Learning Representations from Temporal Data

Ross Goroshin

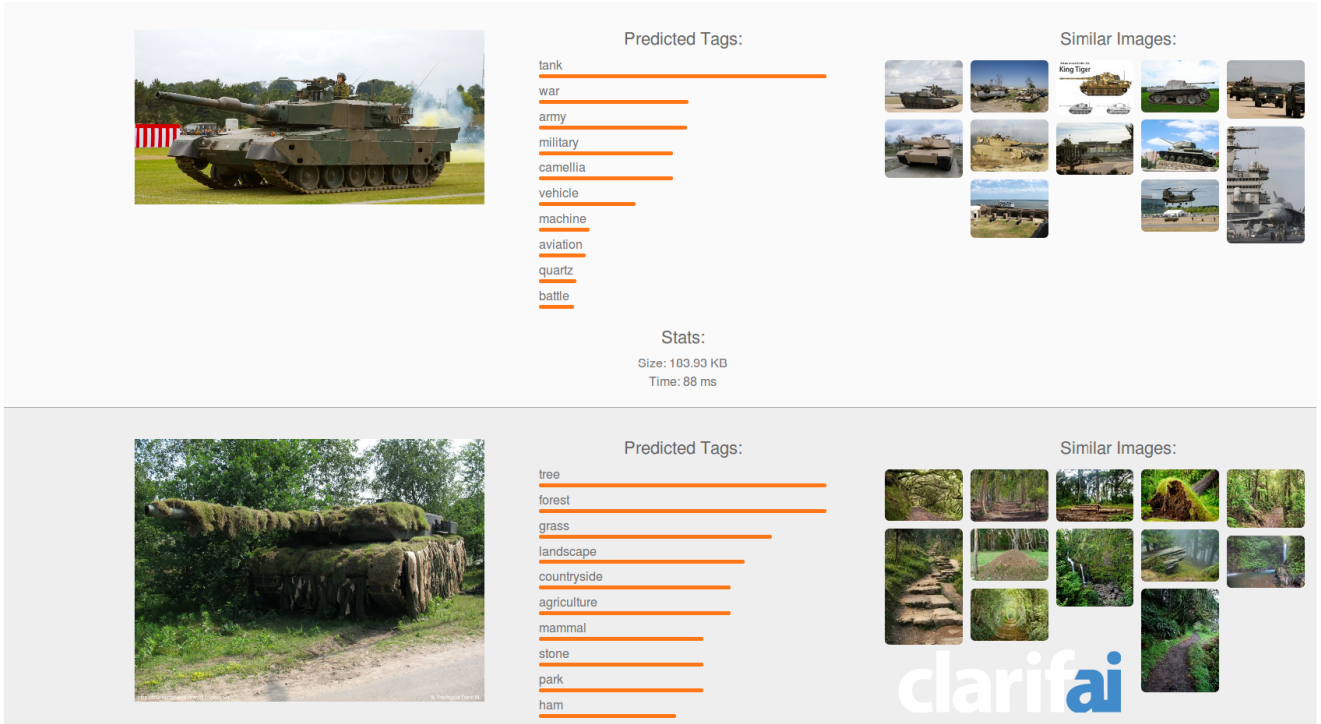
October 29, 2014

NEW YORK UNIVERSITY

Objectives of Research

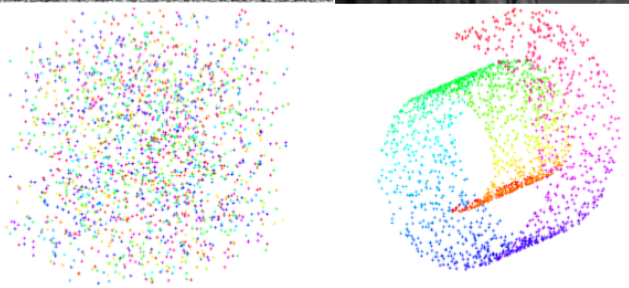
- Characterize *generically* useful properties of features
- Design loss functionals which explicitly or implicitly promote desirable feature characteristics
- Self contained evaluation criteria, requiring no external oracle

