A MECHANISM OF VISUAL CONCEPT LEARNING VIA NEURONAL RANDOM PROJECTION

A Thesis
Presented to
The Academic Faculty

By

Peter S. Koplik

In Partial Fulfillment
of the Requirements for the Research Option in the
School of Computer Science

Georgia Institute of Technology

October 2018

A MECHANISM OF VISUAL CONCEPT LEARNING VIA NEURONAL RANDOM PROJECTION

Approved by:

Dr. Rosa Arriaga
School of Computer Science
Georgia Institute of Technology

Dr. Santosh Vempala
School of Computer Science
Georgia Institute of Technology

Date Approved: October 23, 2018

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

We explore the effectiveness of the random projection method, a biologically plausible, computationally efficient, and data-independent method of dimensionality reduction in distinction between categories of visual stimuli. We observe that a neural network tasked with approximating the original stimulus from the reduced domain generally excludes information not useful in distinguishing visual categories. This suggests that random projection may be useful in the efficient recall and recognition of visual concepts even though the projections only contain small fractions of the original information. Our findings indicate that the reconstruction of visual stimuli from the random projected domain preserves best the features most typical of that particular category of stimuli.

CHAPTER 1
INTRODUCTION

1.1 Introduction

Humans are able to organize massive amounts of data into concepts that they can later recognize and recall. Infants can form categorical representations of dogs and cats, for example, from just a few exposures to images of these animals [7]. In order to process massive amounts of information from visual stimuli as quickly as they do, humans must process the stimuli in a manner that reduces the complexity of those stimuli. Previous work in the field of neuroscience has developed models for the processing of high-dimensional natural stimuli through low-dimensional representations and corresponding neural responses. However, these methods are typically computationally expensive and data-dependent [1].

The Johnson-Lindenstrauss lemma [2] is the basis for a method of dimensionality reduction called random projection, which projects points from a high-dimensional vector space into a random, low-dimensional, orthogonal vector space. It does this by drawing an k-by-n matrix A from a Gaussian distribution. This matrix is then multiplied with all samples x, of the form Ax = b. b is then a low-dimensional representation of the original samples. The lemma guarantees that the relative distances between the points is approximately preserved during this process. Furthermore, the random projection method is completely independent of the structure of the data being projected.

Arriaga and Vempala proposed random projection as a means by which *robust* concept classes can be learned efficiently and from few examples with a biologically plausible neuronal mechanism called neuronal random projection [3]. This is simply a version of the random projection method that can be implemented with a neural network layer of static weights. The existence of a simple neuronal structure that can perform random projection on an input suggests that the method could be a biologically plausible means of the reduction of dimensionality of stimuli. The biological plausibility of this method has been previously explored. [5].

1.2 Sample Robustness

A robust concept is one that is immune to attribute noise to some extent. Arriaga and Vempala define a concept *C* as ℓ-robust over some distribution $\mathcal{D}$ if

$$Pr_{\mathcal{D}}[x \,|\, \exists y \,:\, label(x) \neq label(y), \|x - y\| \leq \ell] = 0 \tag{1.1}$$

If we consider the visual domain to be an array of pixels of varying intensities, concepts of natural visual stimuli do not strictly adhere to this definition, as there may be no real, positive ℓ for which this definition applies. For example, it is possible that a picture of a donkey and a picture of a horse are nearly or completely indistinguishable in the visual domain. To quantify this, Arriaga et al. introduced the notion of *sample robustness* [4]. Given two categories A and B, the sample robustness of a particular sample x with respect to *A* and *B* is defined as

$$R_{AB}(x) = \frac{\|x - \mu_A\|}{\|x - \mu_B\|} - \frac{\|x - \mu_B\|}{\|x - \mu_A\|} \tag{1.2}$$

where μ_A and μ_B are the means of categories A and B, respectively. This metric tends towards $+\infty$ when the sample x is closer to μ_B, and towards $-\infty$ when the sample x is closer to μ_A. Euclidean and Manhattan distance are both acceptable metrics to use. This paper uses Euclidean distance throughout.

1.3 Previous Human Experiment

Arriaga et. al (2015) explored the binary categorization performance of humans and simple neural networks on novel, artificially generated stimuli with respect to sample robustness. They found that both humans and neural networks were more likely to correctly classify samples with greater $|R_{AB}(x)|$, both before and after the application of visual structure preserving random projection methods. Additionally, the performance of humans in classification did not significantly degrade after random projection [4].

The methods chosen for the Arriaga et. al (2015) experiment were not random projection in the literature sense as described above. In order to present the reduced stimuli to humans, the projections maintained the visual structure of the image and did not simply vectorize its pixel intensity values. For example, one method that was used performed random projection on a sliding window across the image to generate macro-pixels that were a random combination of colors in the window.

This suggests that random projection could be responsible for efficient categorization of visual stimuli by humans, but does not suggest anything about the influence of random projection on visual concept learning, or how humans cognitively perform that categorization.

1.4 Stimulus Reconstruction Experiment

The effectiveness of a simple neural network in the task of approximating the original signal over a dataset of natural visual signals after random projection was explored. For samples strongly belonging to a particular category, the robustness of those samples generally increased when reconstructed from the reduced domain by the neural network. This may suggest that random projection enables computationally efficient learning of natural visual concepts by reducing the complexity of the visual stimulus. By applying random projection, the distinction between two categories of visual stimuli is efficiently represented in a low-dimensional neuronal encoding and the neural network's weights by minimization of the error in the reconstruction of those stimuli.

