The Natural Tendency of Feed Forward Neural Networks to Favor Invariant Units

Anonymous Author(s)
Affiliation
Address
email

Abstract

A central goal in the study of the primate visual cortex and hierarchical models for object recognition is understanding how and why single units trade off invariance versus sensitivity to image transformations. For example, in both deep networks and visual cortex there is substantial variation from layer-to-layer and unit-to-unit in the degree of translation invariance. Here, we provide theoretical insight into this variation and its consequences for encoding in a deep network. Our critical insight comes from the fact that rectification simultaneously decreases response variance and correlation across responses to transformed stimuli, naturally inducing a positive relationship between invariance and dynamic range in the subsequent layer. Invariant units then tend to drive the network more than those sensitive to small image transformations. We discuss consequences of this relationship for AI: deep nets naturally weight invariant units over sensitive units, and this can be strengthened with training, perhaps contributing to generalization performance. Our results predict a signature relationship between invariance and dynamic range that can now be tested in future neurophysiological studies.

1 Introduction

Invariances to image transformations, such as translation and scaling, have been reported in single units in visual cortex, but just as often sensitivity to these transformations has been found (El-Shamayleh and Pasupathy, 2016, Sharpee et al. 2013, Rust and DiCarlo, 2012). Similarly, in deep networks there is variation in translation invariance both within and across layers (Pospisil et al., 2018, Shen et al., 2016, Shang et al., 2016, Goodfellow et al., 2009). Notionally, information about the position of the features composing objects may be important to category selectivity. For example, the detection of eyes, nose, and lips are not sufficient for face recognition, the relative positions of these parts must also be encoded. Thus it is reasonable to expect some balance between invariance and sensitivity to position. We empirically observe that in a popular deep network, in both its trained and untrained state, invariant units tend to have higher dynamic range than sensitive units (Figure 1B and C). This raises the possibility that the effective gain on invariant units into the subsequent layer is stronger than that of sensitive units. Here we provide theoretical insight to show that rectification in a deep network naturally biases networks to this difference between invariant and sensitive units. We do this by examining how co-variance of a multivariate normal distribution is influenced by rectification, and we then test these insights in a deep neural network.

2 Statistical model

The response of a unit in a feed forward neural network is: $r = \vec{w} \cdot g(S)$ where S is the response of all n subunits in the previous layer, g the non-linearity of rectification $g(x) = max(0, x)$, $\vec{w}$ is the $n \times 1$ vector of weights, and r is the response of the unit. Randomly sampling from a distribution of input images, the response S takes on a distribution with some expectation and covariance across these images: $E[S] = \mu$ (an $n \times 1$ vector), and $Cov[S] = \Sigma$ (an $n \times n$ matrix). The application of the non-linearity transforms these moments: $E[g(S)] = \tilde{\mu}$, $Cov[g(S)] = \tilde{\Sigma}$. Let S_1 be the responses to randomly sampled input images and S_2 the responses to a transformation of those same images. So the moments of the full distribution are:

$$E\begin{bmatrix}\tilde{S}_1\\ \tilde{S}_2\end{bmatrix} = \begin{bmatrix}\tilde{\mu}_1\\ \tilde{\mu}_2\end{bmatrix}, \quad \text{Cov}\begin{bmatrix}\tilde{S}_1\\ \tilde{S}_2\end{bmatrix} = \begin{bmatrix}\tilde{\Sigma}_{1,1} & \tilde{\Sigma}_{1,2}\\ \tilde{\Sigma}_{2,1} & \tilde{\Sigma}_{2,2}\end{bmatrix},$$

where $\tilde{\Sigma}_{1,1}$ is the covariance of rectified subunits responding to the original images, $\tilde{\Sigma}_{2,2}$ is the covariance of rectified subunits to the transformed images and $\tilde{\Sigma}_{1,2}$ is the covariance between rectified subunits responding to the reference and transformed

