An elastic net-regularized HMAX model of visual processing

Ali Alameer[1,†], Ghazal Ghazaei1[1], Patrick Degenaar[1,2] and Kianoush Nazarpour[1,2,†]

[1] *School of Electrical and Electronic Engineering, Newcastle University, Newcatsle NE1 7RU, UK*
[2] *Institute of Neuroscience, Newcastle University, Newcatsle NE2 4HH, UK*
† *Email for correspondences: {A.m.a.alameer,Kianoush.Nazarpour}@Newcastle.ac.uk*

Keywords: Elastic-net regularization, HMAX, object recognition, sparse coding

Abstract

The hierarchical MAX (HMAX) model of human visual system has been used in robotics and autonomous systems widely. However, there is still a stark gap between human and robotic vision in observing the environment and intelligently categorising the objects. Therefore, improving models such as the HMAX is still topical. In this work, in order to enhance the performance of HMAX in an object recognition task, we augmented it using an elastic net-regularised dictionary learning approach. We used the notion of sparse coding in the S layers of the HMAX model to extract mid- and high-level, i.e. abstract, features from input images. In addition, we used spatial pyramid pooling (SPP) at the output of higher layers to create a fixed feature vectors before feeding them into a softmax classifier. In our model, the sparse coefficients calculated by the elastic net-regularised dictionary learning algorithm were used to train and test the model. With this setup, we achieved a classification accuracy of 82.6387% $\mp$3.7183% averaged across 5-folds which is significantly better than that achieved with the original HMAX.

1 Introduction

Object recognition has attracted a great deal of interest during the last decades. It is an essential step in the direction of building machines that can recognise and interact meaningfully with their surroundings [1]. There are plenty of applications that require fast classification of images; based on features extracted from their pixel content [2]. Current image search and characterization platforms depend on image meta-data and watermarks rather than the image pixel values. Platforms that do make use of pixel values normally depend on previously obtained image features instead of generating and obtaining new features in real-time [3]. A growing body of evidence support the proposition that biological systems are able to recognize an object with different positions and scales after examining it for the first time [4].

The primate brain handles visual information in a parallel and hierarchical structure [5, 6]. Neurons at various stages of the brain ventral pathways have various response characteristics [7]. For instance primary visual area (V1) neurons are responsive to bars at certain orientations, while corners can be detected by neurons in V2 [6].

Inspired by these outcomes, hierarchical models have been suggested to simulate the visual recognition method in the brain. One of the first models in which position invariance and feature complexity emerged into its hierarchical layers is the Neocognitron [8]. Similarly, several computational methods are used to achieve invariance and specificity [5].

It has been shown that hierarchical architectures, such as the HMAX model, outperform the single template object recognition systems in various object recognition tasks [9]. The HMAX model consists of four different brain layers, e.g. V1, V2, and V4. Each of these layers is sub-divided into two layers: a simple and a complex layer. Invariance can be achieved by utilizing the max pooling operation, where each stage of the visual hierarchy receives its input from a pooled units from the previous layer. For this reason, this operation is called pooling. Max pooling is applied to the afferent (i.e. conveying towards the centre) units applied to the same feature but with different scales and positions [10]. If the afferent simple unit activated within the same pooling range, then the complex unit will produce an equal response. If a number of the simple units are active, the response of a complex unit will be equal to the response of the simple unit with the maximum value. This indicates that complex units attain some extent of invariance to spatial position and scale [10]. Selectivity in the original HMAX can be achieved using the template-matching approach over a set of selected prototypes by implementing a radial basis function network [11]. During the training operation, a dictionary of S2 features is produced. As a result, each simple unit is turned to be a specific feature in the S2 dictionary and then selecting the maximum response within all the S2 dictionary features. The response is modelled by a Gaussian function which is a measure of the similarity between the input and the prototype.

Inspired by the brain ventral pathway in terms of selectivity and invariance, the HMAX model provides useful insight of both of these merits. However, the model has some drawbacks. Firstly, S2 template matching is all based on selecting random patches from the C1 layer, which are unlikely to have an explicit resemblance with receptive fields of any neuron [6].

Secondly, the HMAX model only provides a static description in terms of the recognition process. Therefore, the same response can be achieved every time from the same input image. This obviously does not offer the dynamics, complexity and the sparse firing of neuronal populations in the cortex. Thirdly, the object recognition mechanism entirely depends on a hierarchical feed-forward structure, discarding many essential connections which happen to exist across the visual cortex [10], for example the long-range horizontal connections that are responsible for integrating information across the cortical regions [12]. To which extent these are implicated in early stages of immediate object recognition is an open question [13, 14].

