Automatic Neuron Detection in Calcium Imaging Data Using Convolutional Networks

Noah J. Apthorpe1* Alexander J. Riordan2* Rob E. Aguilar1 Jan Homann2
Yi Gu2 David W. Tank2 H. Sebastian Seung12
1Computer Science Department 2Princeton Neuroscience Institute Princeton University
{apthorpe, ariordan, dwtank, sseung}@princeton.edu
*These authors contributed equally to this work

Abstract

Calcium imaging is an important technique for monitoring the activity of thousands of neurons simultaneously. As calcium imaging datasets grow in size, automated detection of individual neurons is becoming important. Here we apply a supervised learning approach to this problem and show that convolutional networks can achieve near-human accuracy and superhuman speed. Accuracy is superior to the popular PCA/ICA method based on precision and recall relative to ground truth annotation by a human expert. These results suggest that convolutional networks are an efficient and flexible tool for the analysis of large-scale calcium imaging data.

1 Introduction

Two-photon calcium imaging is a powerful technique for monitoring the activity of thousands of individual neurons simultaneously in awake, behaving animals [1, 2]. Action potentials cause transient changes in the intracellular concentration of calcium ions. Such changes are detected by observing the fluorescence of calcium indicator molecules, typically using two-photon microscopy in the mammalian brain [3]. Repeatedly scanning a single image plane yields a time series of 2D images. This is effectively a video in which neurons blink whenever they are active [4, 5, 6].

In the traditional workflow for extracting neural activities from the video, a human expert manually annotates regions of interest (ROIs) corresponding to individual neurons [5, 7]. Within each ROI, pixel values are summed for each frame of the video, which yields the calcium signal of the corresponding neuron versus time. A subsequent step may deconvolve the temporal filtering of the intracellular calcium dynamics for an estimate of neural activity with better time resolution. The traditional workflow has the deficiency that manual annotation becomes laborious and time-consuming for very large datasets. Furthermore, manual annotation does not de-mix the signals from spatially overlapping neurons.

Unsupervised basis learning methods (PCA/ICA [8], CNMF [9], dictionary learning [10], and sparse space-time deconvolution [11]) express the video as a time-varying superposition of basis images. The basis images play a similar role as ROIs in the traditional workflow, and their time-varying coefficients are intended to correspond to neural activities. While the basis learning methods are useful for some datasets, they have not fully replaced the traditional workflow. Many neuroscientists still annotate ROIs manually.

The reason for the popularity of manual annotation is not altogether clear. At first glance, the videos produced by calcium imaging seem simple (neurons blinking on and off). Yet automating image analysis has not been trivial. One difficulty is that images are corrupted by noise and artifacts due to brain motion. Another difficulty is variability in the appearance of cell bodies, which vary in shape, size, spacing, and resting-level fluorescence. The most subtle difficulty is that neuroscientists vary in
their criteria for ROI selection. Some may want only cell bodies, while others want dendrites. Some may want only active cells, while others want both active and inactive cells. Some may wish to reject slightly out-of-focus neurons. For all of these reasons, a neuroscientist may end up spending hours or days tuning the parameters of nominally automated methods, or may never succeed in finding a set of parameters that produces satisfactory results. Even after parameters are tuned, the neuroscientist may end up inspecting the results of a basis learning algorithm to verify accuracy.

As a way of dealing with these difficulties, we focus here on a supervised learning approach to automated ROI detection. An automated ROI detector could be used to replace manual ROI detection by a human expert in the traditional workflow, or could be used to make the basis learning algorithms more reliable by providing good initial conditions for basis images. However, the usability of an automated algorithm strongly depends on its attaining high accuracy. Supervised learning has become the dominant approach for attaining high accuracy in many computer vision problems.

We assemble ground truth datasets consisting of calcium imaging videos along with ROIs drawn by human experts, and employ a precision-recall formalism for quantifying accuracy. We train a sliding window convolutional network (ConvNet) to take a calcium video as input, and output a 2D image that matches the human-drawn ROIs as well as possible. The ConvNet achieves near-human accuracy and exceeds that of PCA/ICA [8].

The prior work that is most similar to ours used supervised learning based on boosting with hand-designed features [12]. Other previous attempts to automate ROI detection did not employ supervised machine learning. For example, hand-designed filtering operations [13] and normalized cuts [14] were applied to image pixel correlations.

The major cost of supervised learning is the human effort is required to create the training set. As a rough guide, our results suggest that on the order of 10 hours of effort or 1000 annotated cells are sufficient to yield a ConvNet with usable accuracy. Furthermore, the marginal effort required to create a training set is essentially zero for those neuroscientists who are already using the traditional workflow.

