User Manual

Julienne LaChance
Daniel Cohen
Mechanical and Aerospace Engineering, Princeton University

Contents.

1. Introduction
2. Process Overview
3. Data Collection.
4. Workflow.
5. Complete List of Python Library Dependencies.
6. References.

1. Introduction

This manual is intended to guide cell biologists through the code we have provided at:

https://github.com/CohenLabPrinceton/hello-world/tree/master/Code_For_Distribution

Our goal is to clarify the steps involved with training and testing a U-Net neural network, with Python code for every step provided in its entirely. Users should be able to collect their own raw data and utilize this code with minimal modifications. Ideally, this will give new users the flexibility to quickly train and utilize a model with their own cell types, equipment, and labeled features.

1.1 Scope

This manual will describe each module of the code in detail. All of the required code is written in Python [1], due to its readability and widespread use within the machine learning community. We assume that many users will also have basic familiarity with FIJI/ImageJ [2,3], an open-source image processing toolbox, so we provide alternate versions of some pre-processing tools in the form of ImageJ macros (in addition to the Python scripts). For neural network training, we use the libraries TensorFlow/Keras [4,5], and we use Jupyter notebooks [6] to assist in demonstrating to new users how to use the model to process large images. We assume only basic familiarity with Python, as most modifications to the code will involve basic changes, such as changing paths to the user's local folders.

All TensorFlow/Keras code in this distribution has been tested on the NVIDIA GeForce GTX 1070 Ti GPU and the NVIDIA Tesla P100 GPU.

2. Process Overview

The code workflow is provided in a series of folders which indicate the order of the processes which comprise the workflow. Users may choose to use the code in these folders (named "0_Helper_Functions", "1_Splitting_Images_Into_Patches", etc.), or may find those same scripts grouped together in the folder entitled "All_Files". Regardless of whether the user wishes to utilize individual scripts or utilize all the code together from "All_Files", the user needs to ensure that the "user input variables" in each script are modified correctly, and that paths are set correctly to the helper functions. We describe these steps in more detail below.

In order to use the Python scripts, users will need to install Python 3, plus the versions of TensorFlow/Keras which are compatible with their GPU drivers. The complete list of Python libraries and the versions we used are provided at the end of this document. Additional libraries to be installed are indicated in each Python script and their corresponding helper scripts. Additional requirements will be noted in each subsection of the Workflow below.

3. Data Collection

In order to run this code, the user must collect a dataset of matched image pairs. We will not describe the data collection process in detail here, but we assume users have transmitted-light input images paired with high-quality images of labeled fluorescent features, such as cell nuclei or junctions. We recommend using image registration so that fluorescent features align spatially with transmitted-light features, and for ease of processing downstream in the workflow, we recommend cropping matched pairs into height- and width- values which are multiples of 256 (the height and width of one patch, and therefore the input/output size of images to/from the model). While this cropping step is not necessary, the failure to crop matched pair images will require the user to modify the scripts provided in Step 1. Here we assume the image pairs are stored as image sequences of 16-bit TIFF files in two separate folders, and that all images are of the same size.

Sample matched image pairs in their respective folders are provided in the folder "1_Splitting_Images_Into_Patches" within "Sample_Images". In this example, we wish the "Phase_Image" image to become inputs to the network, and for the "DAPI_Image" nuclei images to become the ground truth for the model output.

4. Workflow

Introduction.
The workflow consists of the following steps:
Step 1: Split Images Into Patches.
Step 2: Perform Test/Train Splits.
Step 3: Get Data Statistics.
Step 4: Train the Network.
Step 5: Test the Model.

Step 6: Get Additional Prediction Statistics.
Step 7: Process New Images.

We now will discuss each step of the workflow in more detail. Initially, the raw images must be split into many patches of size 256x256 pixels^2. Then, the dataset must be split into a training set and a test set, and dataset statistics are collected from the training set for normalization of the data. Next, the network is trained, and evaluated on the test set, with relevant statistics such as test accuracy reported to a CSV file. Finally, we demonstrate how to use the network to process a new large input image.

