Accelerating Two-Photon Calcium Imaging Segmentation with Convolutional Neural Networks

Frank DeGuire III
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
fdeguire@stanford.edu

Anthony Riley
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
anthonyriley@stanford.edu

Emanuel Herberthson
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
herberthson@stanford.edu

Abstract

Recent advances in two-photon microscopy technologies allow simultaneous recording of over 200,000 neurons. While automatic segmentation algorithms designed to find the cells in these recordings exist, they tend to greatly overestimate the number of cells. If the new two-photon imaging technologies are to be fully utilized, researchers will no longer be able to rely on manual sorting of true cell masks versus non-cell masks in the outputs of these segmentation algorithms. This paper presents a novel automated classifier designed to distinguish between 'true cell masks' and 'non-cell masks' derived from the two largest automatic segmentation algorithms for two-photon microscopy data. Using datasets from the Allen Brain Observatory, which have been manually annotated by neuroscience experts, this study benchmarks the accuracy of existing segmentation algorithms against human annotations and develops a specialized convolutional neural network (CNN) to enhance the precision of neuronal segmentation. Initial results show promising improvements in the accuracy of cell identification, with a significant reduction in the number of non-cells being classified as true cells. This advancement not only improves the efficiency of two-photon imaging studies, but also sets a precedent for future developments in automated neuron segmentation.

1. Introduction

Mammalian brains, composed of billions of neurons, are one of the most complex organic structures in the world. Recent advances in two-photon microscopy now enable more in-depth studies of the brain by allowing the simultaneous recording of activity from over 200,000 neurons [4, 1]. However, to fully realize the potential of these advancements, effective methods for interpreting these vast datasets must be developed.

One key task in understanding these datasets involves segmenting the locations of cells in the recordings. Current segmentation algorithms, designed for lower throughput datasets, often overestimate cell counts, necessitating manual post-processing to correct their errors [5, 7]. The massive size of datasets from new two-photon microscopy methods make this manual segmentation step infeasible. Consequently, an automated method for sorting 'true cell masks' and 'non-cell masks' is necessary for full utilization of these exciting advancements in two-photon microscopy.

In this project, we developed a deep learning framework that enhances existing brain imaging segmentation methods by distinguishing true cells from non-cells. We developed separate convolutional neural networks that are specialized to operate on the two most popular existing cell segmentation algorithms, Suiet2p and Caiman. By leveraging existing human-annotated data to train and validate our model [2], both of our models are able to sort true cell masks with almost 90% precision. This development helps to unlock the potential of recent advances in two-photon microscopy by accelerating scientists' workflow, providing them with more time to achieve revolutionary discoveries.

2. Related Works

The two main existing two-photon imaging segmentation algorithms are Suite2p [7] and Caiman [4]. Both algorithms rely on signal processing methods to analyze large (> 10 TB) recordings in an efficient manner. However, these algorithms still suffer from segmenting too many items,

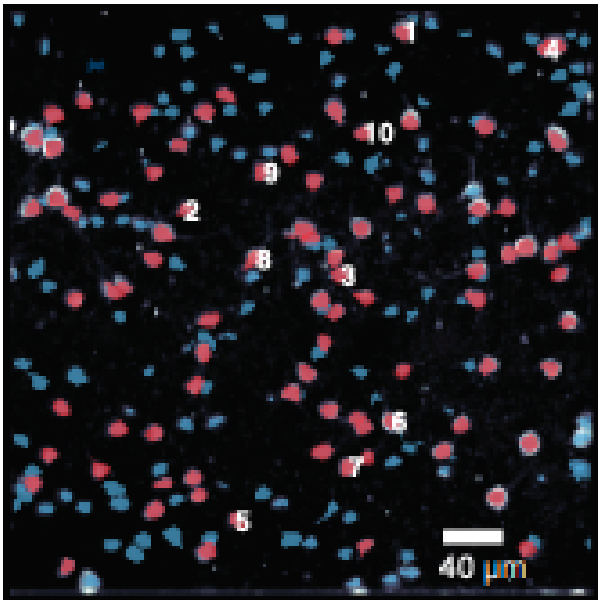

Figure 1. Example of an output from Suite2p automatic cell segmentation algorithm (blue masks) compared to a human-annotated database (red masks). Existing automatic cell segmentation algorithms tend to greatly overestimate the number of cells in recordings.

achieving much lower precision scores than recall scores [3, 7].