From the deep learning perspective, an interesting aspect of our work is that a ConvNet that processes a spatiotemporal (2+1)D image is trained using only spatial (2D) annotations. Full spatiotemporal annotations (spatial locations and times of activation) would have been more laborious to collect. The use of purely spatial annotations is possible because the neurons in our videos are stationary (apart from motion artifacts). This makes our task simpler than other applications of ConvNets to video processing [15].

2 Neuron detection benchmark

We use the precision-recall framework to quantify accuracy of neuron detection. Predicted ROIs are classified as false positives (FP), false negatives (FN), and true positives (TP) relative to ground truth ROIs. Precision and recall are defined by

\[
\text{precision} = \frac{TP}{TP + FP} \quad \text{recall} = \frac{TP}{TP + FN}
\]

Both measures would be equal to one if predictions were perfectly accurate, i.e., higher numbers are better. If a single measure of accuracy is required, we use the harmonic mean of precision and recall, \(1/F_1 = (1/\text{precision} + 1/\text{recall})/2\). The \(F_1\) score favors neither precision nor recall, but in practice a neuroscientist may care more about one measure than the other. For example, some neuroscientists may be satisfied if the algorithm fails to detect many neurons (low recall), so long as it produces few false positives (high precision). Other neuroscientists may want the algorithm to find as many neurons as possible (high recall), even if there are many false positives (low precision).

For computing precision and recall, it is helpful to define the overlap between two ROIs \(R_1\) and \(R_2\) as the Jaccard similarity coefficient \(|R_1 \cap R_2| / |R_1 \cup R_2|\) where \(|R|\) denotes the number of pixels in \(R\). For each predicted ROI, we find the ground truth ROI with maximal overlap. The ground truth ROIs with overlap greater than 0.5 are assigned to the predicted ROIs with which they overlap the most. These assignments are true positives. Leftover ROIs are the false positives and false negatives.

We prefer the precision-recall framework over the receiver operating characteristic (ROC), which was previously used as a quantitative measure of neuron detection accuracy [12]. This is because...
precision and recall do not depend on true negatives, which are less well-defined. (The ROC depends on true negatives through the false positive rate.)

**Ground truth generation by human annotation** The quantitative measures of accuracy proposed above depend on the existence of ground truth. For the vast majority of calcium imaging datasets, no objectively defined ground truth exists, and we must rely on subjective evaluation by human experts. For a dataset with low noise in which the desired ROIs are cell bodies, human experts are typically confident about most of their ROIs, though some are borderline cases that may be ambiguous. Therefore our measures of accuracy should be able to distinguish between algorithms that differ widely in their performance, but may not be adequate to distinguish between algorithms that are very similar.

Two-photon calcium imaging data was gathered from both the primary visual cortex (V1) and medial entorhinal cortex (MEC) from awake-behaving mice (Supplementary Methods). All experiments were performed according to the Guide for the Care and Use of Laboratory Animals, and procedures were approved by Princeton University’s Animal Care and Use Committee.

Each time series of calcium images was corrected for motion artifacts (Supplementary Methods), average-pooled over time with stride 167, and then max-pooled over time with stride 6. This downsampling in time both reduced noise and made the dataset into a more manageable size. Then human experts annotated ROIs using the ImageJ Cell Magic Wand Tool [16], which automatically generates a region of interest (ROI) based on a single mouse click. The human experts found 4006 neurons in the V1 dataset with an average of 148 neurons per image series, and 538 neurons in the MEC dataset with an average of 54 neurons per image series.

Human experts used the following criteria to select neurons: (1) the soma was in the focal plane of the image—apparent as a light doughnut-like ring (the soma cytosol) surrounding a dark area (the nucleus), or (2) the area showed significantly changing brightness distinguishable from background and had the same general size and shape expected from a neuron in the given brain region.

After motion correction, downsampling, and human labeling, the V1 dataset consisted of 27 16-bit grayscale multi-page TIFF image series ranging from 28 to 142 frames per series with 512x512 pixels per frame. Human annotation time was estimated at one hour per image series. The MEC dataset consisted of 10 image series ranging from 5 to 28 frames in the same format. Both datasets had corresponding human-labeled ROIs as 512x512 pixel binary masks. Human annotation time was estimated at 40 minutes per image series.

### 3 Convolutional network

**Preprocessing of images and ground truth ROIs.** Microscopy image series from the V1 and MEC datasets were preprocessed prior to network training (Figure 1). Image contrast was enhanced by clipping all pixel values above the 99th percentile and below the 3rd percentile. Pixel values were then normalized to [0, 1]. We divided the V1 series into 60% training, 20% validation, and 20% test sets and the MEC series into 50% training, 20% validation, and 30% test sets.

