CORE
Metadata, citation and similar papers at core.ac.uk
Provided by DSpace@MIT

Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2011-040
CBCL-302

September 10, 2011

Learning and disrupting invariance in visual recognition

Leyla Isik, Joel Z Leibo, and Tomaso Poggio

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

CSAIL

Learning and disrupting invariance in visual recognition

Leyla Isik[1,2], Joel Z. Leibo[1,2], and Tomaso Poggio[1]

[1]CBCL, McGovern Institute for Brain Research, Massachusetts Institute of Technology
[2]These authors contributed equally to this work

September 9, 2011

Abstract

Learning by temporal association rules such as Foldiak's trace rule [1] is an attractive hypothesis that explains the development of invariance in visual recognition. Consistent with these rules, several recent experiments have shown that invariance can be broken by appropriately altering the visual environment but found puzzling differences in the effects at the psychophysical [2, 3] versus single cell [4, 5] level. We show a) that associative learning provides appropriate invariance in models of object recognition inspired by Hubel and Wiesel [6], b) that we can replicate the "invariance disruption" experiments using these models with a temporal association learning rule to develop and maintain invariance, and c) that we can thereby explain the apparent discrepancies between psychophysics and singe cells effects. We argue that these models account for the stability of perceptual invariance despite the underlying plasticity of the system, the variability of the visual world and expected noise in the biological mechanisms.

1 Introduction

Temporal association methods are attractive solutions to the problem of invariance development or learning [1, 7, 8, 9, 10, 11, 12, 13, 14]. However, these algorithms have mainly been examined in idealized situations that do not contain the complexities present in the task of learning to see from natural vision or, when they do, ignore the imperfections of a biological learning mechanism. This paper presents a model of invariance learning that predicts the invariant object recognition performance of a neural population can be surprisingly robust, even in the face of frequent temporal association errors.

Experimental studies of temporal association and the acquisition of invariance involve putting observers in an altered visual environment where objects change their identity across saccades. Cox *et al.* showed that after a few days of exposure to this altered environment, the subjects mistook one object for another at a specific retinal position while preserving their ability to discriminate the same objects at other positions [2]. A subsequent physiology experiment by Li and DiCarlo using a similar paradigm showed that individual neurons in primate anterior inferotemporal cortex (AIT) change their selectivity in a position-dependent manner after less than an hour of exposure to the altered visual environment [4]. It is important to note that the stimuli used in the Cox et al. experiment were difficult to discriminate "greeble" objects, while the stimuli used by Li and DiCarlo were extremely easy to discriminate e.g., a teacup versus a sailboat.

This presents a puzzle, if the cells in AIT are really underlying the discrimination task, and exposure to the altered visual environment causes strong neural effects so quickly, then why is it that behavioral effects do not arise until much later? The fact that the neural effects were observed with highly dissimilar objects (the equivalent of an easy discrimination task) while the behavioral effects in the human experiment were only observed with a difficult discrimination task compounds this puzzle.

The physiology experiment did not include a behavioral readout, so the effects of the manipulation on the monkey's perceptual performance is not currently known; however, the human evidence suggests it is highly unlikely that the monkey would really be perceptually confused between teacups and sailboats after such a short exposure to the altered visual environment.

In this paper, we present a computational model of invariance learning that shows how strong effects at the single cell level, like those observed in Li et al.'s experiments do not necessarily cause confusion on the neural population level, and hence do not imply perceptual effects. Our simulations show that a population of cells is surprisingly robust to large numbers of mis-wirings due to errors of temporal association. In accord with the psychophysics literature [2, 3], our model also predicts that the difficulty of the discrimination task is the primary determiner of the amount of exposure necessary to observe a behavioral effect rather than the strength of the neural effect on individual cells.

2 The cortical model

We examine temporal feature learning with a class of cortical models inspired by Hubel and Wiesel's discoveries of cells in the visual cortex [6]. These models contain alternating layers of simple S cells or feature detectors to build specificity, and complex C cells that pool over simple cells to build invariance. [15, 16, 17]. We will focus on one particular such model, HMAX [17]. The differences between these models are likely irrelevant to the issue we are studying, and thus our results will generalize to other models in this class.

