Proc. Int. Conf. Art. Neur. Netw. ICANN 2003, pp 385-392 (errors corrected)

Sparse Coding with Invariance Constraints

Heiko Wersing, Julian Eggert, and Edgar Körner

HONDA Research Institute Europe GmbH
Carl-Legien-Str.30, 63073 Offenbach/Main, Germany
{*heiko.wersing,julian.eggert,edgar.koerner*}*@honda-ri.de*

Abstract. We suggest a new approach to optimize the learning of sparse features under the constraints of explicit transformation symmetries imposed on the set of feature vectors. Given a set of basis feature vectors and invariance transformations, from each basis feature a family of transformed features is generated. We then optimize the basis features for optimal sparse reconstruction of the input pattern ensemble using the whole transformed feature family. If the predefined transformation invariance coincides with an invariance in the input data, we obtain a less redundant basis feature set, compared to sparse coding approaches without invariances. We demonstrate the application to a test scenario of overlapping bars and the learning of receptive fields in hierarchical visual cortex models.

1 Introduction

Redundancy reduction has been proposed as an important processing principle in hierarchical cortical networks [1]. Following this concept, wavelet-like features resembling the receptive fields of V1 cells have been derived either by imposing sparse overcomplete representations [9] or statistical independence as in independent component analysis [2]. Extensions for complex cells [7, 6] and spatiotemporal receptive fields were shown [5]. Lee & Seung [8] suggested the principle of nonnegative matrix factorizations to obtain sparse distributed representations. Simple-cell-like receptive fields and end-stopping cells were also found using a predictive coding scheme [10]. Learning algorithms used in these approaches normally get their input from local, isolated regions, and perform a local reconstruction using the gained representation from a single group of feature vectors. As a consequence of invariances in the input ensembles, the obtained feature sets are usually redundant with respect to translation, rotation or scale. In certain architectures like e.g. convolutional networks, it is, however, highly desirable to obtain only a single representative for each structurally different feature.

Here, we consider a setting with several families of feature vector sets, where each family is obtained from one feature of a basis set of feature vectors via invariance transformations. Reconstruction of an input vector is achieved by overlapping contributions of each of these transformation-dependent groups of feature vectors. We illustrate the approach with the example of 2-dimensional inputs and translation in the 2D plane. In this case the feature vectors within a family are translated versions of each other, so that we may sample a large input space by repeating the "same" feature vectors at every position, imposing a translational symmetry on the feature vector set. (This is similar

to weight-sharing architectures, however, the approach presented here can be used with any type of transformations and we are using the weight-shared representation for the unsupervised learning of the feature vectors instead of being a processing constraint.) This has a series of consequences. First, the input reconstruction is achieved by considering the contributions of feature vector groups anchored at different positions in an overlapping manner. Second, after learning we have gained a translation-independent common set of feature vectors, and third, every input image is reconstructed independently of its position, i.e., in a translationally invariant way. The result is a compact representation of encoding feature vectors that reflect transformation-invariant properties of the input. The work thus addresses two problem domains. On the one hand, it proposes a learning and encoding scheme for feature vectors in the case of a "patchy" reconstruction scheme that uses not only a single local input region but several regions that interact with each other. On the other hand, it takes advantage of specific transformation properties of the input (that may be known in advance, e.g. for a neural network that is supposed to detect input vectors at various degrees of translation, scaling, rotation, etc.) to select the best representation subject to the transformation constraints, which can be used afterwards for a transformation invariant postprocessing stage.

In Section 2 we introduce the standard sparse coding approaches and formulate our extension to transformation-invariant encodings. In Section 3 we derive an explicit learning algorithm for the case of nonnegative signals and basis vectors. We give two application examples in Sections 4 and 5 and discuss our results in Section 6.

