Class-specific Sparse Coding for Learning of Object Representations

Stephan Hasler, Heiko Wersing, Edgar Körner

2005

Preprint:

This is an accepted article published in Proceedings of the 15th International Conference on Artificial Neural Networks ICANN. The final authenticated version is available online at: https://doi.org/[DOI not available]

Class-Specific Sparse Coding for Learning of Object Representations

Stephan Hasler, Heiko Wersing, and Edgar Körner

Honda Research Institute Europe GmbH,
Carl-Legien-Str. 30, 63073 Offenbach am Main, Germany
stephan.hasler@honda-ri.de

Abstract. We present two new methods which extend the traditional sparse coding approach with supervised components. The goal of these extensions is to increase the suitability of the learned features for classification tasks while keeping most of their general representation performance. A special visualization is introduced which allows to show the principal effect of the new methods. Furthermore some first experimental results are obtained for the COIL-100 database.

1 Introduction

Sparse coding [4] searches for a linear code representing the data. Its target is to combine efficient reconstruction with a sparse usage of the representing basis, resulting in the following cost function:

$$P_{\mathrm{S}} = \frac{1}{2} \sum_i \left(\mathbf{x}_i - \sum_j c_{ij} \mathbf{w}_j \right)^2 + \gamma \sum_i \sum_j \Phi\left(c_{ij}\right) . \tag{1}$$

The left reconstruction term approximates each input $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})^T$ by a linear combination $\mathbf{r}_i = \sum_j c_{ij} \mathbf{w}_j$ of the weights $\mathbf{w}_j = (w_{j1}, w_{j2}, \ldots, w_{jK})^T$, where $\mathbf{r}_i$ is called the reconstruction of the corresponding $\mathbf{x}_i$. The coefficients c_{ij} specify how much the j^{th} weight is involved in the reconstruction of the i^{th} data vector. The right sparsity term sums up the c_{ij}. The nonlinear function Φ (e.g. $\Phi(\cdot) = |\cdot|$) increases the costs, the more the activation is spread over different c_{ij}, and so many of them become zero while few are highly activated. The influence of the sparsity term is scaled with the positive constant γ.

An adaptation of the sparse coding is the nonnegative sparse coding [7]. In this approach the coefficients and the elements of the weights are kept positive. This forces the weights to become more distinct and produces a parts-based representation similar to that obtained by nonnegative matrix factorization [2] with sparseness constraints [1].

In Sect. 2 we introduce two new class-specific extensions of the nonnegative sparse coding. We visualize their principal effect with a simple 2D example to analyze the influence of certain parameters of the cost functions. Furthermore some first experimental results are obtained for the COIL-100 database [3] in Sect. 3. Section 4 gives a short conclusion of the results.

W. Duch et al. (Eds.): ICANN 2005, LNCS 3696, pp. 475–480, 2005.

2 Class-Specific Sparse Coding

The sparse coding features are useful for general image representation but lack the property of being class-specific, and so their use in classification tasks is limited. Our two new approaches extend the nonnegative sparse coding with supervised components. In the first approach the class information has direct effect on the coefficients and it will therefore be referred to as coefficient coding:

$$P_{\mathrm{C}} = \frac{1}{2} \sum_i \left(\mathbf{x}_i - \sum_j c_{ij} \mathbf{w}_j \right)^2 + \gamma \sum_{i,j} c_{ij} + \frac{1}{2} \alpha \sum_j \sum_{\substack{i,\tilde{\imath} \\ Q_i \neq Q_{\tilde{\imath}}}} c_{ij} c_{\tilde{\imath}j} \; . \tag{2}$$

