Suraj Khetarpal

APMA S-115 Final Project

Early Development of the Macaque Visual Cortex

8/5/16

Abstract

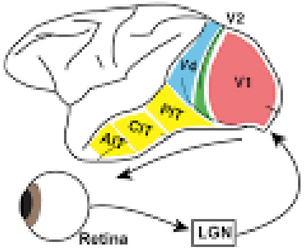
Scientists do not yet fully understand the development timeline of the visual cortex during infancy. In this analysis, I use machine learning techniques to explore the early development of the visual cortex in baby macaque monkeys. Specifically, I perform a multivoxel pattern analysis (MVPA) on fMRI data collected during a longitudinal study involving visual experiments. This analysis reveals that the visual cortex is highly responsive by the age of 3 months and undergoes rapid development between the ages of 3 and 9 months.

Introduction

For decades, neuroscientists have struggled to understand the early development of the primate brain. Such an understanding would unlock the mysteries behind how human cognition works, and could have a world changing effect on the design of artificial intelligence systems. Unfortunately, researchers have made little progress in this area because babies are nearly impossible to study. Not only are they unable to communicate (which rules out most neurocognitive testing), but they also do not perform well in passive tests such as brain scans (for a variety of reasons). Human babies are especially challenging, so some researchers have turned to non-human primate babies, since they possess a very similar brain architecture. Without much hard evidence, it is generally assumed that the visual cortex is very immature during the first year of life, especially in the high level regions.

The Livingstone Laboratory at the Harvard Medical School has recently performed a groundbreaking longitudinal study in which they successfully collected fMRI data from infant macaque monkeys that were subjected to vision experiments. The monkeys were scanned at different points in time, ranging from 3 months to 15 months of age. During the experiments, the monkeys were made to hold still in an fMRI machine and watch a screen that displayed either nothing (the null condition) or an image of a face or object (the stimulus). The lab is now analyzing this fMRI data using standard univariate statistical techniques. However, they are very interested in discovering whether a multivariate machine learning classifier analysis might reinforce their findings, or even provide new insights. A classifier might be able to pick up patterns in the brain activity that go undetected by a univariate analysis.

I have conducted the requested multivariate classifier analysis on one of the monkeys. In this paper, I describe my process and the results. My analysis covers data from 9 experiments, the first done at the age of 3 months, and the last at 15 months. I focus on the development of the dorsal stream of the visual cortex, which is the part of the brain used to identify what is being looked at. Especially interesting are three functionally distinct brain regions: V1, PIT, and AIT (see figure to the left). When visual information enters the visual cortex, it is first processed by the V1 region, which handles low level visual information such as colors, spatial frequencies, and visual orientations. It is then passed downstream for more complex and abstract processing. The PIT region helps to turn visual information into conceptual categories. For example, it tells the monkey whether it is seeing a face or an object. Eventually, visual information reaches the AIT, which is a highly abstract region that is associated with the identification of individuals. It is this region that helps the monkey to recognize other monkeys, or to identify specific objects.



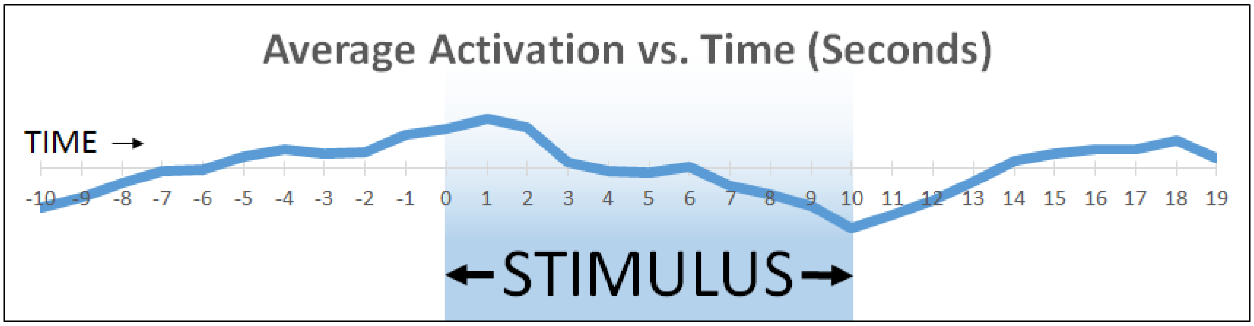
The experiments cycled through 20 second long time blocks during which an image was displayed on the screen for 10 seconds and then nothing was displayed for 10 seconds. This was repeated between 8 and 50 times, depending on the experiment. The fMRI machine measured blood flow in the brain, which correlates with neuronal activity. The measurements occur on a spatial scale of cubic millimeters of brain tissue, which are called "**voxels**". A monkey brain contains up to 100K voxels, and so taking an fMRI scan every second results in massive datasets. This is why data wrangling and computational limitations played a major role in my analysis.

Model

A classifier considers information from multiple voxels simultaneously (which is why it is called "multivariate"). It can detect patterns between the activity of different voxels, which allows for greater sensitivity than a univariate analysis can provide. The purpose of the multivariate classifier analysis is to determine whether or not the brain and its subregions are responding to visual stimuli. An undeveloped brain would not respond well, and therefore it would provide only noise to a classifier. A well developed brain would respond in a clear and consistent manner, and would therefore provide good data to a classifier. The underlying assumption of the analysis is that if a classifier can be trained with a brain's fMRI data to make accurate predictions about a vision experiment, then this is proof that the brain is being "**activated**" by the visual stimuli in the experiment. In other words, if the classifier's accuracy is statistically significant, this is evidence that the brain is responding to the visual stimuli. It is important to note that the term "activated" differs from the term “activation value”. A voxel is only considered “activated” if its fMRI activation value is dependent upon the presence of the experimental stimulus.