2 Transformation-invariant sparse coding

Olshausen & Field [9] demonstrated that by imposing the properties of input reconstruction and sparse activation a low-level feature representation of images can be obtained that resembles the receptive field profiles of simple cells in the V1 area of the visual cortex. The feature set was determined from a collection of local image patches $\mathbf{I}^p$, where p runs over patches and $\mathbf{I}^p$ is a vectorial representation of the array of image pixels. A set of sparsely representing features can then be obtained from minimizing

$$E_1 = \frac{1}{2} \sum_p ||\mathbf{I}^p - \sum_i s_i^p \mathbf{w}_i||^2 + \sum_p \sum_i \Phi(s_i^p), \tag{1}$$

where $\mathbf{w}_i, i = 1, \ldots, B$ is a set of B basis representatives, s_i^p is the activation of feature $\mathbf{w}_i$ for reconstructing patch p, and Φ is a sparsity enforcing function. Feasible choices for $\Phi(x)$ are $\log(1 + x^2)$, and $|x|$ [9]. The joint minimization in $\mathbf{w}_i$ and s_i^p can be performed by gradient descent in the cost function (1).

Symmetries that are present in the sensory input are also represented implicitly in the obtained sparse feature sets from the abovementioned approach. Therefore, the derived features contain large subsets of features which are rotated, scaled or translated versions of a single basis feature. In order to avoid this redundancy, it may be desirable to represent the symmetries explicitly by using only a single representative for each family of transformations. For example in a translational weight-sharing architecture which pools over degrees of freedom in space only a single representative is needed

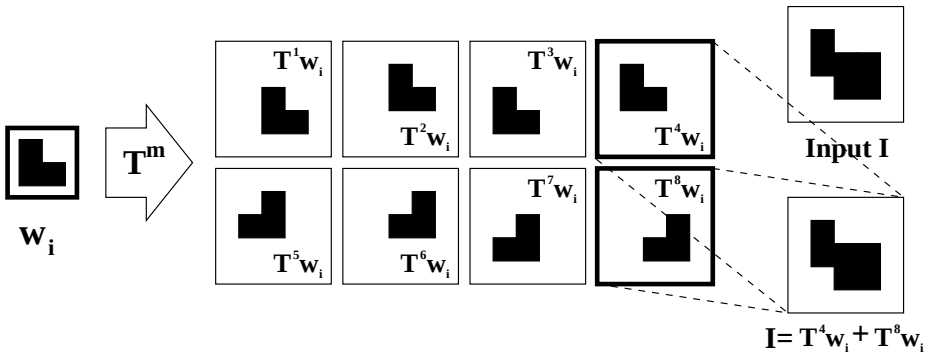

Fig. 1. Transformation-invariant sparse coding. From a single feature representative $\mathbf{w}_i$ a feature family is generated using a set of invariance transformations T^m. From this feature set, a complex input pattern can be sparsely represented using only few of the transformed features.

for all positions. From this representative, a complete set of features can be derived by applying a set of invariance transformations. To deal with shift invariance, nonlinear generative models with steerable shift parameters were proposed [9]. Using this concept as a direct coding model for natural images, Hashimoto & Kurata [4] obtained complex-patterned feature representations which were, however, difficult to interpret.

In the following we formulate the proposed invariant generative model in a linear framework. For a better understanding, we will consider the case of 2D images so that the model works using ensembles of local image patches. Let $\mathbf{I}^p \in R^{MN}$ be a large image patch of pixel dimensions $M \times N$. Let $\mathbf{w}_i \in R^{M'N'}$ (usually with $N' \leq N$ and $M' \leq M$) be a reference feature. We can now use a transformation matrix $\mathrm{T}^m \in R^{MN \times M'N'}$, which performs an invariance transform like e.g. shift or rotation and maps the representative $\mathbf{w}_i$ into the larger patch $\mathbf{I}^p$ as visualized in Figure 1 (departing from the notation in Eq. 1 now $\mathbf{w}_i$ and $\mathbf{I}^p$ have a different dimensionality). For example, by applying all possible shift transformations, we obtain a collection of features with differently placed receptive field centers, which are, however, characterized by a single representative $\mathbf{w}_i$. We can now reconstruct the larger local image patch from the whole set of transformed basis representatives by minimizing

$$E_2 = \frac{1}{2} \sum_p ||\mathbf{I}^p - \sum_i \sum_m s_{im}^p \mathrm{T}^m \mathbf{w}_i||^2 + \sum_p \sum_i \sum_m \Phi(s_{im}^p), \qquad (2)$$

