Learning Local Receptive Fields in Deep Belief Networks for Visual Feature Detection

Diana Turcsany and Andrzej Bargiela

School of Computer Science, The University of Nottingham
Nottingham, United Kingdom

Abstract. Through the introduction of local receptive fields, we improve the fidelity of restricted Boltzmann machine (RBM) based representations to encodings extracted by visual processing neurons. Our biologically inspired Gaussian receptive field constraints encourage learning of localized features and can seamlessly integrate into RBMs. Moreover, we propose a method for concurrently finding advantageous receptive field centers, while training the RBM. The strength of our method to reconstruct characteristic details of facial features is demonstrated on a challenging face dataset.

Keywords: Visual information processing, neural encoding, deep belief network, receptive fields, unsupervised learning, facial feature detection.

1 Introduction

Despite strong multi-disciplinary interest the highly accurate vision system of humans and other biological systems is still not fully understood and cannot be replicated with computational methods. Important discoveries have been made regarding the morphology and functionality of neural cells and networks, however our knowledge is still far from complete. Computational models of neural circuits in the visual pathway have great importance for improving our understanding of biological visual processing. A more informed background could facilitate the design of computational visual processing units, e.g., retinal implants. Currently, with the amount of unknown details, robust computational models of biological visual processing have to account for uncertainty and unknown details. We believe flexible probabilistic models, e.g., deep belief networks (DBNs) [2] possess great potential for modeling in this uncertain environment.

Deep Networks. To learn a multi-layer generative model of the data where each higher layer corresponds to a more abstract representation of information, Hinton et al. [2] train a DBN layer by layer using unsupervised RBMs. The network parameters are subsequently fine-tuned using backpropagation. Since this efficient training method for deep networks was introduced, there has been increasing research within deep learning. The potential of deep architectures for learning meaningful features has been demonstrated on a number of visual

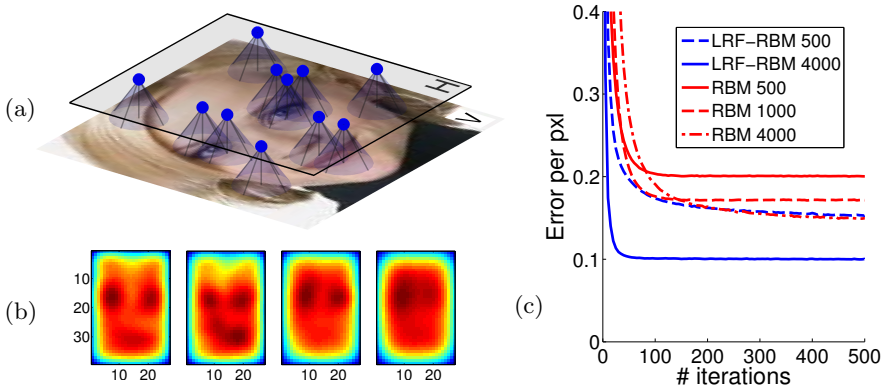

Fig. 1: (a) LRF-RBM model schematic, showing an input image in the visible layer V and local feature detector RFs (blue cones) in the hidden layer H with feature hubs around the eyes and mouth. (b) RF maps of LRF-RBMs run with different learning parameters. Automatically learned feature detector RFs are combined to show areas attracting more detectors. Darker red areas indicate higher feature density. Feature hubs emerged around average eye and mouth locations at pixels (9,15), (20,15), and (15,30). (c) Average squared reconstruction error per pixel is shown on the test set. Method names and hidden node counts are given in the graph. With the same node count LRF-RBMs give significantly lower errors than RBMs. Moreover, a 500 node LRF-RBM performs similarly to a 4000 node RBM.

tasks [3,4,5,6,7], [10], [12]. DBNs have also been shown suitable for modeling feature detection in the retina [11] and visual areas V1 and V2 [8]. Despite this success in neural modeling, primal emphasis has been given to improving performance of deep learning on visual recognition tasks, rather than increasing the fidelity of deep architectures to real neural circuits of the visual pathway. Our aim is to fill this gap by proposing deep network structures that more closely resemble biological neural networks, but still provide flexibility and great performance on visual recognition tasks. Such architectures possess high potential for modeling visual information processing in the retina and visual cortex.