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

images. The joint distribution of a unit with weights $\vec{w}$ on the n rectified subunits is:

$$\begin{bmatrix} \tilde{\sigma}_1^{2\,\prime} = \vec{w}\tilde{\Sigma}_{1,1}\vec{w}^T & \tilde{\rho}'\tilde{\sigma}_1'\tilde{\sigma}_2' = \vec{w}\tilde{\Sigma}_{1,2}\vec{w}^T \\ \tilde{\rho}'\tilde{\sigma}_1'\tilde{\sigma}_2' = \vec{w}\tilde{\Sigma}_{2,1}\vec{w}^T & \tilde{\sigma}_2^{2\,\prime} = \vec{w}\tilde{\Sigma}_{2,2}\vec{w}^T \end{bmatrix}$$

so the correlation between the response of the unit to the reference and transformed images is:

$$\tilde{\rho}' = \frac{\vec{w}\tilde{\Sigma}_{1,2}\vec{w}^T}{\sqrt{\vec{w}\tilde{\Sigma}_{1,1}\vec{w}^T\vec{w}\tilde{\Sigma}_{2,2}\vec{w}^T}}.$$

Below we investigate how the $\tilde{\Sigma}_{i,j}$ depend on $\Sigma_{i,j}$ and μ, which provides insight into a relationship between $\tilde{\rho}'$ and $\tilde{\sigma}^{2\,\prime}$. We begin by examining a model of a single rectified subunit responding to the reference and transformed images.

2.1 Single rectified input

We model the responses, S_1 and S_2, of a single subunit to the reference and transformed inputs, respectively, as a bivariate normal distribution:

$$\begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}\right).$$

When these responses are acted on by rectification, both the variances of the responses and the correlation between the sets of responses is decreased. This observation is analogous to that of de la Rocha et al. (2007) where they investigated the influence of neuronal firing threshold rectification on the pairwise correlations between neurons as a function of firing rate. We extend this observation in the next section to consider how this effect influences invariance in downstream units.

It is instructive to consider a schematic (Figure 2A) of the distribution of responses. The fall in correlation occurs because the variance from the linear relationship $Var(E[\tilde{S}_2|\tilde{S}_1])$ decreases as a result of truncation, whereas the residual $E[Var(\tilde{S}_1|\tilde{S}_1)]$ is not reduced as much. Notionally, the average length of a vertical line in the remaining elipsoid ($E[Var(\tilde{S}_1|\tilde{S}_1)]$) is not reduced as much as the diagonal line of best fit ($Var[E[\tilde{S}_1|\tilde{S}_1]]$). For further demonstration, see de la Rocha et al. (2007).

In the following section, we will write the correlation and variance after rectification explicitly as a function of the relevant parameters: $\tilde{\sigma}(\mu/\sigma)^2$ and $\tilde{\rho}(\mu/\sigma, \rho)$.

2.2 Unit integrating across multiple rectified subunits

Here we extend from the single subunit case to the multi-subunit case by examining the invariance $\tilde{\rho}'$ resulting from taking weighted combinations of rectified subunits. Our key insight is that invariance increases to the degree that directions of maximal variance in the response distribution of rectified subunits are integrated. A helpful geometric perspective comes from decomposing the numerator and denominator of $\tilde{\rho}'$ into their eigenbasis. We assume $\Sigma_{1,1} = \Sigma_{2,2} = \Sigma_{1,2}\epsilon$ (ϵ a small constant) a more reasonable approximation for the case of smaller transformations of the input images. Furthermore we assume that each subunit has the same covariance structure as another (analogous to the covariance of a stationary stochastic process) so that the primary driver in differences post-rectification is the original μ_i/σ. Thus we have:

$$\tilde{\rho}' = \frac{\vec{w}\tilde{\Sigma}_{1,2}\vec{w}^T}{\vec{w}\tilde{\Sigma}_{1,1}\vec{w}^T} = \frac{(\vec{w}^T\vec{u}_1^{(1,2)}\sqrt{\lambda_1^{(1,2)}})^2 + ... + (\vec{w}^T\vec{u}_n^{(1,2)}\sqrt{\lambda_n^{(1,2)}})^2}{(\vec{w}^T\vec{u}_1^{(1,1)}\sqrt{\lambda_1^{(1,1)}})^2 + ... + (\vec{w}^T\vec{u}_n^{(1,1)}\sqrt{\lambda_n^{(1,1)}})^2}.$$

