Learning of Separable Filters by Stacked Fisher Convolutional Autoencoders

Arash Shahriari
arash.shahriari@anu.edu.au;csiro.au

Australian National University (ANU)
Commonwealth Scientific & Industrial Research Organisation (CSIRO)
Canberra, Australia

Abstract

Learning of convolutional filters in deep neural networks proves high efficiency to provide sparse representations for the purpose of image recognition. The computational cost of these networks can be alleviated by focusing on separable filters to reduce the number of learning parameters. Autoencoders are a family of powerful deep networks to build scalable generative models for automatic feature learning. Inspired by their stacked hierarchy, we introduce Fisher convolutional autoencoders to learn separable filters in a distributed architecture. These novel overcomplete autoencoders employ discriminant analysis to impose the highest possible distinction among texture classes whilst holds the minimum separation within each individual class. A distributed network of stacked Fisher autoencoders learns banks of separable filters in parallel and makes an ensemble of deep convolutional features with higher separability for a better classification. This network automatically adjusts depth of each stack with respect to the capability of its correspondent separable filter on extracting higher order convolutional features for the dataset under study. We conduct our experiments on several publicly available datasets varying in number of classes and quality of samples by using a standard implementation. Our results confirm the supremacy of our method on improving the precision of texture understanding in comparison with the recently published benchmarks.

A Mathematics of Projection/Backprojection

We start by formulating of a classical dimension reduction problem and extend its solution to our proposed supervised projection/backprojection paradigms.

Given a sample set $\mathcal{X} = \{X_1, X_2, \ldots, X_{|\mathcal{X}|}\}$ in $\mathbb{R}^d$, we try to find a matrix $\mathbf{B} \in \mathbb{R}^{d\times c}$ that maps the input vector x_i onto the point $y_i = \mathbf{B}^T x_i$ in a lower dimensional space $\mathbb{R}^c$ conditioned on $c \ll d$ by maximizing the separability between and minimizing the scattering within classes of set $\mathcal{X}$.

One of the solutions is supervised learning via linear discriminant analysis (LDA) [2]. The mapping matrix $\mathbf{B}$ is determined to maximize the Fisher criterion given by

$$\mathcal{J}_{\mathcal{F}}(\mathbf{B}) = tr\left(\left(\mathbf{B}\,\mathbf{S}_{\mathbf{w}B}\,\mathbf{B}^T\right)^{-1} \left(\mathbf{B}\,\mathbf{S}_{\mathbf{b}B}\,\mathbf{B}^T\right) \right) \tag{1}$$

which $tr(.)$ is diagonal summation operator. The within/between-class scatterings $\{\mathbf{S_{w}}_B, \mathbf{S_{b}}_B\}$ are defined as

$$\mathbf{S_{w}}_B = \sum_{j=1}^{c} \sum_{x_i \in \mathbf{C}_j} (x_i - \mu_j)(x_i - \mu_j)^T \tag{2}$$

$$\mathbf{S_{b}}_B = \sum_{j=1}^{c} (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T \tag{3}$$

where c, μ_j and $\bar{\mu}$ are number of classes, mean over class $\mathbf{C}_j$ and mean over all dataset, respectively.

The matrix $\mathbf{S_{w}}_B$ can be regarded as the average class-specific covariance, whereas $\mathbf{S_{b}}_B$ can be viewed as the mean distance between all different classes. Thus, the purpose of Equation 1 is to maximize the between-class scatter while preserving within-class dispersion.

The answer is the solution of generalized eigenvalue problem $\mathbf{S_{b}}_B\,\mathbf{B} = \lambda\,\mathbf{S_{w}}_B\,\mathbf{B}$. Since the rank of $\mathbf{S_{b}}_B$ is $c-1$, the solution for c classes is eigenvectors corresponding to the largest $c-1$ eigenvalues of $\mathbf{S_{w}}_B^{-1}\,\mathbf{S_{b}}_B$ for $c \ll d$ [1].