From here on out, I will use the term “**stimulus condition**” to refer to the times during an experiment when the subject (the monkey) was shown a stimulus image, and I will use the term “**null condition**” to refer to times when the subject was shown a blank screen. I will also use the term “**stimulus block**” to refer to the intermittent 10 second long time blocks during which a stimulus was shown, and “**null block**” to refer to the intermittent 10 second time long blocks during which a blank screen was shown.

Before I could begin the analysis, I had to perform some data wrangling. For each voxel in an experiment’s dataset, there are hundreds of fMRI data points (one per second). I needed to boil this down to one average value per each 10 second stimulus block or null block. However, it was hard to know which timepoints to use in taking these averages. There is a time lag between the onset of a stimulus block and the onset of blood flow that would show in an fMRI scan. To solve this problem, I created the below plot. It displays the average fMRI response across all the voxels in the monkey’s V1 brain region. To eliminate noise and creep from the fMRI machine, I had to z-scored the response from each voxel prior to taking the average. Since the average includes multiple stimulus blocks with different starting times, I treated each stimulus block as if it started at time zero.



We see that the fMRI signal is high just before a stimulus is presented. About 2 seconds after the stimulus is presented, the signal begins to decrease. After the stimulus block ends, the signal begins to return to its original level. After seeing this plot, I decided to use the average of the values at time points 8 through 12 to represent the activation of the stimulus block, and the average of the timepoints -1 through 2 to represent the activation of the null block.

After calculating the average activation values for each 10 second stimulus block and null block, I provided them to naive-bayes (supervised learning) classifiers as my “**observations**”. Since each voxel has its own associated set of observations, the classifiers treated each voxel as a separate dimension in its feature space. The classifiers were trained using a data-frame that had one column per voxel, and one row per observation. The classifiers were told which rows represented a stimulus condition (an image) and which represented a null condition (a blank screen). I used a leave-one-out cross validation technique to test the accuracy of the classifier predictions. In actuality, I had to leave two observations out at a time (one stimulus observation and one null observation). Otherwise the number of stimulus observations and null observations would become unequal, which would bias the classifier.

Unfortunately, I only had 16 to 100 observations per experiment, but tens of thousands of voxels to consider. In general, classifiers do not perform well when there are few observations relative to feature dimensions. Furthermore, an excess of feature dimensions can cause a classifier to train slowly and poorly. To mitigate this problem, I used naive bayes classifiers because they are known for their ability to handle a large feature space, and because they train quickly by evaluating a closed form expression (ref 1). To further reduce the feature space, I only analyzed small regions of the brain at a time.

During the first half of my investigation, I performed “searchlight” analyses (ref 3) in which I looked at one 3x3x3 cube of voxels at a time. I allowed each voxel in the brain to act as the center of the cube once, so that I traversed the entire brain. The observations from the 27 voxels contained in each cube were used to train and test a classifier, and the resulting accuracy rate was assigned to the central voxel.

I needed a way to determine whether the accuracy rate at each voxel was high enough to be considered statistically significant. Unfortunately, permutation tests were impossible due to time restrictions (there were 10s of thousands of voxel cubes to analyze for each brain, so doing permutation tests for just one experiment would take a full day). I had no choice but to use the following work around. The null hypothesis of the significance test is that the classifier’s predictions are random guesses, meaning that each prediction has a 50% chance of being correct. Therefore, I used the inverse binomial cumulative distribution function to find the significance threshold for a given P-value (ref 3). The formula to find the threshold is:

$$Accuracy\,Threshold \;=\; binoinv([1 - Pvalue],\ \#observations,\ \%chance)\ /\ \#observations$$

The major flaw here is that this distribution assumes that the observations are independent, which is not necessarily true in this case.

After completing each searchlight analysis, I generated a map of voxels whose accuracies exceeded the statistical threshold. These maps reveal areas where the brain was decidedly responsive to the visual stimuli.

