Emulating Brightness Illusions with Factored 3-Way Boltzmann Machines

Eric Weiss

It is well known that the perceived brightness of a patch of an image depends on its surroundings. For example, White's Illusion (shown below) consists of an arrangement of black, white, and gray bars. Despite having the same luminance, some gray bars appear brighter, while others appear darker. There is a variety of these images, each demonstrating a unique kind of illusion. The existence of these illusions suggests that visual perception is not simply the result of translating the light that enters the eye directly into conscious experience; rather, some internal processing must first be modulating the visual signal before it is perceived. By characterizing the ways in which that internal processing fine-tunes information coming from the eyes, we should be able to learn something about how the human visual system works and gain insight into the nature of perception in general.

There have been a variety of attempts to explain these brightness illusions through reasoning based on our current understanding of early visual cortex, leading to a number of image processing algorithms that “see” some of the same kinds of brightness illusions that humans see. Some of these algorithms are designed according to psychophysical and neurobiological principles, whereas others are based on some analysis of the statistical properties of natural images. There are also a number of high-level models of brightness perception that involve segmentation of the visual scene into distinct perceptual objects, and while high-level mental processing no doubt has a role in brightness perception, here I will be concerned only with low-level processing.

While there is no consensus on what exactly low-level visual processes compute, it is generally thought that low-level visual machinery preconditions the visual signal in some way

that makes it more easily interpretable by higher level cognitive processes. One such operation could be a shifting of the brightness scale in localized regions of the image, as is described by Anchoring theory (Gilchrist et al., 1999). Others involve contrast enhancement or brightness inference as a way to resolve ambiguous stimuli (Corney and Lotto, 2007, Dakin and Bex, 2003). Brightness illusions are then ascribed to “errors” in perceived brightness levels created through these preprocessing steps.

While I will not go too deeply into describing the various models of brightness perception, I will briefly explain one such model, FLODOG (Robinson et al., 2007) as it has been an important motivating factor in my work. This model works by first convolving an entire image with a set of oriented Difference-of-Gaussian filters, which closely resemble the Gabor-like filters thought to be implemented by neurons in low-level visual cortex. In effect, these filters act as oriented edge detectors that are also sensitive to spatial frequency. After obtaining a set of filter responses, filters that are similar in orientation, spatial frequency, and location are allowed to inhibit each other in a way that mimics lateral inhibition between cortical neurons, producing a “sharpened” representation of the edges that make up the image. The original image is then reconstructed from the adjusted filter activities, and the new brightness values are compared to those perceived by humans viewing the same images. In addition to being able to predict a large number of brightness illusions, every operation involved in the FLODOG algorithm appears to be biologically feasible – the function of the filters appears to closely match that of many early visual cortical neurons, and all interactions between variables are local, in that one could design a topographical implementation of the model that involved no global operations. The locality-of-interactions constraint is motivated by the limited size of the dendritic trees of cortical neurons. Still, there are a few known brightness illusions for which FLODOG makes incorrect predictions. It seems likely that this could be due to incorrect assumptions about the shapes of the filters and the patterns of inhibition (that is, they might exhibit different behaviors than their human analogues), which are hard-coded into FLODOG.

Ranzato and Hinton (2010) have developed a model for natural scenes that bears several similarities to the FLODOG model. Their model is called the Factored 3-Way Boltzmann Machine, but I will refer to it as the RBM from now on. It is designed with the goal of object recognition in mind, and it performs very well on standard visual recognition tasks. Presumably, object recognition is the ultimate goal of the human visual system, so a model of brightness

perception should compute something that is at least somewhat useful to object recognition. This, the structural similarities that the RBM bears to FLODOG, and the fact that the Boltzmann machine *learns* those structures by analyzing natural images makes it a very attractive candidate for explaining brightness illusions. In addition, the RBM algorithm appears to be biologically plausible, and the expert model ("expert" in that it has been trained on real data) exhibits several similarities to early visual cortex. I will first describe the model and explain how it learns, and then illustrate how the expert RBM resembles the early visual cortex. Then I will offer a way to apply the expert RBM in a FLODOG-like manner to explain brightness illusions.