CHAPTER 2
LITERATURE REVIEW

2.1 Neuronal Random Projection

In 1999, Arriaga and Vempala proposed random projection as a biologically plausible method for dimensionality reduction [3]. This method takes a set of high-dimensional data, and embeds it into a low-dimensional vector space. It does this by drawing an k-by-n matrix A from a Gaussian distribution. This matrix is then multiplied with all samples x, of the form Ax = b. b is then a low-dimensional representation of the original samples. Johnson and Lindenstrauss proved that random projection preserves the Euclidean distances between all pairs of points in x within some error bound [2]. In other words, the random projection of a set of data is a noisy sketch of the original set, but is represented with far less information. Arriaga and Vempala additionally proposed a method called neuronal random projection. This is simply a version of the random projection method that can be easily implemented with a neural network of static weights. The existence of a simple neuronal structure that can perform random projection on an input indicates that the method could be a biologically plausible means of the reduction of dimensionality of stimuli. Allen-Zhu, Gelashvili, Micali, and Shavit, for example, explored a sparse and sign-consistent version of the Johnson-Lindenstrauss transform, with the motivation that sparsity is necessitated by the large majority of real neurons being inhibitory, and sign-consistency is necessitated by the existence of a minimum neuronal activation potential [5]. Much more analysis has been performed on variations of this method that better lend themselves to biological plausibility.

2.2 Relation to Compressed Sensing

This work is related to the well-explored method of *compressed sensing*, described by Olshausen and Field [6]. They posit that the compressed sensing method shares some important properties with certain groups of visual receptive neurons. In this method, a set of images is reduced into a sparse dictionary which represents a set of distinct, identifying visual features. They are found by attempting to reconstruct the images from the set of features with the constraints that information is preserved, and the coefficient matrix multiplied by the feature dictionary for a given image is as sparse as possible. Essentially, the compressed sensing method represents a set of images with individual features that span the image space but are combined to recreate the image with only a fraction of those features. This work demonstrates that the result of random projection is a partially sufficient dictionary for recreation of prototypical samples, without actually considering the distribution or sparseness of features of the stimuli. In this sense, random projection is very computationally efficient and data-independent. While it does not necessarily meet the hypothesized sparsity criteria that arises from Olshausen and Field's analysis of the mammalian visual system [17], it does take some advantage of the redundancy of natural visual scenes by preserving inter-category variance across distinct visual categories by projecting onto non-redundant dimensions.

2.3 Human Categorization of Stimuli After RP-based Methods

In 2015, Arriaga et al [4] performed an experiment on human subjects relating to random projection. The participants were shown a sample image from two distinct categories. These categories were novel and natural, meaning that they contained features consistent with real objects such as lines and contiguous groups of similarly colored pixels, but they were not anything the participant could have seen before. They were artificially generated with this goal in mind. After presentation of one sample from each category, the participant was then asked to categorize unlabeled examples from the same two categories. The experiment was also performed after the application of two random projection-based methods: corner projections, which randomly scale and average the pixel values surrounding all detected corners in the image, and sliding-window projections, which slides across the image and generates a mega-pixel from a random combination of the colors in the window. Once the human results were collected, a neural network was trained to perform the same task. Arriaga et al found that classification performance did not degrade with the use of these information removing random projection-based methods. Furthermore, human classification performance almost exactly mirrored the performance of neural network on the same samples. Lastly, and most importantly, the performance of both humans and neural networks had a strong positive correlation with $|R_{AB}(x)|$, defined in equation (2). Samples with greater absolute value of sample robustness strongly belonged to one of the categories more than the other. Both humans and neural networks were able to categorize these samples more accurately. This indicates that the intensities of the pixels representing natural images are inherently correlated with the categories to which they belong, for both human and machine recognition tasks. However, this is not a particularly interesting result by itself; in order to distinguish objects visually, they must have distinct visual features.

This experiment had a few drawbacks that this work attempts to circumvent. Firstly, the stimuli presented to the subjects were artificially generated, so even though they consisted of natural features, it was unclear whether the results would extend to naturally occurring stimuli. Secondly, the specific random projection methods used (sliding-window and corner) are not random projections in the commonly used sense in the literature. My experimentation will deal almost entirely with neuronal random projections that are true projections of the data into a random lower-dimensional space, as opposed to algorithmically based sketches with a random component.

2.4 Concept Learning via Random Projection

We look to find computationally efficient methods involving random projection and simple neural networks that indicate some form of category learning of natural visual stimuli without the use of semantic labels. What inherent structures in the stimuli lend themselves to be easily learned with random projection as a preprocessing step?

CHAPTER 3
METHODS

3.1 Training

A simple feed-forward neural network, shown in Figure 3.1, was implemented. The first layer of the network executes neuronal random projection [3] on the visual input in the pixel-intensity domain, with static, random weights drawn from an approximately Gaussian distribution. This reduces the stimulus from n dimensions to k dimensions. The remaining neurons in the network were then trained to reconstruct the original image via backpropagation, using a mean squared error loss function. This trained portion of the network weights can be an arbitrarily complex generator, operating on a low-dimensional, sample-specific coding that is the result of random projection. This model is somewhat similar to a standard generator network, except that the samples in a generative network would be multivariate unit Gaussian instead of the direct result of random projection. All neurons in the network used a sigmoid activation function. The network attempts to learn the identity function with an internal random projection step.

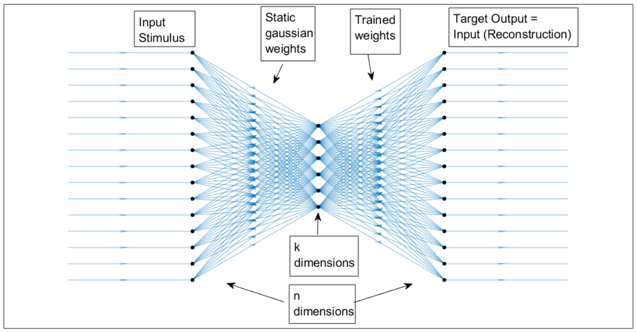


Figure 3.1: Neuronal-RP neural network architecture.

