Universität Hamburg
Department Informatik
Knowledge Technology, WTM

Cortical Computing for Invariant Object Recognition

Seminar Paper

Biologically Inspired Artificial Intelligence

Leena Chennuru Vankdara
Matr.Nr. 6641141
4chennur@informatik.uni-hamburg.de

25.12.2014

Abstract

In this paper, we try to provide a bird's eye view of biologically inspired models of visual recognition. We consider both HMAX models and convolutional networks as representing a single class of models. The paper does not discuss any model in great details bu rather focuses on

Contents

1 Introduction

Invariant object recognition and coherent object representation have been major hurdles in developing efficient artificial vision systems. Invariant object recognition forms the core of the problem of developing egocentrical object based representations of the external world. Given the incredible ability of the biological vision systems which exhibit invariance to a considerable amount of deformation, it is natural to take inspiration from biological vision systems to build artificial vision systems capable of performing invariant object recognition. In this paper, we evaluate different hierarchical models which attempt to mimic the cortical circuit design and the cortical architecture found in the neocortex of the brain to achieve object recognition. In the first section, a brief summary of design principles of cortical computing is presented along with a model [18] linking the designs of cortical computing to behavioral properties of various forms of biological intelligence. In the second section, we define the problem of invariance in object recognition[32] and the challenges faced by Artificial Vision systems which try to achieve this goal. In the third section, We evaluate biologically inspired models for invariant object recognition[14][27][40][36][46][52]. In the fourth section, we discuss various parameters that determine the efficiency of models[6][7][8][13][26][41]. In the fifth section, future directions of research in the field of object recognition is presented.

2 Principles of Cortical Computing

There is increasing concurrence that biological vision systems adopt a hierarchical and complementary computing paradigm [17][18] to achieve high level vision. The ventral and the dorsal streams of the visual cortex compute complementary properties. The ventral stream also called the *what stream* is responsible for visual recognition and it comprises of cortical areas V1-V2-V4-IT-PMC[35]. From V1 to IT, increased receptive field sizes and complexity of the incoming stimulus has been observed. Experiments[28] show the presence of IT cells which respond selectively to objects. These cells are hypothesized as being invariant to transformations of objects. With experiments done on the cat's striate cortex by Hubel and Weisel [10]revealing several properties of the simple and complex cells of the Primary visual cortex and several following anatomical studies, the visual cortex is one of the most extensively studied areas of the brain. The huge amount of behavioral, pyscho-physical and anatomical data collected by studies performed on the Visual Cortex gives rise to a massive hypothesis space. Many computational models

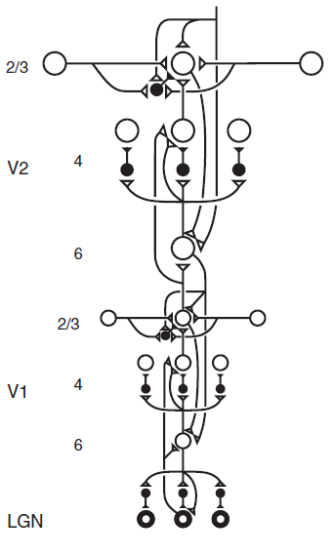

Figure 1: Canonical cortical circuit

built on the broad principles of a set of popular and
narrowed hypothesis space have shown state of the art performance on all the bench mark data sets. These models help further narrowing the hypothesis space and are used in complement to anatomical studies to understand the underlying mechanisms of vision. Grossberg 2007[18] presents the LAMINART model which gives a comprehensive overview of the basic principles of cortical computing. A brief summary of two of the principles relevant to this paper presented in LAMINART, Hierarchical feedforward computing and feedback computing(attention modeling)[18] is presented below . Figure 1 shows the structure of the canonical cortical circuit(which repeats itself across all the cortical layers[18]). Each cortical area can be broadly segregated into 6 hierarchical layers.

Feedforward Computing: Layers 2/3 of the lower cortical areas with smaller receptive fields pool together to provide bottom up activation into layer 4 of the next layer up the hierarchy with larger receptive fields both directly and indirectly through layer 6 - 4 route as shown in Figure 2. The input connections through the layer 6 -4 route provide inhibitory activation to the surrounding neurons increasing sparseness while providing both inhibitory and excitatory activation to the neuron to which a direct connection is made providing a modulatory effect and resulting in gain control and contrast divisive normalization. This form of processing in the visual cortex is believed to be used by the visual cortex in recognition of simple and unambiguous scenes in which processing speed is very high[44].