We also assume that the eigenvectors $\vec{\mu}_i^{(1,2)} \approx \vec{\mu}_i^{(1,1)}$. Thus the geometric picture (Figure 3A) is of the numerator as an eigenvector-aligned ellipsoid (blue) nested inside of the larger ellipsoid formed by the denominator (red). The correlation being determined by the relative distance from the origin of these ellipses in the direction of $\vec{w}$ (pink). The ellipsoid is nested because ($\tilde{\rho} \leq 1$). The widths of these ellipsoids along the eigenvectors are determined by the eigenvalues. In general, small eigenvalues are paired with linear combinations of low-variance sensitive subunits, whereas large eigenvalues are paired with high-variance invariant subunits.

Performing simulations of a few simple subunit covariance structures shows that the $\tilde{\rho}'$ to $\tilde{\sigma}^{2\,\prime}$ relationship is maintained, though its form changes (Figure 3B). Integrating over a population of subunits, the form of the relationship changes from the single subunit case (Figure 3B black dashed and dotted line). Averaging over subunits with different μ_i/σ shifts the relationship because of the curvature of the non-linear transforms $\tilde{\sigma}(\mu/\sigma)^2$ and $\tilde{\rho}(\mu/\sigma, \rho)$. Finally, in the case where subunits have non-zero correlation (i.e. shared tuning, off-diagonals of $\Sigma_{i,j}$ non-zero; Figure 3B red and cyan) overall variance increases because correlated subunits are being added.

Variation in the neighborhood being integrated over in terms of ρ and μ/σ would weaken the strength of the $\tilde{\rho}'$ to $\tilde{\sigma}^{2\prime}$ relationship. For example, sampling randomly from the different traces in Figure 3B, the relationship would appear less systematic. This leads us to empirically test whether in a well known deep neural network the relationship emerges.

3 Experimental results

Here we analyze the covariance structure of the inputs of a popular deep neural network (AlexNet) for translations of input images. We first tested the network in its *untrained* state by presenting a collection of 500 image patches drawn from the 2012 ImageNet validation set. References images were cropped enough to allow the original and translated images to fit within the maximal receptive field of the units being tested. We included a small and a large translation, and at each convolutional layer we measured the correlation and variance. We find a positive relationship at all layers with significant Spearman's ranked correlation for both transformations (Table 1 Untrained). The strength of the relationship tended to be stronger for the smaller translation (Figure 1C, orange). Thus in a popular deep network with no training, units which tended to have greater invariance also had higher dynamic range. We repeated the same analysis in AlexNet after it was fully trained for object recognition. Again we observed a significant positive relationship (Table 1 Trained; Figure 1B). The relationship was somewhat weaker than in the trained network. Thus training weakens but does not remove the bias of the network to associate higher dynamic range with higher translation invariance.

Layer	Distance	Untrained	Trained
Conv2	Far	0.44	0.35
	Close	0.92	0.58
Conv3	Far	0.67	0.56
	Close	0.94	0.75
Conv4	Far	0.65	0.48
	Close	0.92	0.65
Conv5	Far	0.64	0.18
	Close	0.86	0.64

Table 1: Table of Spearman's ranked correlation coefficient (r_s) between ρ and σ^2 in AlexNet layers and for the trained and untrained network. All values $p < 0.001$.