3 Temporal association learning with the cortical model

Temporal association learning rules provide a plausible way to learn transformation invariance through natural visual experience [1, 7, 8, 9, 10, 11, 12, 13, 14]. Objects typically move in and out of our visual field much slower than they transform, and based on this difference in time scale we can group the same object under different transformations.

We learn this translation invariance from a series of continuously translated images. We use a training phase to learn S2 to C2 connections from a continuously translating image sequence, as shown in Figure 1, left. A video sequence is presented, and over a certain temporal period and highly active S2 features/template locations are grouped so they all pool to a C2 cell. One C2 cell is learned during each "association period" or image sweep across a screen, and correct temporal association should group similar features across spatial locations, as illustrated in Figure 1, left. The C2 cell then pools over the learned features. Potential effects of a temporally altered image sequence are illustrated in Figure 1, right.

3.1 Learning rule

In Foldiak's original trace rule, shown in Equation 1, the weight of a synapse w_{ij} between an input cell x_j and output cell y_i is strengthened proportionally to the input activity and the trace or average of recent output activity at time t. The dependence of the trace on previous activity decays over time with the δ term [1].

Foldiak trace rule:

$$\begin{aligned} \Delta w_{ij}^{(t)} &\propto x_j \bar{y}_i^{(t)} \\ \bar{y}_i^{(t)} &= (1-\delta) y_i^{(t-1)} + \delta y_i^{(t)} \end{aligned} \tag{1}$$

In the HMAX model, connections between S and C cells are binary. Additionally, in our training case we want to learn connections based on image sequences of a known length, and thus for simplicity should include a hard time window rather than a decaying time dependence. Thus we employed a modified trace rule that is appropriate for learning S2 to C2 connections in the HMAX model.

Modified trace rule for the HMAX model:

$$\begin{aligned} &\text{for } t \text{ in } \tau : \\ &\qquad\qquad \text{if } x_j > \theta,\ w_{ij} = 1 \\ &\qquad\qquad \text{else},\ w_{ij} = 0 \end{aligned} \tag{2}$$

With this learning rule, one C2 cell is produced for each association period. The length of the association period is τ.

4 Experiments

4.1 Training for translation invariance

We model natural invariance learning with a training phase where the model learns to group different representations of a given object based on the learning rule in Equation 2. Through the learning rule, the model groups continuously-translating images that move across the field of view over each known association period τ. An example of a translating image sequence is shown at the top, left of Figure 1. During this training phase, the model learns the domain of pooling for each C2 cell.

4.2 Accuracy of temporal association learning

To test the performance of the HMAX model with the learning rule in Equation 2, we train the model with a sequence of training images. Next we compare the learned model's performance to that of the hard-wired HMAX [17] on a translation-invariance recognition task. In standard implementations of the HMAX model model, the S2 to C2 connections are hard-wired, each C2 cell pools all the S2 responses for a given template globally over all spatial locations. This pooling gives the model translation invariance and mimics the outcome of an idealized temporal association process.

The task is a 20 face and 20 car identification task, where the images in a given class are similar (but not identical) for different translated views[1]. We collect hard-wired C2 units and C2 units learned from temporal sequences of the faces and cars. We then used a nearest neighbor classifier to compare the correlation of C2 responses for translated objects to those in a given reference position. The accuracy of the two methods (hard-wired and learned from test images) versus translation is shown in Figure 2. The two methods performed equally well. This confirms that the temporal associations learned from training yield accurate invariance results.

4.3 Manipulating the translation invariance of a single cell

To model the Li and DiCarlo physiology experiments in [4] we perform normal temporal association learning described by Equation 2 with a translating image of one face and one car. The S2 units are tuned to the same face and car images (see Figure 1 caption) to mimic object-specific cells that are found in AIT. Next we select a "swap position" and perform altered training with the face and car images swapped only at that position (see Figure 1, top right). After the altered training, we observe the C2 response (of one C2 cell) to the two objects at the swap position and another non-swap position in the visual field that was unaltered during training.

[1] The training and testing datasets come from a concatenation of two datasets from: http://www.d2.mpi-inf.mpg.de/Datasets/ETH80, and http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

As shown in Figure 4.3 the C2 response for the preferred object at the swap position (but not the non-swap position) is lower after training, and the C2 response to the non-preferred object is higher at the swap position. As in the physiology experiments performed by Li and DiCarlo, these results are object and position specific.

