Brief Communication

Machine learning in agriculture: from silos to marketplaces

Philipp E. Bayer and David Edwards*

School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, WA, Australia

Keywords: deep learning, data silos, data-driven agriculture, data sharing, open data.

Introduction

The increasing global population combined with climate change present a major challenge for agriculture. Most crops have been bred to perform in specific environments, and with the time required to produce new varieties, it is unlikely that breeders will be able to adapt varieties to the changing climate (Abbenton et al., 2016). There is an urgent need to develop new approaches to accelerate the production of high performing resilient crop varieties. Crop breeding has seen many changes in recent decades, from the application of molecular markers, through to genetically modified, and more recently, genome-edited crops (Scheben et al., 2017). However, these approaches are often limited by our lack of understanding of the genomic basis for complex traits, even with the deluge of data being generated by new genome sequencing and phenomics technologies. New approaches are required to translate this explosion of data into improved crop varieties.

Deep learning represents a set of machine learning approaches that are not new, but have been termed major advances in the last few years, with the adoption of deep learning in robotics, smart cars, smart homes, and agriculture. The breakthroughs in deep learning have not been driven by major advances in deep learning methods but rather by the increasing availability of large labelled training data as well as advances in computational hardware, especially graphics processors (GPUs). With the continued increase in agricultural phenotype and genotype data, there are opportunities to apply deep learning to accelerate crop breeding and agricultural productivity (Figure 1).

Deep learning approaches are most advanced in the field of image recognition (Kamliras and Prenafeta-Boldú, 2018); for example, Convolutional Neural Networks (CNNs, a class of deep neural network) have been developed to count wheat spikes and spikelets with 95.91% and 99.66% accuracy (Pound et al., 2017). For training, a novel dataset with 520 images of wheat plants was generated. The data were annotated by an expert who manually counted the 4100 ears and 48 000 spikelets, and whether the image contains an awned phenotype.

In the field, images taken by aerial drones of lettuce fields were used to accurately count lettuce plants and predict yields at a fraction of the cost of manual counting (Bauer et al., 2019). This was a larger dataset than the wheat spikelet counting data: 60 subsections of lettuce-growing fields had their pictures taken from light aircraft. Each subsection contained between 300 and 1000 lettuce heads. For training purposes, each lettuce head was manually labelled with a red dot. This work therefore required the labelling of up to 60 000 lettuce heads. The complete training set, including labelled regions containing no lettuce heads, with 100 000 manually generated and curated data points.

Another example is SeedGerm, a breeding platform that combines cost-effective custom hardware implementation with a graphical user interface that uses three different machine learning approaches to automatically phenotype commercial seeds in the growth chamber (Colmer et al., 2020). The performance of SeedGerm matches that of experts when asked to score radicle emergence.

Outside of image recognition, there are few examples of the application of deep learning in crops. The prediction of phenotypes from genotypes is increasingly being applied in crop breeding but remains challenging due to the complexity of interactions in the data; however, this may be addressed using deep learning approaches. DeepGS is a CNN trained with data from 2403 Iranian bread wheat landraces from CIMMYT’s wheat gene bank (Ma et al., 2018). The data consisted of 33 709 DArT markers per individual, and each individual was assessed for 8 phenotypes of agronomic importance such as grain hardness and plant height. Trained using the DArT markers to predict these eight phenotypes, DeepGS led to improvements in prediction accuracy between 1% and 65% greater than the state-of-the-art RR-BLUP approach. A later publication used a CNN similar to DeepGS along with many other state-of-the-art machine learning approaches on six plant datasets (Azodi et al., 2019). These datasets ranged from 332 178 markers for 391 maize individuals to 4420 markers for 5014 soybean individuals and included three phenotypes per species phenotyped for height, flowering time, yield, grain moisture, time to R8 developmental stage, diameter at breast height, wood density, and standability. Six linear and six nonlinear approaches, including Artificial Neural Networks (ANNs) and CNNs, were tested for accuracy in phenotype prediction. None of these approaches, including the more complex ANN and CNN, outperformed all other approaches consistently, with no clear link between different traits and how their prediction performed in different algorithms. This supports a large body of work that machine learning experts need to consider and assess...
multiple diverse approaches. At the same time, these results make clear that machine learning methods will not replace skilled plant breeders. Instead, these methods will support their work, making it more accurate and reliable.

A major limitation to the advancement of deep learning in agriculture is the availability of suitable high-quality labelled data. As outlined above, deep learning approaches need tens of thousands of data points to make highly accurate predictions possible. Deep learning is also somewhat vulnerable to unseen conditions and therefore requires data from as many distinct conditions as possible. Vast quantities of genomic and phenotypic data are generated each year; however, this is predominantly maintained in data silos, and training labels are either missing or not curated. Remaining unlabelled data can be either manually labelled by experts or labels can be imputed using well-trained machine learning solutions. Commercial systems exist to automatically capture farm-scale data, but due to the lack of agreed standards these data often remain within the data silos of the company which built the data capture solution. However, there is an increasing culture of sharing and integrating crop data, and some bodies such as the AgBioData consortium, the wheat information system and international rice informatics consortium are establishing best practices for integrating crop data across data silos (Harper et al., 2018; Scheben et al., 2018).