Finally, we asked whether the network may compensate for this imbalance by placing weights of higher magnitude on low dynamic range units, thus effectively removing this bias. We measured whether the percent of weight magnitude on a given subunit across afferent units was greater for subunits with higher variance. We found Conv3 ($r_s = 0.34$) and Conv4 ($r_s = 0.19$) tended to have higher weights on higher variance subunits while There was no correlation in Conv2 and and Conv5. This indicates the network does not compensate for the imbalance in dynamic range between invariant and sensitive units but actually sometimes emphasizes it.

4 Discussion

We have documented an empirical relationship between the dynamic range of unrectified units in a deep network and their invariance. We provided a simple statistical model to explain this effect in which rectification caused the population representation to primarily vary in dimensions that were invariant to small image perturbations, whereas small perturbations were represented in directions of lower variance. Further work can investigate whether this imbalance improves generalization because of the emphasis placed on invariant over sensitive units.

To extend insights from simplified, artificial networks to neurobiology, it will first of all be important to test whether cortical neurons showing more invariance also tend to have a higher dynamic range. If they do, this will establish a fundamental theoretical connection between computations of deep networks and the brain.

Acknowledgements

References

Pospisil, D. A., Pasupathy, A., Bair, W. (2018). "Artiphysiology" reveals V4-like shape tuning in a deep network trained for image classification. ELife, 7, e38242. https://doi.org/10.7554/eLife.38242

El-Shamayleh, Y., Pasupathy, A. (2016). Contour Curvature As an Invariant Code for Objects in Visual Area V4. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 36(20), 5532–5543. https://doi.org/10.1523/JNEUROSCI.4139-15.2016

Sharpee, T. O., Kouh, M., Reynolds, J. H. (2013). Trade-off between curvature tuning and position invariance in visual area V4. Proceedings of the National Academy of Sciences of the United States of America, 110(28), 11618–11623. https://doi.org/10.1073/pnas.1217479110

Rust, N. C., DiCarlo, J. J. (2012). Balanced Increases in Selectivity and Tolerance Produce Constant Sparseness along the Ventral Visual Stream. Journal of Neuroscience, 32(30), 10170–10182. https://doi.org/10.1523/JNEUROSCI.6125-11.2012

Goodfellow, I., Lee, H., Le, Q. V., Saxe, A., Ng, A. Y. (2009). Measuring Invariances in Deep Networks. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems 22 (pp. 646–654).

Shen, X., Tian, X., He, A., Sun, S., Tao, D. (2016). Transform-Invariant Convolutional Neural Networks for Image Classification and Search. Proceedings of the 2016 ACM on Multimedia Conference, 1345–1354. https://doi.org/10.1145/2964284.2964316

Shang, W., Sohn, K., Almeida, D., Lee, H. (2016). Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units. ArXiv:1603.05201 [Cs].

de la Rocha, J., Doiron, B., Shea-Brown, E., Josić, K., Reyes, A. (2007). Correlation between neural spike trains increases with firing rate. Nature, 448(7155), 802–806. https://doi.org/10.1038/nature06028