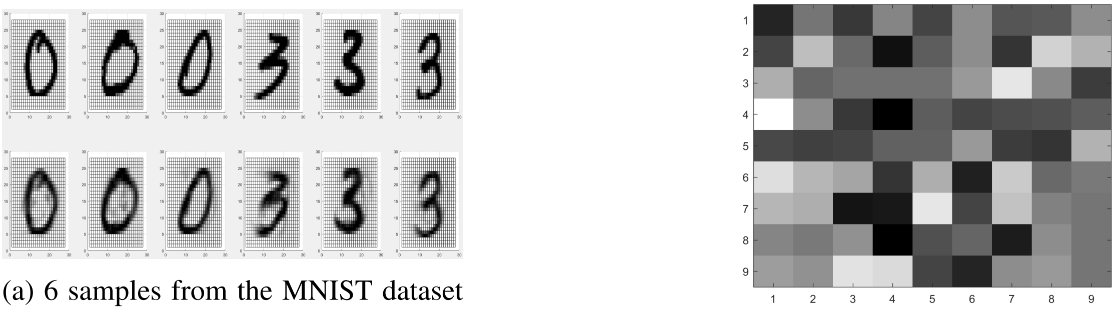
In the context of modeling human concept learning, it is clear that presenting a human with an object or other real category produces many 2-dimensional ”stimuli” from which to learn. The existence of a neuronal structure similar to the one presented is biologically plausible. It has been suggested that neuronal noise assists in the processing of visual stimuli [10]. Work by Quinn, Eimas, and Tarr [7] has found that infants are able to perceptually categorize real visual stimuli such as dogs and cats from very few examples without prior category knowledge. Gallant has proposed, as evidenced by fMRI scans, that much of the activity across the anterior visual cortex is influenced by learned categories of natural visual scenes, and that the brain ”capture[s] the co-occurrence statistics of objects in the world” [9].

3.2 MNIST Dataset

The network was trained to reconstruct 2-subsets of the MNIST handwritten digit dataset categories [8], containing 1000 samples of one handwritten digit, and 1000 samples of another digit. The two specific categories chosen are analogous to A and B from the 2015 human categorization experiment by Arriaga et al [4], except the categories are naturally generated and, in practice, would be familiar to humans. Using only two categories at a time allowed for direct analysis of the sample robustness metric.

3.3 Reconstruction of MNIST Samples After Random Projection

The images, which consisted of 784 pixels/dimensions (28 x 28), were projected into $k = 78$ dimensions by the first layer of the network in Figure 3.1. The rest of the network was trained to reconstruct the original $n = 784$ dimensions from the result of the random projection. Some samples are shown in Figure 3.2a with the original stimulus on the top and the reconstructed stimulus on the bottom. The experiment was repeated on all pairs of digit categories (0 vs 1, 0 vs 2, 7 vs 9, etc) with relatively consistent results. For the sake of brevity, the examples presented in this paper contain stimulus samples from the experiments on the ”0”s and ”3”s of the MNIST set.



(a) 6 samples from the MNIST dataset (top) and their corresponding reconstruction (bottom).

(b) A 78 dimensional neuronal encoding of a single ”0” stimulus.

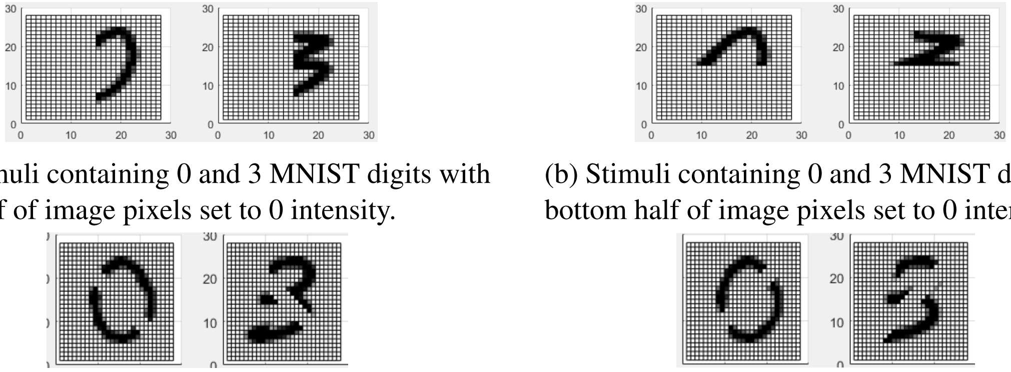
Figure 3.2: The network, in this case, was trained to reconstruct both ”0”s and ”3”s. The visual structure of a ”0” is not preserved in the neuronal random projection encoding .

This dataset was chosen specifically because the categories consist of completely distinct symbols. In other words, there is an explicit, natural, visual concept to be learned. We would expect these symbols to be approximately drawn from a bimodal multivariate non-Gaussian distribution [16]. It was confirmed that, for all pairs of MNIST digit categories, there was a 783-dimensional separating hyperplane between the two categories' members.

These results generally extend to concept learning of other types of visual stimuli, such as animals and inanimate objects, whose categories in the pixel domain are clearly distinct and are drawn from approximately bimodal distributions.

3.4 Reconstruction of Perturbed Samples

After training the network to reconstruct 2-subsets of MNIST digits, variants of the original MNIST data set were fed forward through the network and the output was observed. First, the stimuli were modified such that significant contiguous portions of the image's pixels were set to 0 intensity. Examples of these stimuli are shown in Figure 3.3.



(a) Stimuli containing 0 and 3 MNIST digits with left half of image pixels set to 0 intensity.

(b) Stimuli containing 0 and 3 MNIST digits with bottom half of image pixels set to 0 intensity.

(c) Stimuli containing 0 and 3 MNIST digits with middle 3 main diagonals set to 0 intensity.

