Visibility of eigen-distortions of hierarchical models

Alexander Berardino, Valero Laparra, Johannes Ballé & Eero P. Simoncelli

We compare several models of visual representation in terms of their ability to predict human judgements of visual distortion. Each model is defined by a differentiable mapping from image inputs to a response vector, which is then corrupted by additive Gaussian noise. We use the Fisher Information matrix to predict discrimination thresholds for the visibility of arbitrary distortions, up to an unknown scale factor. We test this by generating, for each model, a pair of specific distortions corresponding to the largest and smallest eigenvectors of the Fisher Information matrix, which represent the model-predicted most and least noticeable changes to the image, respectively. We distort the image by adding multiples of each vector, and measure detection thresholds for human subjects in a two-alternative forced-choice task. Results are quantified using the difference, D, of the log vector length at threshold for the two extremal eigenvectors, $T(\vec{v}_n)$ and $T(\vec{w}_n)$. Two random perturbation vectors would yield of value of approximately $D = 0$, and larger values of D indicate that the sensitivity of the model is better aligned with that of humans. We used this methodology to test three models: a simple model of LGN neurons that includes local luminance and contrast normalization[1, 4], and two different 4-stage convolutional neural networks. Each model was trained on a database of human perceptual judgments[5]. We found that, despite performing slightly better in terms of cross-validated correlation with the database, both artificial neural networks performed much worse than the LGN model on this test. We conclude that in this situation, where data are somewhat scarce, cross validation is not powerful enough to expose failures of a particular model class. Our method provides a more general form of cross-validation, based on the synthesis of model-optimized stimuli and comparison with human judgments on these targeted stimuli, and ensures that the model generalizes beyond curated data.

Model details

The first model, which we call "On-Off", mimics computations carried out in the retina and LGN[4], and is similar to that described in [1]. In brief, the model has two channels (On and Off), each with the same architecture. The image is convolved with a difference of gaussians (DoG) filter, and divisively normalized by a measure of the local pixel luminance. The outputs of this stage are then divisively normalized by a measure of local contrast within each channel, and rectified with a "softplus" nonlinearity, $f(x) = ln(1 + e^x)$. The outputs of the two channels are concatenated into a single response vector. Model parameters – the variance of each gaussian in the DoG filter, as well as the filters used to compute local luminance and contrast – are optimized to maximize the correlation of the euclidean distance between model responses to distorted images and their corresponding original images, and human judgments of the perceptual distance between these same image pairs, as reported in the TID 2008 database [5]. In addition, we trained two 4-layer neural networks that shared the same overall architecture (each layer is composed of convolution with multiple filters, followed by softplus rectification), but utilized different forms of regularization. The first network includes batch normalization (BN), which normalizes the outputs of each hidden layer by the global mean and variance of the outputs at that layer on each batch of training data[2]. After training, these normalization parameters are fixed to the global mean and variance across the training set. The second network utilizes a fixed form of local response normalization (LRN) based on the weighted Euclidean norm of neighboring coefficients, computed with a fixed weight kernel[3]. We trained the filters of both convolutional neural networks on the same objective function and data as above.

Synthesizing Eigen-distortions from the Fisher information matrix

The Fisher information (FI) matrix of a model, $J(\vec{x}_n)$, for even moderately sized images, is too large to be stored in memory. Nevertheless, the FI can be applied to an image, and we can solve for the maximum and minimum eigenvectors using the power iteration method. Starting with a noise vector $\vec{v}_n{}^{(0)}$, we iteratively apply the FI, and renormalize the resulting vector, until it converges to the largest eigenvector, $\vec{v}_n$. We then perform a second iteration, subtracting the eigenvalue of the maximum eigenvector, λ_n on each iteration, until convergence to the smallest eigenvector, $\vec{w}_n$:

$$\vec{v}_n^{(k+1)} = \frac{J(\vec{x}_n)\vec{v}_n^{(k)}}{||J(\vec{x}_n)\vec{v}_n^{(k)}||}, \qquad \vec{w}_n^{(k+1)} = \frac{(J(\vec{x}_n) - \lambda_n I)\ \vec{w}_n^{(k)}}{||(J(\vec{x}_n) - \lambda_n I)\ \vec{w}_n^{(k)}||} \tag{1}$$

$$D = \frac{1}{N}\sum_{n=1}^{N} log_2||T(\vec{w}_n)||_2 - log_2||T(\vec{v}_n)||_2 \tag{2}$$

Results

	On-Off	BN	LRN
R	0.83	**0.86**	**0.86**
D	**8.57**	3.68	2.19

Table 1: First Row: Cross-validated Pearson correlation between model distance and human mean opinion scores for the TID 2008 dataset[5]. **Second Row:** Average log ratio of vector lengths at threshold for the least noticeable over the most noticeable eigenvectors.

