The Wilson Machine for Image Modeling

Saeed Saremi SAEED@SALK.EDU
Terrence J. Sejnowski TERRY@SALK.EDU
Salk Institute, 10010 N Torrey Pines Rd, La Jolla, CA 92037

Abstract

Learning the distribution of natural images is one of the hardest problems in machine learning. We break down this challenging problem by mapping images into a hierarchy of binary images (bit-planes). In this representation, the top bit-plane is *critical*, having fluctuations in structures over a vast range of scales. The ones below go through a gradual stochastic heating process to disorder. We turn this representation into a *directed* probabilistic graphical model, transforming the learning problem into the unsupervised learning of the distribution of the critical bit-plane and the supervised learning of the conditional distributions for the remaining bit-planes. We learnt the conditional distributions by logistic regression in a convolutional architecture. Conditioned on the critical binary image, this simple architecture can generate large natural-looking images with many shades of gray, without the use of hidden units.

1. Introduction

Learning the distribution of natural images remains one of the hardest problems in unsupervised learning. Structures in natural images are complex, varied, and most importantly span a vast range of scales, from the smallest within a few dozen pixels to large structures the size of the image itself. In this respect, they are very similar to critical points, realized for physical systems near continuous (second-order) phase transitions (Wilson, 1979). There is a deep connection underlying this similarity as critical large-scale fluctuations emerge after mapping natural images to a stack of binary images (Saremi & Sejnowski, 2013; 2014; 2015). In this work, we introduce an algorithm to utilize this criticality in learning the distribution of natural images.

The binary representation of (Saremi & Sejnowski, 2013) contained ordered and disordered phases, with a critical bit-plane close to a phase transition. It was interpreted that the existence of the critical bit-plane underlies the large-scale fluctuations of the structures in natural images. Here, we build on those results and exploit the binary representation, turning it into a directed probabilistic graphical model by placing the critical bit-plane at the root of the directed graph. This transforms learning the distribution of natural images into the *unsupervised* learning of the distribution of the critical bit-plane and the *supervised* learning of the conditional distributions of all other bit-planes in a parent-child hierarchy outlined below. The network architecture for learning is simple. We approximate the conditional distributions by logistic regression, where the weights to the children nodes are learnt in a convolutional architecture by weight-sharing.

Turning unsupervised learning into supervised learning has a rich history and goes back to the wake-sleep algorithm and the Helmholtz machine, which trained inference and generative models against each other (Hinton et al., 1995; Dayan et al., 1995). Here, in contrast, the supervision is given by the input itself. Instead of learning hierarchical representations with hidden units (LeCun et al., 2015; Schmidhuber, 2015), we show that the "hierarchy" of structures "hidden" in the data itself can be utilized for learning.

The starting point is to construct an exact binary representation so that the top layer is critical and the bottom ones go through a stochastic heating process to disorder. In the bit representation that was studied in (Saremi & Sejnowski, 2013; 2015), the critical bit-plane was in the middle of the hierarchy, with different types of structures emerging in relation to the critical bit-plane, "ordered" towards the top, and "disordered" towards the bottom. It is not clear in that representation how the layers just above and just below the critical layer could influence each other in a directed graph. However in the new representation, the causal direction is clear: the stochastic heating process from criticality (top) to disorder (bottom).

2. Binary Representations for Images

In this section, we present a new binary representation for images compared to what was presented in (Saremi & Sejnowski, 2013). The new representation is a tweak on the original but it turns out to be key in how we map the binary representation to a directed graphical model. We first give a brief review of the original representation.

The analog values I (assumed to be non-negative integers for simplicity) were mapped to a binary representation $\{B_1, B_2, \cdots, B_\Lambda\}$ by the following decomposition:

$$I = \sum_{\lambda=1}^{\Lambda} 2^{\Lambda-\lambda}\, B_\lambda, \tag{1}$$

where $B_\lambda \in \{0, 1\}$ was found iteratively by evaluating $\lfloor (I - \sum_{l=1}^{\lambda-1} 2^{\Lambda-l}\, B_l)/2^{\Lambda-\lambda} \rfloor$ starting from $\lambda = 1$. The visual representation of the map from analog pixel values to its corresponding bits is given in Fig. 1. For gray-scale images $\mathcal{I}$, Eq. 1 is replaced with matrices of binary images:

$$\mathbb{B} = \{\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_\Lambda\}, \tag{2}$$

mapping $\mathcal{I}$ to a stack of bit-planes $\mathbb{B}$.