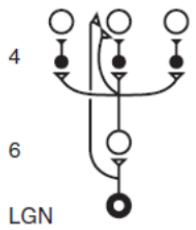

Figure 2: Feedforward

Feedback Computing: Higher cortical areas provide both excitatory and inhibitory activation to layer 4 through the same layer 6-4 route as in feed forward processing as shown in Figure 3. This form of activation is said to be used in recognizing objects in a visual clutter where processing time is comparatively much higher than in unambiguous object recognition. Increased competition in the network decreases all the cell activities and the connections from higher cortical areas which share the same connections as in feed forward processing selectively increase the cell activities while inhibiting the activities of the surrounding cells. The network behaves like a dynamical systems which rapidly increases the contrast between the winner(determined by feedback) and the rest of the neurons in the competition until the system reaches an attractor, the cell activity of the winner neuron crosses a threshold and provides activation to higher cortical areas. These properties are selectively adopted in several biologically inspired artificial vision models[14][27][36][41][46][52] with varying parameters that model the different features mentioned above.

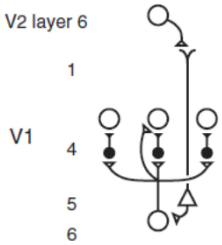

Figure 3: Feedback

3 Problem of Invariant Object Recognition

The major hurdle in building artificial vision systems has been to achieve recognitions invariant to translation, rotation, scaling and minor distortions and to build efficient 3D object representations.The problem of invariance in object recognition forms the core of the problem of object recognition[26]. Images of any real world object can occur in nearly an infinite number of forms in terms of position, scale, orientation, illumination, geometrical deformations etc[41]. Leibo et al.[32] define two forms of invariance in object recognition, generic invariance which corresponds to invariance to transformations of a single object and Class specific invariance which corresponds to constructing a system which can read out class identities of objects belonging to a particular class. This paper talks only about generic invariance as defined above[32] and invariance to any other properties is beyond the scope of this paper. Experiments show that the ventral stream gains invariance slowly from V1 to IT. IT populations are known to show generic invariance while demonstrating selectivity to object identities[24]. When different transformations of objects are presented to a visual recognition system, the system should essentially be able to define a decision function that can separate the images belonging to transformations of the same object from images belonging to transformations of different objects[2] . The learning of this decision function can be supervised, unsupervised or a semi supervised manner. To enable this the architecture transforms the representation of the input image. At each level, the representations are projected onto a higher dimensional space[26]. This problem can be seen as the problem of untangling object manifolds[26](each manifold corresponding to different transformations of an object plotted together in the representation space [12]) in stages such that each stage learns the function when applied on its input untangles local distortions. A classifier which takes the output of the highest layer in the hierarchy produces a discriminant function. A standard model of the visual cortex used in computational models for object recognition consist of a hierarchy of layers of S and C cells corresponding to the simple and complex cells[10] found in the visual cortex. The simple cells show selectivity to their optimal stimulus and the complex cells which have greater receptive fields pool over the simple cells and achieve invariance to local transformations. The weights of the connections across the layers, number of layers in the hierarchy and input representation are the parameters that affect the efficiency of recognition in the visual system. In the next section, we discuss models for object recognition inspired by the Visual cortex.

4 Cortically inspired Vision systems

Many biologically inspired visual recognition models have their foundations in the Neocognitron model.[14]. This paper discusses the following models. SMF model (Serre et al.,) [46], a revised model of HMAX[37], SLF model (Mutch et al.,) [27], Sparsity Regularized HMAX model (Xiaolin et al.,) [52], V1 like model (Pinto et

al.,)[41], Lessmann et al., [36]

SMF Model: The architecture of this system consists of 4 layers (S1,C1,S2,C2) arranged in a hierarchical structure. S1 cells are modeled after the simple cells of the primary visual cortex whose receptive field is mathematically described[25] by a Gabor Function(see equation below)[15]. A Gabor function is given by a convolution of a sinusoidal with a Gaussian and hence it behaves like a localized cosine transform detecting the spatial frequency in the neighborhood.

$$F(x,y) = \exp(-\frac{x_0^2 + \gamma^2 y_0^2}{2\sigma^2}) * \cos(\frac{2\Pi}{\lambda} x_0), \quad where \tag{1}$$

$$x_0 = x\cos(\theta) + y\sin(\theta) \quad and \quad y_0 = -x\sin(\theta) + y\cos(\theta)) \tag{2}$$