We propose to contribute in mitigating the first and third limitations by completely reformulating the HMAX model, by using an elastic net regularizer to perform sparse coding in both higher and lower layers of HMAX. Additionally, we used long-range horizontal connections in all complex layers, in which it contains important features that reinforce the final decision corresponding the object recognition task.

2 Sparse coding and elastic net regularizer

Given an image patch $\mathbf{x_i} \in R^m$ (m denotes the size of the image patch), sparse coding search for a set of bases $\mathbf{d_i} \in R^m$, so that $\mathbf{x_i} = \sum_{j=1}^{p} \mathbf{d_i}\, s_j$, p denotes the number of the coefficients in $\mathbf{s_i} \in R^p$. The coefficients s_j are expected to be sparse (only a limited number of them are non-zero). In the matrix notation, the equation converts to:

$$\mathbf{X} = \mathbf{DS}, \tag{1}$$

where each column of $\mathbf{X}$ is a patch $\mathbf{x_i}$, each column of $\mathbf{D}$ is a basis $\mathbf{d_i}$ and each column of $\mathbf{S}$ is a vector $\mathbf{s_i} \in R^p$ containing the coefficients of the p bases for reconstructing $\mathbf{x_i}$. An elastic net formulation is:

$$\begin{aligned} &\text{minimize} && \|\mathbf{X} - \mathbf{DS}\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 + \lambda_2 \|\mathbf{S}\|_F^2 \\ &\text{subjected to} && \|\mathbf{d_i}\|_2 \leq 1, \forall i = 1, ..., p. \end{aligned} \tag{2}$$

Where $\|.\|_F$ denotes the Frobenius norm. For a given signal $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_n}]$ in $R^{m\times n}$ and a dictionary $\mathbf{D}$ in $R^{m\times p}$, effectively, for every input parameter $\mathbf{x}$, a matrix of coefficients $\mathbf{S} = [\mathbf{s_1}, ..., \mathbf{s_n}]$ in $R^{p\times n}$ will be reproduced. For each column $\mathbf{x}$ in $\mathbf{X}$, the equivalent column $\mathbf{s}$ in $\mathbf{D}$ is the solution of (2). λ_1 and λ_2 are the regularization parameters. They control the trade-off between sparsity and fitting goodness. Where λ_1 increases, the bases become aggressively sparse and more silent. However the reconstruction squared error function becomes incredibly large, which leads to a false description of the bases. On the other hand, λ_2 is used to control the sensitivity in atoms selection [15]. Larger λ_2 will lead to larger reconstruction error with more sparsity.

3 En-HMAX

Inspired by the performances of linear dependencies presented by max pooling, it is suggested to generate linear bases by sparse coding using an elastic net dictionary learning on all of the S layers of HMAX as shown in Fig. 1. The S layer feature maps are produced by a particular number of learned S bases. Each C layer comprised feature maps generated by spatially max pooling over the previous S layer and the elastic net was used to learn the S bases. An over-complete dictionary can be generated, which is often beneficial for image classification because it creates a considerable dictionary size [6]. The elastic net was utilized on small patches sampled from arbitrary positions of C maps. Extracting a patch on a C layer involves sampling the same size of the patch on each map at similar positions. This infers that the bases produced at C layer have dimensions $r \times r \times m_p$, where r denotes the length of the patch on each C map (patches are presumed to be square) and m_p denotes the number of C maps. For image classification, the C1, C2 and C3 maps (in the full En-HMAX model) are concatenated. For SPP a grid resolution of $\{2, 3, 4\}$ was used, so that each S basis in the ultimate S layer of the model produces 29 features. Notice that dissimilar to the S2 codes generated by the original HMAX, the S codes gathered by elastic net can be positive or negative.

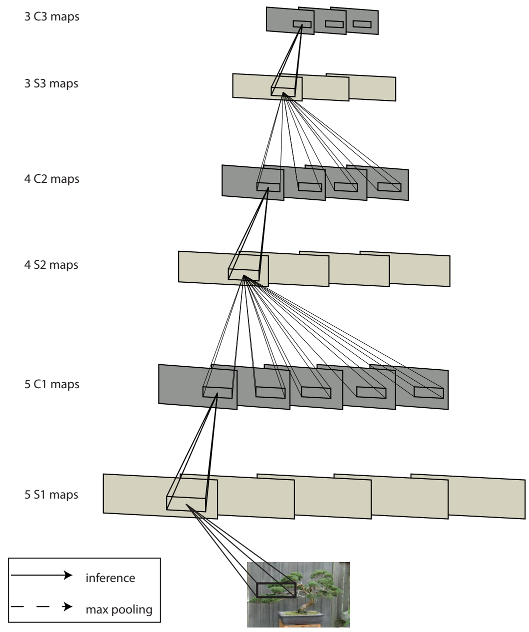

Fig. 1: Illustration of the proposed En-HMAX model.