Many other algorithms have attempted to address the precision shortcomings of Caiman and Suite2p, but they often suffer from comparatively lower recall [9], have been tested mainly on synthetic data [8], or suffer from prohibitively long run-times [5]. For this reason, scientists continue to use Caiman and Suite2p and rely on manual post-processing to sort the true cell masks from the non-cell masks.

In the past, this manual curation step has been feasible for experiments because two-photon microscopy experiments had relatively low throughput. As a result, no automatic classifier for true cell masks versus non-cell masks currently exists [1]. However, due to recent advances in two-photon microscopy techniques, manual curation is completely infeasible. We propose a novel automatic cell segmentation classifier based on convolutional neural networks.

3. Data

The Allen Brain Observatory is a database that catalogs the functional responses of neurons in the mouse visual cortex to various visual stimuli based on 2-photon fluorescence imaging [2].

[1]The authors note that some individual labs may possess their own methods for automatic mask sorting, but no widespread classifier for this task currently exists.

The dataset includes eight different *in vivo* two-photon recordings of the mouse visual cortex. Each dataset is manually annotated by 3–4 independent labelers that were instructed to select active neurons in the recording [4]. For this paper, these manual expert annotations will be treated as the ground truth against which we will benchmark our classifier.

We processed these recordings through two popular existing segmentation algorithms, Suite2p and Caiman. Each of these signal-processing based algorithms produced a set of segmentation masks for each recording. All segmentation masks have the same dimensions as the recordings from which they were generated (ranging from 200x200 to 800x800). For every pixel and mask, the segmentation algorithms assign a value between 0 and 1 that signifies how much each pixel contributes to each mask. For example, if two cells overlap on one pixel, the algorithm may assign a value of 0.5 to the masks for each cell. Similarly, if a pixel lies near the edge of one cell but the center of another, the algorithm may assign a value of 0.1 to the cell edge but 0.9 to the cell center. If the segmentation algorithm does not find the pixel to be a part of any masks, the pixel will have zero weight in all masks.

In order to label each mask as a 'true cell' or 'non-cell', we compared the outputs of the algorithms to the human-annotated masks. We first binarized the algorithm outputs by setting each mask value to 1 if the algorithm output was greater than 0.2, and 0 otherwise. This threshold was selected to maximize matches between nearby cells without accounting for low-weighted pixels. If the binarized algorithm mask overlapped with a human-annotated mask by at least 70% by the Jaccard index (intersection over union), it was considered a 'true cell'. Otherwise, it was considered a 'non-cell'.

4. Methods

4.1. Data pre-processing

To perform the mask classification task, we developed two convolutional neural networks trained on the human-labeled datasets from [2] and the outputs from the Suite2p and Caiman algorithms. We decided to train separate classifiers for each of the segmentation algorithms because it was thought that each algorithm would have nuanced patterns in its outputs that a dedicated model would be able to learn. In order to be able to input the data into our models, we had to first pre-process the data.

We developed a data pre-processing pipeline to standardize the inputs to our deep learning models. First, the area around each mask was cropped to a 64x64 square, with the mask centered in the frame. All other masks in the field of view were removed during this cropping. The authors decided to remove other masks from the field of view to sim-

plify image processing and eliminate the variable of spatial locality. Our objective is for the model to recognize cells independent of their spatial location within the image, accommodating the variable positioning of images captured throughout the brain. An example of a few cell masks after cropping and centering is included in Figure 2.