where s_{im}^p is the activation of the representative $\mathbf{w}_i$ transformed by T^m, with $m = 1, \ldots, C$ indicating the C chosen transformations. The task of the combined minimization of (2) in s_{im}^p and $\mathbf{w}_i$ is to reconstruct the input from the constructed transforms under the constraint of sparse combined activation. For a given ensemble of patches the optimization can be carried out by gradient descent, where first a local solution in s_{im}^p with $\mathbf{w}_i$ fixed is obtained. Secondly, a gradient step is done in the $\mathbf{w}_i$'s, with s_{im}^p fixed and averaging over all patches p. For a detailed discussion of the algorithm see Section 3. Although the presented approach can be applied to any symmetry transformation, we restrict ourselves to spatial translation in the examples. The reduction in feature

complexity results in a tradeoff for the optimization effort in finding the feature basis. Whereas in the simpler case of equation (1) a local image patch was reconstructed from a set of B basis vectors, in our invariant decomposition setting, the patch must be reconstructed from $C \cdot B$ basis vectors (the B basis vectors, each considered at C displacement positions, respectively) which reconstruct the input from overlapping receptive fields. The second term in the quality function (2) implements a competition between the activations of the entire input field. This effectively suppresses the formation of redundant features $\mathbf{w}_i$ and $\mathbf{w}_j$ which could be mapped onto each other via one of the chosen transformations. Note also that, if the set of transformations T^m maps out the whole input space, it is necessary to have a positive sparsity contribution. Otherwise, the input can be trivially reconstructed from a simple "delta-peak" feature that can be used to represent the activity independently at each point of the input.

There exist different approaches for the general sparse coding model as expressed by the cost function (1). Due to the multiplicative coupling of the $\mathbf{w}_i$ and s_i^p either the norms of the $\mathbf{w}_i$ or the s_i^p must be held constant. If the $\mathbf{w}_i$ and s_i^p have no sign restriction, the minimization of (1) can be rephrased in the standard independent component analysis framework. If the basis vectors and activations are constrained to be nonnegative, one obtains the NMF framework for $\Phi(x) = 0$ for all x or the nonnegative sparse coding framework for $\Phi(x) > 0$ for $x > 0$. The general invariant sparse coding approach outlined in (2) is applicable to all these models. In the following examples we will, however, concentrate on nonnegative sparse coding.

3 Sparse Decomposition Algorithm

The invariant sparse decomposition is formulated as minimizing (2) where $\mathbf{I}^p, p = 1, \ldots, P$ is an ensemble of P image patches to be reconstructed, $\mathrm{T}^m, m = 1, \ldots, C$ is a set of invariance transformation matrices applied to the B feature representatives $\mathbf{w}_i, i = 1, \ldots, B$ which are the target of the optimization. We assume nonnegativity for the $\mathbf{w}_i$, i.e. $\mathbf{w}_i = 0$ componentwise. We choose the sparsity enforcing function as $\Phi(x) = \lambda x$, with $\lambda = 0.5$ as strength of the sparsity term [6]. We use $*$ to denote the inner product between two vectorially represented image patches. The algorithm consists of two steps [9]. First for fixed $\mathbf{w}_i$'s a local solution for the s_{im}^p for all patches is found by performing gradient descent. In the second step a gradient descent step with fixed stepsize is performed in the $\mathbf{w}_i$'s with the s_{im}^p fixed. The first gradient is given by

$$-\frac{\partial E_2}{\partial s_{im}^p} = b_{im}^p - \sum_{jm'} c_{jm'}^{im} s_{jm'}^p - \lambda, \tag{3}$$

where $b_{im}^p = (\mathrm{T}^m \mathbf{w}_i) * \mathbf{I}^p$ and $c_{jm'}^{im} = (\mathrm{T}^m \mathbf{w}_i) * (\mathrm{T}^{m'} \mathbf{w}_j)$. A local solution to $\frac{\partial E_2}{\partial s_{im}^p} = 0$ subject to of $s_{im}^p \geq 0$ can be found by the following asynchronous update algorithm:

1. Choose i, p, m randomly.
2. Update $s_{im}^p = \sigma\big(b_{im}^p - \sum_{(jm') \neq (im)} c_{jm'}^{im} s_{jm'}^p - \lambda\big) / c_{im}^{im}$. Goto 1 till convergence.