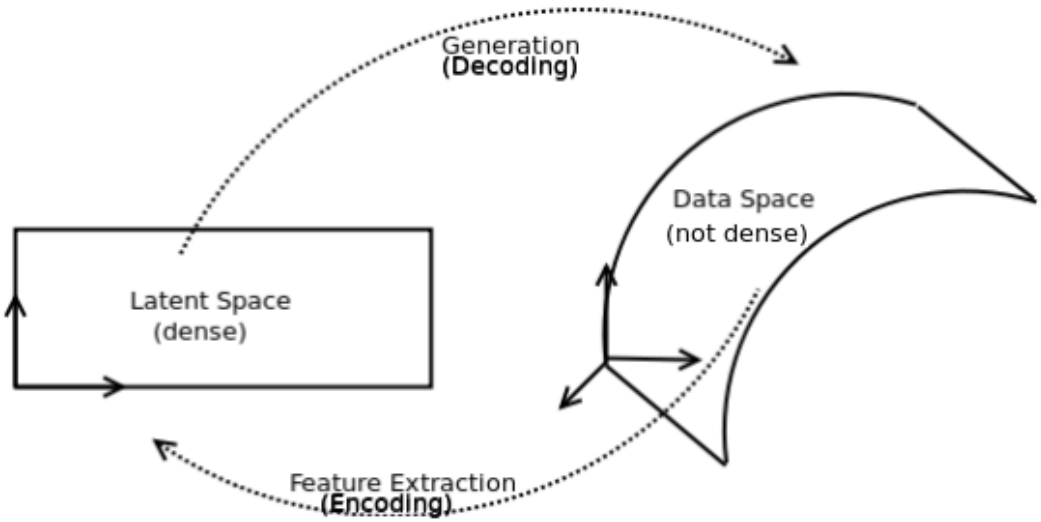
Supervised Deep Feature Learning



Works in practice, scales in depth, stable, immediate results

Data in High Dimensions





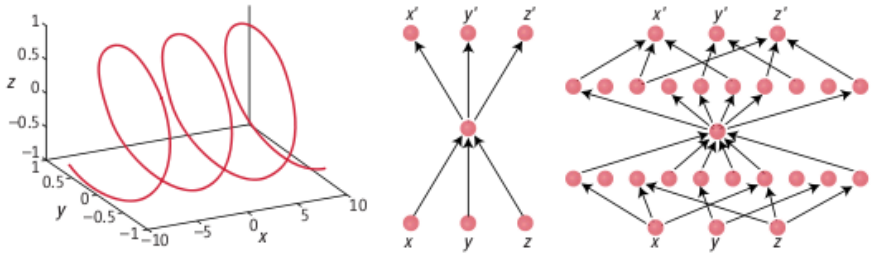
Dependencies among variables imply a low dimensional structure

Models of Data

Algorithm	Model	Encode	Decode	Learning
PCA	Linear	✓	✓	$W_D = W_E^T$
ICA	Linear	✓	✓	$W_D = W_E^T$
Sparse Coding	Locally Linear	✓($)	✓	inference, W_D
CBP	Osculating Sphere	✓($)	✓	inference, W_D
PSD & LISTA	Locally Linear	✓	✓	hybrid, W_D
Metric Learning	Nonlinear	✓	X	W_E only
Auto-Encoders	Nonlinear	✓	✓	W_E & W_D

	De-correlation/Independence
	Sparsity
	Metric Learning/Restricted Metric Learning
	All of the Above

Preserving Information



Searching for structure. **(Left)** Three-dimensional data that are inherently one-dimensional. **(Middle)** A simple "autoencoder" network that is designed to compress three dimensions to one, through the narrow hidden layer of one unit. The inputs are labeled x, y, z, with outputs x', y', and z'. **(Right)** A more complex autoencoder network that can represent highly nonlinear mappings from three dimensions to one, and from one dimension back out to three dimensions.