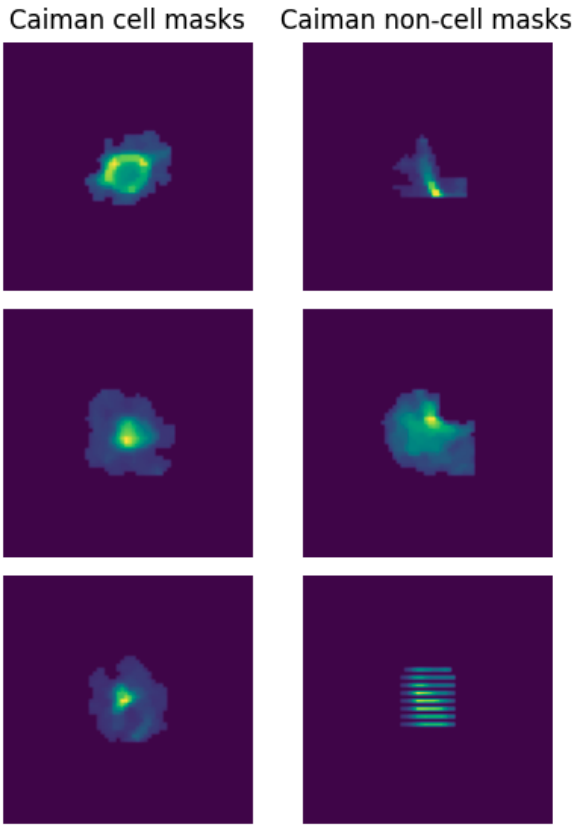

Figure 2. Example of Caiman masks after cropping and centering the area around the mask to a 64x64 frame.

Because the non-cell masks greatly outnumbered the true cell masks in some recordings (in one recording, there was almost a 10:1 ratio of non-cell masks to cell masks), a set of augmented data was introduced to the dataset. The augmented data was produced by selecting a random mask, applying a 0°, 90°, 180°, or 270°degree rotation with equal probability. Then, a horizontal and/or vertical flip was applied with 50% probability each. These data augmentations were repeated until there were 10,000 true cell and non-true cell masks in the Suite2p dataset and 5,000 true cell and non-true cell masks in the Caiman dataset (the Caiman dataset started with a lower number of masks, so fewer augmented data points were needed to equilibrate the true and non-true mask counts).

It should be noted that these augmented images are biologically relevant because true cells can occur in any orientation of the brain, and rotations and flips can easily occur if a research orients a camera differently on any given day. For this reason, the model should be trained on and be able to accurately classify this augmented data.

Finally, the dataset was split into training, validation, and testing sets, with 80% of data reserved for training, 15% reserved for testing, and 5% reserved for validation. This split was performed before the augmented data was introduced to the dataset to ensure that the test set did not include any augmented images.

4.2. Model architecture and hyperparameters

After data pre-processing, we were able to input the data into our CNNs. We tested a few different model architectures on each algorithm. We started with a simple two-layer CNN with 3x3 filters and max pooling layer. To test deeper networks, we also tested modified versions of Resnet18 and Resnet50. We modified the ResNet models from their original form in three small ways. First, because our input images have only one channel, we modified the first convolutional layer to accept single channel images. Second, we again modified the fist convolutional layer to use a smaller filter size (5x5 instead of 7x7) and removed the max pooling layer because our input images were significantly smaller than the input images to the original ResNet (64x64 versus 224x224). Finally, a sigmoid layer was added to the end of the model to convert the model output to a value between 0 and 1 that can be roughly interpreted as the likelihood of a mask being a true cell.

Training was conducted using the Adam optimizer and binary cross entropy loss function. Adam was chosen due to its success on a variety of deep learning applications, and the binary cross entropy loss function was chosen because we have a binary classification task.

Hyperparameters and final model architecture were selected by a random search over regularization strengths, learning rates, batch size, and learning rate decay. First, a crude random search was performed over a wide range of the hyperparameters. These hyperparameters were used to train each of the proposed model archiectures for 5 epochs. After this first round of random search, the best-performing hyperparameters (based on validation set F1 score) were selected, and a finer random search was performed near their values. The proposed model architectures were then trained for 25 epochs on these hyperparameters. The set of hyperparameters and model architecture with the highest validation set F1 score on this final test was used to train the final model. This procedure was repeated for the Suite2p and Caiman models so that separate hyperparameters were developed for each model.

For our Caiman model, we found that the deeper ResNet models tended to severely overfit the data, even with high regularization strengths. The ResNet 50 architecture very quickly achieved 100% prediction accuracy on the training

set, while still predicting the validation set with less than 70% accuracy. As can be seen in Figure 3, the two-layer CNN had a higher area under the curve (AUC) on the validation set than all other models. We believe this overfitting on the larger models might have occurred because of the relatively low amount of Caiman data. This data scarcity issue could be addressed in future work by introducing more augmented data into the dataset or by obtaining more human-labelled datasets. The final learning rate and regularization strengths used for training were $1.1\text{x}10^{-4}$ and $1.3\text{x}10^{-5}$, respectively.

