On spatio-temporal feature learning

Roland Memisevic
Department of Computer Science, University of Frankfurt
ro@cs.uni-frankfurt.de

Unsupervised feature learning and sparse coding have recently gained attention in the vision community because they can learn local image descriptors that work well for object recognition. To utilize learned features for recognition, one crops small patches from images and projects these onto learned basis functions. Pooling over multiple image locations can be used to get a descriptor for a whole image which can then be fed to some classifier.

Although feature learning works well for recognition, in many vision tasks it is not the content of any single image but the relationship between images that carries the most relevant information. Examples include tracking, stereopsis, optical flow, motion and action understanding, or strongly invariant recognition. To address this issue, recently a variety of sparse coding models for learning relations have been suggested (for example, [6, 5, 3]). Common to all these methods is that they deploy either *multiplicative interactions* or *squaring non-linearities*, thus, they can be viewed as instances of *complex cell* models.

In this work we analyze the role of multiplication and of squaring non-linearities in learning relations. We restrict our attention to linear transformations L in pixel space, known as a linear *warps*. To simplify the analysis further, it is convenient to consider *orthogonal warps* [1], for which $LL^{\mathrm{T}} = I$ where I is the identity matrix. These still subsume most common image transformations, including rotation, translation or affine transformations, as well as more complex transformations like arbitrary permutations of pixels. It is well-known that an orthogonal matrix L acts on a vector $\boldsymbol{x}$ by performing a set rotations in two-dimensional subspaces, known as the *invariant subspaces* [2] which are spanned by the eigenvectors of L.

Consider the following inference task: Given a large set of orthogonal transformations $L_1, \ldots, L_N$ and an image pair $(\boldsymbol{x}, \boldsymbol{y})$ related through one of the transformations, infer the transformation. Solving this task is easy when the transformations *commute*, so that they share the same eigenvectors: In this case, inferring the transformation is equivalent to *determining the rotation angles* between the projections of $\boldsymbol{x}$ and $\boldsymbol{y}$ onto the invariant subspaces.

The central observation of our analysis is that the activity of a hidden unit, z_k, in a multiplicative sparse coding model (for example, [5]) is well-suited to encoding these rotation angles: It is given by a weighted sum over products of linear projections of two images $\boldsymbol{x}$ and $\boldsymbol{y}$ (cf., Figure 1):

$$z_k = \sum_f u_{kf} (\boldsymbol{v}_f^{\mathrm{T}} \boldsymbol{x})(\boldsymbol{w}_f^{\mathrm{T}} \boldsymbol{y}) \tag{1}$$

where the u_{kf} are typically constrained, such that a hidden variable computes the sum over a small number of products. Each hidden variable activity is given by a (weighted) inner product of linear projections of $\boldsymbol{x}$ and $\boldsymbol{y}$. Since the inner product is equal to the cosine of the angle between the projections of $\boldsymbol{x}$ and $\boldsymbol{y}$, we can think of z_k as a rotation detector. When images are contrast-normalized, z_k will attain the maximal response for a particular image pair $\boldsymbol{x}, \boldsymbol{y}$, that is compatible with a set of "preferred angles" of that hidden unit. The subspaces and preferred angles of each hidden unit depend on the choice of u_{kf} and of features $\boldsymbol{v}_f, \boldsymbol{w}_f$, which are learned from data [5, 3]. As learning amounts to identifying the invariant subspaces, we can think of learning as approximately performing a joint diagonalization of a set of image warps [4].

An analogous argument holds for the activity of a hidden unit in a "square-pooling" model applied to the concatenation of two images $\boldsymbol{x}$ and $\boldsymbol{y}$ (for example, [3]), which is given by

$$z_k = \sum_f u_{kf} (\boldsymbol{v}_f^{\mathrm{T}} \boldsymbol{x} + \boldsymbol{w}_f^{\mathrm{T}} \boldsymbol{y})^2 = 2 \sum_f u_{kf} (\boldsymbol{v}_f^{\mathrm{T}} \boldsymbol{x})(\boldsymbol{w}_f^{\mathrm{T}} \boldsymbol{y}) + \sum_f u_{kf} (\boldsymbol{v}_f^{\mathrm{T}} \boldsymbol{x})^2 + \sum_f u_{kf} (\boldsymbol{w}_f^{\mathrm{T}} \boldsymbol{y})^2 \tag{2}$$

The activity is the same as in a multiplicative sparse coding model (Eq. 1) up to the square terms. The square terms can be shown to make the rotation detector more conservative but otherwise do not affect the hidden unit's ability to detect subspace rotations [4].