(d) Stimuli containing 0 and 3 MNIST digits with middle 3 antidiagonals set to 0 intensity.

Figure 3.3: The stimuli shown above are examples of inputs given to the network pre-trained to reconstruct their unperturbed counterparts.

Additionally, the original MNIST samples were fed forward through the network, but before reconstruction, some fraction (specifically, half) of the $k = 78$ random-projected neuron values were chosen to be set to 0.

CHAPTER 4

RESULTS

For all of the results hereinafter, the portion of the network that was trained contained one hidden layer of 100 fully connected neurons, fully connected to a second hidden layer of $n = 784$ neurons, all with sigmoid activation functions.

4.1 Reconstruction of Samples Projected into RP Domain

The samples that already had relatively high sample robustness generally had increased sample robustness after projection and reconstruction. The distribution of each category was concentrated towards the mean by this process. The sample robustness of the samples in each category is shown in Figure 4.1, both before and after the reconstruction process. The samples are sorted by sample robustness along the x-axis.

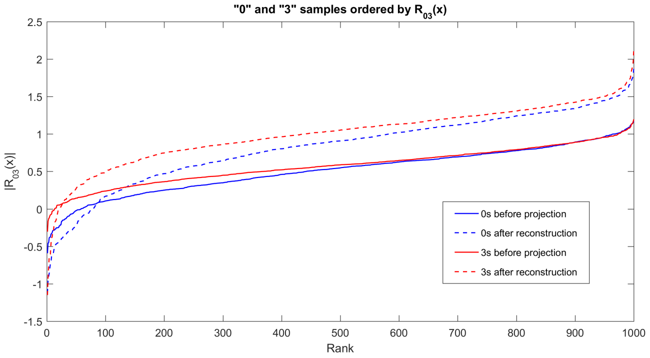


Figure 4.1: Reconstruction increased most samples' robustness.

Figure 4.2 is a projection of the samples onto the first two principal components of the dataset. The arrows indicate the coordinate change of each sample due to the reconstruction process. The arrows are colored by the change in sample robustness from the reconstruction. Red dots indicate the means of the two original categories, while blue dots indicate the means of the reconstructed categories.

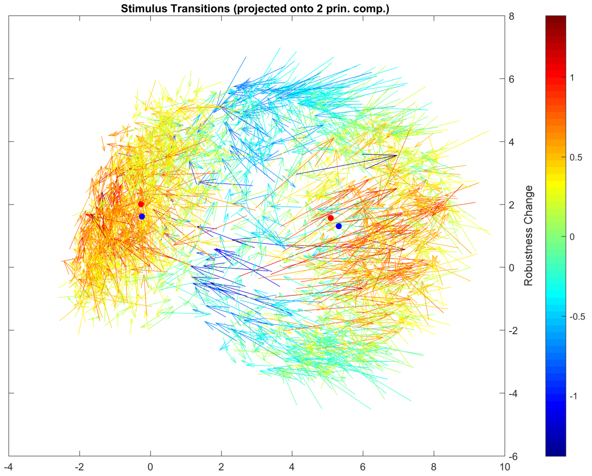


Figure 4.2: There is a clear separation between two groups of samples whose robustness increased after random projection and reconstruction.

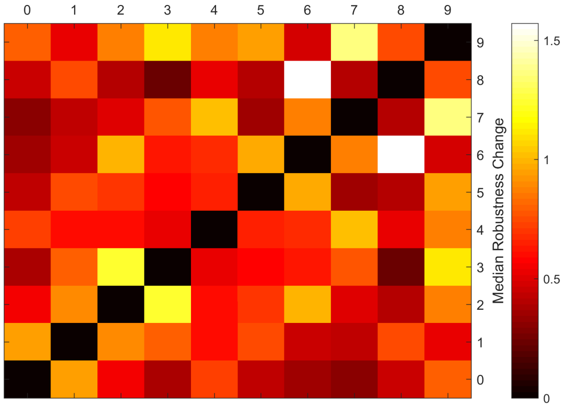


Figure 4.3: Positive median change in $R_{AB}(x)$ for all possible pairs of A and B in the set of MNIST categories

Figure 4.3 illustrates the median change in sample robustness over all samples, for all pairs of MNIST digit categories. The network was trained to reconstruct all pairs of digit categories, and consistently concentrated most of the individual samples towards the mean of their respective category, indicated by the positive median robustness change over the corresponding samples.

4.2 Addition of Gaussian Noise to RP-Neural Encoding

Figure 4.4 demonstrates how the addition of Gaussian noise to the projected stimulus domain affects the median change in sample robustness (over ”0”s and ”3”s, and averaged over 10 trials). In a standard Generative Adverserial Network model, the generator draws from a multivariate unit Gaussian to construct plausible stimuli of a certain category, and is competitively trained to fool a learner that is trained to discriminate real versus generated stimuli [14]. This result demonstrates that the prior distribution for a GAN or other generative model could potentially be manipulated by sampling from a lower-dimensional multimodal distribution to construct multiple classes of stimuli simultaneously, as this model does. GAN models that can construct multiple classes have recently been explored by additionally providing the generator and discriminator with a conditional input. [15]. Estimating a distribution for sampling that is the result of random projection of real data may achieve a similar result.