Assuming $\mathbf{S_{w}}_B \neq \mathbf{I}$ and $\mathbf{S_{b}}_B \neq \mathbf{I}$, by cyclic permutation of trace operator and imposing orthogonality through $\mathbf{B}\,\mathbf{B}^T = \mathbf{I}$ (identity matrix), the Equation 1 holds

$$\begin{aligned}
\mathcal{J}_{\mathcal{F}}(\mathbf{B}) &= tr\Big(\big(\mathbf{B}\,\mathbf{S_{w}}_B\,\mathbf{B}^T\big)^{-1}\big(\mathbf{B}\,\mathbf{S_{b}}_B\,\mathbf{B}^T\big)\Big) \\
&= tr\Big(\big(\mathbf{B}^T\big)^{-1}\mathbf{S_{w}}_B^{-1}\,\mathbf{B}^{-1}\mathbf{B}\,\mathbf{S_{b}}_B\,\mathbf{B}^T\Big) \\
&= tr\Big(\mathbf{B}^T\big(\mathbf{B}^T\big)^{-1}\,\mathbf{S_{w}}_B^{-1} \times \mathbf{I} \times \mathbf{S_{b}}_B\Big) \\
&= tr\Big(\mathbf{S_{w}}_B^{-1}\,\mathbf{S_{b}}_B\Big)
\end{aligned} \tag{4}$$

To come up with our proposed projection for $c > d$, we again consider the Fisher criterion in Equation 1 and redefine inter-class scattering $\mathbf{S_{w}}_A \in \mathbb{R}^{c\times c}$ such that it satisfies

$$tr(\mathbf{S_{w}}_B) = tr(\mathbf{S_{w}}_A) \tag{5}$$

Note that in Equations 2, we sum over all classes (c) and hence, to satisfy Equation 5, we can consider $\mathbf{S_{w}}_A$ as a square matrix of size $c \times c$ with all zeros except main diagonal entries

$$\mathbf{S_{w}}_A(j,j) = tr\Big(\sum_{x_i \in \mathbf{C}_j} (x_i - \mu_j)(x_i - \mu_j)^T\Big) \quad \forall j \in [1,c] \tag{6}$$

From Equations 5 and similarity invariance of trace operator, $\mathbf{S_{w}}_B$ and $\mathbf{S_{w}}_A$ are similar matrices [3] which implies, there should exist a non-singular matrix $\mathbf{\Gamma_w}$ such that

$$\mathbf{S_{w}}_B = \mathbf{\Gamma_w}^{-1}\,\mathbf{S_{w}}_A\,\mathbf{\Gamma_w} \tag{7}$$

By minor matrix operations, Equation 7 can be formed as

$$\mathbf{\Gamma_w}\,\mathbf{S_{w}}_B - \mathbf{S_{w}}_A\,\mathbf{\Gamma_w} = \mathbf{0} \tag{8}$$

which is a special case of Sylvester equation [4] and can be solved for $\mathbf{\Gamma_w}$ by either Kronecker tensor trick or using generalized eigen decomposition because, we define $\mathbf{S_{wB}}$ and $\mathbf{S_{wA}}$ as non-singular matrices. The closed form solution for Equation 8 is

$$vec(\mathbf{\Gamma_w}) = \mathbf{I} \otimes (-\mathbf{S_{wA}}) - \mathbf{S_{wB}^T} \otimes \mathbf{I} \tag{9}$$

which $vec(.)$ is vectorization operator. With the same reasoning, we define $\mathbf{S_{bA}}$ as a square matrix of size $c \times c$ such that

$$tr\big(\mathbf{S_{bB}}\big) = tr\big(\mathbf{S_{bA}}\big) \tag{10}$$

and there should exist a non-singular matrix $\mathbf{\Gamma_b}$ such that

$$\mathbf{S_{bB}} = \mathbf{\Gamma_b^{-1}}\, \mathbf{S_{bA}}\, \mathbf{\Gamma_b} \tag{11}$$

On the other hand, from Equations 7 and 11

$$\begin{aligned} \mathbf{S_{wB}^{-1}}\, \mathbf{S_{bB}} &= \big(\mathbf{\Gamma_w^{-1}}\, \mathbf{S_{wA}}\, \mathbf{\Gamma_w}\big)^{-1} \big(\mathbf{\Gamma_b^{-1}}\, \mathbf{S_{bA}}\, \mathbf{\Gamma_b}\big) \\ &= \mathbf{\Gamma_w^{-1}}\, \mathbf{S_{wA}^{-1}}\, \mathbf{\Gamma_w}\, \mathbf{\Gamma_b^{-1}}\, \mathbf{S_{bA}}\, \mathbf{\Gamma_b} \end{aligned} \tag{12}$$

