput platforms can generate thousands to millions of data points from each sample. Integrating these could provide an unprecedentedly comprehensive data to study the complicated diseases such as cancer (28) or human brain diseases (29). Therefore, this is a golden era for data-driven research, not only due to the huge amount of publicly available datasets, but also because of the rapid development of modern algorithms and giant technology corporations such as Google (TensorFlow and CoLab cloud computing) (30-32), Amazon (Amazon Web Services) (33), and Facebook (PyTorch) (34) and their platforms and cloud computing services. With such favorable conditions and the available open-source environments of the DL

community, it is inevitable that biomedical researchers start to enter the race of DL. Moreover, several databases are available that house a huge number of biomedical images such as the National Cancer Institute's GDC Data Portal (https://portal.gdc.cancer.gov), the National Institutes of Health Database (https://nihcc.app.box.com/v/ChestXray-NIHCC), the Cancer Imaging Archive (https://www.cancerimagingarchive.net), NLM's MedPix database (https://medpix.nlm.nih.gov/home), the Open Access Series of Imaging Studies (OASIS) (http://www.oasis-brains.org), the Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu), and Stanford's AI in Medicine database (AIMI) (https://aimi.stanford.edu/research/public-datasets); all of these could be of immense advantage for the DL community. These databases are maintained and continuously updated with additional samples and data types and play a central role in DL studies due to their well-structured and diverse disease sources. For instance, the GDC data portal can provide whole exome sequencing data, targeted sequencing data, RNA-sequencing data, genotype data, tissue and diagnostic slides, whole genome data, and ATAC-seq data. All of these data are not fully open access, but researchers can apply for access to the controlled portions of the data. However, model training on such large datasets requires data labeling and annotation, which are time-consuming and sometimes expensive, so there are still barriers to the use of all the available data.

Many DL publications describe well-annotated datasets; however, gaining access to these resources is usually difficult. Access to in-house datasets, pre-annotated by experts, is still in demand, for the benefit of the healthcare research community. As the public domain data are usually specific to ethnic groups or local populations, other in-house datasets from varied ethnicities could serve as an external validation resource to prevent model bias of certain datasets. That would ultimately make the pre-trained model more useful across populations.

Clinical application is the ultimate goal in biomedical research. Therefore, the questions or hypotheses that researchers aim to address with DL, leveraging all ready-to-use data and resources, is of utmost clinical importance. This is what leads to the proper design of models that represent complex real-life data, and potentially provide data-driven information for clinical research. All of this requires close collaboration between laboratory researchers and medical doctors, to understand the current needs in each specific disease and successfully translate findings from the laboratory bench to the clinic.

Acknowledgments

We thank Dr. Melissa Stauffer for editing this manuscript.
Funding: This work has been supported in part by the Center of Genomic and Precision Medicine, National Taiwan University, Taiwan [106R8400]; and the Center for Biotechnology, National Taiwan University, Taiwan [GTZ300].

Footnote

Provenance and Peer Review: This article was a free submission to the journal. The article did not undergo external peer review.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/tcr-20-3163). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Parchomovsky G. Publish or perish. Mich Law Rev 2000;98:926-52.
2. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24-9.
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
4. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018;2:158-64.
5. Horst F, Lapuschkin S, Samek W, et al. Explaining the unique nature of individual gait patterns with deep learning. Sci Rep 2019;9:2391.

6. Belthangady C, Royer LA. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. Nat Methods 2019;16:1215-25.
7. Oustimov A, Vu V. Artificial neural networks in the cancer genomics frontier. Transl Cancer Res 2014;3:191-201.
8. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems; 2012.
9. Sutskever I, Martens J, Hinton GE. Generating text with recurrent neural networks. ICML; 2011.
10. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735-80.
11. Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Available online: http://www.robotics.stanford.edu/~ang/papers/icml09-ConvolutionalDeepBeliefNetworks.pdf
12. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Available online: https://papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
13. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
14. Alipanahi B, Delong A, Weirauch MT, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 2015;33:831-8.
15. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
16. Zou J, Huss M, Abid A, et al. A primer on deep learning in genomics. Nat Genet 2019;51:12-8.
17. Zhao Z, Yang L, Zheng H, et al. Deep learning based instance segmentation in 3d biomedical images using weak annotation. Available online: https://arxiv.org/abs/1806.11137
18. Pawlowski N, Ktena SI, Lee MC, et al. Dltk: State of the art reference implementations for deep learning on medical images. arXiv preprint arXiv:171106853 2017.
19. Lee CS, Tyring AJ, Deruyter NP, et al. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. Biomed Opt Express 2017;8:3440-8.
20. Pan X, Li L, Yang H, et al. Accurate segmentation of nuclei in pathological images via sparse reconstruction and deep convolutional networks. Neurocomputing 2017;229:88-99.
21. Lu S, Lu Z, Zhang YD. Pathological brain detection based on AlexNet and transfer learning. J Comput Sci 2019;30:41-7.
22. Vial A, Stirling D, Field M, et al. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. Transl Cancer Res 2018;7:803-16.
23. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods 2008;5:16-8.
24. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57-63.
25. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001;2:418-27.
26. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198-207.
27. Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrom Rev 2007;26:51-78.
28. Chaudhary K, Poirion OB, Lu L, et al. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res 2018;24:1248-59.
29. Xicota L, Ichou F, Lejeune FX, et al. Multi-omics signature of brain amyloid deposition in asymptomatic individuals at-risk for Alzheimer's disease: The INSIGHT-preAD study. EBioMedicine 2019;47:518-28.
30. Abadi M, Barham P, Chen J, et al. Tensorflow: A system for large-scale machine learning. 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16); 2016.
31. Rampasek L, Goldenberg A. TensorFlow: Biology's Gateway to Deep Learning? Cell Syst 2016;2:12-4.
32. Bisong E. Google Colaboratory. Building Machine Learning and Deep Learning Models on Google Cloud Platform. Berlin: Springer, 2019:59-64.
33. Miller FP, Vandome AF, McBrewster J. Amazon web services. Alpha Press, 2010.
34. Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems; 2019.

Cite this article as: Phan NN, Chattopadhyay A, Lu TP, Tsai MH. Leveraging well-annotated databases for deep learning in biomedical research. Transl Cancer Res 2020;9(12):7682-7684. doi: 10.21037/tcr-20-3163