4 Experimental results

The proposed HMAX model was trained on various classes of Caltech 101 [16] data set with 40 S1 bases of dimensions 10×10, 40 S2 bases of dimensions 12×12×40 and 36 S3 bases of dimensions 13×13×40. The bases were learned by elastic net regularization with 50,000 patches arbitrarily extracted from images or C patches. The training set consisted of 15 and 30 images per category. While the testing set consist of 108 and 93 images per category. Four classes were selected:

Model Architecture	Training Size	
	15	30
En-HMAX	79.340∓1.976	82.365∓4.043
Model 1 (C1+C2)	82.638∓3.718	82.849∓4.038
Model 2 (C2)	69.837∓9.167	74.032∓9.938
Original HMAX [5]	34.143∓16.43	47.311∓4.725

Table 1: Mean classification accuracy (i.e. the ratio between the number of correctly classified testing examples to the number of testing Examples) ∓ standard deviation (SD).

faces-easy, bonsai, planes and car-side. The faces- easy category comprises images of eight different persons in both the training and the testing set.

Results in (Table 1) are averaged across 5-folds. The whole experiments are performed with the default settings. A rough calculation of 140 image resolution, a pooling size of 2 in both first layer and the second layer, and SPP grid resolution of $\{2, 3, 4\}$. Pooling was implemented to each grid to generate the concatenated features. Consequently each feature vector had (40 + 40 + 36)×29 = 3364 dimensions (full model). Regularization parameters λ_1, λ_2 were 0.15.

All models were implemented in Matlab; a softmax classifier [17] was used to perform classification. Table 1 summarizes the results of other models extracted from the full model of the En-HMAX (model 1 and model 2). Model 1 has the same number of layers as the original HMAX [5, 9]. It outperformed previous HMAX models by a large margin. It is also outperformed the En-HMAX due to the small number of S3 bases utilized on this particular experiment.

The generated model was tested with the same parameters and conditions. Fig. 2 shows the classification accuracy of the individual categories when using 15 and 30 images for training the model using the En-HMAX. It can be noted that car-side achieve the highest accuracies while planes categories attains the lowest even when the chance level used in all experiments is even. The average time for one single-threaded operate within our architecture was about 90 minutes. Including parameter initialization, training and testing on Caltech-101 with 15 or 30 training images and up to 108 testing images per class. A Large proportion of time is spent on training the dictionaries (about 60 minutes) to extract features. However, less time is needed to perform cross-validation for testing the samples. The PC used for these experiments was a dual-core i5 processors (3.4 GHz) with 16 G RAM and all timings were calculated on a single thread.

5 Discussion

The new model structure implemented in this work improved the original HMAX [5] in two ways. Firstly, the En-HMAX configuration is scalable, as part of the model, for instance, low-level layers can be exploited to perform a particular recognition task (less challenging one). For challenging tasks, for instance, 15 training and 108 testing images per category, the full model (see experiments) can be exploited to do the task. Secondly, the En-HMAX replaced the template matching (based on selecting random patches as prototype or bases in the S2 layer) proposed in the original HMAX, with sparse coding which is deeply rooted in neuroscience [6]. The En-HMAX model suggests the use of the low level features e.g. corners and edges form different feature maps. Mid-level layers calculate more complex pattern counterparts, the same applies to the last stage where it is mixing all these pairs to make the last decision, and it similarly informs us that positions of the features are vitally reinforced by the $\{2, 3, 4\}$ sub-regions of the spatial pyramid pooling. Using an elastic net-regularizer for dictionary learning in higher layers of the En-HMAX encourages the grouping effect when the atoms in the dictionary are highly correlated.