The plan is to make this representation into the following probabilistic graphical model,

$$P(\mathbb{B}) = P(\mathcal{B}_1)P(\mathcal{B}_2|\mathcal{B}_1)\cdots P(\mathcal{B}_\Lambda|\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_{\Lambda-1}). \tag{3}$$

However, to find good estimates for the conditional distributions, the structure of the directed graph should be dictated by a "physical" causal process. This is what we achieve by changing the original binary representation.

One of the main conclusions of (Saremi & Sejnowski, 2013; 2015) was that there exists a critical bit-plane in the bit representation hierarchy, which underlies the scale invariance of natural images. Therefore in building the directed graph the critical bit-plane must be the root and the parent of other bit-planes. In the representation $\mathbb{B}$, the critical bit-plane is in the middle ($\mathcal{B}_6$ in the van Hateren database) and is surrounded by ordered/cold, and disordered/hot phases. That representation cannot be made into a directed graph, since the causal direction from the bit-planes just above and just below the phase transition cannot be deduced. In a simple tweak, we change the binary representation so that $\mathcal{B}_1$ is critical and the bit-planes below go through a gradual heating process to disorder. The parent-child hierarchy becomes clear: from criticality to disorder.

The change in the representation is illustrated in Fig. 1; we simply change all the half-point dividing lines to the median values of their corresponding intervals. This hierarchical median-thresholding is *exactly* equivalent to applying *half-point* thresholding of Fig. 1 after first performing

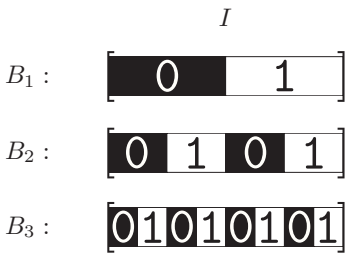

Figure 1. The binary representation for the $\Lambda-$bit integer value I in the range $[0, 2^\Lambda - 1]$. The first three bits is shown schematically here, where each bit divide the interval in half iteratively starting from the most significant bit B_1.

histogram equalization by the cumulative integral of the marginal distribution of the pixel intensities $P(I)$:

$$I \to I' = \int_{-\infty}^{I} P(x)dx, . \tag{4}$$

By definition, I' is in the range $[0, 1]$. The half-point thresholding of the histogram-equalized image is the same as median-thresholding of the original image because $P(x) \geq 0$ and therefore the transformation of Eq. 4 does not change the rank of entries: $\mathcal{I}_k > \mathcal{I}_l \Leftrightarrow \mathcal{I}'_k > \mathcal{I}'_l$ for two pixel locations k and l. The illustration of this new binary representation for an image in the Geisler database (Geisler & Perry, 2011) is given in Fig. 2.

The first bit in the new representation $\mathcal{B}_1 = \mathcal{I} > \texttt{median}(\mathcal{I})$ is the median-thresholded image, which was studied at length in (Stephens et al., 2013; Saremi & Sejnowski, 2014). The median-thresholded images were also analyzed using percolation theory, which studies connected clusters on graphs (Saremi & Sejnowski, 2015). It was shown that they are at the onset of a percolation transition, and their criticality is governed by percolating clusters. In addition, it was demonstrated that $\mathcal{B}_1$ contains connected clusters on all length scales and are governed by scaling laws. These rich connected clusters at top of the graphical model provide the fuel to the algorithm presented here. We refer the reader to (Stephens et al., 2013; Saremi & Sejnowski, 2014; 2015) for detailed analyses on the richness of the median-thresholded natural images.

What we gain in the new representation is what we planned to achieve in that the bit-planes $\mathcal{B}_\lambda$ go through a gradual stochastic heating process to disorder as λ increases (see Fig. 2), which naturally places them in the parent-child hierarchy of the directed graphical model of Eq. 3. In summary, in the new representation $\mathcal{B}_1 = \mathcal{I} > \texttt{median}(\mathcal{I})$ is critical, at the root of the graph, and $\mathcal{B}_\lambda$ is the parent of $\mathcal{B}_{\lambda'}$ for $\lambda < \lambda'$. This hierarchy is the stochastic process to disorder, seen in Fig. 2 by islands of connected clusters gradually dissolving in a sea of noise. With this construction

of the graphical model, rooted in statistical physics and in the theory of critical phenomena, we can hope to find good estimates for conditional distributions of Eq. 3.