The right coefficient term causes costs if coefficients belonging to the same weight $\mathbf{w}_j$ are active for representatives $\mathbf{x}_i$ and $\mathbf{x}_{\tilde{\imath}}$ of different classes Q_i and $Q_{\tilde{\imath}}$ respectively. Q_i stands for the class of a data vector $\mathbf{x}_i$. The influence of the coefficient term is scaled with the positive constant α. In the second approach the class information has a more direct effect on the weights and it will therefore be referred to as weight coding:

$$P_{\mathrm{W}} = \frac{1}{2} \sum_i \left(\mathbf{x}_i - \sum_j c_{ij} \mathbf{w}_j \right)^2 + \gamma \sum_{i,j} c_{ij} + \frac{1}{2} \beta \sum_j \sum_{\substack{i,\tilde{\imath} \\ Q_i \neq Q_{\tilde{\imath}}}} \left(\mathbf{x}_i^T \mathbf{w}_j \right) \left(\mathbf{x}_{\tilde{\imath}}^T \mathbf{w}_j \right) \; . \tag{3}$$

The right weight term is similar to a linear discriminator and causes costs if a $\mathbf{w}_j$ has a high inner product with representatives $\mathbf{x}_i$ and $\mathbf{x}_{\tilde{\imath}}$ of different classes Q_i and $Q_{\tilde{\imath}}$ respectively. The weight term is scaled with the positive constant β.

The minimization of both cost functions is done by alternately applying coefficient and weight steps as described in [7]. In the coefficient step the cost function is minimized with respect to the c_{ij} using an asynchronous fixed-point search, while keeping the $\mathbf{w}_j$ constant. The weight step is a single gradient step with a fixed step size in the $\mathbf{w}_j$, keeping the c_{ij} constant.

In Fig. 1 we introduce our special visualization schematically and apply it to the nonnegative sparse coding, and in Fig. 2 it is used to compare coefficient and weight coding.

The coefficient coding restricts the use of features through different classes. This means that each feature concentrates on a single class and so the influence of other classes is weakened. There is no increase in discriminative quality, because this would require a strong interaction of different classes. The weight coding instead has a more direct effect on the features and removes activation from them, that is present in different classes. So the features represent more typical aspects of certain objects and so their suitability for object representation and recognition is increased. The cost function shows similarity to Fishers linear discriminant and the MRDF approach [5] but does not minimize the intra class variance. The advantage of the weight coding is that it can produce an overcomplete representation while the number of features in the Fisher linear discriminant is limited by the number of classes and in the MRDF by the num-