Some of the scripts in this workflow rely on helper scripts: namely, "**data_loader.py**", "**models.py**", and "**utils.py**". These can be found within the folder "0_Helper_Functions" (or "All_Files"). These scripts serve the following functions:

"**data_loader.py**": This Python file contains helper functions which load in the user's dataset. This is necessary, for example, to read in the input/output images for training or testing the model.
"**models.py**": This contains the U-Net architecture implementation. If the user wishes to test different models, this Python file may be modified.
"**utils.py**": This contains helper functions, including the implementation of the Pearson correlation coefficient loss function, and also stores information about the user's dataset, including normalization statistics and paths to folders containing the user's dataset. This script should be modified as dataset statistics are collected and experimental conditions and datasets are added.

The Python scripts in the following sections of the workflow expect these helper files to be in the same folder, or for paths to the files to be set appropriately in the import statement. The user should be especially wary of this if attempts are made to utilize individual scripts, rather than modifying code from within the "All_Files" folder.

4.1 Step 1: Split Images Into Patches.

The main script to perform this step is entitled "**Splitting_Macro.ijm**", and can be found in either folder: "1_Splitting_Images_Into_Patches" or "All_Files". This portion of the code is provided as a FIJI/ImageJ script. Users can test the functionality of this script on the test images provided in the subfolder, "Sample_Images". (We assume most biologists are more comfortable with this ImageJ macro, but for the Python version, check out the script "Python_Splitting_Script.py).

The sole purpose of this macro is to split a sequence of images within a specified input directory into a number of 256x256 pixel^2 patches, and save these patches into a specified output directory.

To use this macro, four variables must be set by the user, which are detailed in the macro comments. These variables are paths to the input and output directories, an additional string to set the name of the output patch files, and an integer *n*, which indicates how many patches to produce from each input image. Although *n* can be automatically determined, we chose to manually set this value to ensure that the input image size would be verified by the user.

The user should apply this script to each of the image sequences comprising their dataset: for example, on both the "phase" and "DAPI" image sequences. For standard applications, this results in the transmitted light image patches saved in one folder, with the fluorescent label image patches saved in a separate folder.

This script has no dependencies.

4.2 Step 2: Perform Test/Train Splits.

The main script to perform this step is entitled "**run_test_train_split.py**", and can be found in either folder: "2_Perform_Test_Train_Splits"or "All_Files". This portion of the code is provided as a Python 3 (.py) file, as will be most of the subsequent scripts. Sample images from the previous section may be used to test code from any of these sections; however, the model will not be very accurate when trained with so few images!

Once the complete dataset has been pre-processed, it must be split into a training set and a test set. The training set will consist of the data that is used to train the network, while the test set will consist of images which are "held out" (not seen by the network during training), and are only used to validate the accuracy of the model later on. Matched pairs of images must be randomly selected to be sorted into these two categories.

To use this code, four inputs must be specified by the user. These are the path to the folder containing the entire dataset ("base_path"), the names of the sub-folders containing the input and output images, respectively, and a float named "split", which determines the percentage of the dataset to use in the training set. The script creates new sub-folders within the base directory to store the training and test subsets, respectively. Input/output images are stored in separate folders within these training and test subset folders.

This script has the following library dependencies: os, random, math, shutil.

4.3 Step 3: Get Data Statistics.

The main script to perform this step is entitled "**run_get_mean_stdev.py**", and can be found in either folder: "3_Get_Data_Statistics" or "All_Files". Users must also modify "**utils.py**", which can be found in either "0_Helper_Functions" or "All_Files".

Once the data is collected, pre-processed, and split into training and test sets, the last step before training is to collect the mean and standard deviation of the input and output data training sets. These values are used to normalize the data which is input into the neural network. In many computer vision applications, normalization/scaling is essential to ensure that the neural network converges in fewer epochs. So, the user will collect the relevant data statistics prior to training the network.

To use this code, the user will need to modify "run_get_mean_stdev.py" and "utils.py" to configure a new experiment and set the path to the training data. One example is provided in "run_get_mean_stdev.py" involving an experiment with keratinocyte-type cells, at 10 magnification, with imaging in the DAPI channel. The function "get_normalization_factors()" in "utils.py" takes in these parameters and determines the path to the image data as set by the user. Then, data is loaded and the relevant statistics are printed out. Users should think carefully how to organize their data, especially if planning future experiments.