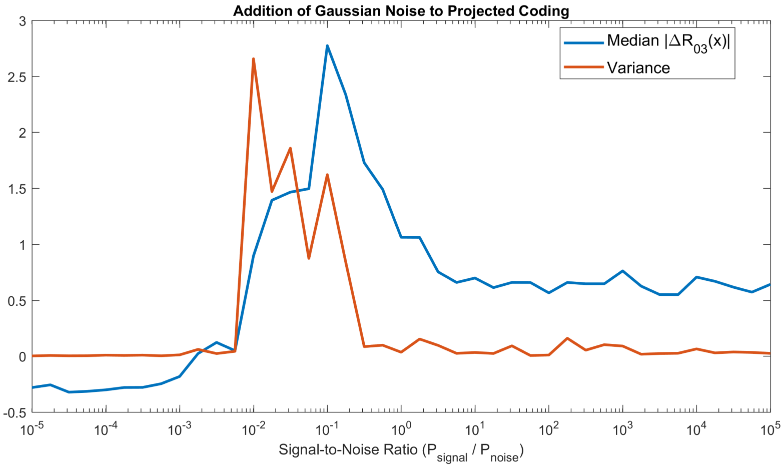
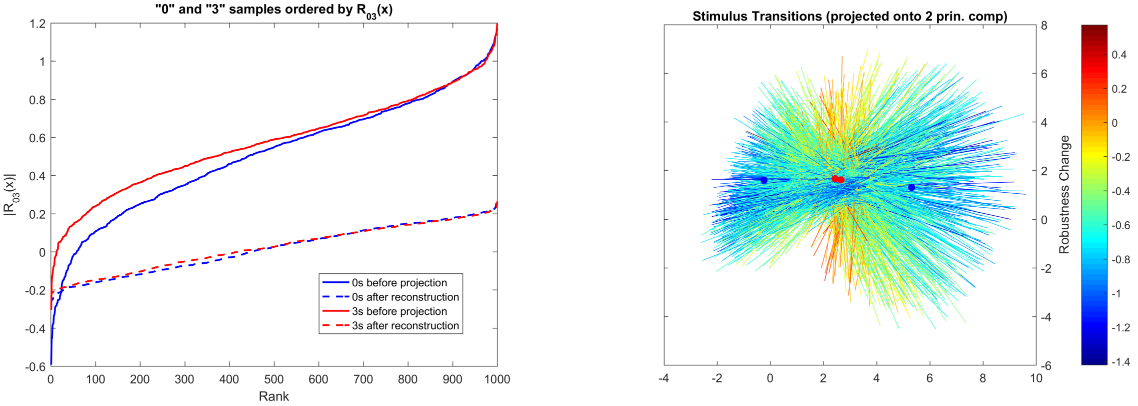


Figure 4.4: Median $\Delta R_{03}(x)$ is negative when Gaussian noise completely overwhelms the random projection. When the signal dominates the added noise, $\Delta R_{03}(x) > 0.5$ with nearly zero variance over the 10 trials. When noise largely, but not completely overwhelms the random projection $\frac{P_{signal}}{P_{noise}} \approx 10^{-1}$, the increase in median $\Delta R_{03}(x)$ is positive on average, but not reliably, as the variance over the 10 trials is high. For a few trials in this range, $\Delta R_{03}(x) < 0$.

The bimodal nature of the random projected neural coding is what causes the network to reconstruct stimuli towards their corresponding mean when multiple categories are presented without labels. As noise collapses the projections to a unimodal Gaussian distribution, the distribution of reconstructed stimuli also collapses to a single mean, which is displayed in Figure 4.5. These plots are analogous to Figures 4.1 and 4.2, except the projected neural signal was replaced with a multivariate unit Gaussian of the same dimensionality ($k = 78$).



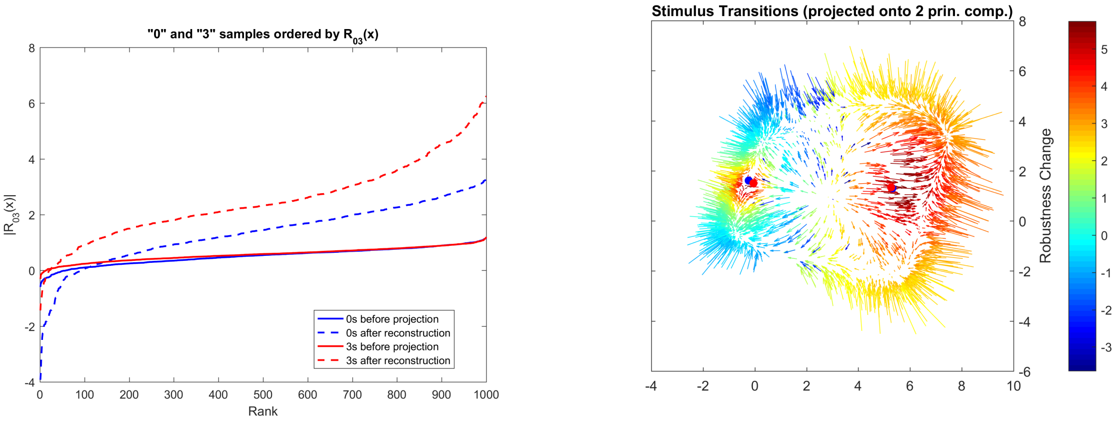
(a) Reconstruction *decreased* individual samples' robustness

(b) Reconstructed stimuli collapse to a single mean

Figure 4.5: Projected samples were replaced with multivariate Gaussian of same dimensionality after the network was trained to reconstruct the "0"s and "3"s.

4.3 Reconstruction of Samples Projected onto Principal Components

When random projection is replaced with Principal Components Analysis, it becomes clear that the preservation of inter-category variance is what enables the reconstructing network to map samples to their category means. Instead of performing random projection to $k = 78$ dimensions, PCA was performed on the "0"s and "3"s, and only the first $k = 2$ principal components were taken and reconstructed. They represented 28.91% of the total original variance. In Figure 4.6, it is clear that inter-category variance dominated intra-category variance, and this caused the network to reconstruct samples based mostly on categorical information instead of sample specific information. The robustness increase effect was much more prominent in this case (note the y-axis scale in Figure 4.6a).