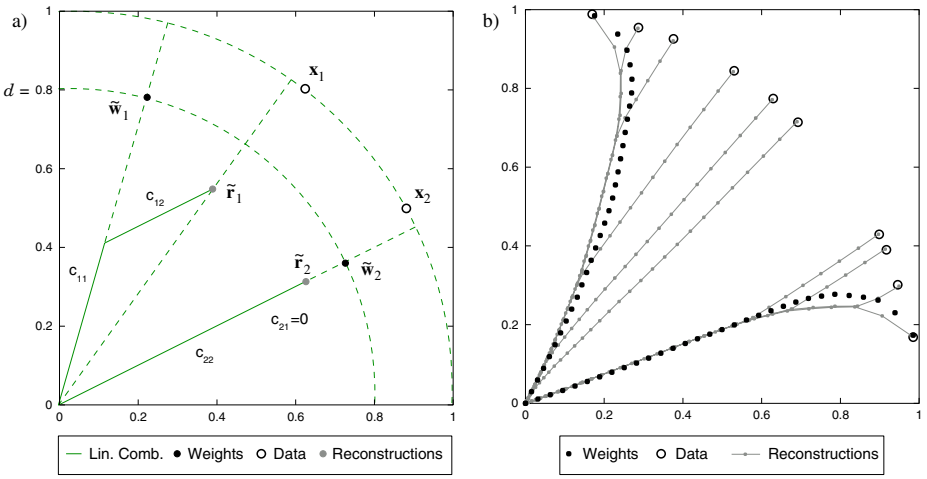

Fig. 1. a) Schematic description of visualization. In the visualization the positions of the data vectors, their reconstructions and the weights are plotted for different values of a control parameter of the cost function, e.g. $\gamma = \gamma_{min} \dots \gamma_{max}$. The data vectors $\mathbf{x}_i$ lie on the unit circle. The shown $\widetilde{\mathbf{w}}_j = d\mathbf{w}_j$ are the weights which have been scaled by a factor $d = (\gamma_{max} - \gamma)/(\gamma_{max} - \gamma_{min})$. Similarly the $\widetilde{\mathbf{r}}_i = d\mathbf{r}_i = d\sum_j c_{ij}\mathbf{w}_j$ are the reconstructions scaled by the same factor. The scaling causes the $\widetilde{\mathbf{r}}_i$ and the $\widetilde{\mathbf{w}}_j$ to move towards the origin with increasing γ. This simply should increase the ability to distinguish the points belonging to different values of γ (see b). Because the weights $\mathbf{w}_j$ are normalized, the distance of the $\widetilde{\mathbf{w}}_j$ from the origin is as big as the scaling factor d. The distance of the $\widetilde{\mathbf{r}}_i$ from the origin is also influenced by the cost function itself. In the nonnegative case it is always shortened, since low values of the basis coefficients c_{ij} are enforced. From the position of the $\widetilde{\mathbf{r}}_i$ and the $\widetilde{\mathbf{w}}_j$ it is possible to determine the coefficients c_{ij} and hence to judge the sparsity of the reconstructions of certain $\mathbf{x}_i$. For example $\mathbf{x}_2$ is reconstructed sparsely, because it only uses $\mathbf{w}_2$. In the visualization the linear combinations will not be plotted and there will be points for each $\widetilde{\mathbf{r}}_i$ and $\widetilde{\mathbf{w}}_1$, $\widetilde{\mathbf{w}}_2$ for each value of γ. The points belonging to the same value of γ could be determined by counting, preferably from the unit circle to the origin. b) Visualization of the influence of the parameter γ on the nonnegative sparse coding. Two scaled weights and the scaled reconstructions of 10 data vectors are plotted for 31 different influences γ of the sparsity term. The scaling factor is $d = (\gamma_{max} - \gamma)/(\gamma_{max} - \gamma_{min})$. The reconstructions belonging to successive values of γ are connected. For $\gamma = \gamma_{min} \to 0$ the algorithm searches for the sparsest perfect reconstruction. And since $d = 1$ in this case the $\widetilde{\mathbf{r}}_i$ lie directly on the $\mathbf{x}_i$. The corresponding $\widetilde{\mathbf{w}}_j$ are aligned with the outermost $\mathbf{x}_i$. With increasing γ the points move towards the origin. Each $\widetilde{\mathbf{r}}_i$ gives up the use of the less suitable weight and therefore the $\widetilde{\mathbf{r}}_i$ unite to two main paths. Each $\widetilde{\mathbf{w}}_j$ aligns to the center of the $\widetilde{\mathbf{r}}_i$ which are assigned to it.

ber of dimensions in the data. The two parameters have to be chosen carefully. When the influence of the sparsity term is too weak, many features tend to represent the same activation.