Let $\sigma(x) = \max(x, 0)$. This update converges to a local minimum of (3) according to a general convergence result on asynchronous updates by Feng [3] and exhibits fast convergence properties in related applications [11].

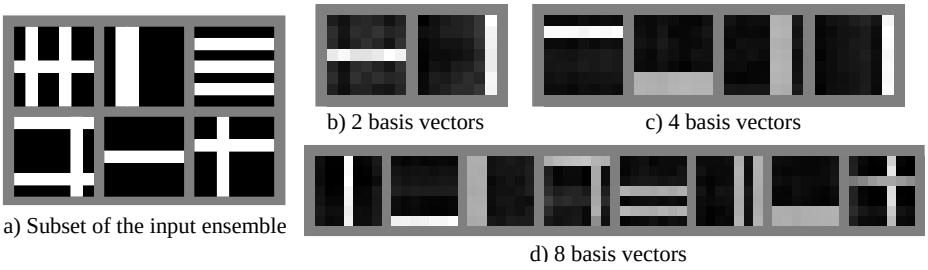

Fig. 2. Bar example. In a), 7 out of 1000 input images generated by overlaying 1-4 bars (randomly horizontal/vertical) are shown. The overlapping coding scheme was used to extract the basis vectors that best encode the input set under consideration of translational invariance. The result for 2 basis vectors are single bars at horizontal and vertical orientations as shown in b). In c) and d) 4 and 8 basis vectors were used, resulting in increasingly complex basis vectors. The basis vectors form an efficient sparse code for the ensemble using translations.

The second step is done performing a single synchronous Euler gradient step in the $\mathbf{w}_i$'s with a fixed stepsize η. For all $\mathbf{w}_i$ set

$$\mathbf{w}_i(t+1) = \sigma\Big(\mathbf{w}_i(t) + \eta \sum_{pm} \Big(s^p_{im}\mathbf{I}^p\mathrm{T}^m - \sum_{jm'} s^p_{im}s^p_{jm'}(\mathrm{T}^m * \mathbf{w}_j(t))\mathrm{T}^{m'}\Big)\Big), \quad (4)$$

where σ is applied componentwise. The stepsize was set to $\eta = 0.0001$. After each update step the $\mathbf{w}_i$ are normalized to unit norm.

4 Example: Translationally invariant bar decomposition

We applied the algorithm of Section 3 to an input image set that was gained by combining horizontal and vertical bars at different positions, thus containing translational symmetries in its components. The expected outcome would be a compact set of basis vectors (the representative set of underlying feature vectors from which the transformed feature vectors for the reconstruction are ultimately gained) that, together with the transformations, encodes the input. The entire input set can be described fairly well with only two basis vectors, and their superposition at different positions.

The input images and feature vectors were images of size 7×7 pixels. We used 1000 input images containing between 1 and 4 bars each, with pixel values 0 (background) and 1 (bar parts), such that the bars do not add linearly at intersections. A subset of the input images is shown in Fig. 2 (a). The transformation set T^m was composed of all possible translations of the basis vectors that influenced the 7×7 input/reconstruction image region. Contributions of the transformed basis vectors were considered only for pixels inside of the input image patch, resulting in a well-defined transformation set $\mathrm{T}^m : R^{M'\times N'} \to R^{M\times N}$ where pixels outside of the input image region do not contribute to measuring the reconstruction error in (2). Effectively, this means that $C = (7+6)^2$ transformations had to be taken into account for full translation invariance.