After running this script, the users should add the data statistics to the file “utils.py”. This file contains all the information about each dataset (from independent data collections), so that statistics and paths can be referenced from one location. The files “data_loader.py” and “models.py” do not require modification.

This script has the following library dependencies: time, numpy. It also utilizes our custom Python files: data_loader.py and utils.py.

The file “utils.py” has the additional dependency: keras.
The file “data_loader.py” has the additional dependencies: os, sys, random, warnings, tqdm, cv2.

4.4 Step 4: Train the Network.

The main script to perform this step is entitled “**run_train.py**”, and can be found in either folder: “4_Train_the_Network”or “All_Files”.

Having gathered the data statistics, the user may proceed to train the network. This portion of the code compiles the model the user wishes to train, loads in and normalizes the data, and proceeds to train the network. This step can take a long time (several hours on an NVIDIA P100 GPU), so it’s helpful to use the most powerful GPU available.

This script requires the user to specify the experimental conditions (as in the previous step, for referencing the correct dataset and normalization factors), the name of a weights file to be saved, the loss function to be used (for example, mean-squared error), and the model type to use (for example, the standard 1-stack U-Net). Please refer to the comments in the script for additional details on setting these parameters.

According to these inputs, the U-Net neural network will be trained using the ADADelta optimization scheme and with early stopping (such that if the validation loss does not decrease in 100 epochs, the training process will end). The script additionally logs the training and validation loss to a CSV file, and saved the finalized model as a YAML file.

Upon completion of this step, users should add the new weight (.h5) file to “utils.py”, so that the weight file can be referenced for each experiment in future steps.
This script has the following library dependencies: time, numpy, and keras. It also utilizes our custom Python files: data_loader.py, models.py, and utils.py.
The file “data_loader.py” has the additional dependencies: os, sys, random, warnings, tqdm, cv2.

4.5 Step 5: Test the Model.

The main script to perform this step is entitled “**run_test.py**”, and can be found in either folder: “5_Test_the_Model”or “All_Files”.

Once the model is trained, the images in the test set may be used to quantify its accuracy. This script loads in the model and weight (.h5) file from the training step, and processes the test set, ultimately printing out the test statistics and saving prediction images into a new folder.

This script requires the user to specify the experimental conditions (as in the previous steps, to normalize the test images appropriately), the path to the output folder into which prediction images will be saved, the loss function to be used (for example, mean-squared error), and the model type to use (for example, the standard 1-stack U-Net). Please refer to the comments in the script for additional details on setting these parameters.

According to these input parameters, the test data will be loaded in, normalized, and passed through the trained model to produce output predictions for each input image. These predictions will be compared to the ground truth images to determine the final accuracy of the model.

This script has the following library dependencies: os, time, numpy, and keras. It also utilizes our custom Python files: data_loader.py, models.py, and utils.py.
The file “data_loader.py” has the additional dependencies: sys, random, warnings, tqdm, cv2.

4.6 Step 6 (optional): Get Additional Prediction Statistics.

The main script to perform this step is entitled “**run_get_signal_stats.py**”, and can be found in either folder: “6_Get_Prediction_Statistics”or “All_Files”.

In previous steps, the user trained and tested the model. In addition to the test loss/accuracy, the user may utilize “run_get_signal_stats.py” to collect additional statistics on the test set, and print them out. Here, we demonstrate how to determine the Pearson correlation coefficient and mean-squared error statistics on the entire test set, and also on a subset of the test set. This subset is determined by specifying an intensity cutoff value, stored in “utils.py” for the corresponding dataset. By utilizing this value, called “signal_thres”, only image pairs are considered if the ground truth image contains pixel intensity values above the threshold, thereby eliminating images which contain only background. Users will manually determine the appropriate intensity cutoff value for each dataset and modify “utils.py” prior to running this script.

This script requires the user to specify the experimental conditions (as in the previous steps, to determine the correct path to the ground truth data), the path to the folder containing the predicted images, the loss function used (for example, mean-squared error), and the model type used (for example, the standard 1-stack U-Net). Additionally, a path called “save_path” is defined to determine where to write out CSV files, which contain the final results. Please refer to the comments in the script for additional details on setting these parameters.