Local Receptive Fields. In focus of this paper is the extension of RBMs with local receptive fields (RFs) in a way that the training process, the final architecture and the inference at test time closely resemble biological neural networks of the visual pathway. We concentrate mainly on early processing stages, the retina and V1. Our contributions are (1) a modification to the contrastive divergence [1] (CD) algorithm that introduces local receptive field constraints for hidden nodes, (2) a method for automatically identifying locations of high importance within the visual input space during RBM training, and (3) by utilizing these locations as RF centers, a compact, yet powerful encoding of visual features. We show using biologically inspired Gaussian shaped local RFs and learning advantageous RF placement improve RBM and DBN based reconstruction of face images.

RFs have been modeled in deep learning through convolutional networks [4, 5,6], [9], however the training methods used do not try to approximate learning in biological neural networks. The main emphasis is on improving the efficiency of learning to scale up deep learning algorithms to high dimensional problems. In these networks weights between the visible and hidden layers are the same across all image locations and the inference procedure can therefore utilize convolution operations. The same feature detectors operate on each part of the image providing translation invariance of feature detection, which can make the learning task easier. On the other hand, spurious detections can often be introduced and in some visual recognition tasks translation invariance may not be advantageous (e.g., for face recognition in aligned images the mouth always appears in the same area, therefore a positive detection of mouth elsewhere will be false).

In contrast, our local receptive field constrained RBM (LRF-RBM) only learns relevant feature detectors at any one image location. As opposed to a fixed grid layout of rectangular RFs [7], [12], our hidden nodes move around the visual space during training to find the best location for their Gaussian RF center. By letting the detectors move to locations of interest *"feature hubs"* can emerge in image regions where the training data has high variation, while more uniform areas will attract less detectors. The resulting network architecture extracts compact representation of visual information, provides very quick inference and by combining local features as building blocks, the network is strong at reconstructing previously unseen images. An illustration of the receptive field learning and RF distributions of our trained models are in shown in Fig. 1(a)-(b).

2 Local Receptive Field Constrained RBMs

The unsupervised phase of Hinton et al. [2]'s DBN training utilizes RBMs for learning each layer of the representation. The energy-based RBM models include a visible and a hidden layer, with connections between hidden and visible nodes but not within layers. This restriction ensures conditional independence of hidden nodes given visible nodes and vice versa, which is key for the efficiency of RBMs. In most vision tasks visible nodes correspond to pixels, while hidden nodes model visual processing neurons and detect image features.

2.1 RBM Training

The energy function of RBMs with binary visible and hidden nodes is given by:

$$E(v,h) = -a'v - b'h - h'Wv\,, \tag{1}$$

where v and h are the states of visible and hidden node, W is the weight matrix describing the symmetric connections between the visible and hidden layer, while a and b are visible and hidden biases respectively. Learning aims at reducing the energy (increasing the log probability) of the training data.

RBMs can be trained with the approximate but very efficient contrastive divergence [1] (CD) algorithm. In each step of CD, (i) visible states are initialized

to a training example, then (ii) hidden states can be sampled parallel, due to the conditional independence properties, according to:

$$p(h_j = 1|v) = \frac{1}{1 + exp(-b_j - \sum_i v_i w_{ij})} , \tag{2}$$

followed by (iii) the reconstruction phase where visible states are sampled using:

$$p(v_i = 1|h) = \frac{1}{1 + exp(-a_i - \sum_j h_j w_{ij})} , \tag{3}$$

finally (iv) the weights are updated according to:

$$\Delta w_{ij} = \epsilon(< v_i h_j >_{data} - < v_i h_j >_{reconst}) , \tag{4}$$

where ϵ is the learning rate, correlation between the activations of v_i and h_i measured after (ii) gives $< v_i h_j >_{data}$, and the correlation after the reconstruction phase (iii) determines $< v_i h_j >_{reconst}$. A similar rule is applied to the biases.