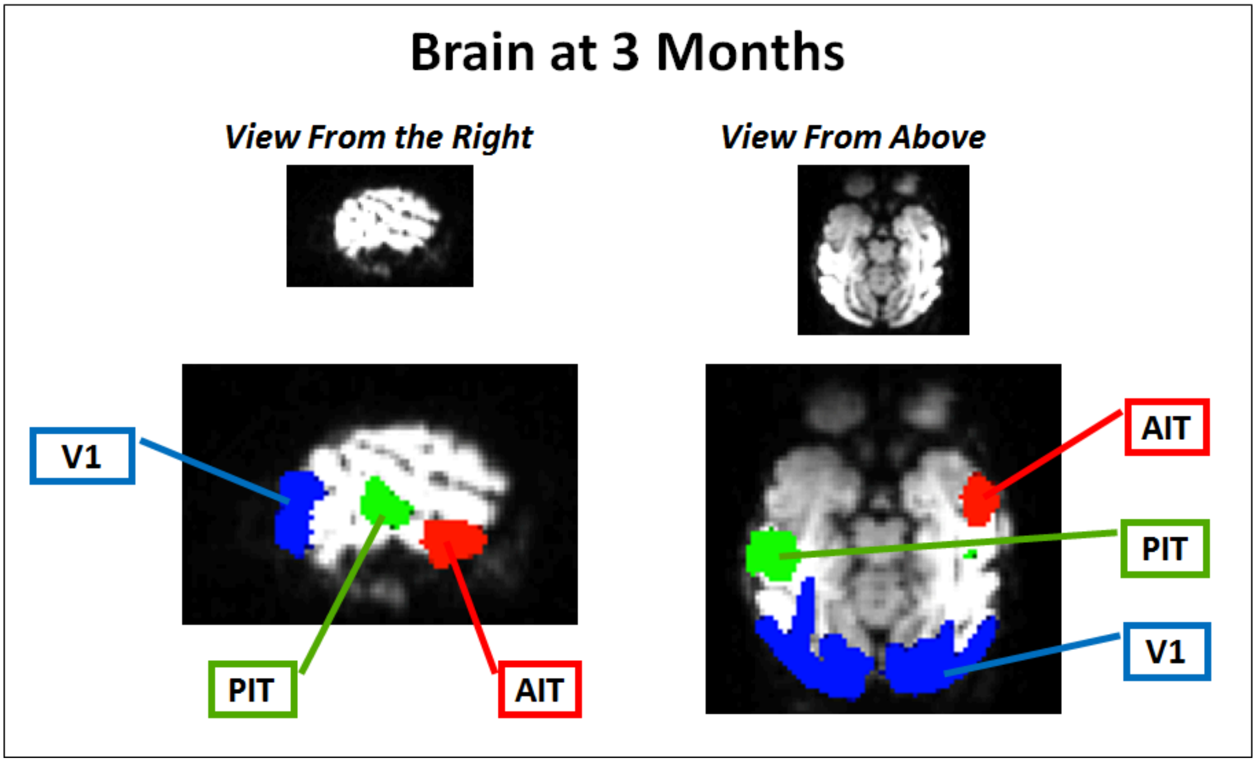
During the second phase of my investigation, I performed a “regional analysis” in which I analyzed the classification performance that comes from brain regions V1, PIT, and AIT. These regions contain huge numbers of voxels (V1 has 10K, PIT has 1.7K, and AIT has 1.4K), which means that the feature space is extremely large. One advantage of the regional analyses was that I could use permutation testing to decide whether the accuracy rates for each region are statistically significant. To find the significance threshold, I did the following process 1000 times:

I randomly permuted the labels (stimulus vs. null) of the observations and reran the classifier to get an accuracy rate. After collecting 1000 accuracy rates, I sorted them and used the 95% (the 950th accuracy rate) as the threshold for significance.

Model variables to be analyzed include the type of test (univariate vs. searchlight classifier vs. regional classifier), the classification P-value, the searchlight size, the number of observations provided to the classifier, and the number of features (voxels) provided to the classifier.

Analysis

I began by analyzing the data from the the earliest experiment, during which the monkey was only 3 months old. The figure below shows the location of the V1, PIT, and AIT regions.

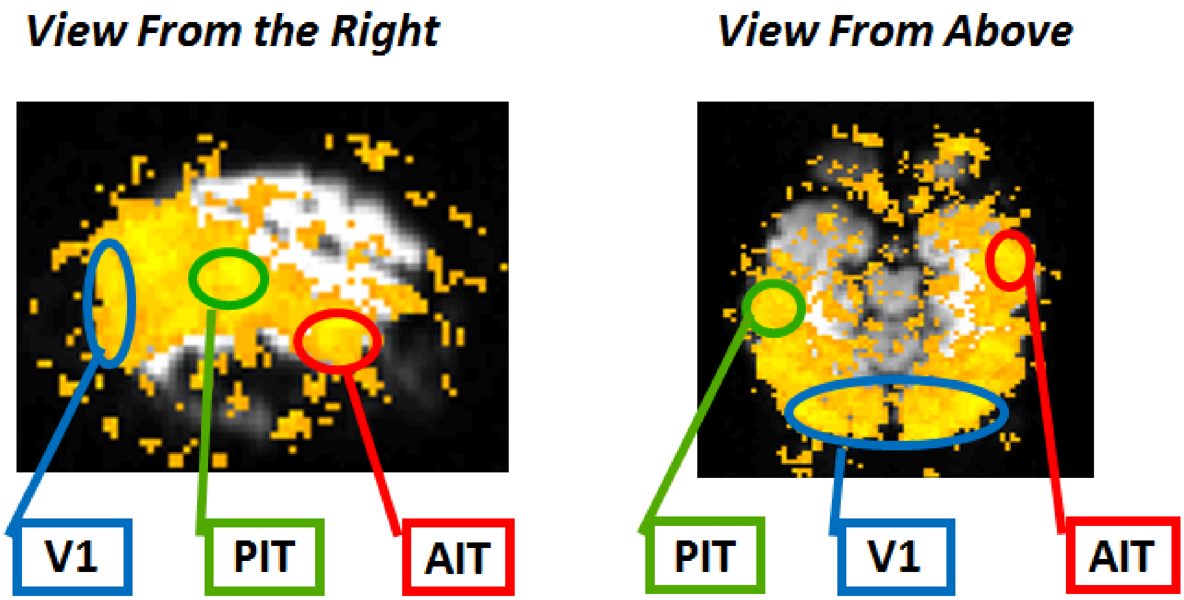


It is important to note that the “View From Above” is misleading. V1 (in blue) is shown on both sides of the brain, while PIT (in green) and AIT (in red) only appear on one side. This is because of the orientation of the brain relative to the fMRI scanner. The PIT and AIT do in fact exist on both sides of the brain, and if the monkey’s head were rotated slightly, we would see the second PIT and AIT regions appear on opposite sides of the brain.

For my first “searchlight” analysis, I examined the 3 month old brain using a threshold P-value of 0.05. The below figure shows in yellow which voxels passed the significance test. V1, PIT, and AIT are circled to help us to see whether voxels in those regions were activated by the stimuli.