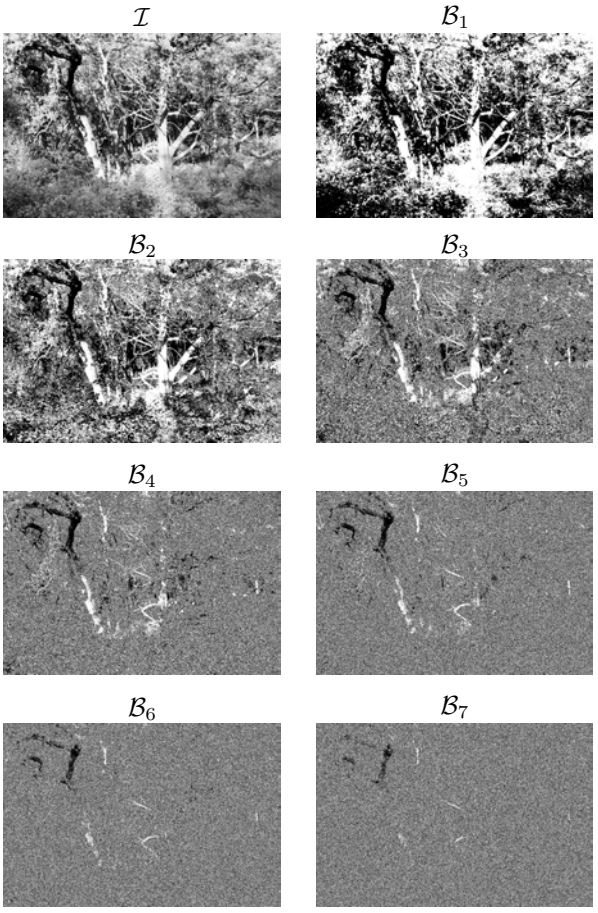

Figure 2. An image $\mathcal{I}$ (2844×4284 pixels) in the Geisler database of natural images and its binary representation up to bit-plane $\mathcal{B}_7$ are shown. The critical bit-plane $\mathcal{B}_1$ goes through a gradual stochastic heating process to disorder.

3. Network Architecture

The network architecture for learning the conditional distributions $P(\mathcal{B}_\lambda|\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_{\lambda-1})$ is very simple. We first assumed that pixels in the bit-plane $\mathcal{B}_\lambda$ are independent conditioned on the layers above. This might appear as a strong assumption for bit-planes close to $\mathcal{B}_1$ but since the critical fluctuations happen at *infinite* scales, large structures in $\mathcal{B}_1$ induce marginal interactions in the layers below. We estimated the conditional distributions by taking $L \times L$ patches and performing logistic regression for bit-plane $\mathcal{B}_\lambda$ taking bit-planes $\{\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_{\lambda-1}\}$ as the input. Since natural images are translation-invariant, we assumed weight-sharing, making the logistic regression convolutional.

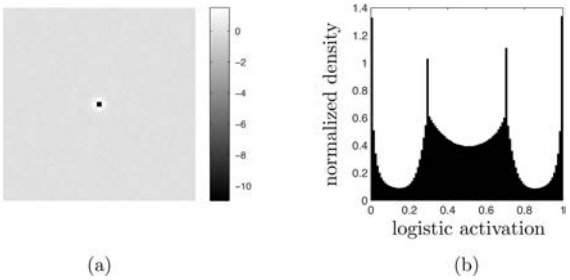

Figure 3. (a) The plot of the receptive field for the conditional distribution $P(\mathcal{B}_2|\mathcal{B}_1)$. The size of the receptive field in what is shown is 41×41. The receptive fields to other bit-planes have the same center-surround form. (b) The normalized histogram of the logistic activation function after applying the convolution filter in (a) to the bit-plane $\mathcal{B}_1$ from Fig. 2. The peaks come from the large clusters in $\mathcal{B}_1$.