For continuous data (e.g., the face images we studied) using Gaussian visible nodes can improve the model. In this case the energy function changes to:

$$E(v, h) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} , \tag{5}$$

where σ_i is the standard deviation at v_i. The probability of hidden node activation and the expected value of a visible node (i.e., the reconstructed value) is then given by:

$$p(h_j = 1|v) = \frac{1}{1 + exp(-b_j - \sum_i (v_i/\sigma_i) w_{ij})} , \tag{6}$$

$$< v_i >_{reconst} = a_i + \sigma_i \sum_j h_j w_{ij} . \tag{7}$$

2.2 Training with Local Receptive Fields

Neurons in early stages of the visual pathway typically only receive input from a small localized area of the previous processing layer. Moving up the layers, receptive field of neurons (the area of the photoreceptor layer in which stimulus can result in neural response) is gradually getting bigger with increasing complexity in structure. As an example, retinal ganglion cell RFs can be closely modeled by difference-of-Gaussians (DoGs), while RFs of V1 simple cells by Gabor filters.

LRF-RBMs include receptive field constraints for hidden nodes to outline the area from which the hidden node is most likely to receive input. These constraints are given in the form of RF masks, denoted by R, that operate on the RBM weights W. Each mask has a center location which corresponds to a hidden node's location in visual space. R describes the likelihood of a connection being present between a visible and a hidden node given their distance in the

visual space, where the likelihood converges to 0 as the distance goes to infinity. During training to avoid the prohibitive process of sampling connections from the likelihood, we will instead use the values of R as additional weights on top of W. R thereby narrows down the scope of hidden nodes to local neighborhoods. Note that from the biological modeling point of view, R provides only a constraint or regularizer on the RF structure, the actual RFs are specified by R and W together. After training, these can show significantly different structures compared to R alone. Still, to keep the description simple we refer to R as RFs.

We found LRF-RBMs with disk or square shaped RFs efficiently learn local features, however Gaussian RFs provide smoother reconstructions with better detail and can be truncated to preserve efficiency. Gaussian RF constraints are also more adequate when modeling biological neurons in early stages of visual processing. From here on we will discuss only the Gaussian case, using fixed standard deviation (SD) for each RF, denoted by σ^{RF}.

The training algorithm described in Section 2.1 can be used for LRF-RBMs with modifications. The energy functions in Eqs. 1 and 5 and also Eqs. 2, 3, 6, 7 can be adapted by substituting w_{ij} with $r_{ij}w_{ij}$, where r_{ij} is the RF constraint on the connection between v_i and h_j. The weight update equation changes to:

$$\Delta w_{ij} = r_{ij}\epsilon(< v_i h_j >_{data} - < v_i h_j >_{reconst}) \,. \tag{8}$$

Learning RF Centers. Hidden nodes can be placed at uniform distances from each other or allocated randomly, but this would not allow the network architecture to adapt to specific properties of the input data. Non-uniform feature detector distributions can be beneficial to obtain compact representations by exploiting patterns in the dataset (e.g., aligned faces have facial features at given locations, most natural images have the center of interest in the middle). When solving a task some areas of the visual space may need representation at better resolution, using many different feature detectors, while other areas do not convey much information. In the human vision system, the retina also has non-uniform ganglion cell distribution between the center (fovea) and the periphery, former being denser, thus better resolution is obtained in the center of the visual space. Our model utilizes non-uniform feature detector distribution by allowing the system to identify areas of the visual input space which need higher number of feature detectors to obtain a good data representation.