The simulations were run for 100 steps and with different numbers of basis vectors $B = \{2, 4, 8\}$. In Fig. 2 (b), the result for 2 basis vectors is shown. The outcome are

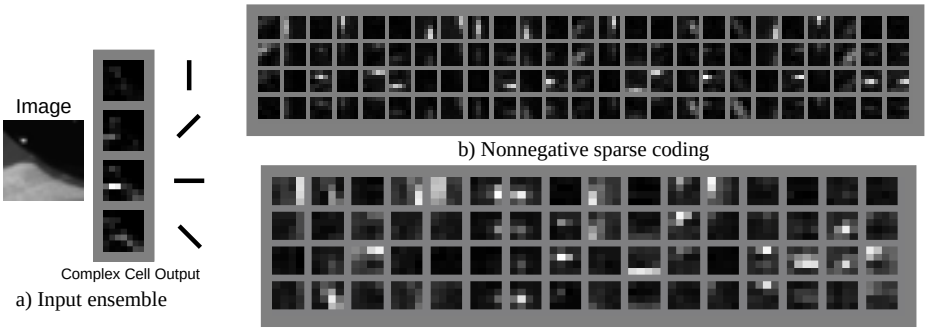

Fig. 3. Learning of receptive fields for complex-cell-integrating neurons. a) shows two examples from the natural image input patch ensemble with the obtained complex cell output, which is aligned vertically for the four considered orientations. In b) we show a subset of 25 features from a set of 144 features learned using the nonnegative sparse coding approach suggested by Hoyer & Hyvärinen. Each column characterizes the weights from the 4 complex cell output patches to the integrating cell. Features are collinear with different lengths and anchor positions. In c) we show the result of applying the invariant sparse coding approach with translation invariance. We optimized a set of 16 features with 4×4 receptive field dimensions. The feature set contains collinear features, but with more structural than positional variation as in b).

the two expected horizontal and vertical bar representatives. When we expand the basis vector set to 4 (see Fig. 2c), the same result is obtained for two of the basis vectors, while the remaining two converge to basis vectors that efficiently describe larger components of the input images, in our example double bars (here the encoding is efficient because the activation is sparser when using a single basis vector for describing an input image with double bars, than when using two horizontal or vertical basis vectors). For 8 basis vectors, yet other statistically frequent components appear, such as 2 aligned bars with a gap and two perpendicular bars forming crosses, as shown in Fig. 2 (d). Note that without the overlapping encoding scheme, the resulting feature vectors are first bars at all different horizontal and vertical positions (i.e., 14 in total), and then the more complex features such as double bars and crosses, again at all different positions.

5 Learning of Visual Cortex Receptive Fields

Recently Hoyer & Hyvärinen [6] applied a nonnegative sparse coding framework to the learning of combination cells driven by orientation selective complex cell outputs. To take into account the nonnegativity of the complex cell activations, the optimization was subject to the constraints of both coefficients s_i^p and vector entries of the basis representatives $\mathbf{w}_i$ being nonnegative. These nonnegativity constraints are similar to the method of nonnegative matrix factorization (NMF), as proposed by [8]. Differing from the NMF approach they also added a sparsity enforcing term like in (1). The optimization of (1) under combined nonnegativity and sparsity constraints gives rise to short and elongated collinear receptive fields, which implement combination cells being

sensitive to collinear structure in the visual input. As Hoyer and Hyvärinen noted, the approach does not produce curved or corner-like receptive fields.

We repeated the experiments with the non-overlapping setup as was done by Hoyer & Hyvärinen and compared this with results gained from our overlapping sparse reconstruction approach for translation invariance. The patch ensemble was generated in the following way (see also [6]). First a set of 24×24 pixel patches was collected from images containing natural scenes. On each patch the response of a simple cell pair which consisted of an even and an odd Gabor filter was computed on a 6×6 grid (see Figure 3). The complex cell output was obtained by a simple sum of squares of a quadrature filter pair. At each grid position, four Gabor orientations were considered. Therefore a total patch activation vector $\mathbf{I}_p$ consisted of $6 \times 6 \times 4 = 144$ components. We generated a set of 10000 patches and applied the nonnegative sparse coding learning rule as described in [6] for a set of 144 feature vectors using the gradient-based relaxation as described in Section 3, but without overlaps. As shown in Figure 3 we could essentially reproduce the results in [6], consisting of collinear features of differing lengths and positions.

