V1-Inspired Features Induce a Weighted Margin in SVMs

Hilton Bristow^{1,2} and Simon Lucey²

¹Queensland University of Technology, ²CSIRO Australia

Queensland University of Technology
Brisbane Australia

Abstract

Image representations derived from simplified models of the primary visual cortex (V1), such as HOG and SIFT, elicit good performance in a myriad visual classification tasks. A central question in the vision, learning and neuroscience communities regards *why* these architectures perform so well. In this work, we offer a unique perspective to this question by subsuming the role of V1-inspired features directly within a linear support vector machine (SVM).

Contributions

- V1-Inspired features can be rewritten as a linear function of the Kronecker expansion of an image
- This expansion can be viewed as a data-independent matrix, inducing a weighted margin in SVMs
- We can approximate the same prior at lower computational cost
- We garner insights into (i) the uniqueness of filters, and (ii) the capacity of the linear SVM tending towards a quadratic kernel SVM

Vision Pipeline

The convolutional square pooling pipeline can be expressed as

$$\Phi(\mathbf{x}) = [\Phi_1(\mathbf{x}), \Phi_2(\mathbf{x}), \dots, \Phi_F(\mathbf{x})]^T$$

where for a single filter,

$$\Phi_f(\mathbf{x}) = \mathbf{b} * [(\mathbf{g}_f * \mathbf{x}) \odot (\mathbf{g}_f * \mathbf{x})] .$$

$$\min_{\mathbf{w}, \xi_i \geq 0} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi_i$$

subject to $y_i \mathbf{w}^T \Phi(\mathbf{x}_i) \geq 1 - \xi_i$

where

- $\mathbf{x}_i$ = input examples
- y_i = class labels
- elephant? $\rightarrow \{-1, 1\}$
- $\mathbf{w}$ = support vectors
- ξ_i = slack variables

Figure : Exemplar V1-Inspired visual pipeline. The input image is projected into an over-complete Gabor basis, squared to increase edge bandwidth and discard step direction, downsampled/pooled to histogram the resulting edge directions, before being vectorised and used as input to a max-margin classifier (SVM).

Methodology

The Hadamard product between any two equal size vectors $\mathbf{x}_i \in \mathfrak{R}^D$ and $\mathbf{x}_j \in \mathfrak{R}^D$ can be written as,

$$\mathbf{x}_i \odot \mathbf{x}_j = \mathbf{M}(\mathbf{x}_i \otimes \mathbf{x}_j)$$

The full response to a bank of filters can be written as,

$$\Phi(\mathbf{x}) = \mathbf{L}(\mathbf{x} \otimes \mathbf{x})$$

where,

$$\mathbf{L} = \begin{bmatrix} \mathbf{BM}(\mathbf{G}_1 \otimes \mathbf{G}_1) \\ \vdots \\ \mathbf{BM}(\mathbf{G}_F \otimes \mathbf{G}_F) \end{bmatrix} .$$

$$\begin{aligned} \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) &= (\mathbf{x}_i \otimes \mathbf{x}_i)^T \mathbf{L}^T \mathbf{L} (\mathbf{x}_j \otimes \mathbf{x}_j) \\ &= (\mathbf{x}_i \otimes \mathbf{x}_i)^T \mathbf{S} (\mathbf{x}_j \otimes \mathbf{x}_j) \end{aligned}$$

“ The square of a linear function is a linear function of quadratic terms ”

Figure : The eigenspectrum of S for three filter representations.

Figure : Two exemplar S matrices. With large support filters (lower triangle), the matrix is reasonably dense.

Figure : Components of the dimensionality reduction viewed spatially

SVM Formulation

Given a support vector machine (SVM) expressed in its primal form,

$$\min_{\mathbf{w}, \xi_i \geq 0} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi_i$$

subject to $y_i \mathbf{w}^T \Phi(\mathbf{x}_i) \geq 1 - \xi_i, \quad i = 1 \dots l$

the S matrix manifests itself as the inverse of the margin [1],

$$\min_{\mathbf{v}, \xi_i \geq 0} \frac{1}{2} \mathbf{v}^T \mathbf{S}^{-1} \mathbf{v} + C \sum_{i=1}^l \xi_i$$

subject to $y_i \mathbf{v}^T (\mathbf{x}_i \otimes \mathbf{x}_i) \geq 1 - \xi_i, \quad i = 1 \dots l$

Figure : An example data distribution with separating planes drawn for different noise samples. With such a limited training set, good generalisation is difficult. Analogously, the margin prior helps guide the hyperplane in noisy or data-limited situations.

Figure : An example linear transform. Shearing the data does not change the margin, but the outer points have better separation

Results

Figure : Average classification performance across all classes of the (a) MNIST, and (b) Caltech 101 dataset for different feature descriptor representations.

Figure : A comparison of dimensionality reduction techniques on the Caltech 101 dataset.

Figure : Classification performance on Cohn Kanade+ broad expressions as a function of increasing registration error.

References

- [1] A. B. Ashraf, S. Lucey, and T. Chen. Reinterpreting the application of Gabor filters as a manipulation of the margin in linear support vector machines. *Pattern Analysis and Machine Learning*, 32(7):1335–41, July 2010.