4.4 Individual cell versus population response

Li and DiCarlo showed, and this simulation reproduces the results, that translation invariant representations of objects can be disrupted by a relatively small amount of altered temporal association learning. However, single cell changes do not necessarily reflect whole population or perceptual behavior. No behavioral tests were performed on the animals in this study. Additionally, it is difficult to perform population decoding for object preference, because while one can find single AIT cells that preferentially fire for a certain object, it is unlikely that multiple cells in the same recording area will have a strong preference for a single object over others.

A cortical model with a temporal association learning rule provides a way to model population behavior with swap exposures similar to the ones used by Li and DiCarlo [4, 5]. A C2 cell in the HMAX model can be treated as analogous to an AIT cell (as tested by Li and DiCarlo), and a C2 vector as a population of these cells. We can thus apply a classifier to this cell population to obtain a model of behavior or perception.

4.5 Robustness of temporal association learning with a population of cells

We next model the response of a population of cells to different amounts of swap exposure. The temporal training sequence with which we train the model replicates visual experience, and thus jumbling varying amounts of these training images is analogous to presenting different amounts of altered exposure to a test subject as in [4, 5]. These disruptions also model the mis-associations that may occur with temporal association learning due to sudden changes in the visual field (such as light, occlusions, etc), or other imperfections of the biological learning mechanism. During the training phase we randomly swap face and car images in the image sequences, and observe the effect on the response of a classifier to a population of C2 cells. The accuracy versus different neural population sizes (number of C2 cells) is shown in Figure 4 for various amounts of altered exposure. We measured altered exposure by the probability of flipping a face and car image.

A small amount of exposure to altered temporal training (0.125 probability of flipping face and car) has negligible effects, and the model under this altered training performs as well as with normal temporal training. A larger amount of exposure to altered temporal training (0.25 image flip probability) is not significantly different than perfect temporal training, especially if the neural population is large enough. With enough C2 cells, each of which is learned from a temporal training sequence, the effects of small amounts of jumbling in training images are insignificant. Even with half altered exposure (0.5 image flip probability), if there are enough C2 cells then classification performance is still reasonable. This is likely because with similar training (multiple translating faces or cars), redundant C2 cells are formed, creating robustness to association errors that occurred during altered training. Similar redundancies are likely to occur in natural vision. This indicates that in natural learning mis-wirings do not have a strong effect on learning translation invariance, particularly with familiar objects or tasks.

4.6 Task complexity versus exposure to altered training

Finally, we aim to reconcile the contradiction posed by the long exposure time needed to alter behavioral effects in a difficult psychophysics task [2, 3] with the relatively short exposure time needed to alter physiology readouts in an much easier discrimination task [4, 5]. We investigate the effect of task complexity on temporal association learning by using a fixed neural population size (number of C2 units) for face recognition tasks of varying difficulty. We can systematically increase task difficulty by adding additional examples of each type of face (again with slight differences such as expression) to the test set. These tasks become more difficult, because they require additional generalization. We measure accuracy on the identification task versus amount of altered training exposure for different complexity tasks: easy (one instance of each face, and thus no generalization needed), medium (two instances of each face, slight generalization needed), hard (ten instances of each face, wide generalization needed). The results are shown in Figure 5.

This simulation shows that although the relative change in performance for each task is roughly the same, increasing task difficulty decreases initial performance thus making small effects in performance more noticeable. These results confirm that the ability to alter invariance learning in psychophysics tasks [2, 3] is not due to significant wiring changes, but rather an overall decrease in performance to near threshold levels due to task difficulty.

5 Discussion

We use a cortical model inspired by Hubel and Wiesel [6], where translation invariance is learned through a variation of Foldiak's trace rule [1] to model the visual response to altered temporal exposure. We first show that this temporal association learning rule is accurate by comparing its performance to that of a similar model with hard-wired translation invariance [17]. This extends previous modeling results by Masquelier *et al.* [9] for models of V1 to higher levels in the visual recognition architecture. Next, we test the robustness of translation invariance learning on single cell and whole population responses. We show that although it is fairly easy to disrupt translation invariance in a single cell with altered temporal training, the whole population response is much more robust. This resolves a puzzling discrepancy between experiments that show how invariance can be altered [2, 4, 5, 3]. Through the use of a cortical model with an appropriate temporal association learning rule, we are able to directly compare the implications of these experimental results and demonstrate that learned errors in translation invariance were likely due to factors besides miswiring. This study shows that despite unavoidable disruptions, models based on temporal association learning are quite robust and therefore provide a promising solution for learning invariance from natural vision.