θ determines the preferred orientation of the Gabor filter and SMF model uses 4 orientations $(0^0, 45^0, 90^0, 135^0)$. The other parameters of Gabor filters are determined based on experimental data generated by the behavior of V1 simple cells[45]. 16 different scales are used in the model to induce invariance to scale. C1 cells with larger receptive fields(depending on the size of the scale band)[46] pool over scales(2) of S1 cells to achieve slight position and scale invariance while maintaining orientation selectivity by pooling over S1 cells with the same preferred orientation. A Max operation is used on the pool of efferent S1 cells to determine the output of C1 cells. S2 units compute the distance(Gaussian of the Euclidean distance) from their inputs to stored prototypes for all scales. The prototypes are stored in the same format as the output of C1 during the learning phase. C2 cells pool over all scales and positions using a MAX operation testing the maximum distance between the features of the image and the prototype vector, i.e essentially C2 units check for the presence of a prototype vector anywhere in the image in any scale. Finally a SVM is applied on the output of C2 units to obtain the decision function.

SLF Model(Mutch et al.,): This model[27] is based on the SMF model described above[46]. In this model, an image pyramid(of scales) is created and is used as input to the network. S1 filters are normalized and an additional normalizing parameter is included in S2 to modulate the effect of varying patch sizes in feature template matching. Apart from these fine tuning changes, a few other changes are made to the model. *Increased sparseness:* As we recall from the SMF model, S2 performs a feature template matching by calculating the distance between each feature and a prototype across all orientations for all scales thus essentially using the combination of distances between the prototype and each of the orientations. In the SLF model, sparseness in increased for inputs to S2 by considering only the distance between the prototype and the strongest orientation for each position. However as the model now computes the presence of an orientation like an on or off function, more number of orientations(12) are used to maintain the accuracy. Additionally to inhibit the output responses of S1 and C1 cells, weakly active neurons are suppressed by defining a global parameter which determines the fraction of cells that are inhibited. Sparsity is further improved at the classification stage

by selecting only the features that are highly weighted. *Partial retention of position/scale information:* Unlike the SMF model which achieves complete position and scale invariance by applying a MAX operation over all positions and scales, the SLF model only pools over a large receptive field(smaller than the size of the image) both in position and scale.

Sparse HMAX[52]: This model modifies HMAX[37] mainly by learning the filters of all S layers using sparse coding methods (Independent Component Analysis[1] and Standard sparse coding[52]). The model allows for layers at levels higher up the hierarchy. A few other fine tuning changes are made to the model which are not relevant to the evaluation performed in this paper and are hence not mentioned.

V1 like Model[41]: (Pinto et al., 2008) designed a baseline model(referred to as a Neuroscience null model in [41]) based on the receptive fields of the simple cells found in V1. In the model, normalized images are convolved with normalized Gabor filters(eq 1) and with 16 different orientations(eq 2) and 6 spatial frequencies(which can model the receptive fields of V1 simple cells[7]) after first applying a local normalizing filter to the input image. This local normalizing step is intended to model the local divisive normalization found in V1 simple cells[21]. The filtered images are subjected to thresholding and saturation and are further subjected to local divisive normalization.

Lessmann et al.,: This model is similar to the Memory Prediction Framework[20], [11], [16] which is based on the ideas of laminar computing, temporal sequence learning and attention modelling to achieve invariance in object recognition. The first primary difference between this model and the ones described above is the use of temporal sequences to learn invariance.

Basic features of the model:

- The computing architecture, similar to HMAX[37] and the other models summarized above takes a laminar form with several stages (3 or 4) with each stage consisting of two sublayers referred to as S and T layers in [36] corresponding to spatial neurons and temporal neurons respectively.
- During the recognition stage, at the lowest level activation is computed through parquet graphs[51] which are Gabor jets[15] placed on a sampled grid and the activation of all other neurons are computed as a hyperbolic tangent of the weighted sum of the afferent inputs converging onto that neuron[36] as given below.

$$Act_i = \tanh \sum_j W_{i,j} Act_j, \tag{3}$$

  where $W_{i,j}$ are the weights between the neurons i and j learned during the

learning phase. They encode a defined similarity between the input to the node and the prototypes stored at each layer.

- For spatial neurons after computing the feedforward input, lateral inhibition is applied to the inputs by setting all but the K most active neurons to zero both to improve sparseness in feature representation and to prevent the system from running into over excitation.