This script has the following library dependencies: numpy, scipy, sklearn. It also utilizes our custom Python files: data_loader.py, models.py, and utils.py.
The file “data_loader.py” has the additional dependencies: os, sys, random, warnings, tqdm, cv2.
The file “models.py” has the additional dependency: keras.

4.7 Step 7 (optional): Process New Images.

The main code to perform this step is entitled “**Process_Directory.py**”, and can be found in either folder: “7_Process_New_Image”or “All_Files”. (The equivalent Jupyter notebook version is also provided here, and is called “**Process_Directory.ipynb**”.)

Finally, the trained model may be used to process new, larger input images. This Jupyter notebook demonstrates how to take in an image sequence (again, assuming width/height are multiples of 256 pixels), and produce a sequence of predicted images, patched together from network outputs. In this demo, we leave boundaries of the predicted images unprocessed to emphasize the fact that edge predictions are assumed to be less accurate. However, users may modify

This notebook requires the user to specify the experimental conditions (as in the previous steps), paths to the input images to be processed and the output directory where predictions are saved, a path to the location of the weights (.h5) file, plus information about the neural network (loss function and model type).

The notebook demonstrates how to read in and normalize the input images, apply the trained model to the data in a sliding-window fashion, and compose and save the predicted images. Central regions of each output patch are averaged in this demo; users may consider modifying to code to utilize the data median instead, as is done in [7].

This notebook has the following library dependencies: os, numpy, Pillow, libtiff, sklearn, keras. It also utilizes our custom Python files: models.py and utils.py.

5. Complete List of Python Library Dependencies.

In sum, all the required Python libraries are listed here:

cv2, jupyter, keras (tensorflow), libtiff, os, math, numpy, Pillow, random, scipy, shutil, sklearn, sys, time, tqdm, warnings

Version 10.1 of CUDA (cuDNN) was used. The specific versions of each library that we used are provided below:

h5py	2.10.0
imutils	0.5.3
ipykernel	5.3.4
ipython	7.16.1
ipython-genutils	0.2.0
ipywidgets	7.5.1
jupyter	1.0.0
jupyter-client	6.1.7
jupyter-console	6.2.0
jupyter-core	4.6.3
jupyterlab	2.2.2
jupyterlab-server	1.2.0
Keras	2.4.3
Keras-Applications	1.0.6
Keras-Preprocessing	1.1.2
libtiff	0.4.2
matplotlib	3.0.2
numpy	1.18.5

opencv-python	4.0.0.21
pandas	0.23.4
Pillow	5.4.1
pip	20.2.1
python-dateutil	2.7.5
PyYAML	3.13
scikit-learn	0.20.2
scipy	1.4.1
tensorboard	2.3.0
tensorflow	2.3.0
tensorflow-estimator	2.3.0
tensorflow-gpu	2.3.0
tqdm	4.29.1

Users may reference the Dockerfile provided in the “Dockerfile” folder on GitHub to rapidly install all Python requirements, or to produce their own image for processing.

6. References.

[1] Python Software Foundation. Python Language Reference, version 3.5. Available at http://www.python.org

[2] Rueden, C. T.; Schindelin, J. & Hiner, M. C. et al. (2017), "ImageJ2: ImageJ for the next generation of scientific image data", BMC Bioinformatics **18:529**, PMID 29187165, doi:10.1186/s12859-017-1934-z (on Google Scholar).

[3] Schindelin, J.; Arganda-Carreras, I. & Frise, E. et al. (2012), "Fiji: an open-source platform for biological-image analysis", Nature methods **9(7)**: 676-682, PMID 22743772, doi:10.1038/nmeth.2019 (on Google Scholar).

[4] Abadi, Martín, et al. "Tensorflow: A system for large-scale machine learning." *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016.

[5] Chollet, François. "Keras." (2015).

[6] Kluyver, Thomas, et al. "Jupyter notebooks." *a publishing format for reproducible computational workflows* 850 (2016): 87-90.

[7] Christiansen, Eric M., et al. "In silico labeling: predicting fluorescent labels in unlabeled images." *Cell* 173.3 (2018): 792-803.