6 Acknowledgements

This work was supported by the following grants: NSF-0640097, NSF-0827427, NSF-0645960, DARPA-DSO, AFSOR FA8650-50-C-7262, AFSOR FA9550-09-1-0606

References

[1] Foldiak, P. (1991) Learning Invariance from Transformation Sequences. *Neural Computation* 3.2: 194-200.

[2] Cox, D., Meier P., Oertelt N. and DiCarlo J.J. (2005) 'Breaking' Position-invariant Object Recognition. *Nature Neuroscience* 8.9:1145-147.

[3] Wallis, G. and Bulthoff H. (2001) Effects of Temporal Association on Recognition Memory. *PNAS* 98.8:4800-804.

[4] Li, N. and DiCarlo, J.J. (2008) Unsupervised Natural Experience Rapidly Alters Invariant Object Representation in Visual Cortex. *Science* 321.5895:1502-507.

[5] Li, N. and DiCarlo, J.J. (2010) Unsupervised Natural Visual Experience Rapidly Reshapes Size-Invariant Object Representation in Inferior Temporal Cortex. *Neuron* 67.6:1062-075.

[6] Hubel, D.H. and Wiesel, T.N. (1962) Receptive Fields, Binocular Interaction and Functional Architecture in the Cats Visual Cortex. *J. Physiology* 160:106-154.

[7] Einhauser, W., Hipp, J., Eggert, J., Korner, E. and Konig P. (2005) Learning viewpoint invariant object representations using a temporal coherence principle. *Biological Cybernetics* 93:79-90.

[8] Franzius, M., Sprekeler, H., and Wiskott, L. (2007)Slowness and Sparseness Lead to Place, Head-Direction, and Spatial-View Cells. *PLoS Comput Biol* 3.8:e166.

[9] Masquelier, T., Serre, T., Thorpe S. and Poggio T. (2007) Learning complex cell invariance from natural videos: a plausible proof. MIT-CSAIL-TR-2007-060.

[10] Masquelier, T. and Thorpe, S.J. (2007) Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity. *PLoS Computational Biology* 3.2:e31.

[11] Spratling, M. (2005) Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27.5:753-761.

[12] Wallis, G. and Rolls. E.T. Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51:167194.

[13] Wiskott, L. and Sejnowski T.J. (2002) Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Computation* 14.4:715-70.

[14] Wyss, R., Konig, P., and Verschur, P. (2006) A Model of the Ventral Visual System Based on Temporal Stability and Local Memory. *PLoS Biol* 4.5:e120.

[15] Fukushima, K. (1980) Neocognitron: A Self Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics* 36: 193-201.

[16] Riesenhuber, M. and Poggio, T. (1999) Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience* 2.11:1019-1025.

[17] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. and Poggio T. (2007) Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.3:411-26.

[18] Leibo, J.Z., Mutch, J., Rosasco, L., Ullman, S. and Poggio, T. (2010) Learning Generic Invariances in Object Recognition: Translation and Scale. MIT-CSAIL-TR-2010-061.