- Additionally the spatial neurons that remain active after applying the lateral inhibition receive feedback from the temporal neurons that have been active in the past T time steps(This information is recorded by storing the activities of all neurons into what is referred to as an activity stack[36]).

- Inputs and activation of temporal neurons is the same as that of spatial neurons except that temporal neurons receive feedback only from the activities of spatial neurons in the higher layer for the last presented image and not the last T images.

- During learning a Codebook of features is learned at each level of spatial neurons first by defining an empty codebook and a similarity function. If the current extracted input features are within the pre defined threshold of the similarity function, the codebook remains unchanged else the codebook is updated with the input features of the current image.

- To form the temporal groups, an adjacency matrix is created with the element (i,j) of the matrix representing, for every node, the probability that the neurons at locations i and j were active at some time in the past T time steps. Spectral clustering[48] is then used to form temporal groups which serve as inputs to the spatial nodes in the next layer.

- Thus essentially, S layer detects the spatial patterns in the input signal and the T layer determines the temporal group of the input signal with an associated probability.

- Also, Each node[36] of the S layer at each stage is connected to one node of the T layer in the same stage and a group of nodes (usually 9) make an afferent convergence into one node in the S layer of the subsequent higher stage in the hierarchy.

- Training is not done simultaneously for all layers but can only be performed layer by layer.

- The system has to be trained from sequences that represent images undergoing a particular transformation in order for the network to develop invariance to that transformation.

- The Activity stack is emptied after presenting each object category to prevent learning associations between different object categories.

Technical details not relevant to the discussion in this paper are excluded and can be referred to in [36]

5 Evaluation

Cortical computing for object recognition is indeed highly complex and the models summarized above belong to a family of hierarchical models which attempt to model certain aspects of the principles of cortical computing summarized in section 2. There are various questions that need to be answered in order to integrate these models into a unifying architecture. Some of the questions that need to be answered are listed below.

- Which principles of cortical computing are relevant to solving the problem of invariant object recognition?
- Which computing architecture is better suited to learn invariance in object recognition?
- Which mathematical models can approximate the processing of the cortex?
- Which mathematical models optimize the behavior of the neurons?
- What is the effect of the parameters like the number of neurons, number of selected features and the parameters used in the hardcoded Gabor filters?
- Is it efficient to use hard coded mathematical models that estimate the behavior of cortical elements or is it more efficient to learn the properties by training on a good statistical sample of images?
- What is a good statistical sample that can efficiently represent the variability of the object in a real world scenario?

The list of questions that can be answered is quite vast and answering all of them is beyond the scope of this paper. Hence this paper focuses on answering a small subset of these questions.

5.1 Feedforward Computing vs Feedback Computing

All the models belonging to the class of hierarchical models adopt a feedforward computing mode. Experimental data that shows that IT populations of neurons respond to visual stimuli within 100ms[9] in an extremely complex network seems to indicate that feedforward processing is mainly used in visual object recognition[47]. The SMF model, SLF model and SPARSE HMAX model, convolutional networks[29],[31] belong to the class of such feedforward models. The authors of these class of feedforward models recognize the limitations of such a feedforward architecture.

Anatomical studies like [5],[50] show that the number of feedback connections in both intra and inter cortical areas far surpass the number of feedforward connections between them. This causes the existence of a different class of models hypothesizing that feedback probably plays an important role in cortical processing[34]. [36],[11] are models that belong to this class.

The feedforward Architectures discussed above can not be categorized as strictly feedforward models as the nonlinearities in the networks such as local divisive normalization or the max operation would need to utilize local feedback connections to be implemented in a strictly neural architecture. This does not contradict the anatomical and physiological evidence of extremely rapid recognition in the first 150ms of the presentation of the stimuli. If any evidence is presented in the future that back projections are utilized in visual recognition in the first 150ms, then these architectures could be restructured to incorporate them. Back projections could be utilized to, for instance, perform segmentation or selective computation of neuronal activities in the lower layers depending on the object category. Lessmann et al.,[36] utilizes feedback from higher cortical layers as well as within the same cortical area to utilize principles of spatial continuity and temporal trace learning to achieve invariance to many transformations. However physiological data shows that synaptic plasticity(Long term potentiation or long term depression) depends on calcium dynamics according to which calcium levels at synaptic sites encode a trace of neural activity. This evidence points towards a different role for the back projections which is yet to be determined.