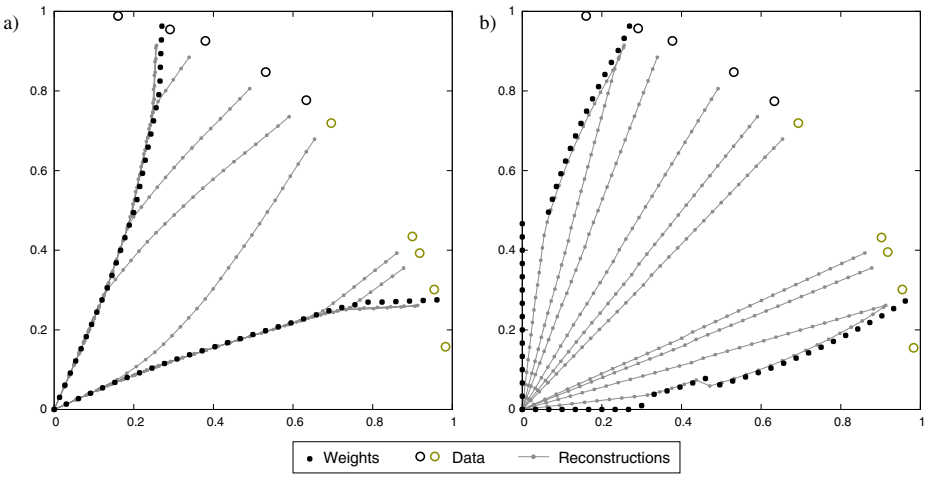

Fig. 2. Visualization of the influence of the parameters α and β on the coefficient and weight coding. The 10 data vectors are now assigned to 2 classes and again are reconstructed using two weights. The influence of the sparsity term is the same constant value $\gamma = 0.05$ in both cases. a) In the coefficient coding the influence α of the coefficient term varies in a defined range. The scaling is $d = (\alpha_{max} - \alpha)/(\alpha_{max} - \alpha_{min})$. With increasing α the points are pulled to the origin. Each weight is forced to specialize on a certain class and therefore moves to the center of this class. There is no gain in discriminative power. b) In the weight coding the influence β of the weight term varies in a defined range. The scaling is $d = (\beta_{max} - \beta)/(\beta_{max} - \beta_{min})$. With increasing β the points are pulled to the origin. The suitability of the weights for different classes is reduced and each aligns to activation which is most typical for the class it is representing. So one weight moves towards the x-axis and the other one towards the y-axis. In the nonnegative case this can be referred to as a gain in discriminative power.

3 Experimental Results

To underpin the qualitative difference between coefficient coding and weight coding both approaches have been applied to a more complex problem. Nine car objects and nine box objects from the COIL-100 database [3], each with 72 rotation views, were combined to two classes (see Fig. 3). Fifty features were trained using the same influence $\gamma = 0.1$ of the sparsity term and relative high

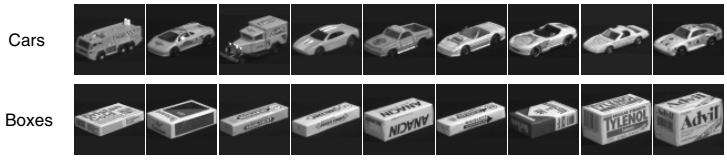

Fig. 3. Two class problem. Nine cars and nine boxes were combined to two classes.

Table 1. The table shows the values of the different terms, neglecting their actual influences (γ, α, and β)to the cost functions. Note that values in brackets were not used for optimization, but are shown to highlight qualitative differences between features.

	Reconstruction	Sparsity	Coefficient	Weight
Nonneg. Sparse Coding	$4.182 \cdot 10^4$	$4.293 \cdot 10^4$	$(1.726 \cdot 10^7)$	$(1.428 \cdot 10^{10})$
Coefficient Coding	$4.682 \cdot 10^4$	$3.872 \cdot 10^4$	$5.709 \cdot 10^5$	$(1.731 \cdot 10^{10})$
Weight Coding	$4.031 \cdot 10^4$	$4.976 \cdot 10^4$	$(2.393 \cdot 10^7)$	$1.035 \cdot 10^{10}$