Neighboring ground truth ROIs often touched or even overlapped with each other. For the purpose of ConvNet training, we shrank the ground truth ROIs by replacing each one with a 4-pixel radius disk located at the centroid of the ROI. The shrinkage was intended to encourage the ConvNets to separate neighboring neurons.

**Convolutional network architecture and training.** The architecture of the (2+1)D ConvNet is depicted in Figure 2. The input is an image stack containing \( T \) time slices. There are four convolutional layers, a max pooling over all time slices, and then two pixelwise fully connected layers. This yields two 2D grayscale images as output, which together represent the softmax probability of each pixel being inside a shrunken ROI.

The convolutional layers were chosen to contain only 2D kernels, because the temporal downsampling used in the preprocessing (§2) was so severe that any neural activity lasted for only a single time frame. If less temporal downsampling had been used in the preprocessing, it could have been advantageous to use 3D kernels in the convolutional layers. Each output pixel depended on a \( 37 \times 37 \times T \) pixel field of view in the input, where \( T \) is the number of frames in the input image stack. \( T \) was equalized
Col 1: Calcium imaging stacks were motion-corrected and downsampled in time. Col 2: Image contrast was enhanced by clipping pixel intensities below the 3rd and above the 99th percentile, then linearly rescaling pixel intensities between those new bounds. Col 3: Human-labeled ROIs were converted into binary masks. Col 4: Networks were trained to detect 4-pixel radius circular centroids of human-labeled ROIs. Primary visual cortex (V1, Row 1) and medial entorhinal cortex (MEC, Row 2) datasets were preprocessed identically.

to 50 for all image stacks in the V1 dataset and 5 for all image stacks in the MEC dataset, using averaging and bicubic interpolation.

The ConvNet was applied in a $37 \times 37 \times T$ window sliding in two dimensions over the input image stack to produce an output pixel for every location of the window fully contained in the input image stack. For comparison, we also trained a 2D ConvNet that took as input the time-averaged image stack and did no temporal computation (Figure 2).

We used ZNN, an open-source sliding window ConvNet package with multi-core CPU parallelism and FFT-based convolution [17]. ZNN automatically augmented training sets by random rotations (multiples of 90 degrees) and reflections of image patches to facilitate ConvNet learning of invariances. The training sets were also rebalanced by the fraction of pixels in human-labeled ROIs to the total number of pixels. See Supplementary Methods for further details.

The $(2+1)$D network was trained with the softmax loss and output patches of size $120 \times 120$. The learning rate parameter was annealed by hand from 0.01 to 0.002, and the momentum parameter was annealed by hand from 0.9 to 0.5. The network was trained for 16800 updates for the V1 dataset, which took approximately 1.2 seconds/update ($\sim 5.5$ hrs) on an Amazon EC2 c4.8xlarge instance (Supplementary Figure 1). The network was trained for 200000 SGD updates for the MEC dataset, which took approximately 0.1 seconds/update ($\sim 5.5$ hrs).

The 2D network training omitted annealing of the learning rate and momentum parameters. The 2D network was trained for 14000 SGD updates for the V1 dataset, which took approximately 0.9 seconds/update ($\sim 3.75$ hrs) on an Amazon EC2 c4.8xlarge instance (Supplementary Figure 1). We performed early stopping on the network after 10200 SGD updates based on the validation loss.

**Network output postprocessing.** Network outputs were converted into individual ROIs by: 1. Thresholding out pixels with low probability values 2. Removing small connected components 3. Weighting resulting pixels with a normalized distance transform 4. Performing marker-based watershed labeling with local max markers 5. Merging watershed regions smaller than $m$ pixels, and 6. Automatically applying the ImageJ Cell Magic Wand tool to the original images at the centroids of the watershed regions.

We plan to release ZNN configuration files and pre- and postprocessing source code for community use. We also intend to create an Amazon EC2 AMI with our pipeline as a publicly available tool.
4 Results

**ConvNets successfully detect cells in calcium images.** A sample image from the V1 test set and ConvNet output is shown in Figure 4. Postprocessing of the ConvNet output yielded predicted ROIs, many of which are the same as the human ROIs (Figure 4c). As described in Section 2, we quantified agreement between computer and human using the precision-recall formalism. Both (2+1)D and 2D networks attained the same $F_1$ score (0.71). Full precision-recall curves are given in Supplementary Figure 1.