Along with the exchange of data, there is a requirement for an exchange of knowledge and skills as technical approaches develop. Choosing which particular deep learning architecture or data-cleaning steps to use in a new project is still an issue of experience which leads to a second type of silo: the silo of skills. Skills are often distributed by informal social media, blogs, forum posts, and communities of practitioners such as Kaggle, which complement the more traditional scientific publications, reference books and presentations at conferences. Practitioners seeking to apply machine learning in agriculture will need to connect to these streams in order to assess models and drive the field forward.

If we are going to continue to accelerate crop production, we need to capture, label and connect as much data as possible, to train complex machine learning models and apply them to crop breeding. This will require that plant breeding companies, universities and other research institutions share large amounts of clean and labelled data from their own data silos in agreed formats, as well as connect the silos of knowledge and skills needed to interpret this data. Connecting these disparate data requires an enormous effort by researchers with experience in handling and normalizing ‘dirty’ data together with enough domain knowledge to assess the quality of agricultural data. This needs to be paired with skilled practitioners on the data-generating side who manually label training datasets with an understanding of what is required of the labels and the data by the machine learning practitioners. Currently, these sets of skills are rare. Over time, some of the skills required may be supplanted by automated machine learning (AutoML) approaches, where parts of the machine learning pipeline are supervised and customized by other learning algorithms, allowing non-experts to perform the required analyses. To our knowledge, no AutoML implementations aimed at non-experts in the agriculture space are available.

Only by merging disparate silos will there be a deep learning revolution in agriculture. Data from different silos and backgrounds will result in models that generalize much better than models trained on data from only one source. Given the increased prediction

Figure 1 An overview of the lifecycle of data in farming. First, data are generated using various observational methods, from cameras to drones to satellites. Then, these data are analysed using different machine learning methods. The predictions of these techniques will then be used to inform business decisions. Outcomes of these business decisions will result in different data being generated in the next growing season, restarting the cycle.
accuracy with larger datasets, we expect that within the next few years, companies and academia will increasingly pool their training data and share their experience, providing a significant advantage over groups which maintain their data and skills within silos.

Acknowledgements

PB acknowledges the support of the Forrest Research Foundation and would like to thank the Perth Machine Learning Group for valuable discussions.

Funding

This work is funded by the Australia Research Council (Projects LP160100030, LP140100537 and LP130100925) and Grains Research and Development Corporation (Projects 9177539 and 9177591).

Conflict of interest

The authors declare no conflict of interest.

Author contribution

PB and DE co-wrote the paper.

References

Abberton, M., Batley, J., Bentley, A., Bryant, J., Cai, H., Cockram, J., Costa de Oliveira, A. et al. (2016) Global agricultural intensification during climate change: a role for genomics. Plant Biotechnol. J. 14, 1095–1098.

Azodi, C.B., Bolger, E., McCarren, A., Roantree, M., de Los Campos, G. and Shiue, S.H. (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3: Genes - Genomes – Genet. 9, 3691–3702.

Bauer, A., Bostrom, A.G., Ball, J., Applegate, C., Cheng, T., Laycock, S., Rojas, S.M. et al. (2019) Combining computer vision and deep learning to enable ultra-scale aerial phenotyping and precision agriculture: A case study of lettuce production. Hortic. Res. 6, 70.

Colmer, J., O’Neill, C.M., Wells, R., Bostrom, A., Reynolds, D., Websdale, D., Shiralagi, G. et al. (2020) SeedGerm: a cost-effective phenotyping platform for automated seed imaging and machine-learning based phenotypic analysis of crop seed germination. New Phytol. 228, 778–793.

Harper, L., Campbell, J., Cannon, E.K.S., Jung, S., Poelchau, M., Walls, R., Andorf, C. et al. (2018) AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. Database, 2018. https://doi.org/10.1093/database/bay088

Kamilaris, A. and Prenafeta-Boldu, F.X. (2018) Deep learning in agriculture: A survey. Comput. Electron. Agric. 147, 70–90.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J. and Ma, C. (2018) A deep convolutional neural network approach for predicting phenotypes from genotypes. Plant. 248, 1307–1318.

Pound, M.P., Atkinson, J.A., Wells, D.M., Priddmore, T.P. and French, A.P. (2017) Deep Learning for Multi-task Plant Phenotyping. bioRxiv.

Scheben, A., Chan, C.K., Mansuetto, L., Mauleon, R., Larmande, P., Alexandrov, N., Wing, R.A. et al. (2018) Progress in single-access information systems for wheat and rice crop improvement. Briefings Bioinform. 20, 565–571.

Scheben, A., Wolter, F., Batley, J., Puchta, H. and Edwards, D. (2017) Towards CRISPR/Cas crops – bringing together genomics and genome editing. New Phytol. 216, 682–698.