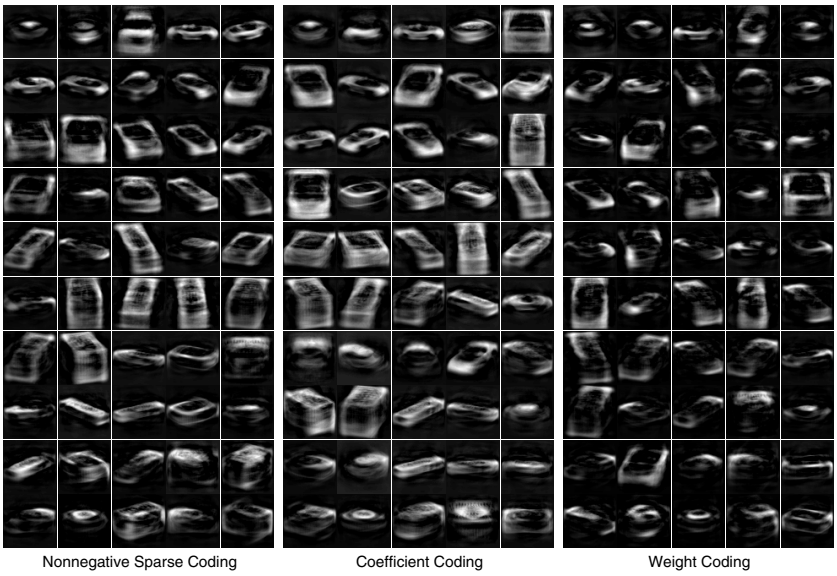

Fig. 4. Features for three approaches. For the visualization, we arranged the features in the following way: Each feature is assigned to the class in which it is most often detected. A feature is detected if its normalized cross-correlation with an image exceeds a feature-specific threshold. This threshold was determined as to maximize the mutual information conveyed by the detection of the feature about the classes (see [6]). The more car-like features start at the top-left and the more box-like features at the bottom-right. The features in one class are arranged by descending mutual information. Therefore the least informative features of the two classes meet somewhere in the middle. The features of the nonnegative sparse coding and the coefficient coding are very similar to each other. The features of the weight coding are sparser and concentrate on more typical class attributes, like certain parts of the cars or the vertices of the boxes.

values for the influence $\alpha = 1 \cdot 10^{-3}$ of the coefficient term and the influence $\beta = 5 \cdot 10^{-7}$ of the weight term.

The resulting features are shown in Fig. 4. Table 1 lists the values of the terms of the cost functions after optimization. These values are useful to interpret the effect of our two new approaches compared to the nonnegative sparse coding:

Since the coefficient term puts a penalty on the use of features across different classes, a splitting into partial class problems is to be observed, leading to a reduced feature basis for each class. As a result, there is an increase of the reconstruction costs and a decrease of the sparsity costs. The demand for sparsity of the coefficients in the nonnegative sparse coding has an opposite effect on the weights, forcing them to become very view-specific and leading to a higher reconstruction cost. In the weight coding the weight term removes activation from the features. They become less view-specific, which causes a decrease of the reconstruction costs, but an increase of the sparsity costs.

4 Conclusion

In this paper two new class-specific extensions of the nonnegative sparse coding were introduced. It was shown that the coefficient coding, by restricting the use of features through different classes, does not increase the discriminative quality of the features, but instead tends to cause a splitting into partial problems, using a distinct feature basis for representing each class. In contrast to that, the weight coding directly penalizes the suitability of features for different classes and so successfully combines representative and discriminative properties. This combination produces features which are more suitable for object representation than features with general representative quality only. The advantage of the weight coding is that it can produce an overcomplete representation, whereas most other approaches are using the covariance matrix directly, and so the number of features is limited by the number of dimensions in the data. The drawback is that two parameters have to be tuned suitably. Also the intra class variance is not reduced as e.g. in the MRDF approach. The evaluation of the usefulness of the weight coding features in object recognition will be subject to further investigations.

References

1. Hoyer, P.: Non-negative Matrix Factorization with Sparseness Constraints. Journal of Machine Learning Research **5** (2004) 1457-1469
2. Lee, D.L., Seung, S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–791
3. Nayar, S.K., Nene S.A., Murase H.: Real-time 100 object recognition system. In Proc. IEEE Conference on Robotics and Automation **3** (1996) 2321–2325
4. Olshausen, B., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature **381** (1996) 607–609
5. Talukder, A., Casasent, D.: Classification and Pose Estimation of Objects using Nonlinear Features. In Proc. SPIE: Applications and Science of Computational Intelligence **3390** (1998) 12–23
6. Ullman, S., Bart, E.: Recognition invariance obtained by extended and invariant features. Neural Networks **17(1)** (2004) 833–848
7. Wersing, H., Körner, E.: Learning Optimized Features for Hierarchical Models of Invariant Object Recognition. Neural Computation **15(7)** (2003) 1559–1588