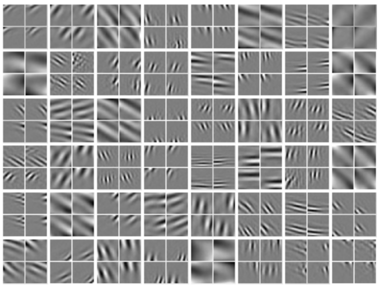
Plan for NIPS 2014

A metric learning-like framework is proposed for unsupervised training of neural networks. As demonstrated in algorithms such as DrLIM, ISOMAP and others, learning a metric can implicitly lead to a representation in which the underlying latent variables are disentangled or linearized. However the approach relies on an oracle to provide semantically meaningful similarity relationships. Temporal coherence can be exploited to provide similarity relationships needed to learn a semantically meaningful metric: temporally neighboring frames in a video are likely to be semantic neighbors. The first objective we considered was to extract slowly varying temporal features which closely resembles slow feature analysis (SFA). Other objectives are possible, such as minimizing the curvature of the trajectories traced out by the features in time (i.e. features that vary linearly).

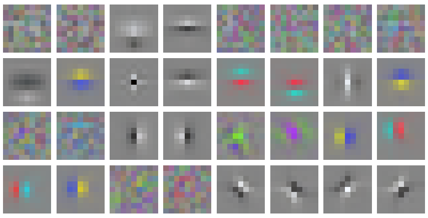
Summary of Results

- We have proposed a method by which to train the standard architecture with unlabeled temporally coherent data. The standard model: (1)linear transform →(2)point-wise nonlinearity→(3)pooling
- Out shunted decoder model proposes to decode after the point-wise nonlinearity, but before the pooling, while maximizing the sparsity of the activations before the pooling and maximizing the slowness of the features after the pooling. *The interpretation of this scheme is that sparsity induces independent features, and slowness causes these features to be pooled together in a way that produces invariant features w.r.t. the transformations contained in the sequence*
- Experiments were performed with fully connected networks on natural image patches (20x20) from high resolution video. Since most of the variation in natural video on small spatial and temporal scales is due to translation, these experiments have an intuitive interpretation



Fully connected → ReLU $rightarrow$ L_2-Pooling in non-overlapping groups of four

- Experiments with a standard convolutional network stage were performed. The architecture used was (1)convolution → (2)ReLU → (3)Volumetric Max-Pooling



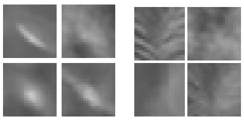
Here is a 32-kernel filter bank consisting of 9x9 kernels. Every two horizontal neighbors constitute a max-pooling group (no overlap). Often semantically similar templates (cyan/red gradients for example) are pooled together.

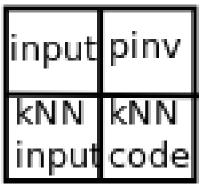
Side Issues

- Dead filters: Initialization with k-means does not help in the convolutional setting. I am not trying to manually detect "dead" filters and revive them manually reinitilizing with k-means.

Experiments to Perform

- Decoding from the pooling: In order to visualize the features in the input space a non-linear decoder can be trained. This is not just a way to visualize the feature space but also a validation that the features are information preserving. If this works, then joint training of encoder and non-linear decoder becomes an option, this would be the first demonstration of a principled non-linear decoder that I know of. We have already seen some success with K-means decoders and this would tie in nicely with Arthur & Joan's recent work on phase recovery.





input	pinv
kNN input	kNN code

- Stacking and Training on Toy Data: I have not yet tried to train a second layer.

- Usefulness of the features: One approach is to show that features are useful in a standard classification task. This is problematic because few temporally coherent datasets contain labels (however NORB and COIL are exceptions). Furthermore very similar experiments have already been performed by in the ICML 2009 paper by Ronan Collobert and Jason Weston. In that work they actually put their own dataset together by photographing common objects in a studio environment and called it a COIL100-Like dataset.