We measured perceptual thresholds for each eigen-distortion, $T(\vec{w}_n)$ and $T(\vec{v}_n)$, for 6 images from the Kodak test set on 1 subject. Specifically, we used a two-alternative forced-choice task to estimate the amplitude of each of the eigenvectors $\vec{v}_n$ and $\vec{w}_n$ that, when added to image $\vec{x}_n$, could be discriminated from the original image with 75% accuracy. We compute D, our measure of how well each model's eigen-distortions align with human perception, from the average difference of log thresholds across images (eq. (2)). Eigen-distortions from the two best-performing models are shown in Figure 1. Results for both traditional cross-validation and our eigen-distortion test are reported in Table 1. We find that despite slightly outperforming the On-Off model (as well as the state-of-the-art) when evaluated on a held out test set from the TID 2008 database, both the BN and LRN networks drastically underperform the On-Off model in predicting human sensitivity. Additionally, the two neural networks achieve equal correlation, while having different values of D, suggesting that even within a model class, traditional cross-validation is not powerful enough to elucidate the failures of particular model assumptions.

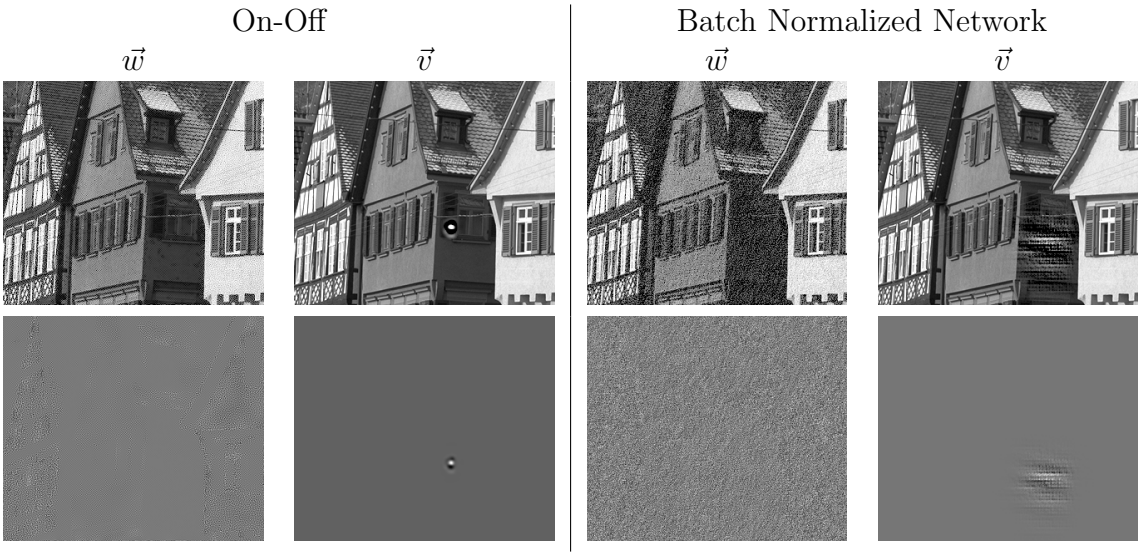

Figure 1:
Left: Least- and most-noticeable Eigen-distortions for the On-Off model, added to the original image (**top**) and alone (**bottom**).
Right: Least- and most-noticeable Eigen-distortions for the BN network, added to the original image (**top**), and alone (**bottom**).

References

[1] A. Berardino, V. Lappara, J. Ballé, and E.P. Simoncelli. Perceptual distortion measured with a gain control model of LGN response. In *Computational and Systems Neuroscience (CoSyNe)*, Feb 2016.

[2] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv.org*, Feb 2015.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS 2012: Neural Information Processing Systems*, pages 1–9, Nov 2012.

[4] Valerio Mante, Vincent Bonin, and Matteo Carandini. Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. *Neuron*, 58(4):625–638, May 2008.

[5] N Ponomarenko, V Lukin, and A Zelensky. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern ...*, 2009.