Our analysis helps explain why Fourier features and circular Fourier features emerge when training complex cell models on translating or rotating images (for example, [5]), and it has a variety of further implications:

- Square-pooling supports selectivity rather than invariance.
- We can train a square-pooling model using a gated sparse coding model and vice versa (Figure 2).

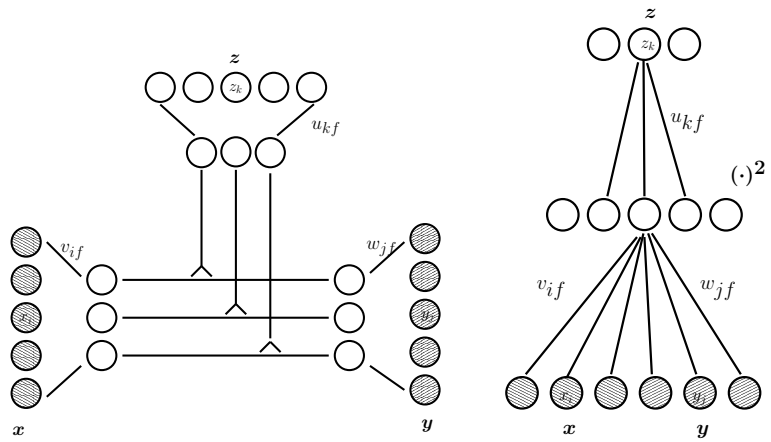

Figure 1: Two ways to model the relationship between two images $\boldsymbol{x}, \boldsymbol{y}$: Using (a) a multiplicative feature learning model, (b) a square-pooling model applied to the concatenation of two images.

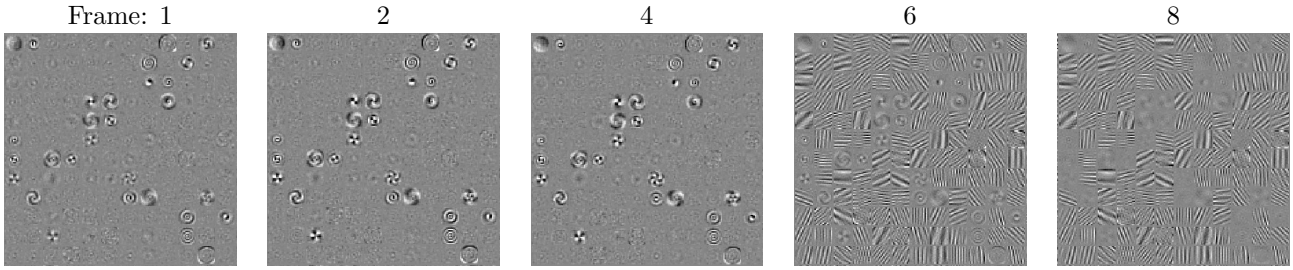

Figure 2: Six frames from the "eigenmovies" of transforming random dots, which rotate for 5 frames, then translate for 5 frames. The figure shows that each filter specializes, such that it encodes either only the first half of a movie or only the second half of a movie.

- Although not common in square-pooling models (for example, [3]), a hidden unit should pool over more than a single subspace to encode motion independently of the content of images.
- The encoding of transformations is itself transformation-invariant, so to perform invariant recognition one should look at video snippets of objects rather than at still images.
- Multiplicative interactions can help extend the applicability of deep learning models towards more interesting tasks than recognizing objects in still images.

References

[1] M Bethge, S Gerwinn, and JH Macke. Unsupervised learning of a steerable basis for invariant image representations. In *Human Vision and Electronic Imaging XII*, Bellingham, WA, USA, 2007. SPIE.

[2] Robert M. Gray. Toeplitz and circulant matrices: a review. *Commun. Inf. Theory*, 2:155–239, August 2005.

[3] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng. Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis. In *Proc. CVPR, 2011*. IEEE, 2011.

[4] Roland Memisevic. Learning to relate images: Mapping units, complex cells and simultaneous eigenspaces. *ArXiv e-prints*, 2011.

[5] Roland Memisevic and Geoffrey E Hinton. Learning to represent spatial transformations with factored higher-order Boltzmann machines. *Neural Computation*, 22(6):1473–92, 2010.

[6] Bruno Olshausen, Charles Cadieu, Jack Culpepper, and David Warland. Bilinear models of natural images. In *SPIE Proceedings: Human Vision Electronic Imaging XII*, San Jose, 2007.

Topic: visual processing and pattern recognition
Preference: oral