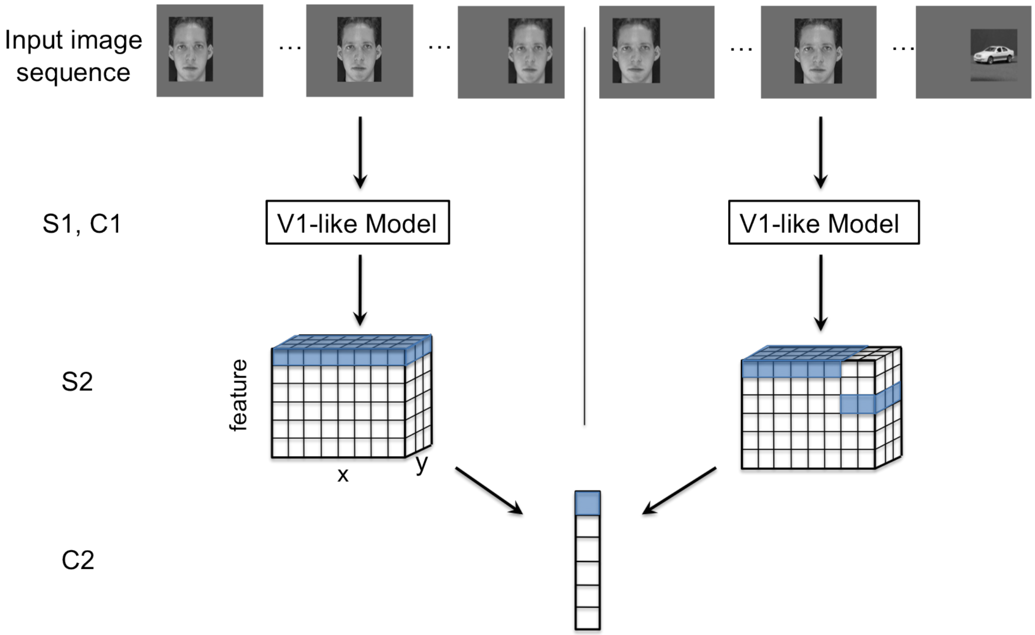


Figure 1: An illustration of the HMAX model with two different input image sequences: a normal translating image sequence (left), and an altered temporal image sequence (right). The model consists of four layers of alternating simple and complex cells. **S1 and C1 (V1-like model)**: The first two model layers make up a V1-like model that mimics simple and complex cells in the primary visual cortex. The first simple cell layer, S1, consists of simple orientation-tuned Gabor filters, and the following complex cell layer, C1, performs max pooling over local regions of a given S1 feature. **S2**: The next simple cell layer, S2, performs template matching between C1 responses from an input image and the C1 responses of stored prototypes (unless otherwise noted, we use prototypes that were tuned to natural image patches). Template matching is performed with a radial basis function, where the responses have a Gaussian-like dependence on the Euclidean distance between the (C1) neural representation of an input image patch and a stored prototype. The RBF response to each template is calculated at various spatial locations for the image (with half overlap). Thus the S2 response to one image (or image sequence) has three dimensions: x and y corresponding to the original image dimensions, and feature the response to each template. **C2**: The final complex cell layer, C2, performs global max pooling over all the S2 units to which it is connected. The S2 to C2 connections are highlighted for both the normal (left) and altered (right) image sequences. To achieve ideal transformation invariance, the C2 cell can pool over all positions for a given feature as shown with the highlighted cells.

