Unlocking Comparative Plant Scoring with Siamese Neural Networks and Pairwise Pseudo Labelling

Zane K. J. Hartley
School of Computer Science
University of Nottingham
NG8 1BB
zane.hartley@nottingham.ac.uk

Rob J. Lind
Syngenta
Jealott's Hill International Research Centre
RG426EY
rob.lind@syngenta.com

Nicholas Smith
Syngenta
Jealott's Hill International Research Centre
RG426EY
nicholas-1.smith@syngenta.com

Bob Collison
Syngenta
Jealott's Hill International Research Centre
RG426EY
bob.collison@syngenta.com

Andrew P. French
School of Computer Science
University of Nottingham
NG8 1BB
andrew.p.french@nottingham.ac.uk

Abstract

Phenotypic assessment of plants for herbicide discovery is a complex visual task and involves the comparison of a non-treated plant to those treated with herbicides to assign a phytotoxicity score. It is often subjective and difficult to quantify by human observers. Employing novel computer vision approaches using neural networks in order to be non-subjective and truly quantitative offers advantages for data quality, leading to improved decision making.
In this paper we present a deep learning approach for comparative plant assessment using Siamese neural networks, an architecture that takes pairs of images as inputs, and we overcome the hurdles of data collection by proposing a novel pseudo-labelling approach for combining different pairs of input images. We demonstrate a high level of accuracy with this method, comparable to human scoring, and present a series of experiments grading Amaranthus retroflexus weeds using our trained model.

1. Introduction

Biological assessment by human observation can have subjective and non-quantitative outcomes which makes downstream decisions to determine relative and comparative scoring difficult, and statistical analysis more challenging. Computer vision approaches have the potential to be non-subjective and truly quantitative, so offer many advantages for this task. Convolutional Neural Networks have proven themselves effective at a range of phenotypic analyses, and have advantages over human scorers due to their efficiency and consistency making them highly applicable to our chosen problem.

Biological assessment needs comparative yardsticks to cope with intra test variability; in experiments, these are primarily represented by the negative control ('untreated') treatments. In this case study, the biological subjects were glasshouse-grown plants whose features, called the phenotype, vary due to the season, husbandry and seed batch. Plants are assessed after being treated with herbicides. Thus, to assess a test, the treatment is compared to the in-test control, and a relative score assigned. Thus, in practical terms, a comparison is generated evaluating the relative difference in appearance of two plants. This is a challenging task for human observers, which requires regular calibration between scorers to try and avoid intra assessor variation and drift over time. A computer vision approach to evaluating this relative difference in appearance

could be much more stable and consistent across time.

In this paper, we conceptualize the problem as one of *similarity* scoring, as it is the relative affect on phenotypes of the treatment and control that is key to determining the toxicity of a given herbicide. To this end, we propose to use a Siamese neural network to analyse pairs of images representing a given treatment and an in-test control, and regress a single score for the two images.

Overall our main contributions are as follows:

1. We present a novel computer vision approach to the problem of comparative plant grading based on a similarity-scoring approach.

2. We enhance performance of our model using a pairwise pseudo-labelling approach, allowing us to significantly increase the overall number of training samples without collecting more data.

3. We demonstrate near human-level performance when testing our approach on an unseen test dataset.

2. Background

To our knowledge this paper presents the first use of deep learning to solve the problem of comparative plant grading. In this section we provide a background on Plant Grading methods, as well as a literature review of comparative deep learning methods and data augmentation for computer vision problems.

