Learning Stereo Features with Stacked Autoencoders

Daniel Jin Hao Chia, Pang Wei Koh, Zhenghao Chen
CS229 Final Project

1 Introduction

Single-layer stacked autoencoders have been shown to be successful in training artificial neurons with receptive fields that are similar to those found in the V1 cortex, but on monocular data. In this project we investigate extending a single-layer stacked autoencoder network to learn receptive fields on stereo data, and evaluate them with respect to their effectiveness as features for object classification and their similarity to neuronal receptive fields characterized in physiological experiments.

The primary motivation for using stacked autoencoders is their intrinsic non-linearity, which allow us to obtain higher-level features by stacking more levels and iterating the same algorithm multiple times. In contrast, stacking linear transformations such as those obtained from ICA does not yield any meaningful results. This higher-level representation would be useful both as a tool for performing image and depth recognition with stereo data, as well as a model for how the brain processes visual information.

2 Methodology

We generate 14x28 receptive fields from a training set of 291,600 14x28 images, sampled from the first training dataset of the NORB dataset [1]. For each of the 29,160 stereo pairs of 108x108 images in the NORB training set, we first whiten the data and then sample 10 pairs of 14x14 images around a 80x80 bounding box centered on the middle of the image, concatenating each pair of corresponding 14x14 images to form a 14x28 image. Within each 14x28 image, the 14 columns on the left represent the image that the left eye sees, while the 14 columns on the right represent the image that the right eye sees.

Training of the stacked autoencoder network is done with the Stanford Deep Learning Network Library, with each of the 200 neurons characterized by a 14x28 receptive field. The network uses a sum-of-squares reconstruction error as the objective term, coupled with sparsity regularization on the bias term and L2 weight decay. Optimal coefficients of the sparsity regularization term and the weight decay term are found with a grid search.

We evaluate the learnt 14x28 receptive fields through supervised training on the original set of 29,160 pairs of 108x108 images and classification on a distinct but similar set. Each 108x108 image is whitened and then cropped to 98x98 to eliminate border effects. Feature vectors are generated by convolution: we extract 8x8 overlapping stereo patches, each of size 14x28 (half from the left image and half from the right image), from each pair of 98x98 images, and run each of these 14x28 patches through the network with the learnt receptive fields, taking the hidden layer activations as our features. Through concatenation, each pair of 98x98 images therefore translates into a feature vector of length 64*200 = 12800.

The stereo receptive fields are contrasted against mono receptive fields learnt on the same dataset. We generate the latter by treating left and right pairs of images as independent, essentially learning receptive fields on 291,600*2 = 583,200 14x14 images, and doing supervised training and classification on sets of 58,320 98x98 images.

We also present results obtained from running the FastICA algorithm [3] on the same set of 291,600 sampled 14x28 images. No dimensionality reduction was done, resulting in 14*28=392 independent components, each of size 14x28, found. Classification results run on the whitened raw data, as well as on the convolution of whitened raw data projected upon the ICA bases, are also shown.

Disparity tuning curves are calculated using the method described by Hyvärinen *et al.*[2]. Each curve plots response against horizontal disparity. To calculate the response of a neuron at a particular horizontal disparity, we use the left side of the receptive field of the neuron itself as the stimulus, translate it horizontally by the required amount, and find the maximum over all vertical translations of the activation of the neuron when presented with that stimulus. We then repeat this process with the right side of the receptive field. The response is taken to be the average of the activations when presented with the stimulus from the left and the right side.

3 Results - Receptive Fields and Classification

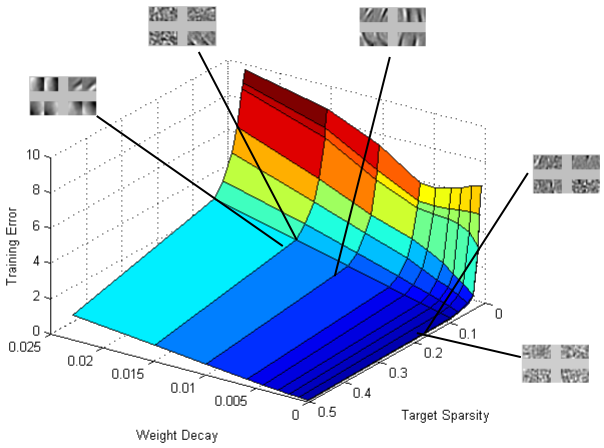
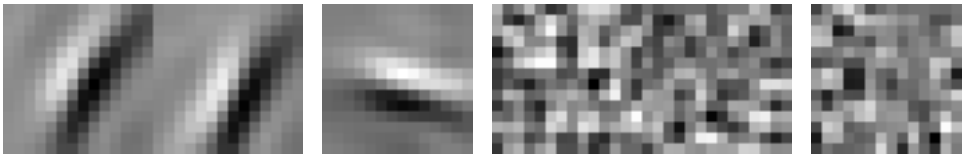