To sum up, feedforward processing has the advantages of being simple and fast but the design principles do not deal with ambiguity in visual scenes, which is an element of real world scenes.Visual attention, focusing and selective learning are hypothesized as utilities of back projections in the neural circuitry. Feedback connections that modulate the responses of the lower layers could be an effective tool to eliminate the competition between competing neurons in the presence of an ambiguous stimulus. The feedforward models could be integrated into a visual system with greater computational capability which include other principles observed in cortical computing such as attention modeling or perceptual grouping[18]. Certain non HMAX like models have attempted to incorporate top down inference in the models to resolve ambiguities by generating predictions.[23] uses undirected Deep belief networks(with stacked RBM's) and probabilistic max pooling[23] to achieve translation invariance and recognition in ambiguous as well as occluded scenarios. In this paper we try to address various parameters of the feedforward mode of computing that affect the performance of recognition.

5.2 Statistics of Natural Images

Strong evidence[7] and several other theories have strongly supported the claim that the visual system(all cortical areas) adapts itself to the statistics of the inputs that it processes. Which means that the change in representation is optimized to suit the statistics of natural images and only focuses on those aspects of data which are essential for further analysis[1]. This implies that the processing of the visual system is influenced both by genetic information and visual experience. In light of this theory it would be a good idea to look at the statistics of the natural images to identify useful representations. If we look at a statistical distribution of natural images the distribution is highly skewed(non normal) i.e, there is a high amount of redundancy in the natural images[1]. This fact has led the vision systems to draw inspiration from Information theory(Image compression). Each natural image consists of much higher amount information than is required to represent it. The objective of the visual system is then to eliminate the redundancy and to represent the image in least number of bits(features). Apart from the number of bits there are also other criterion such as ease of processing. This adds additional constraints to the transformation of representation. The concepts of sparseness of representation, dimensionality reduction, unsupervised learning stem from this approach and this paper discusses each of the aspect in some detail.

5.3 Dimensionality of Input - Output spaces

The principle of HMAX like architectures can be described in the perspective of untangling object manifolds[26]. The objective is to uncover the underlying low dimensional object manifolds which lie in a high dimensional feature space and at the same time untangle the manifolds such that a simple hyperplane(a linear classifier) could separate the classes. This view incorporates two basic ideas of dimensionality explosion and dimensionality reduction. Dimensionality reduction is achieved by mapping similar(determined by a pre defined similarity) images in the input space onto a low dimensional manifold thus extracting the underlying manifolds. This step needs to be followed by change in representation of the inputs by projecting the manifolds onto a much higher dimensional space(overcompleteness) to establish statistical independence between the inputs to achieve untangling. In HMAX like architectures a hierarchical approach to untangling is incorporated where each layer untangles local subspaces[26]. This can be visualized in the figure4. This view supports the view of complementary cortical processing where each cortical area or even each sub cortical area could be viewed as independent processing units computing properties(untangling local subspace manifolds) thus providing an abstraction to enable further processing. Dimensionality reduction techniques identify low dimensional structures in a high dimensional input space. We will discuss this further in the subsection5.6

Several dimensionality reduction techniques exist in literature. Principle Component Analysis(PCA), Local Linear Embedding(LLE), Dimensionality reduction by learning an invariant mapping(DrLIM) and several other linear and non linear

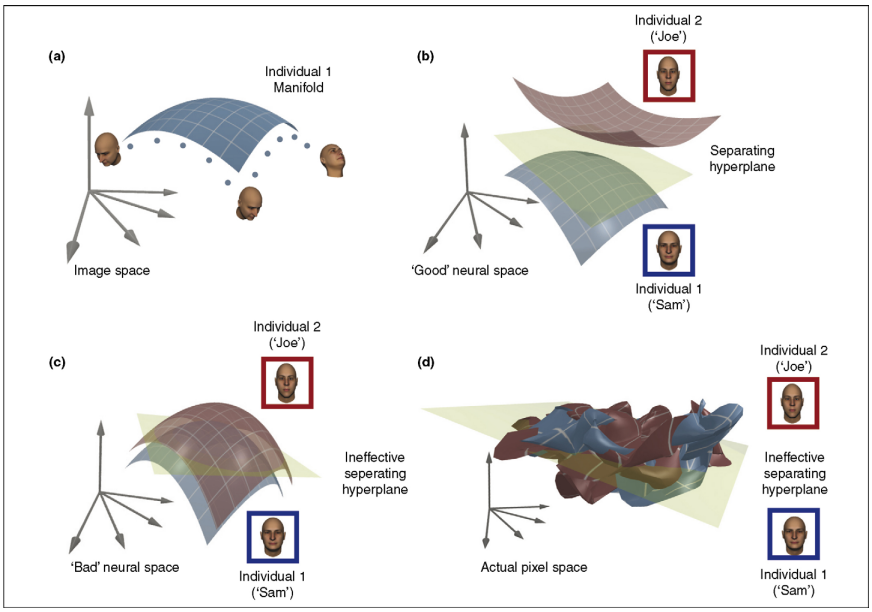

Figure 4: Untangling object manifolds [26]