2.1. Plant Grading

Biological assessment is a fundamental step in bioassay screening, and the outputs drive progression decisions in a screening platform. Thus the quality of the assessment needs to be consistent both temporally and within a test.
To rank treatments, it is preferable to assign a single overarching score to a treatment which is made of many different observations. For plants, observations could typically incorporate the size, shape and colour.
Biological materials are well known to be highly variable from test to test due to differences in biotic and abiotic growing conditions such as seed batch and seasonality, respectively. To compensate for this variability, an untreated control treatment is used as a comparative baseline by human assessors, who must hold an image in their mind of the control whilst evaluating the treatments.
Lastly, human assessors are not truly quantitative, and when scoring on a 0-100 percent scale will often use a banded scoring system with increments of 10 percent, thus having 11 different scores including zero. Furthermore, these semi-quantitative scores are not ideal for a statistical analysis. A more objective, and truly quantitative approach would be beneficial, and is the motivation for the rest of the work in this paper.

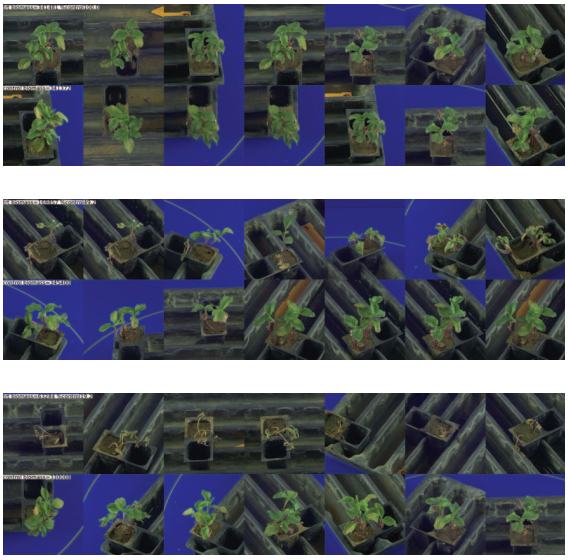

Figure 1. Figure showing 3 different montages created for human evaluation, each montage shows treatment (top row) and control (bottom row) plants from a particular test. Here we can observe the effects of different treatments on plant health. In our examples each subsequent montage shows treatment plants with poorer health.

2.2. Siamese Networks

Siamese neural networks were first proposed in 1993 by Bromley et al [1] as a method of solving the problem of signature verification. The Siamese framework introduced the concept of a pair of feature extractors for the comparison of two input images. Feature extractors with shared weights were trained to generate embeddings for input pairs. Embeddings generated by this pair of networks can then be compared by a distance function (or in later examples a feed forward network [6]) enforcing a latent space representation that groups similar or matching pairs of inputs similar to the intuition behind triplet losses [2]. This comparison of the embeddings can then be used to classify a match or mismatch between the two input images, or in more complex cases such as our own to regress a difference score between the two images.

In the past decade, advances in deep learning have allowed Siamese neural networks to solve a much wider range of problems, including human re-identification [12], COVID-19 diagnosis [10], and text based comparisons as seen in Neculoiu et al [7]. Many of these cases, such as the siamese network for face verification in Facebook's DeepFace [11], use an alignment stage, or assume a spacial

alignment of input images that makes it easier to the network to make a direct comparison between inputs. Due to the variety of angles our images are taken at however, we are unable to assume that images can be spatially aligned in this way, making our task more challenging.

In a similar use-case to our own, Li et al [5] uses Siamese networks to compare disease severity in medical imaging scans, comparing healthy scans against symptomatic scans and predicting scores representing progression and change. Here, we propose to use the twin-networks in a Siamese architecture to extract relevant features from the test and control image pairs, to help derive the comparative score.

2.3. Data Augmentation & Pseudo Labelling

The most common approach to increasing the size of training datasets in the field of Computer Vision is Data Augmentation [8], using image transforms to effectively extend the existing dataset to improve training variety. It is common in plant phenotyping problems to see data augmentation applied because of the diverse variety of plant species each of which can appear different based on varied environmental factors. Common forms of augmentation such as image cropping, rotations and flips, which, for example, are standard practice, such as those seen in Pound et al [9], to increase variety in a relatively homogeneous dataset. Kuznichov et al [3] uses a system of collages of different plant components to generate images of new plants composited from a smaller real set.