Our method learns RF centers during RBM training. In each RBM iteration, a hidden node's RF is allocated to the local area that has the strongest connections to the hidden node and thus give the most well defined feature. This is done by first (i) writing the weights of the hidden node in the shape of the input image data and (ii) applying a transformation, then (iii) filtering the weight image with a Gaussian filter (SD: σ^{RF}) on each channel, (iv) the responses over channels are combined by taking the max, finally, (v) the location with the maximum response is selected as the new RF center and (vi) a Gaussian (SD: σ^{RF}) around this center provides the updated RF. We examined element-wise transformations including identity, absolute and squared value, and found the latter two worked similarly well and superior to identity (results are shown with squared value).

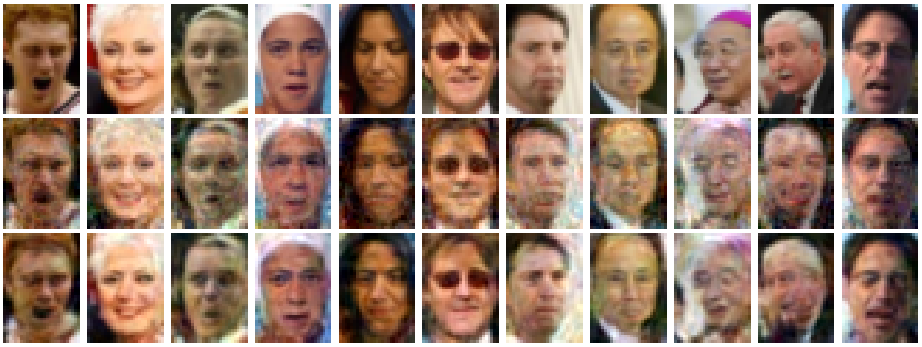

Fig. 2: Samples of the test data are shown in the first row. The second and third rows show their reconstructions produced by an RBM and an LRF-RBM respectively. Note how small details, e.g., eye and mouth shapes or direction of gaze are better retained with the LRF-RBM due to the number of specialized eye and mouth detectors. Note also how images of side facing people can confuse the RBM but not so the LRF-RBM.

LRF-DBNs. According to the DBN training procedure, we can train multiple layers of feature detectors on top of an LRF-RBM hidden layer using either RBMs or LRF-RBMs (e.g, with increasing RF sizes) with binary visible nodes. We call these models LRF-DBNs. If LRF-RBMs are used for training higher layers, hidden node locations are fixed after training a layer and RF constraints of higher layer nodes are applied when training the next layer. Although here we focus on unsupervised training, we note however for classification tasks supervised fine-tuning could subsequently be applied, analogously to DBNs.

3 Experiments

In the followings, we demonstrate how our LRF-RBM can discover important feature hubs in the deep funneled Labeled Faces in the Wild (LFW)[1] [4] face recognition dataset containing aligned faces. We have also experimented on the MNIST handwritten digit dataset using LRF-RBMs/DBNs, which successfully learned feature detectors for digit parts. When trained on the simulated photoreceptor input of [11] our LRF-RBMs were detecting local features including Gabor-like filters. Here we focus on a detailed analysis using the LFW dataset.

Dataset. LFW contains 13233 RGB images of public figures (see examples in Fig. 2 first row). The intended task on the dataset is recognizing whether two face images are taken of the same person, without having seen the person(s) during training. RBMs with rectified linear or binary hidden nodes, first trained unsupervised on single faces and subsequently fine-tuned in a supervised manner on pairs of images, have been shown to achieve good results on this task [10].

[1] Available at http://vis-www.cs.umass.edu/lfw/

Applying supervised methods on pairs of faces is out of the scope of this paper, which focuses on modeling of biological vision systems. Our primal interest is to investigate the capability of LRF-RBMs to identify regions of high importance in LFW images and utilize these hubs to provide compact representation.