dimensionality reduction techniques map similar inputs in a high dimensional input space to neighboring points in a low dimensional manifold. These methods however can not be useful by themselves as they map the training samples to the same low dimensional manifold which does not conform to the view under our consideration. These techniques could however be employed with slight modifications along with additional constraints of sparsity of representation to achieve the required change in representation. Discussion of sparsity of representation will be presented in the later subsections.

5.4 Feature extraction

The class of models that the paper focuses on contain linear and non linear transformations of the input images in several stages. Each stage usually consists of two different kinds of layers. The first layer corresponds to a linear filter and the second stage corresponds to a non linear filter. The processing in the visual system can be viewed as a trade off between selectivity and invariance. The linear filters show selectivity to their desired input configurations and the non linear filtering layer achieves invariance by pooling over the cells in the lower layers. In this subsection, we discuss various properties of the linear filtering layer. The simple cells of V1 which can be modeled as Gabor filters are an inspiration to the linear filtering layers in all the stages.

extracts features in one or more stages thus changing the representation of the object. The filters that extract these features in each of the stages can either be learned in a supervised/unsupervised/semi-supervised manner or they can be built

based on previous knowledge of the statistics of the images that will be encountered by the network. Biologically inspired models[14],[27],[37],[41],[46],[36],[52] popularly use Gabor filters[15] for learning low level features of images which have been proven to be good mathematical descriptors of receptive fields of the simple cells in V1[25]. Other models[29] and single stage feature extraction systems use feature descriptors like Scale Invariant Feature Transform (SIFT)[33], Histogram of Gradients(HOG), Geometric Blur[3],[19] [39] to extract low level features. Several evaluation studies[38] compare low level feature descriptors and there seems to be increasing concurrence that SIFT[33] descriptors with some modifications are the best performing feature descriptors for object recognition. However these studies do not evaluate[42] the performance of these feature descriptors in tackling the so called invariance problem[41]. A screening approach similar to [40] which test the effect of different feature descriptors while maintaining all the other parameters on different datasets can determine the ideal feature descriptor to be used in the first stage of the feedforward models. Other models[29] learn these low level features in an unsupervised manner or supervised manner. The neocortex looks remarkably similar across different processing systems which led to the hypothesis that a fundamental principle of cortical computing underlies all the sensory processing systems[49] of the brain and each processing system develops its abilities by learning through statistical properties of the stimuli fed into that processing unit. Supplementing this claim, Oshlausen et al.,[7] show that unsupervised learning using sparse coding(discussed later in the paper) strategy when trained on natural images generates filters which behave like the receptive fields of simple cells in V1(Gabor Filters)[25],[15]. Using Gabor filters to extract features in the low level stages are effective due to their sparseness(dependent on the parameters) and hence can effectively reduce the statistical dependencies[7] between the input images which is essential to untangle object manifolds[26]. [30] also shows that for a generic set of images where the image statistics are not previously known, learning the filters rather than applying random filters(which may have shown good performance on images which do not follow the same statistical properties) considerably improves the performance of these systems on visual recognition.

Although the mathematical behavior of V1 simple cells is well described by Gabor filters, the behavior of the cortical cells in the remaining layers of hierarchy is not well known and no such mathematical descriptor exists that describes the behavior of these cells. Hence these properties need to be learned according to the same principle as we discussed above. SPARSE HMAX[52] attempts to learn these features for all stages using Independent Component Analysis and Standard Sparse Coding[52]. It shows significant improvement in performance over HMAX[37]. Since, learning the first stages

5.5 Nonlinearity

Feature extraction stages involve more than simple feature descriptors. Hierarchical models use a non linear functions such as application of a non linear activation

function, local divisive normalization/local contrast normalization, lateral inhibition, rectification, pooling etc. As the models attempt to untangle manifolds, linear functions are not sufficient to perform the task. Hence non linear functions need to be used to reduce the statistical dependencies between image representations[4],. Methods such as local divisive normalization, lateral inhibition, pooling are biologically inspired as they are found in cortical computing elements in the neocortex. Biologically inspired models use these non linearities and an improved performance has been recorded after applying some of the non linearities as it was expected.