4. Experimental Results

We performed the learning of the convolutional weights by maximizing the log likelihood with a single batch using the second-order Newton's method. The batch contained 10^5 samples of size 41×41, taken from 1024 images of size 2844×4284 pixels from the Geisler database of natural images (Geisler & Perry, 2011). The learnt center-surround receptive field for $P(\mathcal{B}_2|\mathcal{B}_1)$ is shown in Fig. 3. Receptive fields for other conditional distributions have the same center-surround form. The results reported here was obtained by having ℓ_2 prior on the weights, but setting the prior penalty term to zero did not change the results qualitatively.

After learning the convolutional weights, conditioning on the first bit-plane, we found the logistic activations of the lower bit-planes, up to $\mathcal{B}_8$. A key characteristic of natural images is the hierarchy in object sizes. This appears as connected clusters in the bit-plane representation (Saremi & Sejnowski, 2015). When the convolution filter scans large clusters it will output the same logistic activation. The peaks in Fig. 3 is the signature of those large clusters. To preserve the bigger clusters in lower bit-planes, which are key in natural images, we adopted a winner-take-all strategy, except for the interval $[0.4, 0.6]$. In that interval we took samples according to the logistic activation. We can easily generate binary images by the order dictated in Eq. 3. For example, the image in Fig. 4 was generated by conditioning on the bit-plane $\mathcal{B}_1$ of Fig. 2 and obtaining other bit-planes up to $\mathcal{B}_8$ by the logistic activations from the learnt convolutional filters. The bit-planes were then combined to obtain the gray-scale image.

The model performance was evaluated by the normalized mean square error and the conditional log likelihood scores. The normalized mean square error for the generated im-

ages, averaged over the database is 0.0588. The same error measured for the null model, where the convolutional weights/biases are set to zero, is 0.1244. For the density estimation measure, the average negative log likelihood for the conditional probabilities were evaluated and are given in Table 1. For the null model, corresponding to white noise, the negative log likelihood scores are 1 bit/pixel. Note that in this framework, the log likelihood for all bit-planes $\mathbb{B}$ combined is *not* a good measure to compare models, in part because the analog image is obtained by a *weighted* sum of the bit-planes. In other words, it is (much) more important to have a high log likelihood score for bit-planes closer to the critical bit-plane $\mathcal{B}_1$ than the ones further below.

$\mathrm{NLL}(\mathcal{B}_2\|\mathcal{B}_1)$	$\mathrm{NLL}(\mathcal{B}_3\|\mathcal{B}_{1:2})$	$\mathrm{NLL}(\mathcal{B}_4\|\mathcal{B}_{1:3})$
0.2369	0.2901	0.2749

Table 1. Results on the density estimation of the conditional distributions of the bit-planes, where $\mathcal{B}_{1:\lambda}$ is a shorthand for $\{\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_\lambda\}$. The average negative log likelihood (NLL) is reported in bits/pixel. There were about 10^{10} pixels in the experiments. The null model is characterized by the negative log likelihood of 1 bit/pixel.

The state of the art in natural image modeling has been limited to texture synthesis, dead leaf images, small patches, or larger images that do not look natural — see (Gerhard et al., 2015) for a recent review. That being said, the direct comparison cannot be made to other models at the present stage, since the results here were obtained by conditioning on the critical bit-plane $\mathcal{B}_1$ with its rich structures. Our results, however, point to a new research direction in image modeling, as learning the prior on the critical bit-plane $\mathcal{B}_1$ will transform this algorithm to a fully probabilistic model and perhaps a very powerful model for natural images.