Brain Activation from Classifier Searchlight Test

Threshold at P-value = 0.05



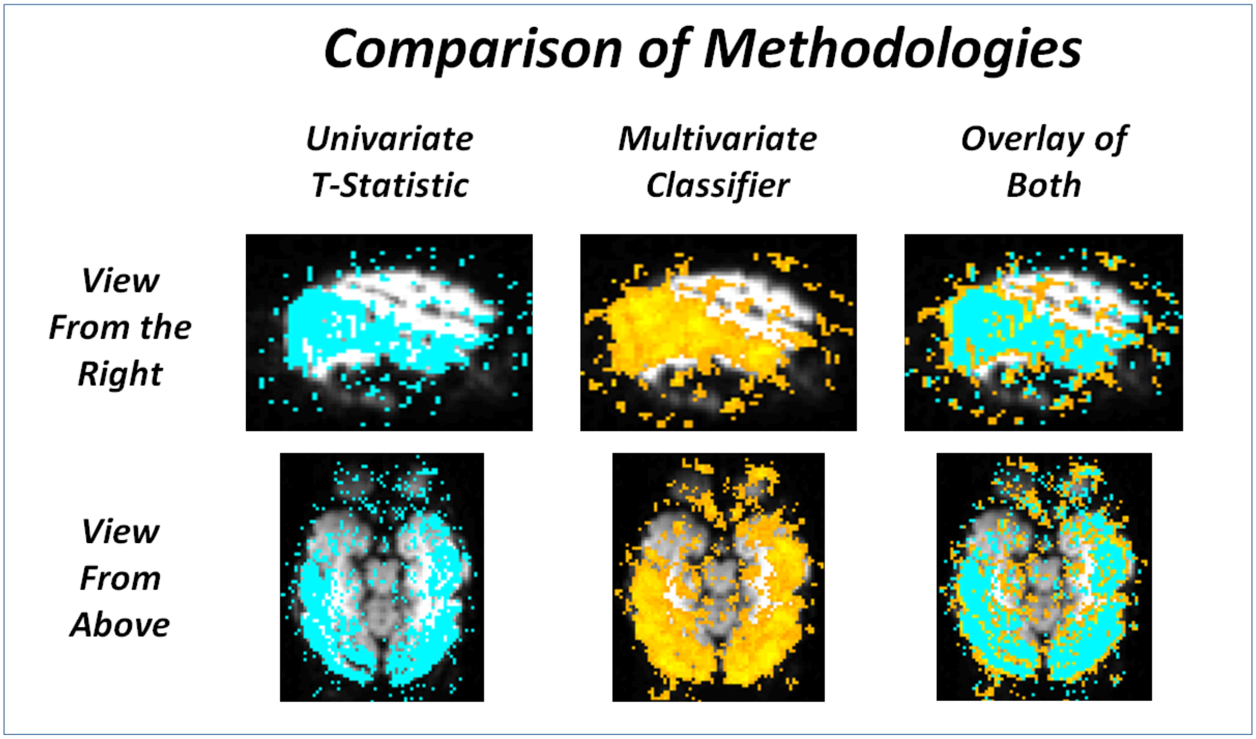
View From the Right

View From Above

Surprisingly, a great deal of the brain was activated even though the monkey was only 3 months old. Even the higher function AIT region appears to have been activated. This may come as a surprise to the neuroscientific community.

Having completed a searchlight analysis, I now had the opportunity to compare it with the results of a univariate analysis. This would address a major concern that I had coming into the project, which is that a classifier analysis might not be appropriate when there are so few observations and so many feature dimensions (each voxel represents a dimension). Classifiers perform best when there is a large pool of observations relative to feature dimensions, and each searchlight cube of voxels contains 27 feature dimensions and only 16 to 100 observations.

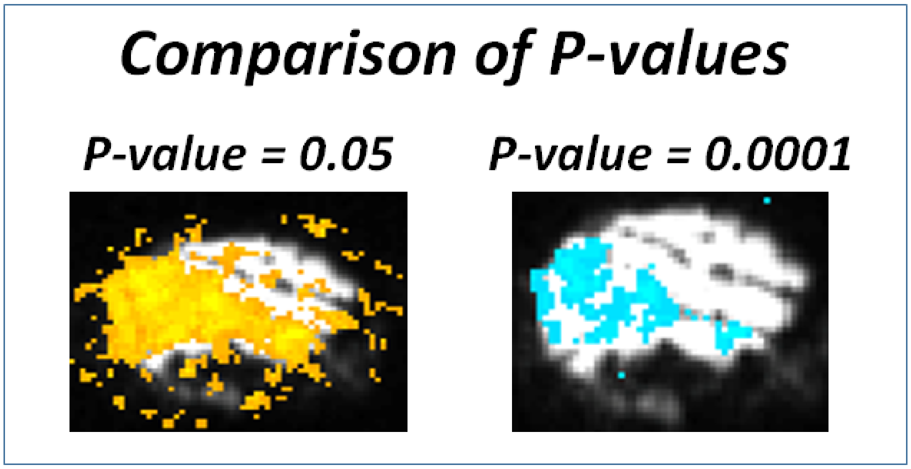
The below figure shows the results of both the univariate and multivariate classifier tests on the 3 month old brain. To do the univariate analysis, I calculated the t-statistic for each individual voxel, comparing its stimulus activations to its null activations. As with the classifier tests, I used a P-value of 0.05 to threshold the result of each voxel.



The images on the left display voxels in blue if they were activated according to the univariate t-statistic test. The images in the middle display voxels in yellow if they were activated according to the multivariate classifier. The images on the right display the result of both tests, with the univariate results overlaid on top of the multivariate classifier results. We see that there is a great deal of overlap, but that the classifier results are more expansive than the univariate results.