Figure 1: Representative receptive fields from different areas of search space



(a) NS - Stereo (b) NS - Mono (c) LR - Stereo (d) LR - Mono

Figure 2: Images of receptive fields

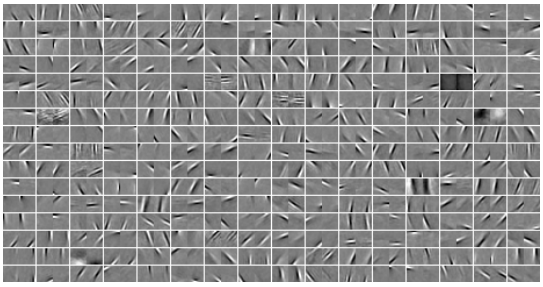


Figure 3: ICA bases obtained from the training set

We find that the most neurologically accurate learnt stereo receptive fields lie in the area of moderate weight decay and sparsity (Fig. 1). These learnt patches resemble, for the most part, a pair of Gabor filters that are phase and/or positionally shifted from each other (Fig. 2a, 2b). We compare these to the 'receptive fields'

(a) Original image patch (b) Reconstruction by NS (c) Reconstruction by LR

Figure 4: Reconstructed patches

that result in the lowest reconstruction error. These correspond to networks that have not been constrained much by sparsity. We term these sets of receptive fields NS and LR respectively. In comparison, Fig. 3 shows the results of running ICA on the same training set.

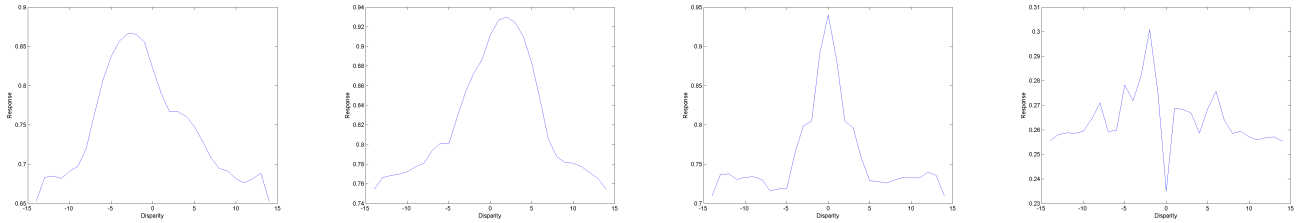
Fig. 4 shows an example of the reconstructed images obtained using the NS and LR stereo receptive fields.

Type	**Classification Accuracy (%)**
Stereo NS	41.58
Stereo LR	43.51
Mono NS	34.06
Mono LR	37.11
Raw pixels	31.27
ICA	31.09

Table 1: Classification results on 29,160 pairs of 98x98 images

Table 1 displays the results of classification through the various methods described.

4 Results - Neurological Comparisons

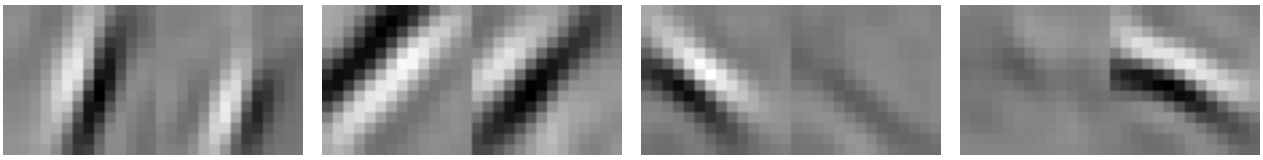


(a) Far (Autoencoders) (b) Near (Autoencoders) (c) Tuned excitatory (Autoencoders) (d) Tuned inhibitory (ICA)

Figure 5: Characteristic disparity tuning curves

Traditionally, binocular neurons have been classified into four different categories: far, near, tuned excitatory, and tuned inhibitory, based on characteristic disparity responses for each category. The receptive fields that the single-layer stacked autoencoder learn show disparity tuning curves from the first three categories (Fig. 5a, 5b, 5c). However, none of our current receptive fields match the response of a tuned inhibitory neuron (Fig. 5d).

In accordance with biological data [4], phase (Fig. 6b) and positional (Fig. 6a) shifts are visible. We also report the presence of ocular dominance (in both the autoencoder (Fig. 6c, Fig. 6d) as well as the ICA receptive fields, though to a significantly greater extent in the latter. To the best of our knowledge, this phenomenon is not well understood within the neurological literature.



(a) Positional shift (b) Phase shift (c) Left ocular dominance (d) Right ocular dominance

Figure 6: Disparity and ocular dominance in autoencoder receptive fields

5 Discussion and Future Directions

We have obtained some encouraging preliminary results as to how stacked autoencoders can produce receptive fields with properties that resemble neuronal receptive fields, in particular when sparsity constraints are set on the network. This supports sparse coding hypotheses of the workings of the visual cortex. However, there are types of receptive fields found in the real neurons that are currently lacking in those that the autoencoders produce, for example, those that result in tuned inhibitory disparity tuning curves. This could be due to the difference in the workings on the human visual system as compared to the system we are implementing: in particular, our eyes are able to vary their focal lengths, and tuned excitatory and tuned inhibitory neurons are thought to be related to this change in fixation length [5].