The Boltzmann Machine

The model I am interested in is a member of the class of models referred to as Boltzmann machines. These models are named so because they were inspired by the Boltzmann distribution of statistical mechanics and borrow the concept of free energy to define the model's behavior. The model is a type of artificial neural network, containing units (or nodes) and connections between the units. Each unit has a defined range of states; some units may be allowed to take on a continuum of values while others may be limited to a discreet set of values. Given a particular configuration of all of the states of the units and the connection weights, we can define the energy of that state as a measure of the "disagreement" between the states of the units, defined by the magnitudes and signs of their bidirectional connection weights. High free energy corresponds to very unlikely states, and the opposite is true for states with low free energy. It is common to separate the units into distinct layers, with no connections between units belonging to the same layer. The model I am interested in has essentially three layers: a visible layer that corresponds to real pixel values, and two hidden layers, with the first hidden layer projecting to the second.

The Factorized 3-Way RBM and FLODOG

The simplest Boltzmann machines only involve interactions between pairs of units. It is possible to write a new energy function for the Boltzmann Machine that involves products of triplets of units (Sejnowski, 1986). This increases the model's ability to model complex patterns, but drastically increases the number of parameters. To keep the number of parameters under control, one can design a network with two modified hidden layers. The first hidden layer

receives “factorized” connections from the visible units, which are effectively the squared responses of linear filters applied to the pixel values. The next hidden layer connects to the factorized units in the usual way, with pairwise connections. (for more information on this type of model see Ranzato 2010). It has been found that after training this kind of model on natural images, the linear filters in the first hidden layer learn Gabor-like filters that detect oriented edges of various spatial frequencies. In addition, the higher level units learn to receive inputs from filters that are similar in orientation, spatial frequency and location in the visual field, introducing a degree of translational invariance to the model's representation of an image. In this way it bears similarity to FLODOG, which also involves interactions between similar Gabor filters. It also resembles the simple cell / complex cell structure of early visual cortex, with the linear filters acting like simple cells to detect oriented edges, and the higher level units acting like complex cells that pool the responses of many similar simple cells.

Emulating Mutual Inhibition with Soft k-Winner-Take-All

As I have explained, there exist several similarities to FLODOG that make the RBM an attractive candidate for modeling brightness perception. The RBM does not, however, implement any kind of inhibition between the filters or “complex” units. In fact, the model explicitly forbids any interaction between units belonging to the same layer. Such a process is certainly relevant to brightness perception, though, as is demonstrated by the success of FLODOG and by the fact that cortical neurons mutually inhibit one another. To emulate this inhibitory system I artificially introduce k-Winner-Take-All (kWTA) to the activities of the complex units, where only the most active k units are allowed to stay fully active. The remainder of the units become nearly inactive. This mimics cortical mechanisms that maintain a constant average level of neural activity. It has also been shown that a single layer in a neural network that implements a kWTA-like algorithm has the same computational power as multiple layers of units that compute simple threshold functions (Maass 2000). Therefore the application of kWTA to the activities of the units in the RBM is not wholly unjustified, despite it not being part of the original model. (I will discuss this more in the section on future work).

In my experiments I use two types of kWTA: “soft” and “hard”. The soft method involves applying a sigmoid function to the total activations that each complex unit receives from the simple units. The sigmoid function is shifted as to be centered on the mean of the activations plus

an offset, which effectively controls “k” by determining how many of the units have high enough activations to fall on the higher side of the sigmoid curve. The sharpness of the sigmoid function is scaled by the standard deviation of the complex units' original activations and by a sharpness parameter. Explicitly, this is:

$$A_{new} = \frac{1}{1 + e^{\frac{-(A_{old} - (\mu + k\sigma))}{s\sigma}}}$$

where A_{new} is the adjusted activation, A_{old} is the pre-kWTA activation, and μ and σ are the mean and standard deviation of the original (pre-kWTA) activations. Positive k will cause some units to become inactive after this function is applied, while negative k may cause some weakly activated units to become more excited. The sharpness parameter, s, determines how many units end up with intermediate levels of activation. Small s causes every unit's activation to converge to either zero or one, and large s makes every unit's activity approach 0.5. By adjusting these two parameters it is possible to partially control the total average activation of the layers of the RBM. (If the distribution of the units' pre-activations is highly skewed, the effect of this method may not resemble that of true kWTA. A better version might center the sigmoid function around the median of the data, rather than the mean.)