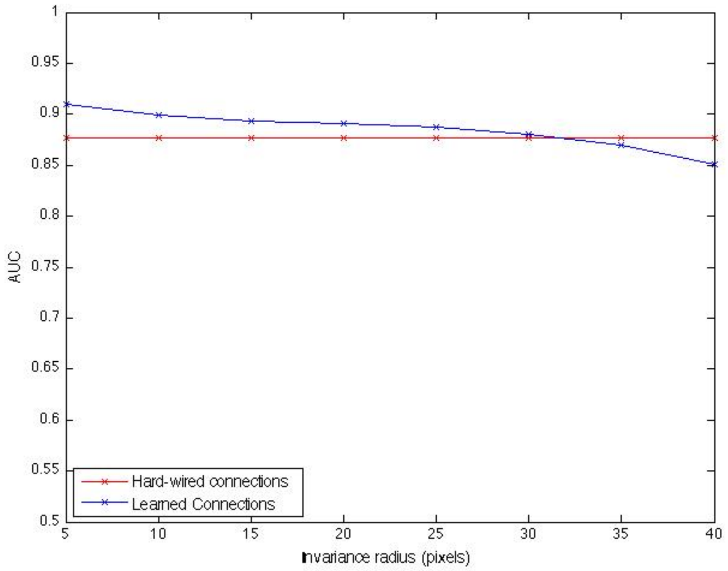
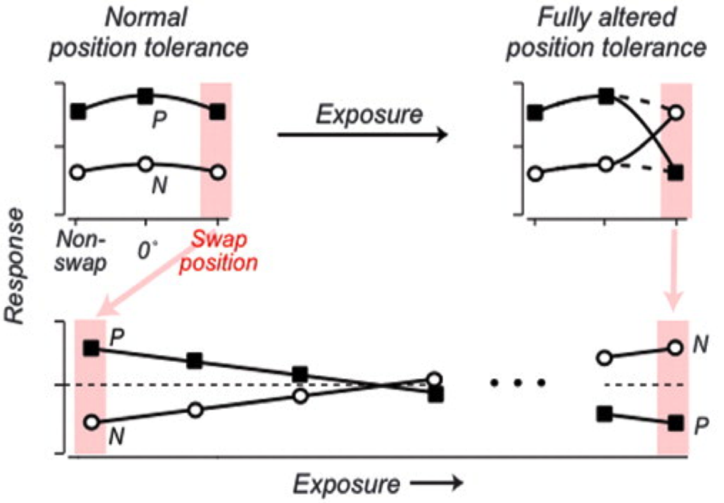
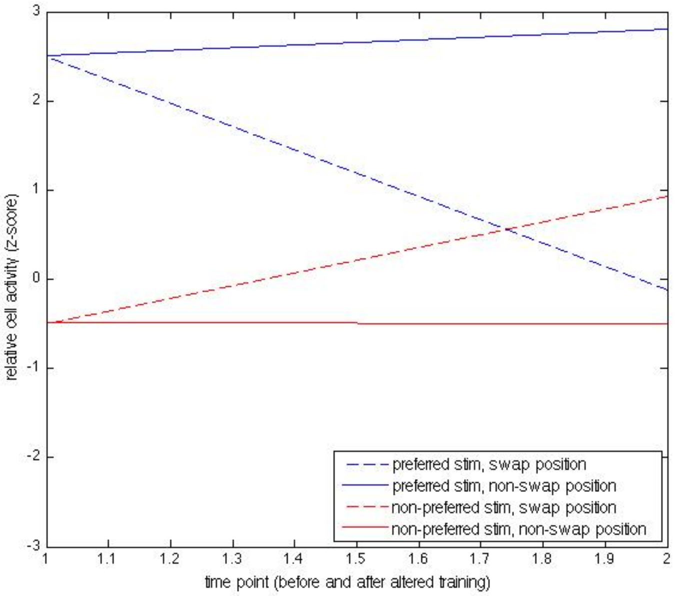


Figure 2: The classification accuracy (AUC for ROC curve) for both hard-wired and temporal association learning model plotted for different degrees of translation compared to a reference position with a nearest neighbor classifier. The model was trained and tested on separate training and testing sets, each with 20 car and 20 face images. For temporal association learning, one C2 unit is learned for each association period or training image, yielding 40 learned C2 units. One hard-wired C2 unit was learned from each natural image that cells were tuned to, yielding 10 hard wired C2 units. Increasing the number of hard-wired features has only a marginal effect on classification accuracy. For temporal association learning, the association period *tau* was set to the length of each image sequence (12 frames), and the activation threshold *theta* was empirically set to 3.9 standard deviations above the mean activation.



(a) Figure from Li and DiCarlo 2008 [4] summarizing the expected results of swap exposure on a single cell. P is response to preferred stimulus, and N is that to non-preferred stimulus.



(b) The response of a C2 cell tuned to a preferred object before (time point 1) and after (time point 2) altered visual training where the preferred and non-preferred objects were swapped at a given position. To model the experimental paradigm used in [4, 5, 2, 3], training and testing were performed on the same altered image sequence. The C2 cell's relative response (Z-score) to the preferred and non-preferred objects at both the swap and non-swap positions are plotted.

Figure 3: Manipulating single cell translation invariance through altered visual experience.

