Ira A. Fulton Schools of Engineering

ARIZONA STATE UNIVERSITY

School of Computing and Informatics Decision Systems Engineering

Software Engineering Capstone Project Proposal

1. Contact Information

Project Sponsor Information

Company	**ASU**	**Division/Unit**	**School of Life Sciences**
Name	**Marek Borowiec**	**Email**	**mborowie@asu.edu**

Project (technical) Contact Information (if different from Proposer). If there are multiple people who will be interacting with the student team, then identify a lead point of contact.

Name	**Gabriele Valentini**	**Email**	**gvalent3@asu.edu**

2. Project Description

a. **Project title:**

“Deep learning for automated identification of insect specimens”

b. **Project description:**

This project is a data science project that aims to develop a system to automatically identify insect specimens. This system would primarily help researchers studying biodiversity, behavior, and ecology of ants but also provide means to support amateur entomologists in the field as well as to generalize to other insect species. The project consists in the development of a system composed by: 1) a deep learning image classification model, 2) a back-end software running on a server to automatically classify pictures of insect specimens using the model, and both 3) a web-based interface and 4) an app-based interface to upload pictures and visualize classification results.

The project requires skills in data science, network programming as well as web and app interface programming. The development of this project will provide students with the opportunity to learn how to obtain large datasets from the public domain, to design and train a production-ready deep learning model on a dedicated hardware, and to interface it using both web-based and app-based platforms.

The project is modeled after a recent application of deep learning in computer vision to automating identification of herbarium specimens (described in Carranza-Rojas et al. 2017, *BMC Evol Biol*, DOI: 10.1186/s12862-017-1014-z). The team will have to download three image datasets: ImageNet, AntWeb, and BugGuide. The first is a curated, generic dataset of approximately one million images classified into 1,000 generic classes and the other two will consist of web-scraped, publicly available images of identified specimens of ants and other insects at AntWeb.org and BugGuide.net. These datasets will then be used to train and test the neural network used for image recognition. The AntWeb data consists of highly standardized, well-curated images of natural history collection specimens taken under optimal conditions and with the aid of a stereomicroscope. In contrast, BugGuide is a much larger set of images of insects taken under a variety of conditions, mostly in nature. The major challenge of the project involves building a deep learning model capable of distinguishing among insect species and higher categories of taxonomic classification from images. Because the aim is to develop a system that will be optimized for classification of either clean standardized images or alternatively noisy images taken in the field, the project may require separate training and optimization of two neural networks. The design of the deep learning model(s) will require programming in TensorFlow, a popular Python interface that allows flexible design of neural network architectures. This part of the process will be iterative and requiring multiple rounds of model training and testing and likely involving augmentation of the original AntWeb and BugGuide image datasets for best performance. The curation of the datasets and model development will need to be well-documented. Model training will be performed on a dedicated high-performance computing unit provided by the sponsors. Once satisfactory model performance is achieved, defined as close to 80% accuracy using good-quality images to classify as one of the 150-200 common ant genera, team members will design and develop a website and both Android and iPhone apps providing a simple interface to the deep learning model running on a server. The interface should allow a user to upload insect images and receive suggestions of most likely classifications. All of the back-end and front-end software developed will be open source and documented following best practices in software engineering.

c. **Deliverables:**

- Web-scraped datasets consisting of images from the public databases AntWeb.org and BugGuide.net. The datasets will be curated and annotated with the help from project sponsors. The datasets need to be augmented using standard techniques in order to increase their size, improve accuracy of the resulting model, and reduce chances of overfitting.
- Deep learning image classification model(s) (Convolutional Neural Networks) trained and tested to recognize insect taxa on the labeled AntWeb and BugGuide datasets as well as appropriate scripts to incrementally re-train the model given new image data.
- Web interface that allows users to upload images to be classified by the model and to visualize the results of the classification process.
- Back-end software running on a web server that allows to query the deep learning model after uploading a new image using previously mentioned interfaces and to return the results of the classification process. The back-end software should come with a clear API explaining the querying protocol.
- Front-end interface Android and iPhone apps that allow users to upload images to be classified by the model and to visualize the results of the classification process.

All programmatic manipulations of the deliverables need to be well-documented to allow post-project maintenance by project sponsors.

d. **Motivation:**

The biological classification system and identification of animal and plant taxa are necessary for any research project involving living organisms. Unfortunately, the task of identification of many species or higher taxa (such as genera or families) requires considerable expertise developed over years of taxonomic research. This is especially true for hyper-diverse groups of organisms, such as insects. The shrinking number of active insect taxonomists and their limited time resources leave many biologists to try and identify their samples on their own, instead of consulting experts. This is usually done using so-called dichotomous keys, which are often cumbersome and challenging to use for non-experts. As a result, the correct identification of study organisms is often difficult. An alternative approach, only recently available due to large amount of digitized biodiversity data and developments in neural network model applications, is using deep learning for automated classification of images of animals and plants. Well-performing tools for automated identification of biological species from images will constitute a significant development in life sciences and allow more effective conservation efforts. This system will be useful for documenting and digitizing natural history specimens from museum collections around the world and will help foster citizen science by empowering amateur naturalists to identify insects in the field.

e. **Student learning experience:**

This project will give the team members the opportunity to develop some of the most sought-after skills in data science and software engineering. The students will gain experience in: 1) developing web crawlers to download and organize large datasets of images suitable for machine learning projects, 2) developing deep learning solutions for cutting-edge computer vision applications, 3) developing a client-server application whereby a server-based computing software provides a standardized interface to query the deep learning model by different on-line interfaces (i.e., website, Android and iPhone apps), 4) working with a high-performance computing unit and a network attached storage unit, and finally, 5) developing an industry-like project that is maintainable and is supported by appropriate documentation.

f. **Required background:**

Students are required to have good programming skills, to know how to design and organize software following the principles of Software Engineering (e.g., model-view-controller architecture) as well as to be familiar with Unix-like operating systems. A basic background in machine learning is also required from the team members. The students will be using the TensorFlow library for Machine Intelligence and should have a good background in Python programming. Experience with other data science tools facilitating project documentation, such as Sphinx or IPython notebook/Jupyter, and version control (e.g., GitHub, subversion), will be useful. Basic web and app development skills will be required to produce different model interfaces.

g. **Anticipated Expenses:**

No expenses anticipated.

Availability to meet with project sponsor (either remotely or face to face) on the Tempe campus if ASU for period meetings.

The sponsors can provide, upon provision of motivating evidence, access to:

1) High-performance computing machine suitable for training models using the TensorFlow library (see https://exxactcorp.com/index.php/solution/solu_detail/225),
2) Network Attached Storage suitable to store large image datasets (see Synology NAS DS1815+).

h. **Non Disclosures and IP:**

No NDA and/or IP documents required.

3. Other

Some background reading on deep learning in biology:

- Carranza-Rojas, J., Goeau, H., Bonnet, P., Mata-Montero, E., & Joly, A. (2017). Going deeper in the automated identification of Herbarium specimens. *BMC Evolutionary B*iology, 17(1), 181.

- Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878.