The previous two figures reveal another important phenomenon, which is the prevalence of false positives in the multivariate classifier data. We see that the yellow highlighting extends outside of the brain, which should not occur since this area is just the skull, hair, or open space. These false positives can be caused by one of two things. Either random noise in the fMRI data happened to correlate with the experiment's testing sequence, or fMRI readings are being influenced by the readings from nearby voxels. One way to reduce these false positives is by tightening up the P-value requirement beyond the 0.05 used here.

In order to test the benefit of tightening up the P-value, I did another searchlight analysis using a P-value of 0.0001. The figure to the left reveals that the more stringent P-value provides much less messy results, with all activation occurring within the brain. We should keep in mind that the tighter P-value reveals only those voxel regions that provide the most robust and consistent response to the visual stimuli. **That is why I performed the remainder of my analysis using a p-value of 0.0001.**

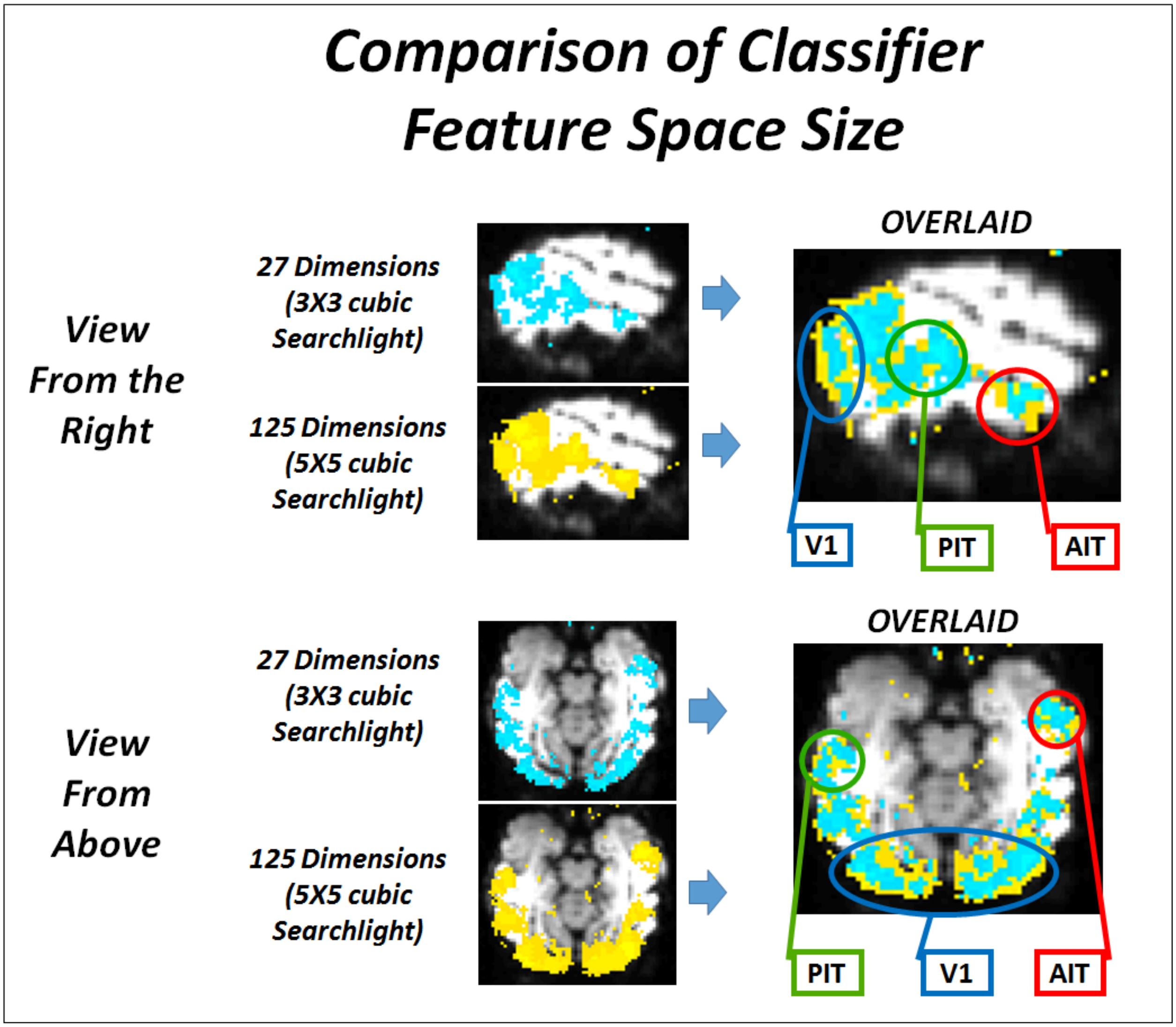


Comparison of P-values

P-value = 0.05

P-value = 0.0001

I next tested to see whether the classifier performance would improve or deteriorate if I enlarged the size of the searchlight cube to be 5x5x5. This would require the classifier to handle a feature space with 125 dimensions. Therefore, I initially expected the results to be poor. The figure below shows the result of the 3x3x3 searchlight in blue, and the 5x5x5 searchlight in yellow, both thresholded with a stringent P-value of 0.0001.