Non linear activation function: All well known and accepted models of neurons like the Hodgkin Huxley Model[22] and even the simplest models of a neuron like the perceptron[43] use thresholding as it is observed in the generation of Action potentials in the neurons. Thresholding is usually modeled by sigmoidal or hyperbolic tangent functions. HMAX inspired models do not explicitly use thresholding as they use a set of prototype features to compare with by defining a similarity function usually using functions with higher degree of kurtosis.

5.6 Sparseness

There are different different kinds of sparseness that we have mentioned in the discussions above. Sparseness of features and sparseness of representation. It is essential to identify the differences between the two.

5.7 Temporal and spatial continuity

As the networks self organize to adapt to the statistics of the natural world and both temporal and spatial continuities are prominent features of the statistics of the visual stimuli, there is no reason for the network to not adapt to both the features. However, if there is redundancy in terms of spatial and temporal continuity then to recreate such a network exploiting such redundancy could lead to higher computational efficiency.

If two objects are presented in close temporal proximity then the question of what is the length of the temporal proximity association arises.

retinotopical organization.

5.8 Miscellaneous observations

This architecture of layer wise unsupervised learning bypasses problems such as local minima which are faced by many deep architectures. Many deep AI architectures struggle with the trade off between depth and viable learning rates.

6 Conclusion

References

[1] Patrik O. Hoyer Aapo Hyvrinen, Jarmo Hurri. *Natural Image Statistics*. Springer London, 2009.

[2] Gott RE. Ashby FG. Decision rules in the perception and categorization of multidimensional stimuli. *Journal of experimental psychology, Learning, memory and cognition*, 1988.

[3] J.Malik Berg A.C. Geometric blur for template matching. *Computer Vision and Pattern Recognition*, 1, 2001.

[4] Matthias Bethge. Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? *Journal of the Optical Society of America*, 23(6):1253 – 1268, June 2006.

[5] Henry G. Boyapati J. Corticofugal axons in the lateral geniculate nucleus of the cat. *Experimental Brain Research*, 52(2):335 – 340, 1984.

[6] David J Field Bruno A Olshausen. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14:481–487, July 2004.

[7] David J.Field Bruno A. Olshausen. Emergence of simple cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(13):607–609, June 1996. Complete Code.

[8] Matteo Carandini, David J. Heeger, and J. Anthony Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *JOURNAL OF NEUROSCIENCE*, 17(21):8621–8644, 1997.

[9] James J. DiCarlo Chou P. Hung. Gabriel Kreiman, Tomaso Poggio. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310, November 2005.

[10] T. N. Wiesel D. H. Hubel. Ferrier lecture: Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London*, 198:1–59, May 1977.

[11] Jeff Hawkins Dileep George. Towards a mathematical theory of cortical micro-circuits. *PLoS Computational Biology*, 5, 2009.

[12] Shimon Edelman. *Representation and recognition in vision*. The MIT Press, 1999.

[13] T.Milward Edmund T. Rolls. A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information based performance measures. *Neural Computation*, 12:2547–2572, 2000.

[14] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

[15] D. Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 93(26):421 – 441, November 1946.

[16] Dileep George. *How the brain might work : A hierarchical and temporal model for leaning and recognition*. PhD thesis, Department of Electrical Engineering, 2008.

[17] Stephen Grossberg. The complementary brain: Unifying brain dynamics and modularity. Technical Report CAS/CNS-TR-98-003, Department of Cognitive and Neural Systems, Boston University, 2000a.

[18] Stephen Grossberg. Towards a unified theory of neocortex: laminar cortical circuits for vision and cognition. *Progress in Brain Research*, 165:79–104, 2007.

[19] A.C. ; Maire M. ; Malik J. Hao Zhang, Berg. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. *Computer Vision and Pattern Recognition*, 2:2126 – 2136, 2006.

[20] J. Hawkins and S. Blakeslee. *On Intelligence*. Henry Holt and Company, 2004.

[21] David J. Heeger. Normalization of simple cells in cat's striate cortex. *Visual Neuroscience*, 9:181–197, December 1992.