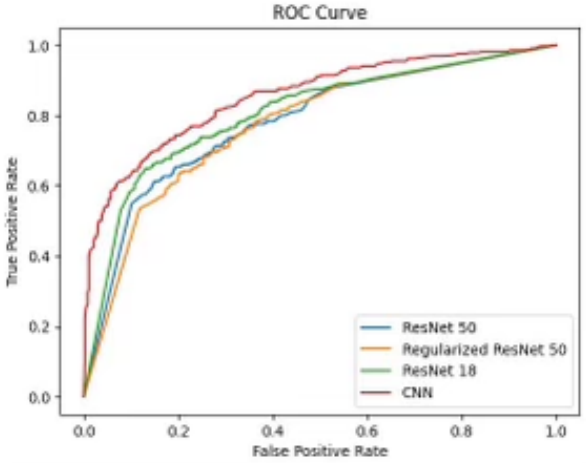

Figure 3. ROC curves for different Caiman model architectures after hyperparameter tuning. The shallower models tended to achieve better AUC than the deeper models for Caiman data.

Conversely, we observed that the Suite2p model performance continued to increase as model depth increased, and we decided to use the ResNet 50 architecture. The final learning rate and regularization strengths used for Suite2p training were $6.8\text{x}10^{-5}$ and $3.9\text{x}10^{-4}$, respectively.

4.3. Evaluation

Because the segmentation algorithms tend to overestimate the number of cells compared to a human annotator (often about 40-90% of identified masks are not true cells), accuracy alone does not provide a good estimate of the quality of our classifiers (e.g. if 90% of the masks in a recording are non-cells, then the model could achieve 90% accuracy by predicting 'non-cell' for every mask). Instead, we evaluated our classifier based on the precision and recall of the classifier's ability to categorize each mask as a true cell or a non-cell. As mentioned above, the F1 score, a common measure that combines precision and recall into a single metric, was also used to evaluate model training.

After training, several qualitative measures, including saliency maps and occlusion maps, were also used to evaluate training success. These qualitative measures also have the benefit of providing additional insights into the key features of the masks of the Suite2p and Caiman algorithms. A better understanding of what discerns a true cell mask from a non-cell mask in each algorithm could be used to guide better algorithm development in the future.

5. Experiments

After completing training, our Caiman model achieved 76.6% accuracy on the Caiman test dataset. Our Suite2p model achieved 93.3% accuracy on the Suite2p test dataset. These accuracy scores were calculated using a threshold of 0.5 on the sigmoid output of the models. However, because removing false cells is much more important than keeping all true cells (to prevent introducing false data that can cause erroneous scientific conclusions into datasets), a model that achieves high precision is preferred over a model that achieves high recall. For this reason, we plotted ROC curves and precision-recall curves for both models and selected a threshold that would achieve good precision at the expense of a slightly lower recall.

As one can see on the ROC curves in Figure 4, our Caiman model achieved a moderately high area under the curve, indicating successful training. The Suite2p model, on the other hand, achieved almost perfect AUC, demonstrating the benefit of having more training data and a deeper network. Based on these ROC curves and precision-recall curves, a nominal threshold of 0.8 was selected to classify masks as true cells cells for the Caiman model, while a threshold of 0.99 was selected for the Suite2p model. These strict thresholds ensure that very few non-true cell masks are automatically classified into the true cell group, while accepting that some true cells may be filtered into the non-cell group.

Misclassifying some true cells as non-cells can potentially lead to discarding good data. To make up for this, our model can return the scores for all masks, allowing researchers to manually inspect masks that have high scores but were not automatically classified as true cells by our models. Despite the high throughput of new two-photon imaging methods, this manual curation step once again becomes possible because our models sort through the bulk of masks, leaving only a few masks left for researchers to manually classify if they want to ensure they retain all good data.