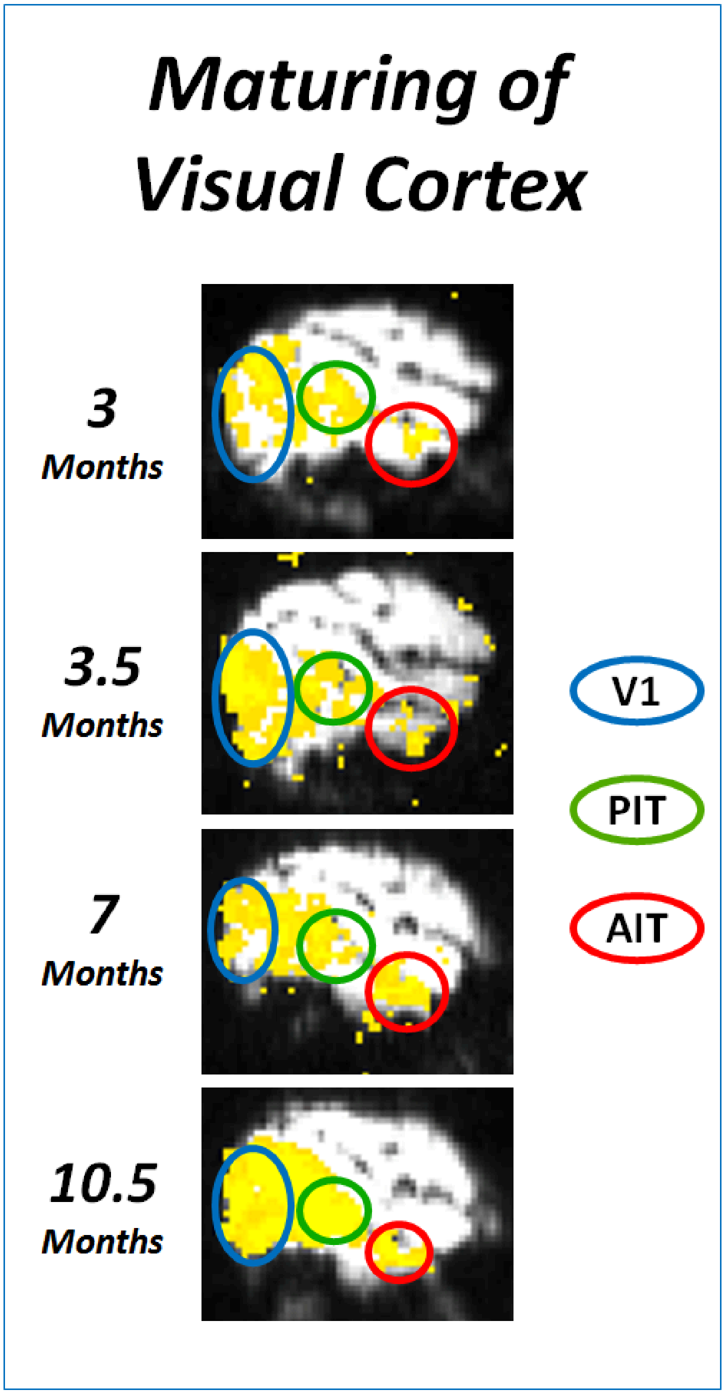


Comparison of Classifier Feature Space Size

View From the Right
27 Dimensions (3X3 cubic Searchlight)
125 Dimensions (5X5 cubic Searchlight)
OVERLAID
V1
PIT
AIT
View From Above
27 Dimensions (3X3 cubic Searchlight)
125 Dimensions (5X5 cubic Searchlight)
OVERLAID
PIT
V1
AIT

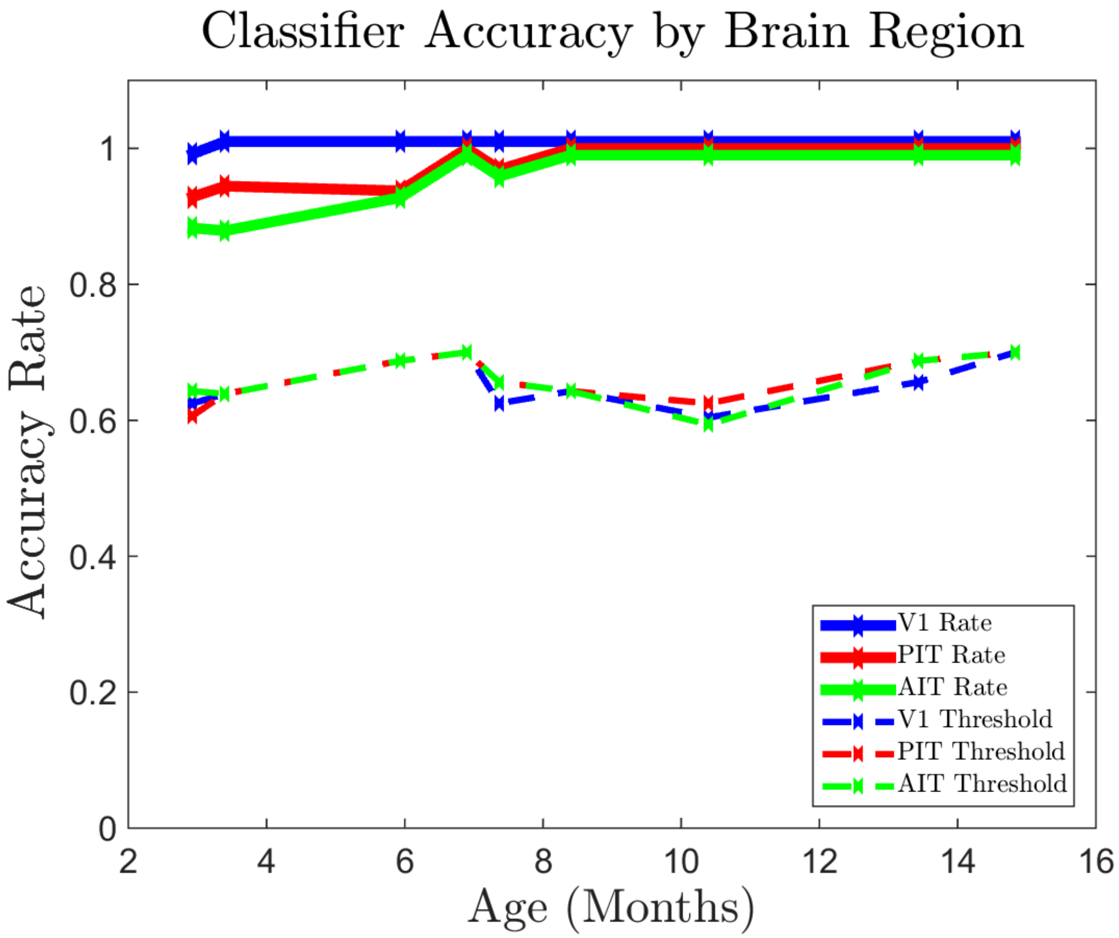
Surprisingly, we see that the larger searchlight size seems to have improved performance throughout the entire visual cortex, including V1, PIT, and AIT!