We then investigated the invariant sparse coding scheme by considering all possible translations of the features on the complex cell output patches for a basis of 16 features. Due to computational performance constraints we reduced the patch size to 4×4 pixels (resulting in $C = (4 + 3)^2 = 49$) and reduced the number of patches to 2000. In this setting the optimizations takes about 1-2 days on a standard 1 GHz CPU. The result shows that we obtained a less redundant feature set with respect to translations. The set contains collinear features of differing lengths (or width), with a greater emphasis on the vertical and horizontal directions for long edges. There are some rather local features which combine two local neighboring orientations, and which may be used to capture a local residual that is not covered by the other features.

6 Summary and Discussion

We have demonstrated how to exploit explicitly formulated transformation symmetries for the overlapping reconstruction of input vectors and the learning of an optimal encoding scheme subject to given transformation constraints. Although we have shown these capabilities using an example with translational symmetries and transforms only, the presented algorithm can be used for any transformation set T^m. One could think of rotational, scaling and other transforms in addition to the translational transform. This would result (after learning) in a network that is able to reconstruct an input stimulus equally well for different translations, rotations and scaling operations. Such a transformation invariant preprocessing could well be a necessary step to achieve transformation invariant classification/detection in a hierarchical system. Posterior stages could take advantage of the known transformation properties to achieve a transformation invariant response, like e.g. pooling over transformed variants of the same feature. Future work on the subject may therefore include the extension of the shown principle to incorporate additional transforms and its application in a larger, hierarchical network.

We have shown in the two examples that the invariant sparse coding approach allows to describe an input ensemble using fewer features than a direct sparse encoding. This reduction of free parameters is achieved by using a more powerful representa-

tional architecture, which in turn is paid by a greater effort of estimating the model. For the receptive-field learning example the representational effort scales linearly with the number of pixels in the input patches. The same scaling, however, also holds for the direct sparse encoding, since the number of representing features must be large enough to carry all possible translated versions of a particular local feature.

Could the implemented encoding scheme be part of a biological neural system, e.g. for feature learning and input representation in the early visual pathway? On a first glance, the spread of receptive field profiles suggests a sampling of the visual input with inhomogeneous "feature" vectors, conflicting with the idea of general basis vectors from which the individual feature vectors are drawn. Nevertheless, on a semi-local basis, the brain has to deal with transformation symmetries that could be exploited in the proposed way. There is no reason to specifically learn all feature vectors at every position anew, if they turn out to be translated/transformed versions of each other. The biological feasibility of the basis vector learning rule (4) is, therefore, a matter of debate. On the other hand, one could certainly devise neural-like mechanisms for the activity adjustment rule (3), since the equation for the s_{im}^{p} can be seen as a shortcut of a relaxation dynamics of graded-response type with a positive rectifying nonlinearity. The constraint of the activations to positive values even adds to the biological plausibility of this rule, since a biologically plausible rate coding implies positive activation variables.

Acknowledgments: This work was partially supported by the Bundesministerium für Bildung und Forschung under LOKI project grant 01IB001E.

References

1. H. B. Barlow. The twelfth Bartlett memorial lecture: The role of single neurons in the psychology of perception. *Quart. J. Exp. Psychol.*, 37:121–145, 1985.
2. A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.
3. J. Feng. Lyapunov functions for neural nets with nondifferentiable input-output characteristics. *Neural Computation*, 9(1):43–49, 1997.
4. W. Hashimoto and K. Kurata. Properties of basis functions generated by shift invariant sparse representations of natural images. *Biological Cybernetics*, 83:111–118, 2000.
5. J. H. Van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. London B*, 265:2315–2320, 1998.
6. P. O. Hoyer and A. Hyvärinen. A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, 42(12):1593–1605, 2002.
7. A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neur. Comp.*, 12(7):1705–1720, 2000.
8. D. L. Lee and S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.
9. B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1 ? *Vision Research*, 37:3311–3325, 1997.
10. R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosc.*, 2(1):79–87, 1999.
11. H. Wersing, J. J. Steil, and H. Ritter. A competitive layer model for feature binding and sensory segmentation. *Neural Computation*, 13(2):357–387, 2001.