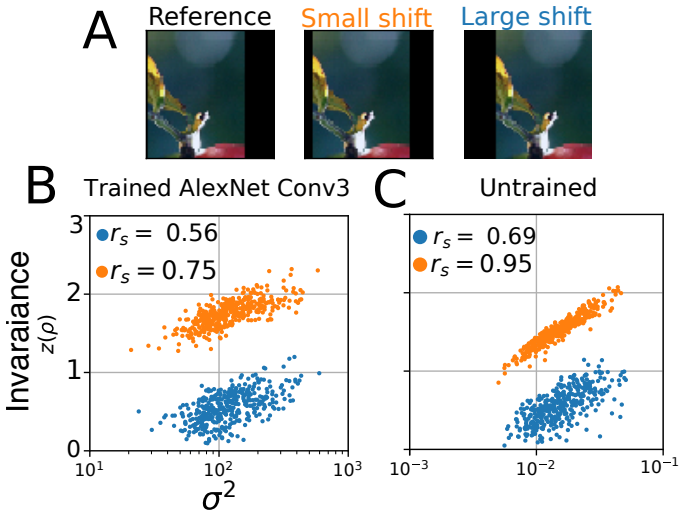

Figure 1: Empirical relationship between invariance and response variance. **(A)** Example stimuli conditions: reference position images cropped to 3/4 of unit maximal receptive field, then consider the correlation of responses across images for a small shift (1/8 of RF) and large shift (1/4 of RF). **(B)** The variance of unrectified responses from Conv3 units in trained AlexNet to the reference position plotted against Fisher's z of reference images and shift images (orange small shift, blue large shift). Both show strong positive relationship with Spearman's ranked correlation coefficient (r_s) listed in legend. **(C)** Same plot for the untrained AlexNet showing little change in the form of the relationship except for its strength.

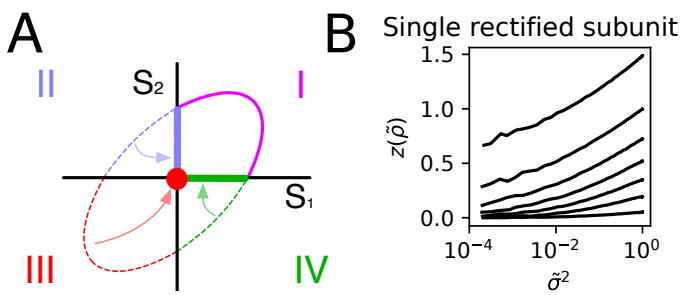

Figure 2: Influence of rectification on variance and correlation in bivariate normal. **(A)** Transformation of bivariate normal by rectification. Distribution in quadrant I is preserved (pink), quadrant II, IV collapsed onto S_1 and S_2 axis respectively (thick green, blue lines), and III mapped onto origin (red). **(B)** Plotting $\tilde{\sigma}^2(\mu/\sigma)$ against $\tilde{\rho}(\mu/\sigma, \rho)$ there is a positive relationship because both are increasing with μ/σ. $\tilde{\rho}$ is transformed to Fisher's z and $\tilde{\sigma}^2$ plotted on log axis revealing an approximate relationship: $a(\tilde{\sigma}^2)^b = z(\tilde{\rho})$.

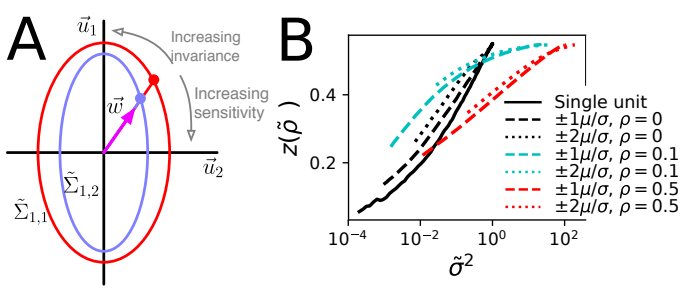

Figure 3: $\tilde{\rho}'$ to $\tilde{\sigma}^2{}'$ relationship in unrectified unit resulting from integrating over rectified subunits. **(A)** Schematic intuition of $\tilde{\rho}'$ to $\tilde{\sigma}^2{}'$ relationship. In blue $\tilde{\Sigma}_{1,2}$, in red $\tilde{\Sigma}_{1,1}$, pink vector is $\vec{w}$. Axes are eigenvectors with vertical axis having larger associated eigenvalue. Ratio of $\vec{w}^T\tilde{\Sigma}_{1,2}\vec{w}$ and $\vec{w}^T\tilde{\Sigma}_{1,1}\vec{w}$ increases as $\vec{w}$ points in direction of greater variation of both. **(B)** Simulation of integrating over populations of rectified subunits varying average μ/σ (within a trace), subunit correlation (ρ across colors), and neighborhood size of μ/σ averaged over (line pattern).