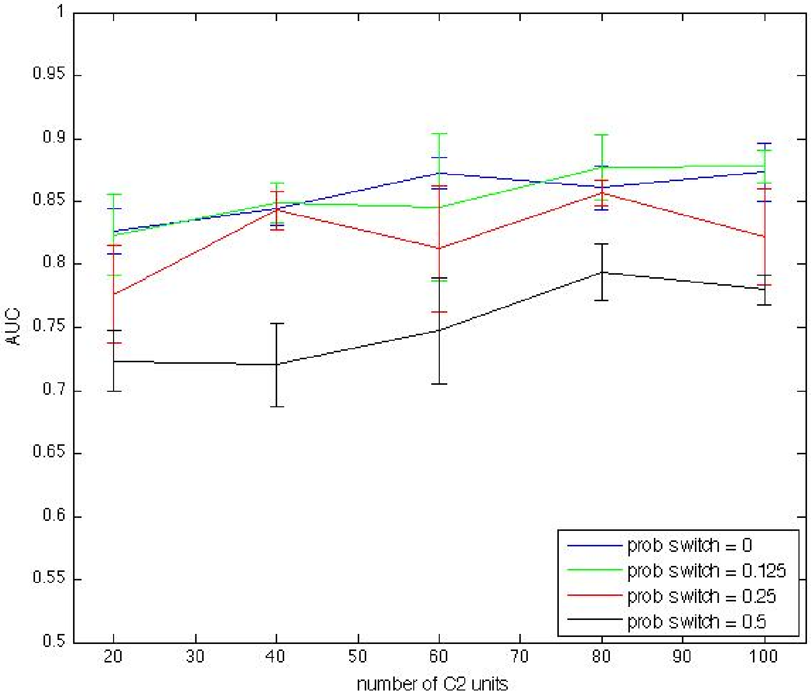


Figure 4: Results of a translation invariance task (+/- 40 pixels) with varying amounts of altered visual experience. To model the experimental paradigm used in [4, 5, 2, 3], training and testing were performed on the same altered image sequence. The accuracy (AUC for ROC curve) with a nearest neighbor classifier compared to center face for a translation invariance task versus the number of C2 units. Different curves have a different amount of exposure to altered visual training as measured by the probability of swapping a car and face image in training. The error bars show +/- one standard deviation.

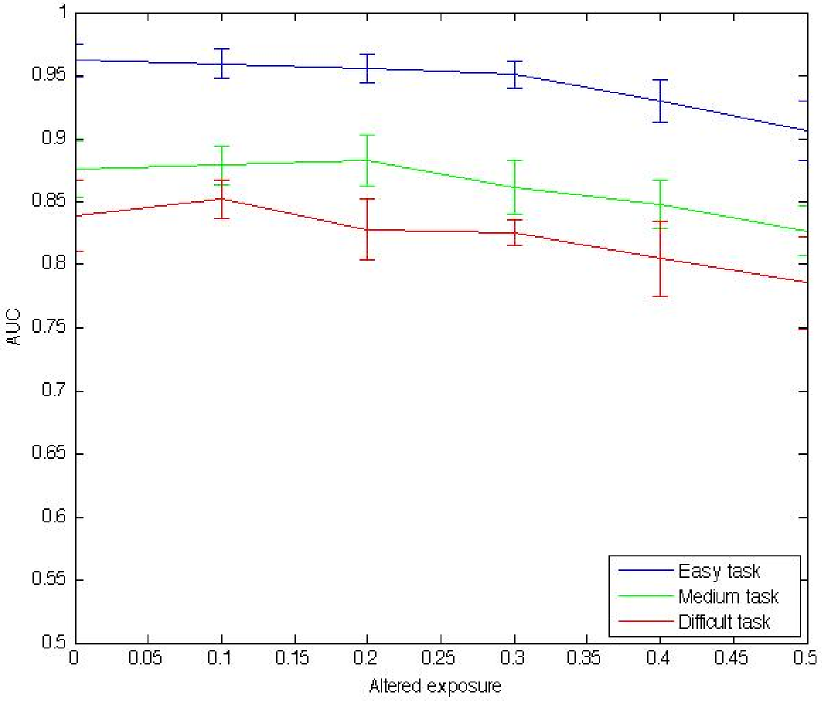


Figure 5: Results of a translation invariance task (+/- 20 pixels) for tasks of varying difficulty. The accuracy (AUC for ROC curve) with a nearest neighbor classifier compared to center face for a translation invariance task versus the amount of altered training (measured by probability of flipping two training images). The easy task includes one instance of each face, medium includes two instances of each face, and difficult includes 10 instances of each face. The error bars show +/- one standard deviation.