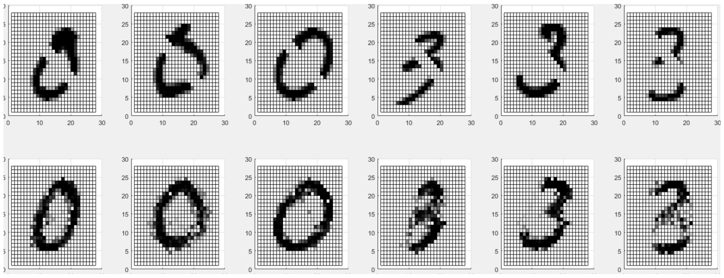
(a) Increase in robustness more consistent than with random projection

(b) Stimuli more precisely transition towards corresponding category mean. Few samples remain in between-category region of visual space.

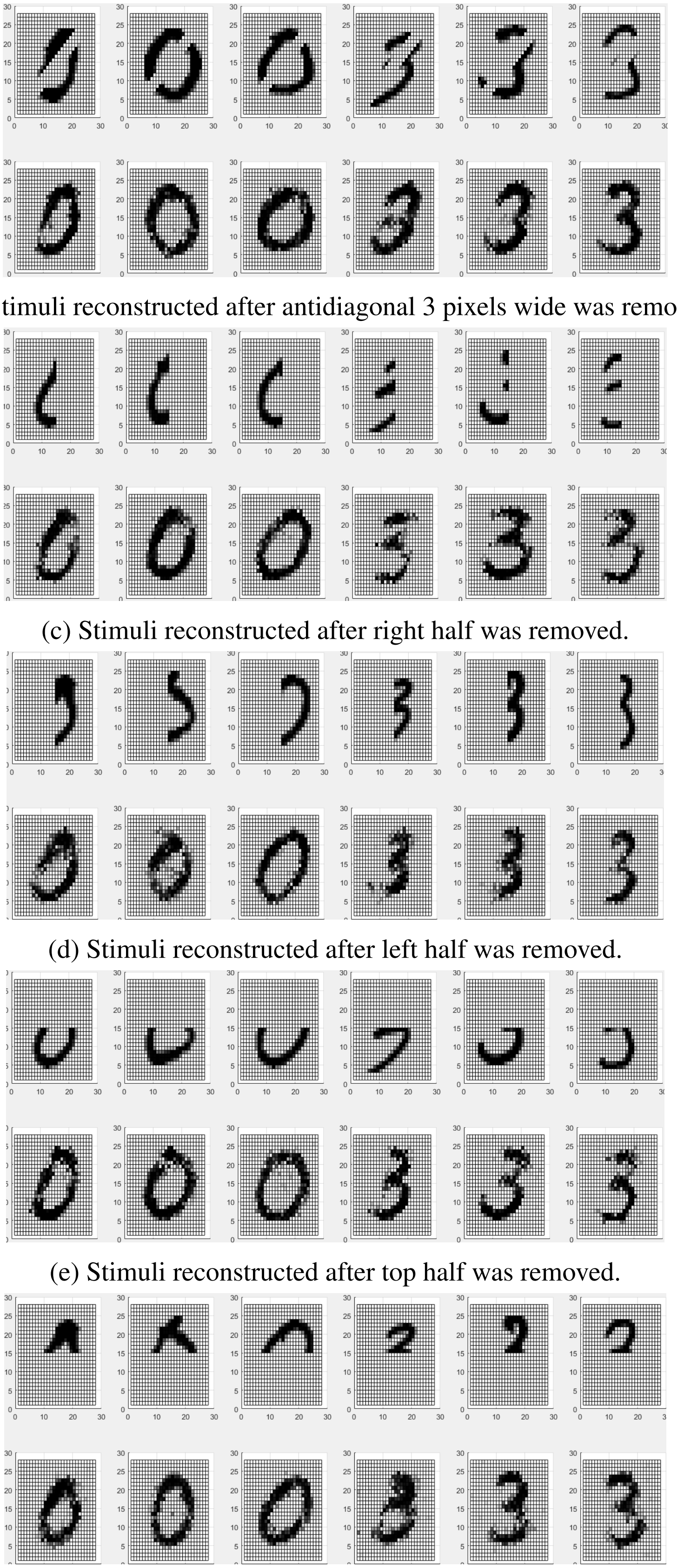
Figure 4.6: Projected samples replaced by first two principal components

4.4 Reconstruction of Perturbed Samples

As outlined in section 3.3, samples of the MNIST dataset were modified such that contiguous groups of pixels were set to 0 intensity, and then fed into the network trained to reconstruct original samples. The network was able to 'fill-in-the-blank' on these perturbed samples, and produced prototypical representations of the perturbed sample's original category. Some examples of this phenomenon are visible in Figure 4.7.



(a) Stimuli reconstructed after main diagonal 3 pixels wide was removed.



(b) Stimuli reconstructed after antidiagonal 3 pixels wide was removed.

(c) Stimuli reconstructed after right half was removed.

(d) Stimuli reconstructed after left half was removed.

(e) Stimuli reconstructed after top half was removed.

(f) Stimuli reconstructed after bottom half was removed.

Figure 4.7: The network output contained parts of the digit that were removed from the stimulus fed in.

4.5 Reconstruction of Partially Zeroed Random Projections

The network was trained to reconstruct the original random projected samples, as previously described. The same samples were fed through the network, but after random projection, one half of each stimulus' random projection's values were randomly chosen to be set to 0. These samples were then fed through the pre-trained network and reconstructed.The results of reconstructing these zeroed random projections are shown in Figure 4.8 along with their original sample.