I next tested to see how the searchlight results progress over the timespan of the longitudinal study. The figure on the left shows a 3x3x3 voxel searchlight analysis with a P-value of 0.0001 for the brain at 4 timepoints in its cortical development. Voxels that pass the significance test are shown in yellow. We can see that the 3 month old brain shows activation throughout the visual cortex, but that activation is patchy. By the time the brain is 10.5 months old, the activation is filled out and uniform across the visual cortex. Note that the performance in V1 appears to change dramatically between 3 months and 3.5 months. It seems unlikely that V1 changed very much over 2 weeks. It would take further investigation to determine what has caused this abnormality in the results. Moving on, we can notice that both the 3 and 3.5 month scans show very little activation in AIT. But by 7 and 10.5 months, the AIT region appears to be very active.



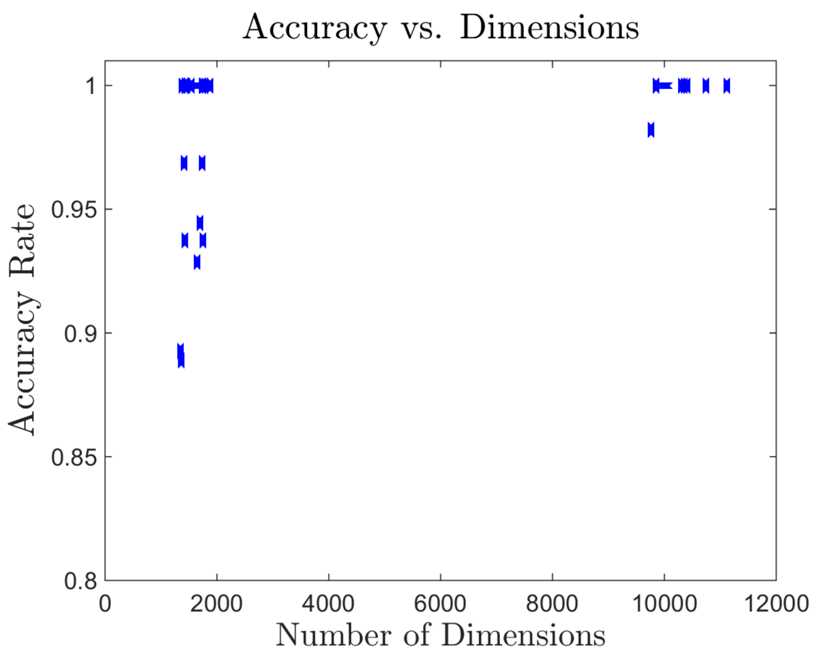
I next proceeded to the second phase of my analysis, in which I tested the classification performances of the entire V1, PIT, and AIT regions. Even though the searchlight analyses revealed activated voxels in all of these regions, it was still important to assess these regions as a whole, and to use permutation testing to decide whether they were truly activated. This may seem slightly redundant, but it can only serve to reinforce the results of the previous analysis.

For each of the three regions, I used a naive bayes classifier to determine a prediction accuracy rate. I then performed a permutation test to find the 95% threshold accuracy, and compared this to the actual measured accuracy. I repeated this process for all the 9 experiments conducted over the span of the longitudinal study. The results are shown below.



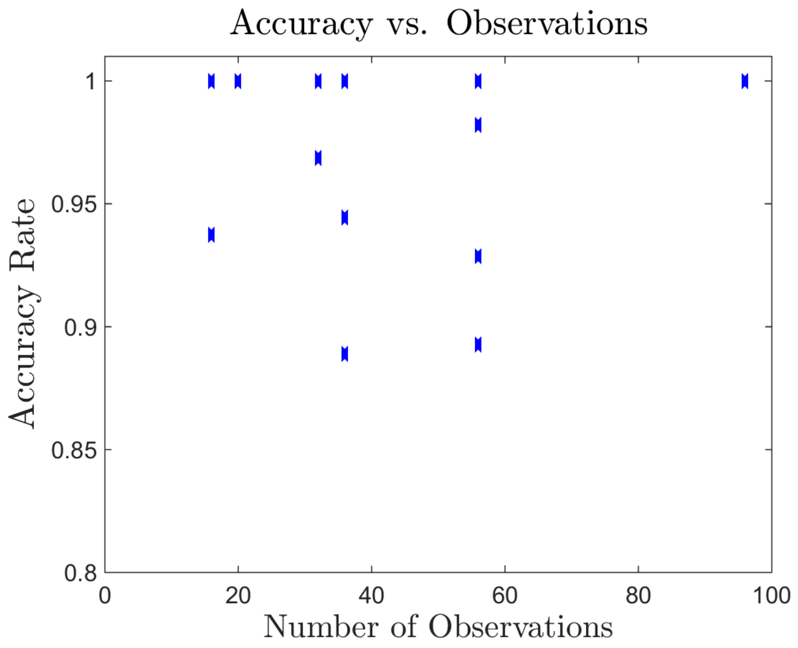
V1 is shown in blue, PIT in green, and AIT in red. The solid lines represent the actual accuracy rates, while the dashed lines represent the P=0.05 threshold accuracy rates. We see that at all ages, and for all regions, the measured accuracy rate is well above the significance threshold. V1 reaches 100% accuracy by 4 months of age, and PIT and AIT by 9 months. During the first 3 months of the monkey's life, PIT performs slightly better than AIT, and V1 performs the best.