A second type of kWTA may be implemented by simply allowing the most active k units to become fully active and inactivating the rest – I will call this hard kWTA. This method guarantees that exactly k units will be fully active, regardless of the distribution of the units' pre-activations. It is unclear whether or not such a rigid normalization scheme will enhance or diminish the model's explanatory power, so in my experiments I investigate the effects of both types of kWTA: the soft type introduced in the previous paragraph, and the hard type I just described.

Predicting Brightness Illusions

Before we can start simulating human brightness perception, we must first define what characterizes a visual percept. Does visual awareness consist of an enormous array of pixel-like points, or does it consist of a high-level scene containing a small number of complex objects? Presumably the former is not true, as it would require an incredible capacity for attention to attend to every pixel in our visual field. The latter seems far more plausible, but is not always the

case; we certainly have the power to focus on the minute details of a small region of our visual field if we choose to do so. This suggests that the types of percepts that are accessible to high-level cognitive mechanisms can vary depending on one's attentional state. It is not unreasonable, then, to assume that when a subject is presented with a brightness judgment task, they are tapping into a relatively low-level area in the visual processing hierarchy. But perhaps there is a limit: one probably cannot gain conscious access to the activities of single retinal ganglion cells, and maybe even not to single simple or complex cells in V1. While I am not aware of any neurobiological evidence that supports such a claim, the fact that our everyday tasks almost always deal with high-level representations of the objects in our vision, and very rarely with details as minute as those detected by individual simple cells, it is not unreasonable to assume that we only partially develop the ability to consciously access the earliest levels of our visual processing stream. We don't normally have much use for such information; it is better kept “under the hood” as are many subconscious processes. The question of where exactly this perceptual boundary lies, however, is not one that can be easily addressed through high-level reasoning about vision. In the next section I propose a way in which the RBM-based model (trained on real images) might help to resolve this ambiguity.

The RBM-based model decomposes images into two levels of abstraction: the Gabor-detecting “simple” units and the “complex” units that pool the activities of the simple units. There are a few ways one might then use the model to emulate brightness illusions. The simplest method might be to propagate activity to the complex units and then reconstruct the image with a downward pass. This corresponds to a model of visual perception where the activations of the simple cells do not directly influence visual consciousness, in that the activities of the complex cells are the only things that would matter in some visual discrimination task such as the judgment of brightness. It is important to note that to make any reconstruction with the RBM model, one must save the signs of the Gabor responses and use them again while making a reconstruction. The complex units receive input proportional to the square of the raw Gabor activations (which are dot products, and can therefore be negative), so some information about the original image is lost while propagating activity to the complex units. Any model of brightness perception based on the RBM that involves reconstructing images must therefore assume that we are, at very least, consciously aware of the signs of the raw Gabor responses. Another method might involve an upward pass where the original activations of the simple units