Also, stacked autoencoders can be used to produce stereo feature sets that achieve better linear classification accuracy with a smaller number of features as compared to using raw pixel data, or linear transformations thereof. However, the neurologically similar features that we find are not those that result in the highest classification accuracy. We can think of three possible reasons for this: firstly, the NORB dataset is not perfectly stereo in that the images chosen for background noise are placed at a fixed disparity between the left and right patches, which is not a totally realistic representation of natural images. Secondly, the advantage of the neurologically similar features might lie not so much in achieving the greatest classification accuracy, but rather in optimizing a tradeoff between the amount of data required versus the classification accuracy. In this sense, the sparser feature set would be more amenable to compression via thresholding of each neuron's activity, for example. Thirdly, it could be that the neurologically similar features lend themselves better to the finding of higher-level features through the stacking of additional autoencoders.

With these in mind, we propose the following future directions:

- **Quantitative comparison to biological data.**

  The methods used to generate disparity tuning curves vary from paper to paper, and we have not yet found a way to statistically and non-qualitatively measure how similar the disparity responses of our neurons are to those of real, biological neurons. For example, the comparison in Hyvärinen*et al.*[2] is done by eye. There is also a lack of clear statistics on factors such as the degree of ocular dominance present within a set of receptive fields. We would like to investigate how we might quantitatively measure such similarity, in order to make more confident claims about the usefulness of the stacked autoencoders as a biological model.

- **Learning and evaluating on a different dataset.**

  We would like to see if other datasets give us similar results. In particular, the next step would be to run the same algorithm on a dataset comprising true stereo images, with pictures taken at differing focal lengths. We would also like to see if the stacked autoencoders, particularly the NS feature set, perform comparatively better on classification tasks with a greater number of classes (NORB has 5).

- **Using non-linear classifiers.**

  Our current restriction to linear classifiers is a likely cause of the low classification accuracy that we are seeing across the different feature sets, and might impede proper analysis of the effectiveness of the features generated by the stacked autoencoders. Memory and time restrictions due to the large size of the feature sets prohibits more complex methods, but we would like to explore alternative methods for classification that might result in better classification accuracy.

- **Extension to multi-layered stacked autoencoders and addition of stochastic elements.**

The learnt features can also be improved, and higher-level features obtained, by stacking more autoencoders on top of our single layer. This is the main long-term aim of our work. Additionally, improvements to the learnt features can also be made through the addition of random elements into the neural network, either through using denoising autoencoders, or by making the activation of each neuron binary and probabilistic, instead of continuous and deterministic as it is now.

- **Modifying the objective function.**
  The current objective of sum-of-squares reconstruction error might not be optimal in terms of producing receptive fields that compare favorably with neuronal receptive fields. In particular, the best receptive fields are not the ones with the lowest reconstruction error, and we are currently judging the quality of a receptive field by eye. We would like to investigate how a modified objective function, perhaps involving kurtosis or a similar convex function to encourage sparsity, would affect the results we get. In particular, we would like to see if a modified objective function could help in the automation of hyperparameter tuning, as well as in the driving of the learnt receptive fields to match the ICA bases and the neurological data.
- **Using ICA to train the neural network.**
  Another possible and related way to obtain receptive fields closer to the ICA bases is to fix the decoder weights as the ICA bases, and then train encoder weights using the autoencoder algorithm. This should have the effect of indirectly modifying the objective function to result in learnt receptive fields that resemble the ICA bases more closely. Coming up with a non-linear approximation to the linear ICA bases in this manner might allow for higher-level features to be learnt by repeating the algorithm.

6 Acknowledgements

We would like to thank our TAs, Andrew, Ian and Quoc for the huge amount of help that they have given us in this project.

References

[1] Y. LeCun, F.J. Huang, L. Bottou, Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting , CVPR (2004).

[2] A. Hyvärinen, J. Hurri, P. O. Hoyer, Natural Image Statistics, Springer (2009).

[3] A. Hyvärinen, Fast and Robust Fixed-Point Algorithms for Independent Component Analysis, IEEE Transactions on Neural Networks 10(3):626-634 (1999).

[4] A. Anzai, I. Ohzawa, R. D. Freeman, Neural Mechanisms For Encoding Binocular Disparity: Receptive Field Position vs. Phase, Journal of Neurophysiology 82(2):874890 (1999).

[5] B. Fischer, J Kruger, Disparity Tuning and Binocularity of Single Neurons in Cat Visual Cortex, Exp. Brain Res. 35, 1-8 (1979).