Confusion matrices for both the Caiman and Suite2p models after applying their respective thresholds are included in Figure 5. As one can see, the Caiman model achieved precision, recall, and F1 scores of 0.904, 0.582, and 0.717, respectively. The Suite2p model achieved precision, recall, and F1 scores of 0.865, 0.915, and 0.893, respectively. With both models achieving precision scores near or above 0.9, both models do a fantastic job at sorting out the non-cell masks. The Suite2p model, which has a

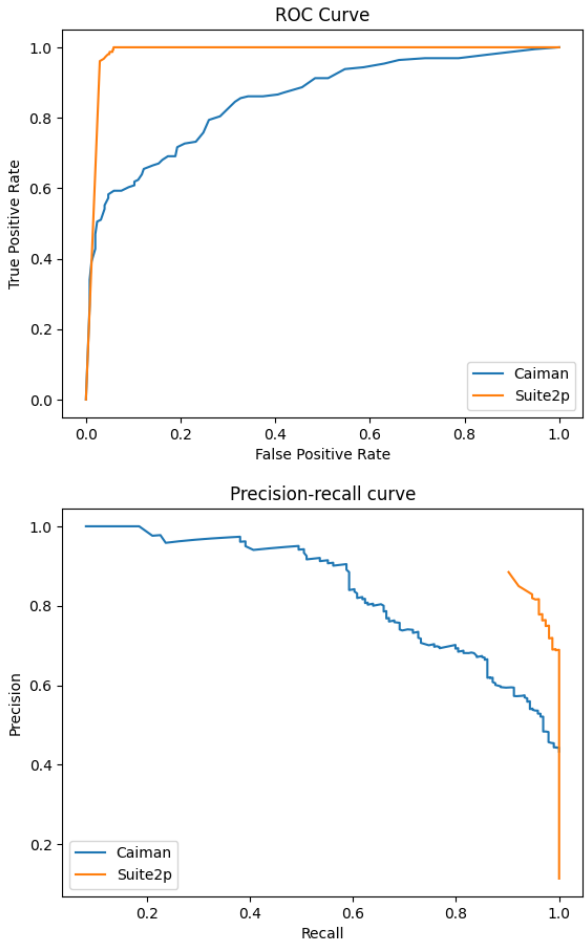

Figure 4. ROC curves (top) and precision-recall curves (bottom) for the Caiman (blue) and Suite2p (orange) final models. The Suite2p model, utilizing a deeper network and more training data, tended to achieve better AUC and more even trade-offs between precision and recall.

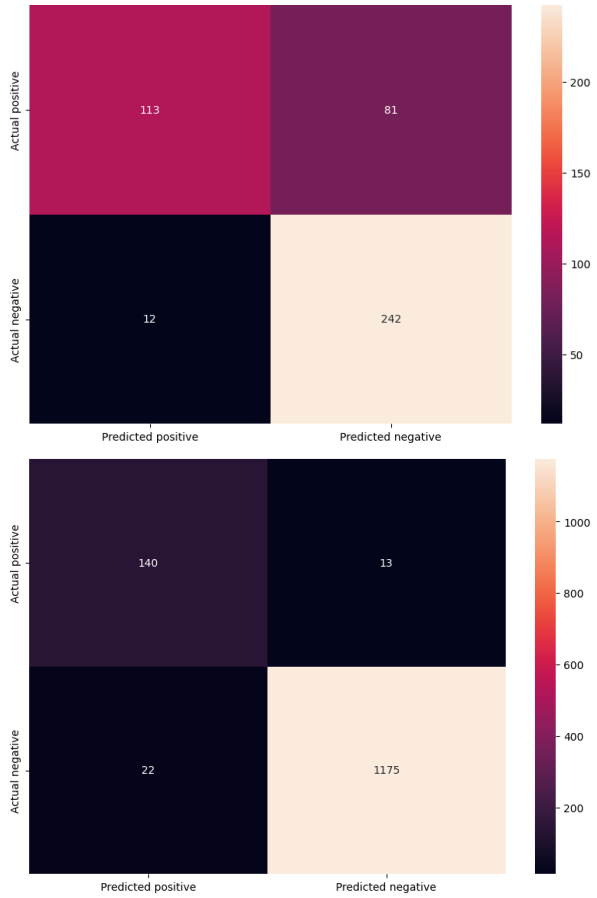

Figure 5. Confusion matrices of predictions for both the Caiman (top) and Suite2p (bottom) models

recall score of over 0.9, also does a fantastic job in preventing the need for any post-processing curation, as almost all true cells are classified correctly. The Caiman model, on the other hand, with a recall of about 0.6, will require some manual post-processing if researchers want to utilize all true cell masks (but, once again, even without this manual post-processing, researchers can still be confident that their results are not being corrupted by non-cell masks!). As stated above, even though models that achieve even higher accuracy would be preferred, given the trade-off between precision and recall, these models represent great advancements in expediting mask sorting tasks for researchers.