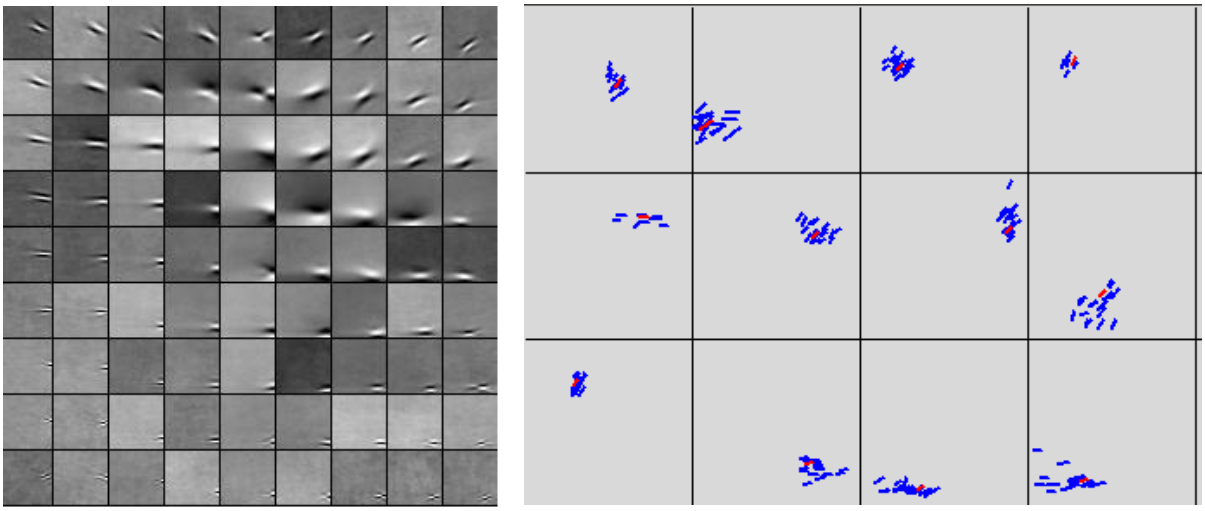
(in addition to their signs) are combined with the activities of the complex units while making the reconstruction. This corresponds to a model of brightness perception in which we do indeed have conscious access to the activities of individual simple cells. Determining which of these methods most accurately predicts brightness illusions may provide a way to address the question raised in the previous section: what is the lowest level of the visual processing stream to which we can gain conscious access? One could also implement either of the previous methods but apply soft kWTA to the activities of either the simple units, the complex units, or both before making reconstructions. This may provide insight into the role of kWTA-like neural competition in the formation of visual perception and in cortical processing in general.

Details of Model and Data

In my work, I train a Factored 3-Way Boltzmann Machine nearly identical to the one described by Ranzato and Hinton on a dataset consisting of 500,000 natural images of resolution 32 by 32 pixels, sampled randomly from the Berkeley Segmentation Dataset (Martin et al., 2001). I preprocess the images with ZCA whitening, giving each principal component unit variance. The first hidden layer contains 4096 units, four times the number of units in the visible layer. The second hidden layer contains 1024 units. The weight matrix between the visible layer and first hidden layer is initialized to random values, while the weights between the first and second hidden layers are initialized with a specific topography as in (Osindero, 2005) designed to encourage similar filters to develop near each other (this is not necessary, but aids in visualizing the model). I first train the filters while leaving the higher level inhibitory connections fixed, and once they have converged, I train the higher level connections until they too have converged, while holding the filters constant. (I have found this layer-wise training to be critical to the formation of a good set of Gabor-like filters and localized complex unit connectivity patterns. Due to the nature of the model and learning algorithm, allowing both layers to learn simultaneously causes most of the weights to converge to zero early on during training, resulting in very small learning gradients and therefore very small weight updates. I have not seen any model recover from this state in a reasonable amount of time.) I use the Theano Python library (Bergstra et al.) to accelerate the training process by implementing it on an nVidia Tesla C1060 GPU processor.

After training is complete, it is possible to view the development of a simple/complex

cell-like architecture. As expected, the first layer learns a set of Gabor-like filters of varying orientations, spatial frequencies, and locations. The second layer pools the responses of similar Gabors. This is visualized below. (Compare to the figures presented in Ranzato 2010)



The image on the left shows the connection weights of a small subset of the first hidden layer. It is evident that most of them closely resemble Gabors, with a few having slightly more unique shapes. The image on the right provides a visualization of the connectivity patterns for several units in the second hidden layer. Each square represents one complex unit (in the second hidden layer), and each of the blue bars within those squares represent simple units (in the first hidden layer) to which those complex units are strongly connected. The sizes, orientations, and positions of the lines represent the spatial frequencies, orientations, and locations of the Gabor-like filters computed by those simple units. The red line represents the simple unit to which each of those complex units is most strongly connected. It is evident that the complex units group together simple units that are similar in location, spatial frequency, and orientation.

Experiments

I use the model to create reconstructions of every illusion tested in (Robinson & de Sa 2007) and on the double increment and double decrement versions of White's illusions. Each image is represented by an array of brightness values ranging from 0 to 255, as was used to train the model. The model is defined on a 32-by-32-pixel patch, but all of the test images are at least 64 by 64, so to create reconstructions I compute a reconstruction at every 32 by 32 pixel patch within each test image and average the resulting pixel values. This preserves the model's

biological plausibility, as it limits any interactions between simple or complex units to a small 32 by 32-pixel neighborhood.

