The goal is to train neural networks to identify traits on images of herbarium sheets. We need training data of angiosperm records with both images and potential trait annotations. For this, we are using a data dump of the iDigBio database [1]. We filter this database for records that contain exactly one URI for an image and also contain notations of at least one trait of interest (flowering, fruiting, leaf-out). We download the images and use rule-based spaCy [2] parsers on text fields (reproductive condition, field notes, occurrence remarks, and dynamic properties) to determine the presence or absence of traits. Next, we use this data to bootstrap neural network classifiers (EfficientNet [3] & PyTorch [4]) to recognize traits. Few records contain notations of all of the traits, so we use pseudo-labeling [5] to retrain the models for higher accuracy.

[1] "Integrated Digitized Biocollections (iDigBio) Cloud-Based Cyberinfrastructure” José Fortes, Renato Figueiredo, Andréa Matsunaga, Alex Thompson, Matthew Collins, Jiangyan Xu Advanced Computing and Information Systems Laboratory Department of Electrical and Computer Engineering, University of Florida PO Box 116200, Gainesville, FL 32611-6200, USA."

[2] “spaCy: Industrial-strength Natural Language Processing in Python” Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd. doi: "10.5281/zenodo.1212303" 2020

[3] @misc{tan2020efficientnet,

title={EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks},

author={Mingxing Tan and Quoc V. Le},

year={2020},

eprint={1905.11946},

archivePrefix={arXiv},

primaryClass={cs.LG}

}

[4] title = {PyTorch: An Imperative Style, High-Performance Deep Learning Library},

author = {Paszke, Adam and Gross, Sam and Massa, Francisco and Lerer, Adam and Bradbury, James and Chanan, Gregory and Killeen, Trevor and Lin, Zeming and Gimelshein, Natalia and Antiga, Luca and Desmaison, Alban and Kopf, Andreas and Yang, Edward and DeVito, Zachary and Raison, Martin and Tejani, Alykhan and Chilamkurthy, Sasank and Steiner, Benoit and Fang, Lu and Bai, Junjie and Chintala, Soumith},

booktitle = {Advances in Neural Information Processing Systems 32},

editor = {H. Wallach and H. Larochelle and A. Beygelzimer and F. d\textquotesingle Alch\'{e}-Buc and E. Fox and R. Garnett},

pages = {8024--8035},

year = {2019},

publisher = {Curran Associates, Inc.},

url = {http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf}

[5] “ Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks” Dong-hyun Lee. CML 2013 Workshop: Challenges in Representation Learning (WREPL), Altlanta, Georgia, USA, 2013.