Another common method used for the expansion of limited datasets is pseudo-labelling, often considered a form of semi-supervised training. In Lee et al [4] this approach is introduced, using a limited dataset to train a network, and then labelling a larger unseen dataset using the network's predictions. For comparative problems such as our own, pseudo labels can be created by combining pairs of training samples to create a large number of inputs. Zheng et al [13] use a Siamese network for sequence analysis, noting that the large number of possible pairs *N(N-1)/2*, could be too large with high values of N, and focuses their work on selecting an unbiased selection of the possible combinations.

3. Materials and Methods

In this section we describe our entire pipeline for creating our training dataset as well as describing our Siamese neural network in detail. We also describe the experiments we used to test our approach, results of which are found in section 4. The section begins with an overview of the biological approach, followed by details of the deep learning proposed.

3.1. Bioassays

The chemical substances to be tested as herbicides on plants were dissolved in dimethyl sulfoxide for storage. Sub-samples to be tested were dried down and formulated into spray solution of acetone, water and Tween 20 for application. The compounds were tested for pre- and post-emergence activity against the weed species tested, with the compounds applied at 1000g/ha. The plants were then placed in the glasshouse for 12 days. The weed tested was *Amaranthus retroflexus*. Assessments were made of percent phytotoxicity and converted to a banded score between 0 and 100, where complete control of the target is 100 and 0 is no control.

3.2. Data Capture

Images were captured of the plants with three 5MP industrial cameras at different viewing angles of 0, 45 and 90 degrees. Plants were rotated on a turntable at 6 increments providing a total of 18 images per plant. The imaging box was custom made and was light sealed, and plants illuminated off axis with white LED lighting. Plants were graded by 2 human assessors from sets of digital images rather than observing the plants *in situ*. For pairs of plants, the human scorers would observe sets of images of the same control and treatment plant and give a score. Grading is a process by which the treatments are compared to the in-test control for herbicidal phytotoxicity as described in section 2.1. For every treatment-control pair in the two sets of images observed we create one training pair, allowing 49 samples to be created from every human graded pair.

3.3. Dataset Creation

While data collected and manually annotated by biologists is the current industry standard for highly accurate plant assessment; the wide range of factors that contribute to different plant phenotypes makes even a moderately sized dataset like the one we have gathered fairly limited. This becomes apparent when we consider the extremely wide range of possible combinations of treatment outcomes, and suggests that this challenge would be vulnerable to overfitting.

While treatment plants represent a range of outcomes caused by the different herbicides they have been treated with, we observe that control plants in our dataset are all relatively homogeneous, with virtually all examples being in good health. In practice, we might expect even control plants to have a wide range of appearances, caused by different seed batches, seasonal effects, and other environmental and growing factors. In these cases it is important that the network is able to predict a relative score of the difference in health between the two plants, as such we create additional pairs of images to which we assign pseudo labels we allows us to better replicate these

possibilities.

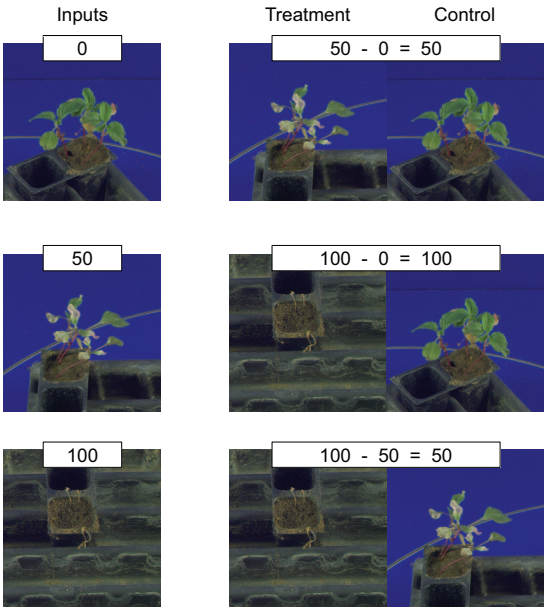

Figure 2. Figure showing how new pairs are given pseudo labels. Human annotated scores are combined to generate new pairs of images with a wider variety of control images.