Inspection of the computer-human disagreements suggested that some were not actually computer errors. To investigate this hypothesis, the original human expert was requested to reevaluate all disagreements with the (2+1)D network. After reevaluation, 131 false positives became true positives, and 30 false negatives became true negatives (Figure 4D). Some of these reversals appeared to involve unambiguous human errors in the original annotation, while others were ambiguous cases (Figure 4E–G). After reevaluation, the $F_1$ score of the (2+1)D network increased to 0.82. The $F_1$ score of the human expert’s reevaluation relative to his original annotation was 0.89. Based on these scores, it seems fair to say that the ConvNet is nearing human performance.

**(2+1)D versus 2D network.** The (2+1)D and 2D networks achieved similar precision, recall, and $F_1$ scores on the V1 dataset; however, the (2+1)D network produced raw output with less noise than the 2D network (Figure 3). Qualitative inspection also indicates that the (2+1)D network finds transiently active and transiently in focus neurons missed by the 2D network (Figure 3). Although such neurons occurred infrequently in the V1 dataset and did not noticeably affect network scores, these results suggest that datasets with larger populations of transiently active or variably focused cells will particularly benefit from (2+1)D network architectures.

**Convolutional network segmentation outperforms PCA/ICA.** The (2+1)D network was also able to successfully locate neurons in the MEC dataset (Figure 5). For comparison, we also implemented and applied PCA/ICA as described by Ref. [8]. The (2+1)D network achieved an $F_1$ score of 0.51, while PCA/ICA achieved 0.27. Precision and recall numbers are given in Figure 5.
ConvNet accuracy was lower on the MEC dataset than the V1 dataset, probably because the former has more noise and larger motion artifacts. Also, the amount of training data for the MEC dataset was much smaller.

PCA/ICA accuracy was numerically worse, but this result should be interpreted cautiously. PCA/ICA is intended to identify active neurons, while the ground truth included both active and inactive neurons. Furthermore, the ground truth depends on the human expert’s selection criteria, which are not accessible to PCA/ICA.

Training and post-processing optimization for ConvNet segmentation took ~6 hours with a forward pass taking ~1.2 seconds per image series. Parameter optimization for PCA/ICA took ~2.5 hours with a forward passes taking ~40 minutes. This amounted to ~6 hours total computation time for the ConvNet and ~9 hours for the PCA/ICA algorithm. This suggests that ConvNet segmentation is faster than PCA/ICA for all but the smallest datasets.

5 Discussion

The lack of quantitative difference between (2+1)D and 2D ConvNet accuracy (same $F_1$ score on the V1 dataset) may be due to limitations of our study, such as imperfect ground truth and severe temporal downsampling in preprocessing. It may also be due to the fact that the vast majority of the neurons in the V1 dataset are clearly visible in the time-averaged image. We do have qualitative evidence that the (2+1)D architecture may turn out to be superior for other datasets, because its output looks cleaner, and it is able to detect transiently active or transiently in-focus cells (Figure 3).

The (2+1)D ConvNet outperformed PCA/ICA in the precision-recall metrics. ConvNets readily locate inactive neurons and process new images rapidly once trained. ConvNets adapt to the selection criteria of the neuroscientist if they are implicitly contained in the training set. They do not depend on hand-designed features and so require little expertise in computer vision. ConvNet speed could enable novel applications involving online ROI detection, such as computer-guided single-cell optogenetics [18] or real-time neural feedback experiments.
Figure 4: The (2+1)D network successfully detected neurons in the V1 test set with near-human accuracy.  
A) Slice from preprocessed calcium imaging stack input to network.  
B) Network softmax probability output. Brighter regions are considered by the network to have higher probability of being a neuron. 
C) ROIs found by the (2+1)D network after post-processing, overlaid with human labels. Network output is shown by green outlines, whereas human labels are red. Regions of agreement are indicated by yellow overlays. 
D) ROI labels added by human reevaluation are shown in blue. ROI labels removed by reevaluation are shown in magenta. Post hoc assessment of network output revealed a sizable portion of ROIs that were initially missed by human labeling. 
E) Examples of formerly negative ROIs that were reevaluated as positive.  
F) Initial positive labels that were reevaluated to be false. 
G) Examples of ROIs that remained negative even after reevaluation.  
H) $F_1$ scores for (2+1)D and 2D networks before and after ROI reevaluation. Human labels before and after reevaluation were also compared to assess human labeling variability. Boxplots depict the variability of $F_1$ scores around the median score across test images.
Figure 5: The (2+1)D network successfully detected neurons in the MEC test set with higher precision and recall than PCA/ICA. A) Slice from preprocessed calcium imaging stack that was input to network. B) Network output, normalized by softmax. C) ROIs found by the (2+1)D network after postprocessing, overlaid with neurons previously labeled by a human. Network output is shown by red outlines, whereas human labels are green. Regions of agreement are indicated by yellow overlays. D) The ROIs found by PCA/ICA are overlaid in blue. E) Quantitative comparison of $F_1$ score, precision, and recall for (2+1)D network and PCA/ICA on human-labeled MEC data.
Acknowledgments