To make a single reconstruction, the 32-by-32-pixel region of the image in question is first preprocessed with ZCA whitening. The whitened image is presented to the model and the activations of the simple and complex units are computed in a deterministic upward pass (saving the signs of the simple units' pre-activations). There are then three ways in which the reconstruction can continue:

Complex units only, no kWTA: The raw activations of the complex units are left unaltered, the precise activations of the simple units are discarded (save their signs) and a downwards pass followed by inverse ZCA whitening creates a reconstructed image.

Complex units only, with kWTA: Exactly the same as the first method, except kWTA (one of the two types described earlier) is used to modulate the activities of the complex units before computing the downwards pass.

Complex and simple units, with kWTA: The exact original activations of the simple units are used while computing a reconstruction. I proceed as in the previous model, adjusting the complex units' activities with kWTA, and then adjust the simple units' activities in proportion to the changes in the activities of the complex units (to which they connect) induced by kWTA, and in proportion to the strength of their connections to those complex units. A downward pass using the simple units' new activities is used to create the reconstruction.

To compute the predicted strength of each illusion, I first take the average reconstructed pixel values of the regions defined by the gray spots in the original images. I then take the difference of those values to determine the illusion's overall strength and direction.

Results

The predictions for each illusion are tabulated below. A positive strength means that the predicted direction of the illusion matched that of humans, and negative implies the opposite. The strength of each illusion is scaled relative to White's illusion, to which I assign strength 1. I

used k=0.1, s=0.02 for the soft kWTA, and k=650 for hard kWTA. Out of a small set of parameter settings (hand chosen by me) these values were found to result in the greatest number of correct predictions.

Illusion	no kWTA	soft kWTA	+simple	hard kWTA	+simple
Anderson	0.21	**-2.81**	**-1.96**	0.10	**-2.25**
Bernary_Cross	0.89	2.78	**-0.24**	0.87	0.15
Bullseye-thick	1.46	7.24	**-0.48**	1.47	0.16
Bullseye-thin	**-0.06**	8.31	6.52	0.02	8.63
Checker-small	2.61	11.15	**-0.78**	2.55	**-0.05**
Checker-medium	0.47	3.94	0.37	0.46	0.33
Checker-large	0.93	5.17	0.14	0.91	0.30
Corrugated Mondrian	0.39	1.80	0.30	0.37	0.50
Double Decrement	**-0.26**	0.50	1.29	**-0.23**	1.17
Double Increment	**-0.05**	1.12	0.79	**-0.04**	0.57
Howe	0.40	8.60	2.70	0.49	2.64
Radial-thin	0.42	1.67	**-0.35**	0.38	**-0.38**
Radial-wide	0.68	2.20	**-0.37**	0.67	**-0.01**
Rings-medium	**-0.63**	**-3.79**	**-0.40**	**-0.64**	**-0.62**
Rings-thick	**-1.17**	**-5.07**	**-0.16**	**-1.16**	**-1.01**
Rings-thin	**-0.19**	**-2.13**	**-0.72**	**-0.20**	**-0.75**
SBC-big	3.82	14.52	1.85	3.70	4.64
SBC-small	6.08	16.48	0.36	5.80	4.54
Todorovic-Bernary_1-2	1.29	3.93	**-0.77**	1.27	0.16
Todorovic-Bernary_3-4	1.61	5.60	**-0.94**	1.56	**-0.06**
Todorovic-equal	1.57	5.46	0.65	1.57	2.20
Todorovic-in-large	**-0.96**	**-3.61**	**-1.23**	**-0.93**	**-2.12**
Todorovic-in-small	0.50	1.12	**-1.21**	0.46	**-1.42**
Todorovic-out	**-0.98**	**-4.04**	**-1.02**	**-0.97**	**-2.11**
Whites	1.00	1.00	**-1.00**	1.00	**-1.00**
Whites-high frequency	2.63	9.16	**-2.08**	2.53	**-2.02**
Zigzag	**-0.06**	**-1.17**	**-0.55**	**-0.06**	**-0.76**

Surprisingly, simply doing an up-down pass without applying kWTA correctly predicts many illusions. Adding kWTA helps a little, with soft kWTA performing slightly better than hard kWTA. Recombining the original simple unit activations with the modified complex unit activations performs significantly worse, failing even to predict White's illusion. No version of the model correctly predicts the ring, zigzag, Todorovic-in-large, nor Todorovic-out illusions.