We applied similar pre-processing to Nair and Hinton [10] and trained RBMs with binary hidden nodes on single faces using their published settings. With this we compare our LRF-RBM model run on the same data. A separate training and test set was used, with 4038 training and 1711 test images. The central 105x153 part of the images was cropped, thereby eliminating much of the background.[2] Images were then subsampled to 27x39(x3). Finally, to simplify training, the data was normalized along each component to have zero mean and unit variance.

Training. RBMs with Gaussian visible nodes and binary hidden nodes were trained on mini-batches of size 100 for 2000 iterations, with hidden node numbers of 500, 1000 or 4000, applying a learning rate (ϵ) of 0.001 (higher learning rates failed) and momentum. LRF-RBMs were trained using the same settings, except for ϵ, where 0.1 was optimal. Results are shown when σ^{RF} of 3 and a filter size of 5 was used during RF center learning. Both RBMs and LRF-RBMs were able to learn good models within a few hundred iterations, after which performance only slightly improved. In the followings, if not stated otherwise, results are displayed for models with 4000 hidden nodes trained for 2000 iterations. We also trained LRF-DBNs to learn a second layer of feature detectors on top of our LRF-RBM features using a 1000 hidden node RBM without RF constraints.

Testing. Reconstructions are obtained by calculating the top-down activations after feeding in an image. Performance was evaluated on the test set quantitatively by comparing the squared reconstruction errors (SRE), i.e., the squared distance between the original data and its reconstruction, and qualitatively by displaying example reconstructions. In the case of LRF-RBMs, the spatial distribution of feature detectors were examined and feature hubs identified. Visualization of features learned by RBMs and LRF-RBMs are obtained by displaying their weight vector in the shape of the input images. Visualization of higher layer hidden nodes is obtained by a linear combination of the strongest connected lower level features with their weights. RBM and LRF-RBM features where compared based on the distinctiveness of their appearance and locations.

4 Results

Reconstruction. SREs are compared on normalized test data in Fig. 1(c), indicating a superior reconstruction capability for LRF-RBMs. SREs shown translate to an average 16 pixel difference on original test images for the LRF-RBM vs. 18 for the RBM. Table 1 demonstrates LRF-RBMs/DBNs give lower pixel errors than the corresponding RBMs/DBNs. The comparison of reconstructed images

[2] These pixels are known to unintentionally provide helpful context for recognition.

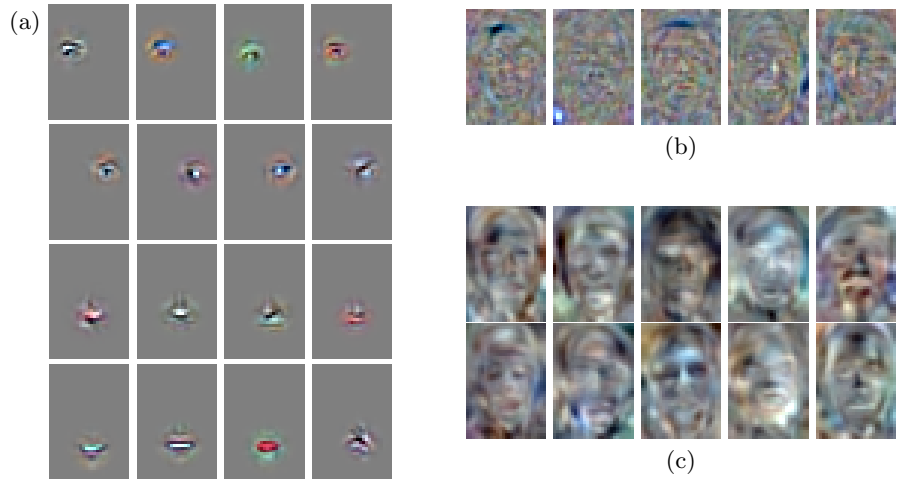

Fig. 3: (a) Distinctive looking detectors located in feature hubs within an LRF-RBM. From top to bottom row: detectors of the persons' right eye, left eye, nose and mouth can be seen. (b) RBM features having global structure. (c) Sample of second layer features learned by an LRF-DBN, corresponding to characteristic looking faces.