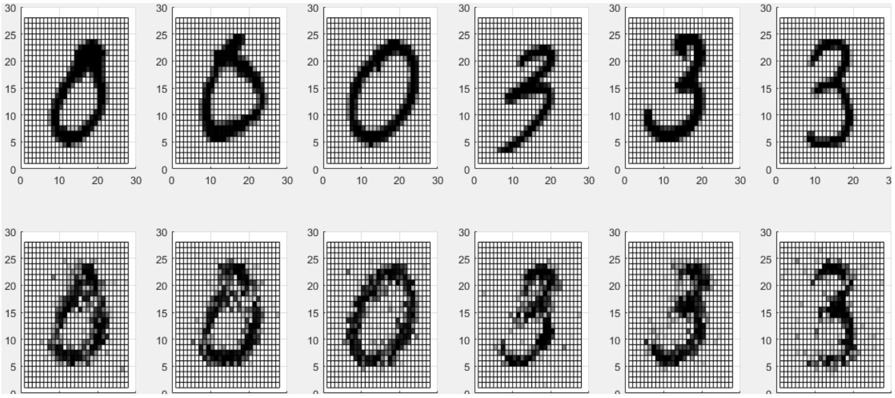


Figure 4.8: Reconstruction of stimuli after $k = 38$ random projection values were set to 0

CHAPTER 5
DISCUSSION

The random projection/reconstruction process is analogous to a synchronization between low-dimensional, neuronal representations of stimuli and the pixel-domain presentation of those same stimuli. In order to minimize the reconstruction mean square error, the neural network tends to map reduced stimuli towards the mean of their ground truth category upon reconstruction. Preservation of inter-category variance through random projection enables this process. This is clearly visible in Figure 4.2, as there is a distinct separation between two groups of stimuli whose sample robustness increased via the random projection and reconstruction, and these groups are centralized near the actual means of the ”0”s and ”3”s.

This phenomenon may suggest that random projection enables computationally efficient learning of natural visual concepts by reducing the complexity of visual stimuli. Random projection provides a noisy sketch of the presented image with missing information in the neuronal domain, and the neural network attempts to decode the original image from just the noisy sketch. The Johnson-Lindenstrauss lemma [2] only guarantees relative preservation of distances between points within some error bound. Since the fine details of the stimulus in the visual space are lost in this sketch, the reconstruction generally contains all of the features specific to the digit (”0” or ”3”, etc.) but few of the extraneous features specific to the individual sample, as is visible in Figure 3.2a. The reconstructed samples match the category of the input, but typically do not look exactly the same as the input and are missing sample specific features. The generalized concepts are learned in an unsupervised manner.

Ell and Ashby (2012) have shown that inter-category and intra-category variance in certain feature spaces have a significant impact on the ability of humans to categorize stimuli in an unsupervised manner[11]. Dasgupta (1999) has demonstrated that random projection makes Gaussians more spherical, and his algorithm used this property to learn mixtures of Gaussians in a computationally efficient manner [12]. An algorithm by Kalai, Moitra, and Valiant (2012) disentangles mixtures of Gaussians into their corresponding parameters, similar to those created by 2-combinations of MNIST categories, through computation in the random projected vector space [13].

It is clear that the low-dimensional result of random projection preserved the bimodal nature of the original distribution, and this caused the reconstructed stimuli to form two distinct categories. We propose this simple mechanism as a method of low-dimensional neuronal concept learning without explicit categorization or presentation of category labels.

When the random projection step is skipped, and the decoded representation is replaced with distinct samples from a multivariate unit Gaussian of the same dimensionality, the reconstruction collapses to a single mean, as shown in Figure 4.5. This indicates that the low-dimensional neuronal representations of the original stimuli contain enough information to both distinguish between samples from the two different categories, as well as to reproduce a prototypical instance of the corresponding category.

While a standard autoencoder or compressed sensing model will find an efficient or non-redundant representation to encode samples, this experiment demonstrates that an unstructured, data-independent low-dimensional random projection encoding of the original stimuli is sufficient to differentiate and reconstruct the categories of stimuli in question. These results are relatively consistent across all pairs of digit categories from the MNIST dataset.

This mechanism is very time efficient. It produces an interpretable neuronal encoding of distinct visual stimuli in $O(kn)$ time with no preprocessing step other than drawing kn weights from a univariate Gaussian.

The most atypical feature of the random projection method is its data-independence.

The algorithm performs exactly the same operation regardless of input. So long as the presented categories of stimuli form a jointly bimodal distribution of features, no matter the feature space, a random projection step beyond that feature space will reduce the stimuli into a low-dimensional variation that preserves the bimodality of the original distribution. Since many natural stimuli of a single category or concept have variations in their features deviating from the mean, this mechanism could be used to categorize naturally occurring stimuli in general.

Another useful property of the random projection method is that, since the entries of the random projection matrix a drawn from an approximate Gaussian distribution with mean 0, the projected samples' values have a mean of approximately 0; the expectation of the product of random variables is the product of their expectations, the expectation of the sum of random variables is the sum of their expectations, and random projection is the sum of the products of the input elements multiplied by normally distributed weights.

The usefulness of this property is apparent in Figure 4.7. When the network was trained on the original stimuli and samples with contiguous regions of pixels set to 0 were fed through, the reconstruction from the random projection domain filled in the removed regions. The effect on the values of the random projection of setting some pixels to 0 is proportional across those values. In other words, since the neuronal random projection weights are normally distributed, the probability that setting a single pixel to 0 increases a value in the random projection is approximately equal to the probability of that value decreasing. Since the result of random projection is not spatially oriented or otherwise ordered in any particular way, and additionally because the random projection values should remain distributed similarly regardless of small portions of pixels being zeroed, zeroing small portions of pixels in the stimulus passed in affects the network output minimally. This results in the random projection-reconstruction network model essentially being able to 'fill-in-the-blank', so long as there are some category distinguishing features that remain in the perturbed stimulus.