5.1. Qualitative measures of model training

After confirming satisfactory precision and recall by our models, we then wanted to evaluate the qualitative outputs of our models. First, we wanted to observe what kinds of true cell masks the model was misclassifying as non-true cells and vice versa. An example from this analysis is included in Figure 6. For the Caiman mis-classifications, it seemed like the model was not consistently able to identify true cells when the strongest mask weight was centered around a few pixels near the edge of the mask. Meanwhile, it classified some non-cells as cells when a strong weight was concentrated in the middle of the mask. These mis-classifications seem reasonable and seem like they could be resolved only by providing the model with temporal information about each of the masks in the recording instead of just spatial information.

When looking at the Suite2p masks, it is first evident that the structure of Suite2p masks was very different than the Caiman masks, as the Caiman masks tended to be round and large, while the Suite2p masks tended to be smaller and take on lots of different shapes. These differences are expected given the different underlying processing techniques in each algorithm. For the Suite2p mis-classifications, it seemed like the model had a difficult time distinguishing the long, thin masks as non-cells. It also had a difficult time

discerning the small, concentrated masks as true cells. Because these types of masks are relatively rare in the datasets, including more examples of these masks in the training dataset, or generating more augmented versions of these types of masks, could help the model be able to better classify these masks.

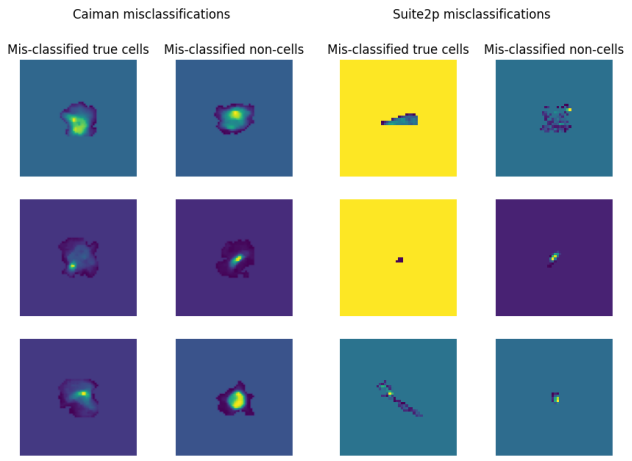

Figure 6. Examples of mis-classified true cell masks and non-cell masks for both Caiman (left) and Suite2p (right)

Next, we plotted saliency maps based on guided backpropagation. By adapting code from [6], we were able to carry out guided backpropagation on the sigmoid outputs of our models. The guided backpropagation works by following gradients from the output image with respect to the input image, with ReLU activations applied in the backwards direction at each activation layer. As [10] demonstrated, applying these ReLUs in the backwards direction prevents backward flow of negative gradients, which correspond to the neurons that decrease the activation of the higher layer unit we aim to visualize.

Two examples from this guided backpropagation analysis are included in Figure 7. As one can see in the examples, the gradients tend to follow the shape of the masks pretty closely, indicating successful model training. In both examples, it seems like the models have the strongest gradients where the mask is most concentrated and the weakest gradients on the edges and center of the mask. It is interesting that the gradients are weakest in these locations because these are not the locations of weakest weights in the input images (as the entire background has zero weight). Instead, it seems indicative of the characteristic annulus of in-plane cells. Because the calcium fluorescent that is being imaged by the two-photon method resides mostly in the cytoplasm of the cell, an annulus around the nucleus of the cell can usually be seen in in-plane cells. These saliency maps seem to trace out the annulus on many of the masks, indicating that the models found one of the defining characteristics of many cells.