We thank Kisuk Lee, Jingpeng Wu, Nicholas Turner, and Jeffrey Gauthier for technical assistance. We also thank Sue Ann Koay, Niranjan Prasad, Cyril Zhang, and Hussein Nagree for discussions. This work was supported by IARPA D16PC00005 (HSS), the Mathers Foundation (HSS), NIH R01 MH083686 (DWT), NIH U01 NS090541 (DWT, HSS), NIH U01 NS090562 (HSS), Simons Foundation SCGB (DWT), and U.S. Army Research Office W911NF-12-1-0594 (HSS).

References

[1] D. Huber, D. A. Gutnisky, S. Peron, D. H. O’Connor, J. S. Wiegert, L. Tian, T. G. Oertner, L. L. Looger, and K. Svoboda. Multiple dynamic representations in the motor cortex during sensorimotor learning. *Nature*, 2012.

[2] Daniel A. Dombeck, Forrest Khabbaz, Anton N.and Collman, Thomas L Adelman, and David W. Tank. Imaging large scale neural activity with cellular resolution in awake mobile mice. *Neuron*, 2008.

[3] W Denk, JH Strickler, and WW Webb. Two-photon laser scanning fluorescence microscopy. *Science*, 1990.

[4] Tsai-Wen Chen, Trevor J. Wardill, Yi Sun, Stefan R. Pulver, Sabine L. Renninger, Amy Baohan, Eric R. Schreiter, Rex A. Kerr, Michael B. Orger, Vivek Jayaraman, Loren L. Looger, Karel Svobodz, and Douglas S. Kim. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 2013.

[5] Christopher D. Harvey, Philip Coen, and David W. Tank. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature*, 2012.

[6] Priyamvada Rajasethupathy, Sethuraman Sankaran, James H. Marshel, Christina K. Kim, Emily Ferenczi, Soo Yeun Lee, Andre Berndt, Charu Ramakrishnan, Anna Jaffe, Maisie Lo, Conor Liston, and Karl Deisseroth. Projections from neocortex mediate top-down control of memory retrieval. *Nature*, 2015.

[7] W Göbel, BM Kampa, and F Helmchen. Imaging cellular network dynamics in three dimensions using fast 3d laser scanning. *Nature Methods*, 2006.

[8] Eran A. Mukamel, Axel Nimmerjahn, and Mark J. Schnitzer. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron*, 2009.

[9] Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, Misha Ahrens, Randy Bruno, Thomas M Jessell, Darcy S Peterka, Rafael Yuste, and Liam Paninski. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 2016.

[10] Marius Pachitariu, Adam M Packer, Noah Pettit, Henry Dalgleish, Michael Hausser, and Maneesh Sahani. Extracting regions of interest from biological images with convolutional sparse block coding. *NIPS*, 2013.

[11] Ferran Diego and Fred A. Hamprecht. Sparse space-time deconvolution for calcium image analysis. *NIPS*, 2014.

[12] Ilya Valmianski, Andy Y Shih, Jonathan D. Driscoll, David W. Matthews, Yoav Freund, and David Kleinfeld. Automatic identification of fluorescently labeled brain cells for rapid functional imaging. *Journal of Neurophysiology*, 2010.

[13] Spencer L Smith and Michael Häusser. Parallel processing of visual space by neighboring neurons in mouse visual cortex. *Nature neuroscience*, 13(9):1144–1149, 2010.

[14] Patrick Kaifosh, Jeffrey D. Zaremba, Nathan B. Danielson, and Attila Losonczy. Sima: Python software for analysis of dynamic fluorescence imaging data. *Frontiers in Neuroinformatics*, 2014.

[15] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. 2014.

[16] T. Walker. Cell magic wand tool. 2014.

[17] Aleksandar Zlateski, Kisuk Lee, and H. Sebastian Seung. ZNN - A fast and scalable algorithm for training 3d convolutional networks on multi-core and many-core shared memory machines. *CoRR*, 2015.

[18] JP Rickgauer, K Deisseroth, and DW Tank. Simultaneous cellular-resolution optical perturbation and imaging of place cell firing fields. *Nature Neuroscience*, 2014.