Also due to this property are the results shown in Figure 4.8. Setting some proportion of the random projection values to zero has a minimal effect on the reconstruction by the neural network. Each neuron essentially corresponds to a random projection into $k = 1$ dimension. Setting half of the $k = 78$ original neurons results in the outputs of a $k = 39$ dimension random projection with 39 additional neurons outputting 0. Since these neurons have an expected mean output of 0 over the data set, this operation produces reconstructions very visually similar to those produced by random projection into $k = 39$ dimensions reconstructed by the same network. The random projection neurons are indistinguishable in the sense that their weights are drawn from the same distribution and they each perform the same operation. Random projection then, could occur into any large number of neurons and subsets of those neurons can be used to reconstruct the original stimulus.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

Neuronal random projection on a visual stimulus produces a low-dimensional encoding of that stimulus. Over a bimodal distribution, the projected samples remain approximately bimodal. However, information is lost during the random projection process. Because a small amount of relative error is introduced to each sample, a neural network trained to reconstruct the original sample from the random projected domain by minimizing mean square error tends to produce samples that have category-dependent features but not features specific to the original sample. Additionally, the neuronal random projection is somewhat flexible. The components of a reconstruction of an original stimulus sample are not strongly dependent on any single neuron's output, and no single neuron's output is strongly dependent on any particular component of the input stimulus. The model proposed can infer category-specific visual features of MNIST digits in their reconstruction when those features are removed from the input to the network after it has been trained.

This mechanism is suitable for learning visual concepts in that it is data-independent and computationally efficient. It can be viewed as a synchronization between a random neural impulse response to the presentation of a visual stimulus and the simultaneous reconstruction of that stimulus.

Future work could explore these properties over more complex datasets, such as those of pictures of objects, as well as over different types and levels of concepts such as directionality, texture, smoothness, sounds, etc. Experiments on concept learning of artificially generated stimuli could be useful. Are there certain types of visual concepts that are inherently easy for neural networks to learn? Do the statistics of natural images lend themselves to be easily learned by humans? How does this relate to the concept learning capabilities of neural networks? These experiments would potentially require the use of convolutional neural networks in either the random projection phase, reconstruction phase, or both. The inclusion of convolutional models and different neural network architectures would certainly allow for more complex stimuli to be learned.

This paper only explores the learning of two visual concepts at a time, specifically, 2-subsets of categories of the MNIST dataset. Changes to the sample robustness metric could allow for analysis of learning many distinct visual concepts at once by a single neural network model and its capacity to store those concepts. Additionally, different random projection models could be explored. Sparsity and positivity constraints on neuronal random projection in this context may yield interesting results.

While this work was performed on a limited dataset and on just a single neural network model, it highlights important properties of random projection in the task of representing visual concepts in a lower-dimensional, neuronal encoding. Neural network models can exploit the relative error introduced by random projection as well as its low dimensionality to efficiently construct neuronal representations of distinct visual concepts that are independent of sample specific features in an unsupervised manner.

BIBLIOGRAPHY

[1] Sharpee, T., Rust, N. C. & Bialek, W. Analyzing Neural Responses to Natural Signals: Maximally Informative Dimensions. *Neural Computation*, 16, pp. 223–250, 2004.

[2] Johnson, W. B. & Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26(189-206):11, 1984.

[3] Arriaga, R. & Vempala, S. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161 182, 2006.

[4] Arriaga, R. I., Rutter, D., Cakmak, M. & Vempala, S. S. Visual Categorization with Random Projection. *Neural Computation*, 2015.

[5] Allen-Zhu, Z., Gelashvili, R., Micali, S., & Shavit, N. Sparse sign-consistent Johnson-lindenstrauss matrices: Compression with neuroscience-based constraints. *Proceedings of the National Academy of Sciences*, 111(47), 16872-16876, 2014.

[6] Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*. 381, 607, 1996.

[7] Quinn, P. C., Eimas, P. D. & Tarr, M. J. Perceptual categorization of cat and dog silhouettes by 3 to 4-month-old infants. *Journal of Experimental Child Psychology*. 79, 78-94, 2001.

[8] Lecun, Y., Cortes, C. & Burges, C. J. MNIST Database of Handwritten Digits. http://yann.lecun.com/exdb/mnist/.

[9] Stanbury, D. E., Naselaris, T. & Gallant, J. L. Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. *Neuron*, 79, 10251034. 2013.

[10] Anderson, J. S., Lampl, I., Gillespie, D. C & Ferster D. The Contribution of Noise to Contrast Invariance of Orientation Tuning in Cat Visual Cortex. *Science*, 5498, 1968-1972. 2000.

[11] Ell, S. W. & Ashby, F. G. The Impact of Category Separation on Unsupervised Categorization. *Attention, Perception & Psychophysics*, 74, 466-475. 2012.

[12] Dasgupta, S. Learning Mixtures of Gaussians. *IEEE Symposium on Foundations of Computer Science*, 634-644. 1999.

[13] Kalai, A. T., Moitra, A. & Valiant, G. Disentangling Gaussians. *Communications of the ACM*, 55, 113-120. 2012.

[14] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. Generative Adversarial Nets. *NIPS*, 2014.

[15] Mehdi, M. & Osinder, S. Conditional Generative Adversarial Nets. 2014.

[16] Weiss, Y. & Freeman, W. T. What Makes a Good Model of Natural Images? *Proc. IEEE CVPR*, 1-8, 2007.

[17] Field, D. J. What Is The Goal of Sensory Coding? *Neural Compuation*, 6, 559-601, 1994.

[18] Xie, H., Li, J. & Xue, H. A Survey of Dimensionality Reduction Techniques Based on Random Projection. *CoRR*, 2017.

[19] Cannings, T. I. & Samworth, R. J. Random Projection Ensemble Classification. *J. Roy. Statist. Soc.* 79, 959-1035, 2015.