[22] A. F Hodgkin, A. L.; Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117(4):500 – 544, 1952.

[23] Rajesh Ranganath Andrew Y. Ng Honglak Lee, Roger Grosse. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[24] Poggio T DiCarlo JJ. Hung CP, Kreiman G. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863 – 866, November 2005.

[25] L Palmer J P Jones. An evaluation of the twodimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 1987.

[26] David D. Cox James J. DiCarlo. Untangling invariant object recognition. *TRENDS in Cognitive Sciences*, 11(8), July 2007.

[27] David G.Lowe Jim Mutch. Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision (IJCV)*, 80(1):45–57, October 2008.

[28] Tanaka K. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109 – 139, 1996.

[29] R. Fergus K. Kavukcuoglu, M. Ranzato and Y. LeCun. Learning invariant features through topographic filter maps. *Computer Vision and Pattern Recognition*, 2009.

[30] Marc' Aurelio Ranzato Yann Le Cunn Kevin Jarret, Koray Kavukcuoglu. What is the best multi stage architecture for object recognition ? *IEEE 12th International Conference on Computer Vision*, pages 2146 – 2153, 2009.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[32] Joel Z. Leibo, Qianli Liao, Fabio Anselmi, and Tomaso Poggio. The invariance hypothesis implies domain-specific regions in visual cortex. *bioRxiv*, 2014.

[33] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[34] Stephen L. Macknik and Susana Martinez-Conde. The role of feedback in visual masking and visual processing. *Advances in Cognitive Psychology*, 3(1-2):125 – 152, 2007.

[35] Michael A. Paradiso Mark F. Bear, Barry W. Connors. *Neuroscience Exploring the Brain*. Williams & Wilkins, 1996.

[36] Rolf P.Wurtz Markus Lessmann. Learning invariant object recognition from temporal correlation in a hierarchical network. *Neural Networks*, 54:70–84, February 2014.

[37] Tomaso Poggio Maximilian Riesenhuber. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019 – 1025, November 1999.

[38] Leibe B. ; Schiele B. Mikolajczyk, K. Local features for object class recognition. *Computer Vision and Pattern Recognition*, 2:1792 – 1799, October 2005.

[39] Bill Triggs Navneet Dalal. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition*, 2005.

[40] David Doukhan James J.DiCarlo Nicolas Pinto, David D.Cox. A high throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, 5(11), November.

[41] James J.DiCarlo Nicolas Pinto, David D.Cox. Why is real world visual object recognition hard ? *PLoS Computational Biology*, 4(1), January 2008.

[42] Youssef Barhomi James J.DiCarlo Nicolas Pinto, David D.Cox. Comparing state of the art visual features on invariant object recognition tasks. *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 463 – 470, January 2011.

[43] Frank Rosenblatt. Perceptron–a perceiving and recognizing automaton. Technical report, Cornell Aeronautical Laboratory, 1957.

[44] Catherine Marlot Simon Thorpe, Denis Fize. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, June 1996.

[45] Maximilian Riesenhuber Thomas Serre. Realistic modeling of simple and complex cell tuning in the hmax model, and implications for invariant object recognition in cortex. Technical report, Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, 2004.

[46] Stanley Bileschi Maximilian Riesenhuber Thomas Poggio Thomas Serre, Lior Wolf. Robust cortex like recognition with cortex like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–425, March 2007.

[47] Tomaso Poggio Thomas Serre, Aude Oliva. A feedforward architecture accounts for rapid categorization. *PNAS*, 104(15):64246429, November 2007.

[48] Mikhail Belkin Ulrike von Luxburg, Olivier Bousquet. *Learning Theory*, chapter On the Convergence of Spectral Clustering on Random Samples: The Normalized Case, pages 457 – 471. Springer Berlin Heidelberg, 2004.

[49] Mountcastle VB. The columnar organization of neocortex. *Brain*, 120(4):701 – 722, April 1997.

[50] Montero V.M. A quantitative study of synaptic contacts on interneurons and relay cells of the cat lateral geniculate nucleus. *Experimental Brain Research*, 86(2):257 – 270, 1991.

[51] Gnter Westphal and Rolf P. Wrtz. Combining feature- and correspondence-based methods for visual object recognition. *Neural Computation*, 21(7):1952–1989, 2009.

[52] Jianmin Li Bo Zhang Xiaolin Hu, Jianwei Zhang. Sparsity regularized hmax for visual recognition. *PLoS ONE*, 9(1), January 2014.