May 8th, 12:00 AM

# Arsenic Classification: Deep Learning Finding Toxin Exposure

Joshua Turner
*University of Southern Maine*, joshua.turner@maine.edu

# Arsenic Classification: Deep Learning Finding Toxin Exposure

UNIVERSITY OF SOUTHERN MAINE
PORTLAND • GORHAM • LEWISTON • ONLINE

Joshua Turner, Mentors: Doug Currie & Bruce MacLeod, University of Southern Maine

## Abstract

With promising results in the use of deep learning to classify arsenic exposure at the cellular level, we sought to derive best practices in model training to optimize results. Using fastai as our software framework, challenges included data loss due to default cropping, discovering best pre-processing practices, and finding optimal training parameters. Through the automation of model training the focus could be turned to numeric analysis of results.

## Background

Studies have shown that long term exposure to non-lethal levels of arsenic can have destructive effects on cognitive growth. In rural areas where well water is often contaminated the need for classification of these effects is critical. This task, which is currently done by hand, can easily be automated. This would allow for the exploration of other critical questions such as at which levels of exposure, and over what length of time, are the effects irreversible.

## Methods

- Automated model training to gather larger amounts of data pertaining to accuracy involving different training methodologies.
- Worked with various crop sizes in attempts to minimize data loss while working with pretrained models.
- Used two separate ensemble techniques to derive predications on entire image from training on sub-images
- Used Shannon entropy filtering to combat the creation of featureless sub-images
- Used transfer learning with pretrained models, as well as trained models from scratch.
- Worked with two separate architectures, including ResNet and SqueezeNet.