Table 1: Average pixel error of (LRF-)RBM reconstructions from 500 and 4000 length encodings, and (LRF-)DBN reconstructions from 1000 length encodings on top layer.

hidden nodes	RBM	LRF-RBM
500	22.31	**19.97**
4000	18.12	**16.06**

hidden nodes	DBN	LRF-DBN
1000-1000	22.76	**18.96**
4000-1000	19.32	**18.19**

in Fig. 2 is even more convincing. Both models can reconstruct main features of test examples with a limited amount of nodes. However, characteristic details, especially around eye and mouth areas, are better retained using LRF-RBMs. Such details can help distinguish persons. This analysis confirms LRF-RBMs outperform RBMs for reconstructing previously unseen data.

Features. Figure 3(a) shows local facial feature detectors learned by an LRF-RBM, while Fig. 3(b) shows a sample of RBM features. All the RBMs we trained have learned features similar in nature to the ones shown, having global structure with an occasional local peak. We could not identify any clear local detector modeling a single facial feature. Our LRF-RBMs on the other hand attracted feature hubs around eye and mouth regions and by focusing on these areas have learned a number of distinctive looking eye, mouth and nose detectors. The spatial arrangement of detectors is shown in the maps of Fig. 1(b). The second map from the left belongs to the LRF-RBM that generated the local features in Fig. 3(a). Alongside these specific eye and mouth detectors, Gabor filters and DoG detectors were also common among the learned features, especially in areas

along the contour of the face. The layout of features with the emergence of feature hubs around key areas within the input space demonstrates how LRF-RBMs can identify important regions within the image data which need a higher density of feature detectors for representing their details. Features learned on the second layer in our LRF-DBN have more global receptive field structures corresponding to well defined, varied looking faces as can be seen in Fig. 3(c).

5 Conclusions

We proposed a modified unsupervised RBM training algorithm, the LRF-RBM, which poses constraint on feature detector RFs and can automatically discover advantageous placement of RF centers. We have shown how feature detectors converge to important areas within face images, e.g., eyes and mouth, forming feature hubs. We have demonstrated the superiority of LRF-RBMs to reconstruct details of test images. In future work we will incorporate RFs of varying sizes and further investigate LRF-DBNs for learning multi-layer representations.

References

1. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Computation 14(8), 1771–1800 (2002)
2. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation 18(7), 1527–1554 (2006)
3. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006)
4. Huang, G.B., Mattar, M.A., Lee, H., Learned-Miller, E.: Learning to align from scratch. In: Neural Information Processing Systems. pp. 773–781 (2012)
5. Kavukcuoglu, K., Sermanet, P., Boureau, Y.L., Gregor, K., Mathieu, M., LeCun, Y.: Learning convolutional feature hierarchies for visual recognition. In: Neural Information Processing Systems. pp. 1090–1098 (2010)
6. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Neural Information Processing Systems. pp. 1106–1114 (2012)
7. Le, Q.V., Monga, R., Devin, M., Corrado, G., Chen, K., Ranzato, M.A., Dean, J., Ng, A.Y.: Building high-level features using large scale unsupervised learning. In: International Conference on Machine Learning. pp. 81–88 (2012)
8. Lee, H., Ekanadham, C., Ng, A.: Sparse deep belief net model for visual area V2. In: Neural Information Processing Systems. pp. 873–880 (2008)
9. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: International Conference on Machine Learning. pp. 609–616 (2009)
10. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: International Conference on Machine Learning. pp. 807–814 (2010)
11. Turcsany, D., Bargiela, A., Maul, T.: Modelling retinal feature detection with deep belief networks in a simulated environment. In: European Conference on Modelling and Simulation. pp. 364–370 (2014)
12. Zhu, Z., Luo, P., Wang, X., Tang, X.: Deep learning identity-preserving face space. In: IEEE International Conference on Computer Vision. pp. 113–120 (2013)