To create our additional dataset we begin with an assumption, based on our observations, that as all control treatments are in extremely good health and show no phytotoxic symptoms, we can therefore consider the human labels to simply be an *absolute* score of the health of the treated plant, rather than a relative score with respect to the control. We then take our treatment images and group them by their scores, values between 0 and 100 in increments of 5, based on the average score of the two human annotators. We then are able to create new pairs of images by combining their individual scores, each image can then be paired with other images and the difference between their scores can then be used as a new pseudo label. As control plants should always be healthier than the treatment plant in our scenario, we ensure that for any new pair of images we only select plants with higher scores to represent the treatment plant compared to the score of the control plant. Because images can be combined in a wide variety of pairs, despite only having a few hundred training image we are now able to generate potentially hundreds of thousands of new training samples. As described below, we create a number of new dataset combinations for our experiments, including different quantities of image pairs and controlling the spread of values to reflect our limited test sample, and explore the effect of these in the results.

3.4. Siamese Network

We employ a Siamese neural network to predict relative scores between pairs of images.

Our network consists of a pair of ResNet-18 feature extractors with shared weights. To combine our feature embeddings we choose to use a concatenation layer, rather than a dot product. This choice reflects the limited spatial alignment between different samples, and should enable the network to learn more complex relationships between input pairs which is likely to be relevant due to the complexity of plant grading.

The second half of the network is a series of fully connected layers that predicts comparative scores for the pair of input images via regression. Due to the banded nature of the grading we also could consider a classification head for our final layer, but opt instead to frame the problem as regression owing to the continuous nature of the grades and the higher precision used in grading at the top end of the scale.

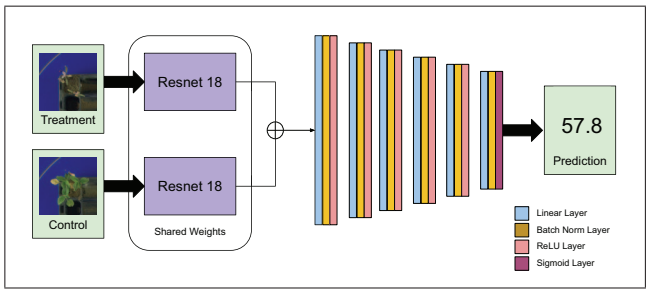

Figure 3. Figure showing our Siamese architecture. Pairs of images are input to ResNet-18 feature extractors with shared weights. These concatenated features are then passed through a feed-forward regression network to predict a score.

4. Results

In this section we present a series of experiments designed around testing the effectiveness of our model and the impact of training with additional pseudo labels. Results are then shown in tables 1 and 2.

4.1. Experiments

We train a number of models in order to evaluate our neural network approach to plant grading. Our evaluation is performed using an unseen test split of our original dataset, including only unseen images and comparisons, with ground truth values being the average score of both human evaluators. For our experiments we aim to compare performance between different versions of our network with the error of an individual human annotator versus the average. By doing so we aim to acknowledge that this particular problem is especially prone to variance in human

annotation, and evaluate our model in accordance with this premise.

First, we perform a series of experiments with different quantities of training images, evaluating the need to build a sufficiently large training set to achieve human level performance. We report our findings for these results in table 1.

(a) Scores of human annotators This baseline is established by comparing the scores of an individual expert human annotator against the average score of both annotators (which we use as our ground truth). This gives us a proxy-MSE score, which in this case can really be thought of as capturing variance between the annotators - similar scores will result in a low human annotator MSE error, divergent scores a larger MSE.