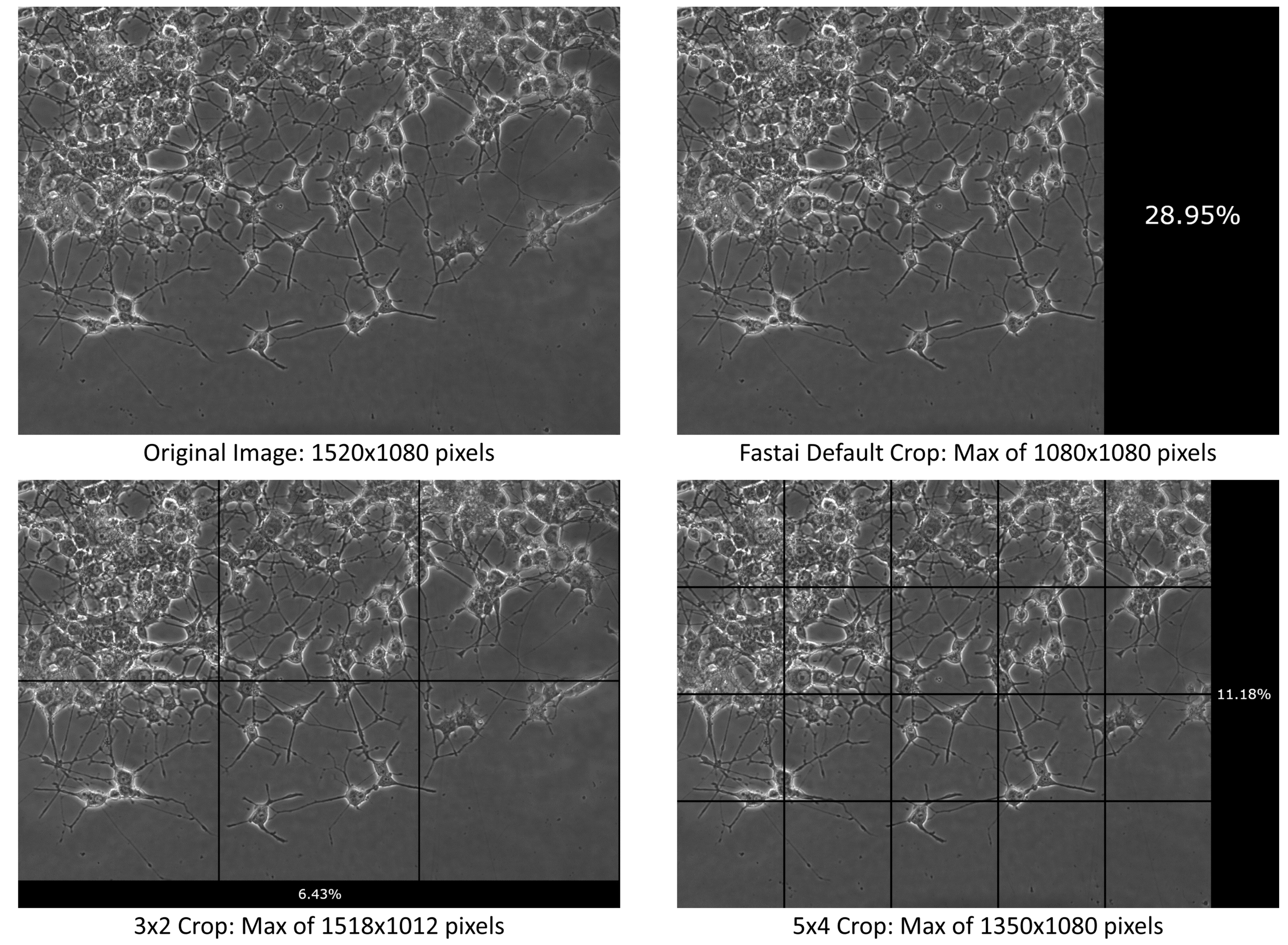


Figure 1. Crop levels as data retention

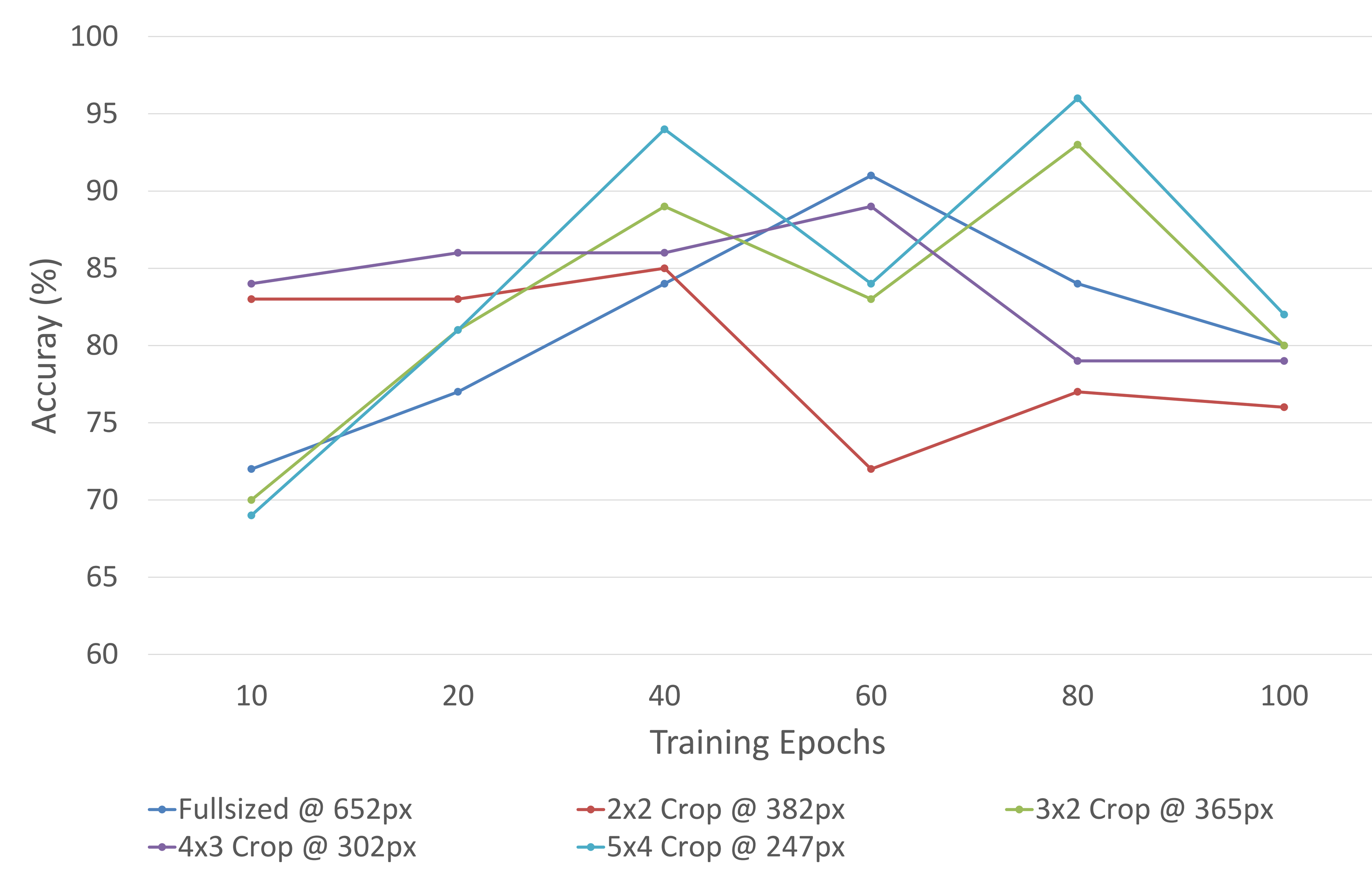


Figure 2. Probability Summation Results

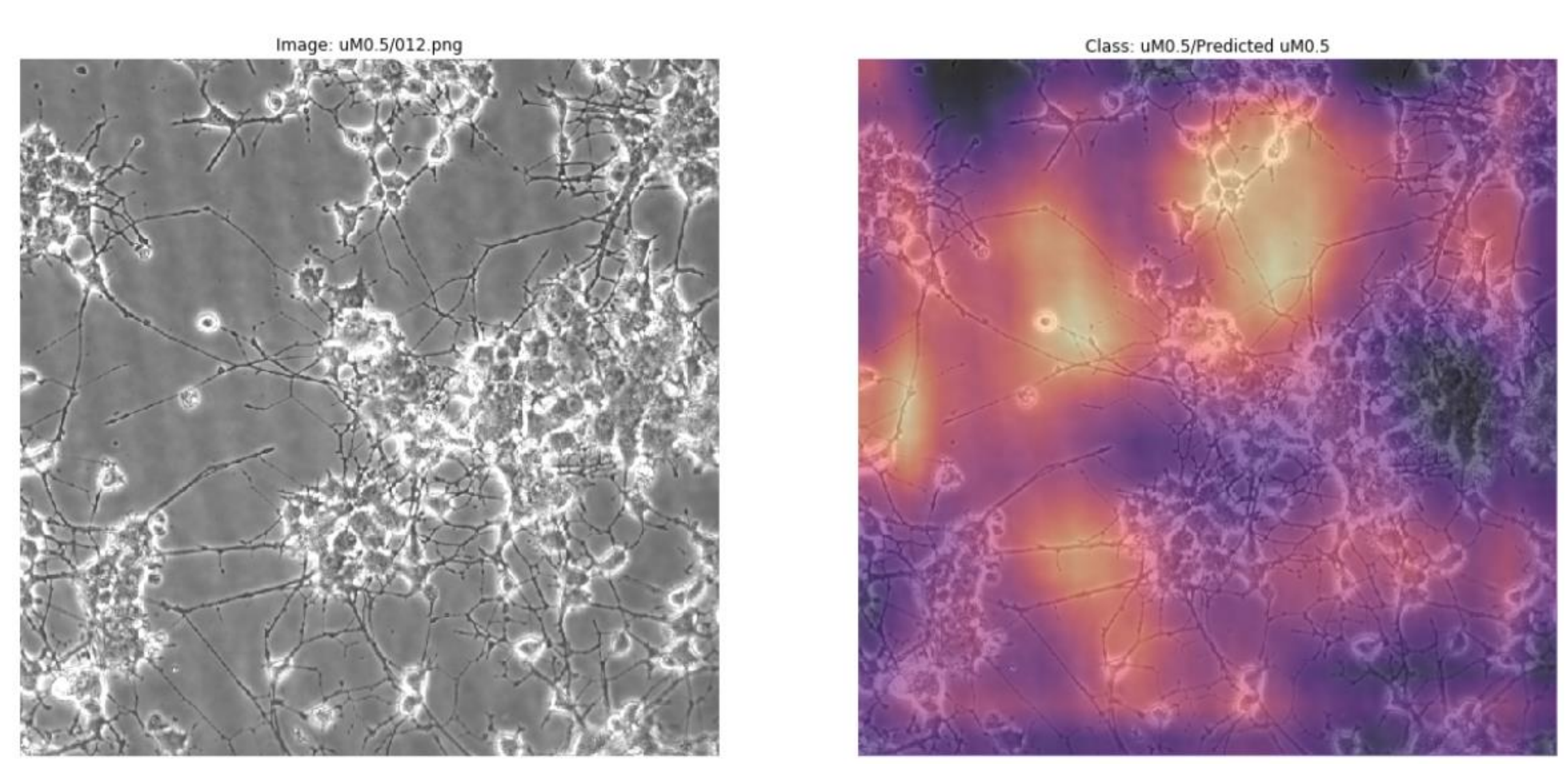

Figure 3. Heat Map of Activations

## Results

Using various crop sizes to decrease data loss due to default image augmentations performed by fastai (as shown in Figure 1), accuracy was able to be increased past the standard achieved with the original images. Due to the data lost when any image is compressed down to the ImageNet standard of 224x224 pixels, full resolution images have been found to perform worse than lower resolutions. Figure 2 displays the results of a middle ground resolution, which performed best in most tests, calculated as: $\frac{Full\ Resolution - 224}{2} + 224$ pixels

## Discussion

- Crop sizes have proven to be an effective means of both decreasing data loss found in standard fastai image augmentation as well as increasing the size of the overall dataset for training.
- Both prediction ensemble techniques used to formulate predictions based on sub-images often outperformed the models trained directly on the full-sized images.
- Shannon entropy filtering proved to be of little help in the case of the given dataset, as many sub-images didn't contain few enough features to be filtered out.
- As research demonstrates, utilizing transfer learning through pretrained models lead to faster and more accurate results than those received with untrained models.
- As expected SqueezeNet couldn't compete with ResNet due to the smaller size of the model.
- Moving forward, the question of the use of random cropping during training has been raised and will be pursued.
- The heat map in Figure 3 demonstrates the model's capability of recognizing the dendrites and cellular structure via its activations in the control image, and it's use of them to make an accurate classification prediction.

Acknowledgements
Thanks to Doug Currie for providing the problem of arsenic classification as well as the datasets used during this research. Thanks to Bruce MacLeod for offering me the opportunity to work on such an exciting project. Thanks to my peers Nick Littlefield and Chris Faulhaber for offering advice in the topic of deep learning.