Due to the similarity invariance in Equations 5 and 10, we consider the cyclic permutation of trace operator and suppose that

$$\mathbf{\Gamma_b} = \mathbf{\Gamma_w} \tag{13}$$

and hence, Equation 11 implies $\mathbf{S_{bA}}$ as

$$\mathbf{S_{bA}} = \mathbf{\Gamma_b}\, \mathbf{S_{bB}}\, \mathbf{\Gamma_b^{-1}} \tag{14}$$

Now, we work out Equation 12 by substitution from Equation 13 as follows

$$\begin{aligned} \mathbf{S_{wB}^{-1}}\, \mathbf{S_{bB}} &= \mathbf{\Gamma_w^{-1}}\, \mathbf{S_{wA}^{-1}} \times \mathbf{I} \times \mathbf{S_{bA}}\, \mathbf{\Gamma_b} \\ &= \mathbf{\Gamma_w^{-1}} \big(\mathbf{S_{wA}^{-1}}\, \mathbf{S_{bA}}\big)\, \mathbf{\Gamma_w} \end{aligned} \tag{15}$$

that proves $\mathbf{S_{wB}^{-1}}\, \mathbf{S_{bB}}$ and $\mathbf{S_{wA}^{-1}}\, \mathbf{S_{bA}}$ are similar matrices such that it holds

$$tr\big(\mathbf{S_{wB}^{-1}}\, \mathbf{S_{bB}}\big) = tr\big(\mathbf{S_{wA}^{-1}}\, \mathbf{S_{bA}}\big) \tag{16}$$

Looking back at Equation 4, we are able to define a new optimization problem for $\mathbf{S_{wA}}$ and $\mathbf{S_{bA}}$ considering the same discrimination power and orthogonality of Equation 1 as

$$\mathcal{J}_{\mathcal{F}}(\mathbf{A}) = tr\Big(\big(\mathbf{A}\, \mathbf{S_{wA}}\, \mathbf{A}^T\big)^{-1} \big(\mathbf{A}\, \mathbf{S_{bA}}\, \mathbf{A}^T\big)\Big) \tag{17}$$

which is aligned with the number of classes (c) instead of the dimension of the input (d). Employing the same eigenvector solution of Equation 1 to maximize Equation 17, gives the projection matrix $\mathbf{A} \in \mathbb{R}^{d \times c}$ for $c > d$.

B Closed Forms of Gradients

Suppose that $\mathcal{H}(\mathbf{A})$ is composed of $\mathcal{H}_1(\mathbf{A})$ and $\mathcal{H}_2(\mathbf{A})$ as follows

$$\mathcal{H}_1(\mathbf{A}) = \frac{tr\big(\mathbf{A}\,\mathbf{S}_{\mathbf{w}A}\,\mathbf{A}^{\mathbf{T}}\big)}{tr\big(\mathbf{A}\,\mathbf{S}_{\mathbf{b}A}\,\mathbf{A}^{\mathbf{T}}\big)} \tag{18}$$

$$\mathcal{H}_2(\mathbf{A}) = \|\mathbf{I} - \mathbf{A}\,\mathbf{A}^{\mathbf{T}}\|_2 \tag{19}$$

According to matrix calculus [5],

$$\frac{\partial\, tr\big(\mathbf{A}\,\mathbf{S}_{\mathbf{w}A}\,\mathbf{A}^{\mathbf{T}}\big)}{\partial \mathbf{A}} = \big(\mathbf{S}_{\mathbf{w}A}^{\mathbf{T}} + \mathbf{S}_{\mathbf{w}A}\big)\,\mathbf{A}^{\mathbf{T}} \tag{20}$$

$$\frac{\partial\, tr\big(\mathbf{A}\,\mathbf{S}_{\mathbf{b}A}\,\mathbf{A}^{\mathbf{T}}\big)}{\partial \mathbf{A}} = \big(\mathbf{S}_{\mathbf{b}A}^{\mathbf{T}} + \mathbf{S}_{\mathbf{b}A}\big)\,\mathbf{A}^{\mathbf{T}} \tag{21}$$