This experiment is meant to highlight inherent inconsistency in using human annotators, and to act as a representative accuracy level that a neural network would need to match or exceed for to be considered as a suitable replacement for human annotators. We evaluate our human annotators across the entire dataset to reduce variance caused by random test split sampling across such a small dataset. The annotators, of course, do not need a train and test split, as they had been asked to score all images using their expertise. So, we note that the annotators error is measured across all images, whilst the networks are evaluated over the dedicated test split.

(b) Neural Network Trained on Human Annotations For experiment (b) we train our Siamese network on the human annotations only. This dataset contains 4404 total pairs of images.

(c-f) Neural Network Trained on Pseudo Labels and human annotations For these experiments we trained our Siamese network on datasets of pseudo labels created using the method described in 3.3 combined with our human annotated dataset from experiment (b). Experiments are performed on datasets containing human annotations combined with additional pseudo label datasets of sizes 500, 1000, 5000 and 10000.

Secondly we perform further evaluation on the performance of our best performing model, considering performance over the range of possible ground truth score bands. This experiment aims to explore how the network performs differently where the ground truth score lies within different ranges of possible scores; as such, we arbitrarily create 5 bands of 20 ground truth values each to represent the entire spectrum of possible scores (0-100). We perform these experiments to highlight the uneven distribution of difficulty of scoring plants at various score ranges.

For all experiments we used both mean squared error (MSE) and mean absolute error (MAE) for our evaluation.

4.2. Training

For our experiments all models were run on NVIDIA A6000 GPUs with a batch size of 24. All networks were trained for 200 epochs with our best performance selected using a validation split of our main training dataset. Hyperparameters were selected empirically, with an Adam optimiser with a learning rate of $1e{-}4$ being selected. During training additional augmentation was added to our input images, including random flips, rotations and color jitter to improve dataset variety and generalization.

Experiment	Accuracy	
	MAE	MSE
(a) Human Annotators	**6.70**	**103.48**
(b) Siamese Real	15.74	547.12
(c) Siamese 500 Combined	31.37	1267.24
(d) Siamese 1000 Combined	16.35	593.99
(e) Siamese 5000 Combined	15.46	444.81
(f) Siamese 10000 Combined	**11.10**	**268.59**

Table 1. In this table we present results of experiments a-f, demonstrating the performance of our Siamese network trained on different training set combinations, and comparing against human annotators. The first row presents an indication of the inter-annotator error calculated using a psuedo-MSE score, comparing each annotator to the mean.

GT Range	Human Annotators		S. 10000 Combined	
	MAE	MSE	MAE	MSE
0-19	**4.29**	**57.14**	9.93	197.30
20-39	10.00	155.56	**9.18**	**120.80**
40-59	**11.33**	**180.00**	19.54	752.31
60-79	**6.96**	**130.43**	16.81	408.26
80-100	4.44	57.78	**4.08**	**52.48**

Table 2. In this section we compare the scores of our best performing Siamese network against the performance of human annotators against different subsets of the training data based on their ground truth score.

5. Discussion

In this section we analyse the results presented above, and discuss the performance of our Siamese network, as well as the importance of our dataset as it relates to human performance.

5.1. Analysis of Results

In table 1 we can see how the overall performance of the network initially decreases as we add a small about of our pseudo labelled data. We then see that the performance improves as we increase the quantity of our pseudo labelled data included in our training set, achieving the highest score of 11.10 mean average error and 268.59 mean squared error. This result reveals two significant insights about our approach to automated grading. First, we can see that our pseudo labelling approach improves on the scores of training solely on the real annotations which score only 15.74 MAE and 547.12 MSE by comparison. We also see that our network error score approaches human annotator level with more pseudo labelled data, even though it doesn't exceed it. We hypothesise that this gap would likely become even smaller if this approach was scaled up for use in an industrial setting as a larger dataset would likely lead to greater inter and intra assessor variation, while our Siamese network would perform consistently at scale. The high consistency of the healthy controls in this study would additionally likely help the annotators to limit their variation, compared to a larger study where controls contained more heterogeneity.