Discussion

The success of the models that use solely the complex unit activations when making

reconstructions supports the hypothesis that visual perception is constructed from the kinds of patterns detected by complex cells. This offers a new kind of explanation for brightness illusions: namely, that they are caused by the loss of information from early levels in the visual processing stream, or rather that some of the information in low level visual processing areas is not accessible to higher level cognitive processes, and it is this lack of information that leads to illusions of brightness. Lateral inhibition still seems to play an important role, though, as is suggested by the improvements in the model's predictions brought about by introducing kWTA. The fact that soft kWTA performs a little better than hard kWTA suggests that having a less rigidly enforced “k” is important for explaining certain illusions, or that it is important for some complex units to take on intermediate levels of activation, in addition to fully “on” or “off”. It is not too surprising that the model fails to predict some illusions that FLODOG is able to explain - the model has no way to implement competition between similar Gabor detectors (kWTA causes *all* of the complex units to compete for activation, making it possible for two very dissimilar filters to inhibit each other), and the failure of the model to predict certain illusions that depend heavily on FLODOG-like competition between similar Gabor detectors (such as the ring illusion) reconfirms the importance of this kind of selective inhibition to producing brightness illusions.

Future Work

There are a few ways one might improve the RBM-based model's predictive power. The model is fairly sensitive to the particular selection of kWTA parameters, and I have little doubt that the values I chose are suboptimal. Integrating the kWTA mechanism into the RBM from the start and allowing the parameters to be learned during training seems like the most reasonable way to find better values. The accompanying changes in the dynamics of the model might also result in subtle changes in the simple unit filters and complex unit connectivity patterns, possibly leading to a better set of Gabor filters and complex units. Also, one might divide the complex units into many “kWTA groups”, small groups of units that share the same kWTA mechanism. This would give the model a way to implement selective competition, as long as similar complex units end up being grouped together in the same kWTA groups.

Acknowledgements

Many thanks to Virginia de Sa and Alan Robinson for guidance and useful discussions. I also thank Rafael Nunez and the Cognitive Science honors students for their helpful suggestions.

References

D. Corney and R.B. Lotto. What are lightness illusions and why do we see them. PLoS Comput Biol, 3(9):e180, 2007.

S.C. Dakin and P.J. Bex. Natural image statistics mediate brightness "filling in". Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(1531):2341, 2003.

A. Gilchrist, C. Kossyfidis, F. Bonato, T. Agostini, J. Cataliotti, X. Li, B. Spehar, V. Annan, and E. Economou. An anchoring theory of lightness perception. PSYCHOLOGICAL REVIEW-NEW YORK-, 106:795– 834, 1999.

G.E. Hinton. Training products of experts by minimizing contrastive divergence. Neural Computation, 14(8):1771–1800, 2002.

A.K. Marc'Aurelio Ranzato and G. Hinton. Factored 3-way restricted boltzmann machines for modeling natural images, 2010.

G.E.H. Marc'Aurelio Ranzato. Modeling pixel means and covariances using factorized third-order boltzmann machines. 2010.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. 2001.

Simon Osindero, Max Welling, and Geoffrey E. Hinton. Topographic product models applied to natural scene statistics. Neural Computation, page 2006, 2005.

A.E. Robinson, P.S. Hammon, and V.R. de Sa. Explaining brightness illusions using spatial filtering and local response normalization. Vision research, 47(12):1631–1644, 2007.

T.J. Sejnowski. Higher-order Boltzmann machines. In AIP Conference Proceedings, volume 151, pages 398–403. Citeseer, 1986.

Maass, W.: On the computational power of winner-take-all, Neural Computation 12(11), volume 12, MIT Press, 2519–2535, 2000

J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-

Farley and Y. Bengio. “Theano: A CPU and GPU Math Expression Compiler”.Proceedings of the Python for Scientific Computing Conference (SciPy) 2010. June 30 - July 3, Austin, TX