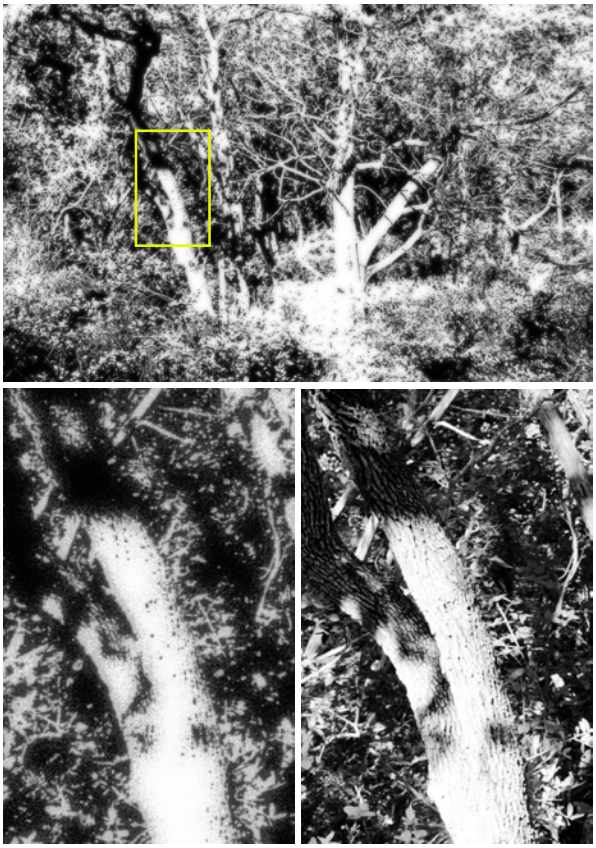

Figure 4. The image on top was generated by conditioning on the bit-plane $\mathcal{B}_1$ of Fig. 2 and obtaining other bit-planes up to $\mathcal{B}_8$ by the logistic activations from the learnt convolutional filters. The bit-planes were combined to obtain the gray-scale image. The area enclosed in the yellow rectangle is blown up on the left and the corresponding region in the original image is given on the right.

5. Discussions

We introduced a novel framework for modeling natural images. It is named after Kenneth G Wilson (1936 – 2013) for his immense contributions to our understanding of the nature of criticality. Infinite correlation lengths are one of the key signatures of critical points, and we utilized that fully in this algorithm. In contrast with the research focus in the deep learning community on learning hierarchical representations, we showed that the "hierarchy" of structures "hidden" in the data itself can be utilized for learning.

The results here were obtained without hidden units, but this algorithm will obviously become more powerful by hierarchical architectures, especially in learning the distribution of the critical bit-plane, but also in finding better estimates for the conditional distributions in the directed graph. Along these lines, there has been a recent work in learning data distributions by turning them into noise through stochastic processes with the use of hidden units (Sohl-Dickstein et al., 2015). Learning in that framework is to figure out how to reverse that process, thus going from noise to data. In contrast to that work, in the Wilson machine described here, the learning happens due to the stochastic heating process to disorder that is already "hidden" in the data and is part of the data itself.

In addition, the results here point to a new direction in image compression as learning a better probabilistic model in this framework might push the (lossy) compression limit to 1 bit/pixel (i.e. the first bit-plane.) Finally, even though we focused on natural images, the framework here is general and we think it will be especially powerful for analog signals with very long correlation lengths/times.

Acknowledgments

We thank conversations with Ruslan Salakhutdinov, and the support of The Howard Hughes Medical Institute.

References

Dayan, Peter, Hinton, Geoffrey E, Neal, Radford M, and Zemel, Richard S. The Helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

Geisler, Wilson S and Perry, Jeffrey S. Statistics for optimal point prediction in natural images. *Journal of Vision*, 11 (12):14, 2011.

Gerhard, Holly E, Theis, Lucas, and Bethge, Matthias. Modeling natural image statistics. *Biologically-inspired Computer Vision: Fundamentals and Applications*, 2015.

Hinton, Geoffrey E, Dayan, Peter, Frey, Brendan J, and Neal, Radford M. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *Nature*, 521(7553):436–444, 2015.

Saremi, Saeed and Sejnowski, Terrence J. Hierarchical model of natural images and the origin of scale invariance. *Proceedings of the National Academy of Sciences*, 110(8):3071–3076, 2013.

Saremi, Saeed and Sejnowski, Terrence J. On criticality in high-dimensional data. *Neural computation*, 26(7): 1329–1339, 2014.

Saremi, Saeed and Sejnowski, Terrence J. Correlated percolation, fractal structures, and scale-invariant distribution of clusters in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. doi: 10.1109/TPAMI.2015.2481402.

Schmidhuber, Jürgen. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

Sohl-Dickstein, Jascha, Weiss, Eric A, Maheswaranathan, Niru, and Ganguli, Surya. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.

Stephens, Greg J., Mora, Thierry, Tkačik, Gasper, and Bialek, William. Statistical thermodynamics of natural images. *Physical Review Letters*, 110:018701, Jan 2013.

Wilson, Kenneth G. Problems in physics with many scales of length. *Scientific American*, 241:140–157, 1979.