Then, in table 2 we can see more clearly how the error rate is lowest for values close to 0 and 100, and highest for values in the middle of this range. For both human and Siamese networks, the scores are best in the 0-19 and 80-100 ranges, and worst for the 40-59 range. Most importantly we see that our network struggles most with the ranges 40-79 while actually outperforming the human assessors in the ranges 20-39 and 80-100. As these worse performing ranges are those with the lowest quantity of training samples it is possible that with a less-homogeneous training set with a greater range of variation we could see our network narrow the gap with human performance further.

5.2. Future Work

In this section we highlight a selection of possible avenues of further research that were not within the scope of this paper.

Significance of Viewing Angles During this paper we have used a dataset containing images of plants taken from a range of angles. While human annotators had access to a range of angles during their inference, our model saw only a single pair of Treatment and Control image for any given inference. Further investigation is needed to determine an optimal approach to image capture for this problem; we hypothesise that a multi-angle approach where the model has access to multiple images of each plant will yield improved results.

Generalizing to other Plant Species *Amaranthus retroflexus*, the species of plant used in our experiments is characterised by its even branching of large leaves, making it relatively easy to capture a significant amount of detail from just a few photographs. Further research looking into other species with more complex structures should be conducted to assess our approach's ability to generalize onto a range of species.

6. Conclusions

In this paper we have demonstrated near-human level performance of comparative plant grading using a novel deep learning approach. We have overcome many of the limitations of human assessment, leveraging computer vision to provide non-subjective analyses, addressing a major bottleneck in the evaluation of new herbicides.

We have shown that automated grading of plants is possible even in the case of extremely limited data. Our results demonstrate that using our novel approach to pairwise pseudo-labelling we can significantly increase the size of our dataset, improving performance and allowing our model to generalize better over a wide variety of test cases.

Overall our approach has made a significant contribution to the field of plant phenotyping, highlighting a significant problem in the space that has previously found very little favour or attention from the research community. We believe that with further research the need for human assessors can be partnered with artificial intelligence, or in some scenarios perhaps replaced in favour of Deep Learning approaches which offer a non-subjective and truly quantitative biological assessment, freeing up valuable expert time for use elsewhere. However, in this case, approaches with more explainability would be welcome to increase confidence in the predictions.

References

[1] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

[2] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer, 2015.

[3] Dmitry Kuznichov, Alon Zvirin, Yaron Honen, and Ron Kimmel. Data augmentation for leaf segmentation and counting tasks in rosette plants. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[4] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural net-

works. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

[5] Matthew D Li, Ken Chang, Ben Bearce, Connie Y Chang, Ambrose J Huang, J Peter Campbell, James M Brown, Praveer Singh, Katharina V Hoebel, Deniz Erdoğmuş, et al. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *NPJ digital medicine*, 3(1):48, 2020.

[6] Xuning Liu, Yong Zhou, Jiaqi Zhao, Rui Yao, Bing Liu, and Yi Zheng. Siamese convolutional neural networks for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1200–1204, 2019.

[7] Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, 2016.

[8] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[9] Michael P Pound, Jonathan A Atkinson, Darren M Wells, Tony P Pridmore, and Andrew P French. Deep learning for multi-task plant phenotyping. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2055–2063, 2017.

[10] Mohammad Shorfuzzaman and M Shamim Hossain. Metacovid: A siamese neural network framework with contrastive loss for n-shot diagnosis of covid-19 patients. *Pattern recognition*, 113:107700, 2021.

[11] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[12] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 791–808. Springer, 2016.

[13] Wei Zheng, Le Yang, Robert J Genco, Jean Wactawski-Wende, Michael Buck, and Yijun Sun. Sense: Siamese neural network for sequence embedding and alignment-free comparison. *Bioinformatics*, 35(11):1820–1828, 2019.