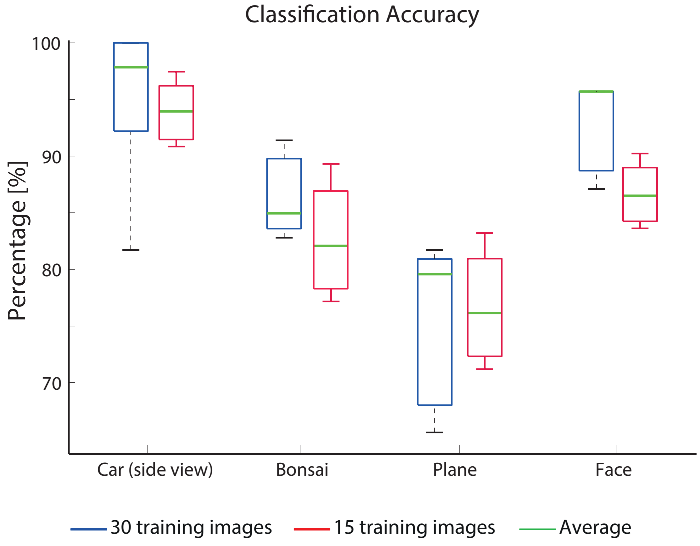

Fig. 2: The accuracy of the individual categories when using 15 and 30 images for training.

6 Conclusion

Conventional HMAX coupled with dictionary learning algorithms with l_1 regularizer have attained good performance for image classification. Nevertheless, while a group of atoms in the dictionary are strongly correlated, the l_1 regularizer have a tendency to choose one atom from the group and neglect the other atoms as explained above. Using elastic net regularizers for HMAX, offers three main benefits. Firstly, the l_2 norm regularizer assists to eliminate the drawback on the amount of selected atoms from dictionary. Secondly, it supports grouping effect, which help the image feature to find atoms tend to match same class of images. We showed that our proposed En-HMAX outperforms the original HMAX model significantly.

Acknowledgements

A. Alameer has a PhD scholarship from the HCED, Iraqi Government. The work of G Ghazaei was supported by a DTA from Newcastle University. The work of K. Nazarpour is supported by EPSRC, UK (grant reference numbers: EP/M025977/1 and EP/M025594/1).

References

[1] G. Orchard, J. G. Martin, R. J. Vogelstein, and R. Etienne-Cummings, "Fast neuromimetic object recognition using FPGA outperforms GPU implementations," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 24, no. 8, pp. 1239–1252, 2013.

[2] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proceedings of the National Academy of Sciences*, vol. 104, no. 15, pp. 6424–6429, 2007.

[3] G. Orchard, C. Meyer, R. Etienne-Cummings, C. Posch, N. Thakor, and R. Benosman, "Hfirst: A temporal approach to object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 1, 2015.

[4] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the seventh IEEE International Conference on*, vol. 2, pp. 1150–1157, 1999.

[5] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[6] X. Hu, J. Zhang, J. Li, and B. Zhang, "Sparsity-regularized HMAX for visual recognition," *PloS One*, vol. 9, no. 1, p. e81813, 2014.

[7] A. Pasupathy and C. E. Connor, "Population coding of shape in area v4," *Nature Neuroscience*, vol. 5, no. 12, pp. 1332–1338, 2002.

[8] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[9] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 3, pp. 411–426, 2007.

[10] S. Dura-Bernal, T. Wennekers, and S. L. Denham, "Top-down feedback in an hmax-like cortical model of object perception based on hierarchical bayesian networks and belief propagation," *PLoS ONE*, vol. 7, p. e48216, 11 2012.

[11] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford university press, 1995.

[12] M. A. Williams, C. I. Baker, H. P. O. de Beeck, W. M. Shim, S. Dang, C. Triantafyllou, and N. Kanwisher, "Feedback of visual object information to foveal retinotopic cortex," *Nature Neuroscience*, vol. 11, no. 12, pp. 1439–1445, 2008.

[13] S. Hochstein and M. Ahissar, "Hierarchies and reverse hierarchies in the visual system," *Perception ECVP abstract*, vol. 29, pp. 0–0, 2000.

[14] T. S. Lee, "Computations in the early visual cortex," *Journal of Physiology-Paris*, vol. 97, no. 2, pp. 121–139, 2003.

[15] B. Shen, B.-D. Liu, and Q. Wang, "Elastic net regularized dictionary learning for image classification," *Multimedia Tools and Applications*, pp. 1–14, 2014.

[16] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[17] K. Duan, S. S. Keerthi, W. Chu, S. K. Shevade, and A. N. Poo, "Multi-category classification by soft-max combination of binary classifiers," in *Multiple Classifier Systems*, pp. 125–134, Springer, 2003.