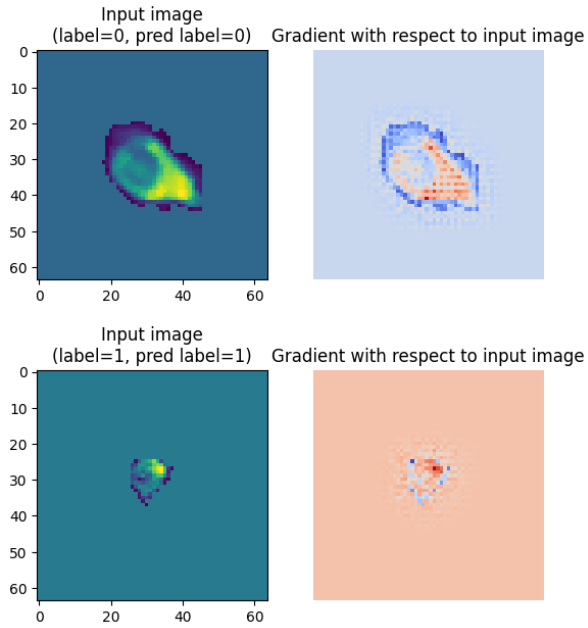

Figure 7. Saliency maps generated by guided backpropagation for two masks.

Finally, we wanted to develop occlusion maps to determine which parts of the masks had the biggest effect on model decisions. To perform this test, we applied a 5x5 sliding window over input images to zero out a small portion of the image. Then, the image was passed through the model, and the output score was recorded for each position of the window. Regions where the model score drops substantially demonstrate which parts of the image the model relied on the most for its prediction. Examples from this analysis are included in Figure 8.

Similar to the saliency maps generated by guided backpropagation, these occlusion maps seem to suggest that the model is influenced most by an annulus around the center of the mask. The model is able to make confident predictions unaffected by the occlusion when the window is in front of the center of the mask (covering the nucleus that does not contain the calcium fluorescent) or near the edge of the frame (not covering the mask). However, when the occlusion blocks the area of the cell corresponding to cytoplasm (that contains the calcium flurescent), the model predictions tend to suffer. These model learnings are very biologically relevant, as the annulus around the center of the cell is one of the defining characteristics of true cells versus non-cells.

6. Conclusion

Using a deep learning approach, we have developed novel classifiers to accelerate two-photon calcium imaging analysis using the two most popular segmentation algorithms. Our model for the Caiman algorithm achieves pre-

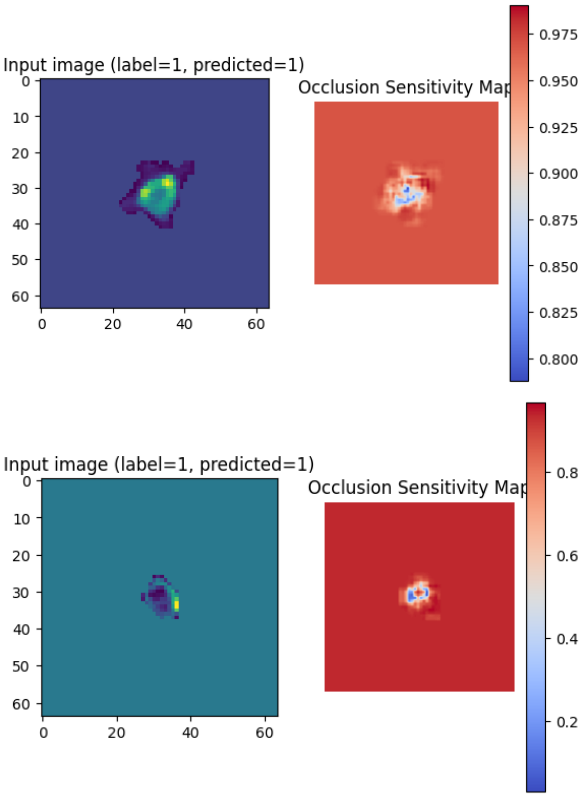

Figure 8. Occlusion maps generated by applying a sliding window over the input image and recording the model output

cision of over 0.9 and recall of almost 0.6, while our model for the Suite2p algorithm achieves precision of 0.87 and recall of 0.92. These precision values indicate robust classification against false positives, which help prevent false data from corrupting scientific conclusions. Furthermore, our models return the output scores of all masks to the researcher, so that he or she can recover some of the true cells that our model misclassified as non-cells. Although this manual curation is not currently feasible because of the size of the datasets being analyzed, the ability of our classifier to confidently sort through the bulk of the data allows researchers to inspect the minority of masks that our models were not as confident on.

Additionally, the qualitative visualizations of the models demonstrate successful training of the models, with both models seeming to identify the characteristic annulus that is present in many cells.