I next explored the possibility that the results of the classifier tests were affected by the number of voxels contained within each region. Again, V1 had around 10K, PIT had around 1.7K, and AIT had around 1.4K. Furthermore, the monkey's brain grew slightly over the time range of the study, causing the number of voxels in each region to vary slightly over time. Below is a plot of the classification accuracy rate vs. the number of voxels fed into the classifier.



We see that the accuracy rate is generally higher for brain regions with more voxels. This is counterintuitive, since we expect a classifier to struggle with an excessive number of feature dimensions. What we see here might be explained by the fact that the data points on the right side of the graph all come from the large V1 region. Since the voxels in V1 are processing low level information, they are the most likely to provide a clear and robust response to a visual stimulus. This might have made it very easy for the classifiers to differentiate between stimulated and unstimulated observations.

In the last part of my analysis, I checked whether the regional accuracy rates were dependent upon the number of observations used to train the classifier.



I expected that an increase in observations would result in a better performing classifier. Again I was surprised, as there does not appear to be a correlation between the accuracy rate the number of observations.

Discussion

My analysis revealed that the monkey's visual cortex develops much earlier than previously thought. The regional analysis revealed that V1, PIT, and AIT were definitely responding to the visual stimuli at 3 months of age. This was especially unexpected for a high level region like AIT. It would be very helpful if the lab could perform similar experiments with even younger monkeys. Then we could test whether they are born with a responsive visual cortex. I would be interested to know whether my results for the 3 month old brain were caused by cortical development that occurred before or after birth.

However, both the searchlight and regional analyses demonstrated that the monkey's visual cortex was continuing to develop during the span of the study. In the searchlight maps, we can see voxel activation spreading in the PIT and AIT regions between the ages of 3 and 9 months. This is confirmed by the regional analysis which shows the PIT and AIT accuracies improve during this same 6 month period.

The analysis also demonstrated that naive bayes classifiers are effective tools for analyzing the brain activation in baby monkeys during visual experiments. There was no need to revert to the univariate statistical analysis. In fact, the multivariate searchlight analyses revealed a more widespread brain response than did the univariate. This difference is partially explained by the fact that the multivariate searchlight uses information contained in an entire neighborhood of voxels. We must keep in mind that some of the voxels on the outer edge of the searchlight map's highlighted region may not have been activated in reality, but showed up on the map anyways because they shared a neighborhood with more activated voxels. Unfortunately, non-activated edge voxels can make the activation region appear larger than it actually is on the searchlight map.

I was surprised to see that classifier performance did not correlate with the number of observations in the dataset, nor did it vary inversely with the size of the feature space. In general, classifiers are known to struggle when there are too many feature dimensions and too few observations. I expected that the sparsity of observations would result in highly biased classifiers that would make poor predictions. On the contrary, my classifiers seemed to perform equally well regardless of the number of observations, and actually performed best with V1 data, which contains huge numbers of feature dimensions. These results demonstrate the ability of naive bayes classifiers to handle a large feature space. Furthermore, they suggest that V1 has a much more differentiable response to visual stimuli than does PIT and AIT, especially during infancy. This makes sense, given that V1 processes visual information on a very low level.

In the future, it would be better to do the searchlight analysis with permutation testing, assuming that I gain access to the computation resources necessary to perform permutation testing with reasonable speed. It was not ideal to use the inverse binomial cumulative distribution equation to test for significance because I was not sure that the experiment's observations were truly independent. Because consecutive blocks are spaced so close together, the blood flow level in one block could have an effect on the blood flow level during the block. This may invalidate the binomial equation as a measure of statistical significance.

It might also be fruitful to do a more thorough analysis using the larger 5x5x5 searchlight size, as it seemed to perform even better than the 3x3x3 searchlight. After seeing the naive bayes classifiers perform so well in the regional analysis, I am no longer surprised that it easily handles the 125 feature dimensions of the larger searchlight. Furthermore, a larger searchlight might provide added sensitivity and capture inter-voxel activation patterns that exist on a larger scale.

A next step towards understanding the baby monkey's visual cortex would be to test whether a classifier can be trained to differentiate between different categories of visual stimuli. For example, we could use the data from the Livingstone Lab and test the categories of faces and objects. I expect that this would have a strong result in the PIT region at an early age. It would also be helpful to do a systematic comparison of the performance of a variety of classifier types, and to try a logarithmic regression.

Conclusion

The longitudinal study conducted by the Livingstone Laboratory has provided a novel glimpse into the cortical development of baby macaque monkeys. This analysis has revealed that the infant monkey has an active visual cortex at the young age of 3 months, and that the activation extends through to the higher level regions such as AIT. These results will serve to reinforce the results of the univariate analysis being done by the Livingstone Lab. Machine learning classifiers have proven to be a useful tool in the analysis of fMRI data.

References

1. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
2. Livingstone Laboratory, Harvard Medical School, Boston, MA.
3. Pereira, Francisco, Tom Mitchell, and Matthew Botvinick. "Machine learning classifiers and fMRI: a tutorial overview." *Neuroimage* 45.1 (2009): S199-S209.