ARIGAN: Synthetic Arabidopsis Plants using Generative Adversarial Network

Motivations and Contributions
- Few of labelled plant data for machine learning purposes
- Dataset augmentation spans a space
- We aim to learn data distribution to sample from it and generated realistic images
- Learning data distribution will allow to extend the limitations of dataset augmentation
- We present a method to artificially synthesise realistic 128x128 RGB images of Arabidopsis using DCGAN.
- We present Ax: The Synthetic Arabidopsis Dataset.

Proposed Methodology
Two models are trained simultaneously
- The Generator (G) outputs 128x128 RGB Images
- The Discriminator (D) is learnt to classify real vs. generated images.

\[ z \sim p_z \quad \text{Random vector drawn from a uniform distribution} \]
\[ y \sim p_y \quad \text{Random vector representing the leaf count.} \]
\[ x \sim p_{data} \quad \text{Random vector representing the training data drawn from an unknown distribution} \]

\( G \) learns to map vectors \( z \) into \( x' = G(z | y) \) s.t. \( x' \) looks like a sample from the unknown distribution \( p_{data} \), given the condition \( y \).

\[
\min_G \max_D V(G, D) = \mathbb{E}_{z \sim p_z} [\log D(G(z | y))]< + \mathbb{E}_{x \sim p_{data}} [\log (1 - D(G(z | y)))]
\]

In this objective function, \( G \) and \( D \) compete with each other in a zero-sum game. As the Generator learns how to produce more realistic images, the Discriminator tries to classify correctly real vs. generated images, until an equilibrium is reached.

Experimental Results

<table>
<thead>
<tr>
<th>Training Error</th>
<th>Testing Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \textbf{Trained on Ax only} )</td>
<td>( \textbf{Trained on Ax and Ax} )</td>
</tr>
<tr>
<td>DiC</td>
<td>0.013 (0.185)</td>
</tr>
<tr>
<td>DCE</td>
<td>0.094 (0.992)</td>
</tr>
<tr>
<td>MSE</td>
<td>0.031</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>1.865</td>
</tr>
<tr>
<td>( \textbf{Trained on Ax and Ax} )</td>
<td>( \textbf{Trained on Ax} )</td>
</tr>
<tr>
<td>DiC</td>
<td>0.029 (0.370)</td>
</tr>
<tr>
<td>DCE</td>
<td>0.156 (1.089)</td>
</tr>
<tr>
<td>MSE</td>
<td>0.996</td>
</tr>
</tbody>
</table>

Conclusions
- Our method, using DCGAN by Radford et al. (2015) is able to generate synthetic Arabidopsis plants.
- The generation of new plants is conditioned on the number of leaves, such that the user has control of the plant size (in terms of leaves).
- We generated a new dataset of synthetic RGB image plants, called Ax, where we gathered 57 candidates.
- Using Ax to augment the training dataset, we found that the leaf count algorithm of Giuffrida et al. (2015) improved testing results, reducing overfitting.

Acknowledgments
This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1, and also by the BBSRC grant BB/P023487/1.