Despite the initial success of our models in distinguishing true cell masks from non-cell masks, there are still a few key points that we would like to develop further in the future. First, the Caiman model achieved relatively low performance compared to the Suite2p model. We believe this is primarily due to the lower amount of training data preventing us from using deeper CNN architectures (ironically, we only had more data for the Suite2p model because the algorithm generated more false masks to begin with). A first step in obtaining more data would be to continue the augmented data generation to higher counts of true cells and non-cells. However, because the augmented images are so similar to the original images, this technique would likely lead to even more overfitting in the long run. Instead, more time should be devoted to manually labelling recordings so that the models can be trained on additional data. Although this manual labelling has a high upfront cost, it would save time in the long run, as better models would decrease the need for manual post-processing curation on every dataset.

Another reason we think more manually labelled datasets are necessary is for additional testing on our models. Because the number of available recordings was low, we used all available data for training for models. However, it is possible that the algorithms produce nuanced mask topologies specific to the recording on which they were generated. It would be very beneficial to reserve all masks from a recording for testing to better understand how the models perform on masks from an unseen recording.

One final key point related to data quality that could be improved is how the labelled datasets are aligned with the Suite2p and Caiman outputs. As discussed in the data section, we considered a mask equivalent to a human-labelled mask if they overlapped by at least 70% by the Jaccard index. It is likely, however, that some of these alignments were not accurate, leading to some masks being mislabelled. For pristine data quality, one should not only label the locations of true cells in the original recording, but also label which masks output by the segmentation algorithms align with the cells. This extra step would again incur a large upfront cost but would also help develop better models that would save more time in the future.

Our models, which were specifically developed for the two most popular existing cell segmentation algorithms, demonstrate the possibility of automatically classifying true cell masks versus non-cell masks in high throughput settings. This novel development helps unlock the potential of the newly developed high throughput two-photon imaging techniques. Our models automatic cell classification abilities help negate the possibility of non-cell masks corrupting scientific data, thereby enabling revolutionary discoveries utilizing new two-photon imaging technologies.

References

[1] J. Demas, J. Manley, F. Tejera, K. Barber, H. Kim, F. M. Traub, B. Chen, and A. Vaziri. High-speed, cortex-wide volumetric recording of neuroactivity at cellular resolution using light beads microscopy. *Nature Methods*, 2021.

[2] A.-I. for Brain-Science. Allen brain atlas, software development kit, 2017.

[3] J. Freeman, M. Rebo, and M. Conlen. Neurofinder challenge. https://github.com/codeneuro/neurofinder, 2018.

[4] A. Giovannucci, J. Friedrich, P. G. Kalfon, B. L. Brown, S. A. Koay, J. Taxidis, F. Najafi, J. L. Gauthier, P. Zhou, B. S. Khakh, D. W. Tank, D. B. Chklovskii, and E. A. Pnevmatikakis. Caiman an open source tool for scalable calcium imaging data analysis. *eLife Neuroscience*, 2019.

[5] H. Inan, C. Schmuckermair, T. Tasci, B. O. Ahanonu, O. Hernandez, J. Lecoq, F. Dinç, M. J. Wagner, M. A. Erdogdu, and M. J. Schnitzer. Fast and statistically robust cell extraction from large-scale neural calcium imaging datasets. *bioRxiv*, 2021.

[6] U. Ozbulak and M. Freidank. Guided backprop. https://gist.github.com/MFreidank/c61f5c762ed9311c5083a04c826396c9, 2018.

[7] M. Pachitariu, C. Stringer, M. Dipoppa, S. Schröder, L. F. Rossi, H. Dalgleish, M. Carandini, and K. D. Harris. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *bioRxiv*, 2017.

[8] Sità, Brondi, Lagomarsino, de Leon, Roig, Curreli, Panniello, Vecchia, and Fellin. A deep-learning approach for online cell identification and trace extraction in functional two-photon calcium imaging. *Nature News*, 2022.

[9] S. Soltanian-Zadeh, K. Sahingur, S. Blau, Y. Gong, and S. Farsiu. Fast and robust active neuron segmentation in two-photon calcium imaging using spatiotemporal deep learning. *PNAS Biological Sciences*, 2019.

[10] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *ArXiv*, 2015.