and hence, we have

$$\begin{aligned}\frac{\partial \mathcal{H}_1}{\partial \mathbf{A}} &= \frac{\big(\mathbf{S}_{\mathbf{b}A}^{\mathbf{T}} + \mathbf{S}_{\mathbf{b}A}\big)\,\mathbf{A}^{\mathbf{T}} \times tr\big(\mathbf{A}\,\mathbf{S}_{\mathbf{w}A}\,\mathbf{A}^{\mathbf{T}}\big)}{\Big(tr\big(\mathbf{A}\,\mathbf{S}_{\mathbf{b}A}\,\mathbf{A}^{\mathbf{T}}\big)\Big)^2} \\ &- \frac{\big(\mathbf{S}_{\mathbf{w}A}^{\mathbf{T}} + \mathbf{S}_{\mathbf{w}A}\big)\,\mathbf{A}^{\mathbf{T}} \times tr\big(\mathbf{A}\,\mathbf{S}_{\mathbf{b}A}\,\mathbf{A}^{\mathbf{T}}\big)}{\Big(tr\big(\mathbf{A}\,\mathbf{S}_{\mathbf{b}A}\,\mathbf{A}^{\mathbf{T}}\big)\Big)^2}\end{aligned} \tag{22}$$

On the other hand,

$$\frac{\partial \mathcal{H}_2}{\partial \mathbf{A}} = \frac{\partial\big(\mathbf{I} - \mathbf{A}\,\mathbf{A}^{\mathbf{T}}\big)}{\partial \mathbf{A}} \times \frac{\mathbf{I} - \mathbf{A}\,\mathbf{A}^{\mathbf{T}}}{\|\mathbf{I} - \mathbf{A}\,\mathbf{A}^{\mathbf{T}}\|_2} \tag{23}$$

which gives

$$\frac{\partial \mathcal{H}_2}{\partial \mathbf{A}} = \frac{-2\,\mathbf{A}^{\mathbf{T}} \times \big(\mathbf{I} - \mathbf{A}\,\mathbf{A}^{\mathbf{T}}\big)}{\|\mathbf{I} - \mathbf{A}\,\mathbf{A}^{\mathbf{T}}\|_2} \tag{24}$$

The derivatives for $\mathcal{Q}(\mathbf{B})$ can be calculated with the same reasoning as

$$\begin{aligned}\frac{\partial \mathcal{Q}_1}{\partial \mathbf{B}} &= \frac{\big(\mathbf{S}_{\mathbf{b}B} + \mathbf{S}_{\mathbf{b}B}^{\mathbf{T}}\big)\,\mathbf{B} \times tr\big(\mathbf{B}^{\mathbf{T}}\,\mathbf{S}_{\mathbf{w}B}\,\mathbf{B}\big)}{\Big(tr\big(\mathbf{B}^{\mathbf{T}}\,\mathbf{S}_{\mathbf{b}B}\,\mathbf{B}\big)\Big)^2} \\ &- \frac{\big(\mathbf{S}_{\mathbf{w}B} + \mathbf{S}_{\mathbf{w}B}^{\mathbf{T}}\big)\,\mathbf{B} \times tr\big(\mathbf{B}^{\mathbf{T}}\,\mathbf{S}_{\mathbf{b}B}\,\mathbf{B}\big)}{\Big(tr\big(\mathbf{B}^{\mathbf{T}}\,\mathbf{S}_{\mathbf{b}B}\,\mathbf{B}\big)\Big)^2}\end{aligned} \tag{25}$$

$$\frac{\partial \mathcal{Q}_2}{\partial \mathbf{B}} = \frac{-2\,\mathbf{B} \times \big(\mathbf{I} - \mathbf{B}^{\mathbf{T}}\,\mathbf{B}\big)}{\|\mathbf{I} - \mathbf{B}^{\mathbf{T}}\,\mathbf{B}\|_2} \tag{26}$$

References

[1] Christopher M Bishop. Pattern recognition. *Machine Learning*, 2006.

[2] Reinosuke FUKUNAGA. Statistical pattern recognition. 1990.

[3] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[4] Sang-Gu Lee and Quoc-Phong Vu. Simultaneous solutions of sylvester equations and idempotent matrices separating the joint spectrum. *Linear Algebra and its Applications*, 435(9):2097–2109, 2011.

[5] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7:15, 2008.