Optimizing desirable properties ('contrastive terms') of the representation alone often leads to degenerate solutions.

- Directly reconstruct the input from it's latent representation (PCA, Auto-encoders, Sparse Coding)
- Assign high probability to your data, low probability everywhere else (Max-Likelihood Models)
- Enforce the features to have unit variance (kurtosis-based ICA, slow-feature analysis)
- Push the representation of samples apart (DrLIM, nuclear norm)

Auto-Encoders/Dictionary Learning

- Auto-Encoders replace the 'inference' step of dictionary learning with a learned feed forward function
- In addition to reconstruction additional terms are included to induce desirable properties in the representation
- Example: **Sparse Coding / Sparse Auto-Encoder**

$$\min_{z,W} \frac{1}{2}\|x - Wz\|_2^2 + \lambda\|z\|_1$$

Dictionary Learning is to alternate between:

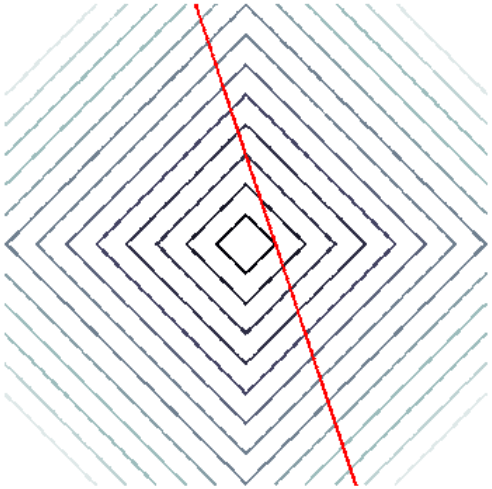
$$z_{k+1} = shrink(z_k - \eta_1 \nabla_{z_k} \frac{1}{2}\|x - W_k z_k\|_2^2) \quad (1)$$

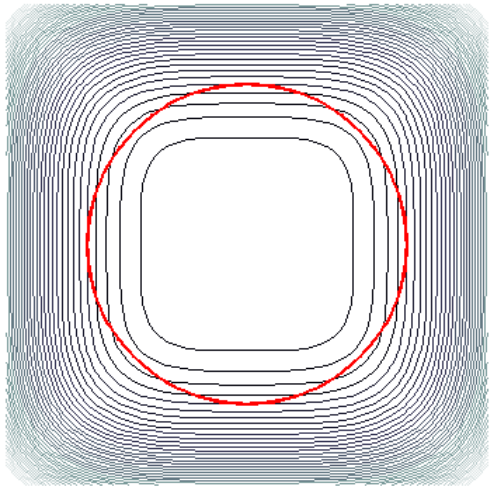
$$W_{k+1} = W_k - \eta_2 \nabla_{W_k} \frac{1}{2}\|x - W_k z_k\|_2^2 \quad (2)$$

Auto-Encoder formulation usually uses SGD to solve:

$$\min_{W_e,W_d} \frac{1}{2}\|x - W_d F_{W_e}(x)\|_2^2 + \lambda\|F_{W_e}\|_1$$

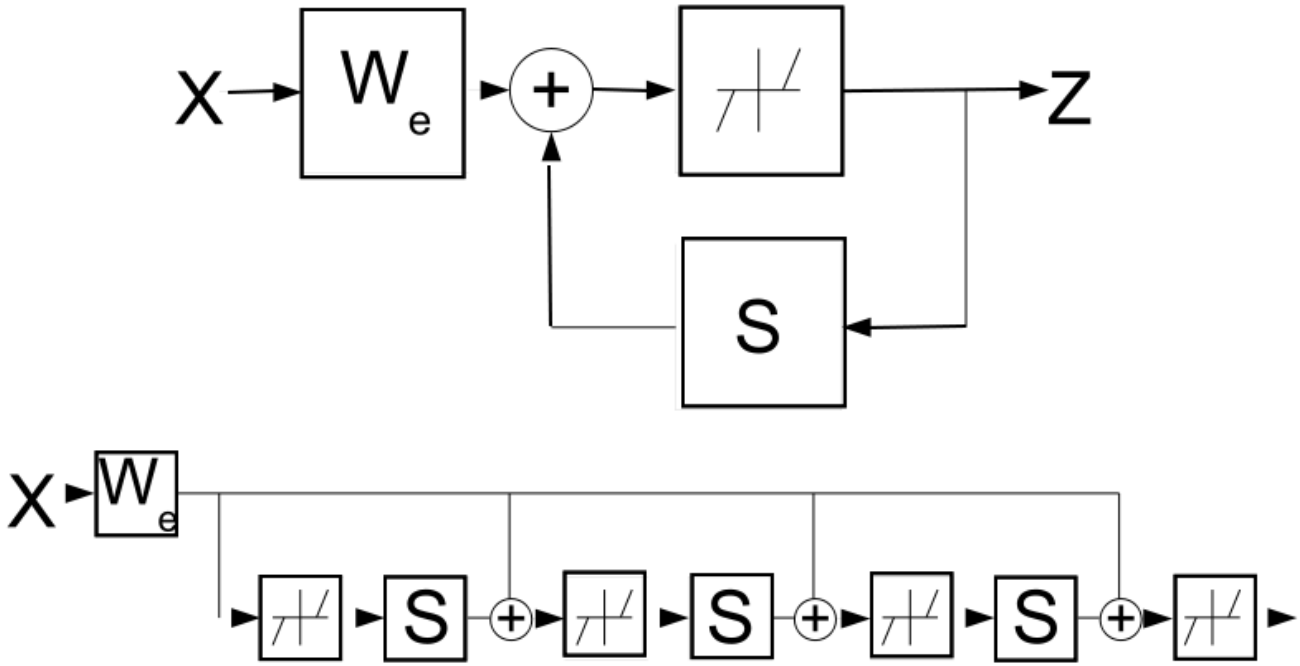
Sparsity and Independence





Left: Minimize L_1, goal: sparsity
Right: Maximize kurotsis, goal: independence
(If individual code elements have low enetropy (sparsity), code variables should be as independent as possible in order to maximize information capacity)

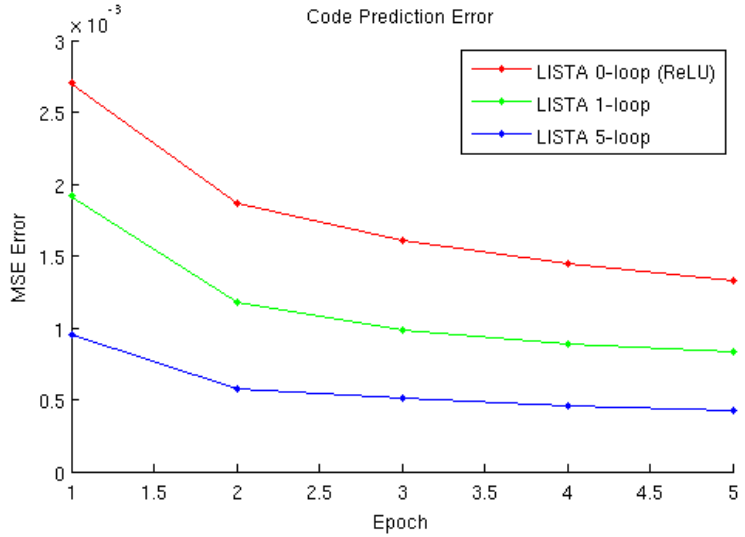
Parameterizing F_{W_e}: LISTA Inference

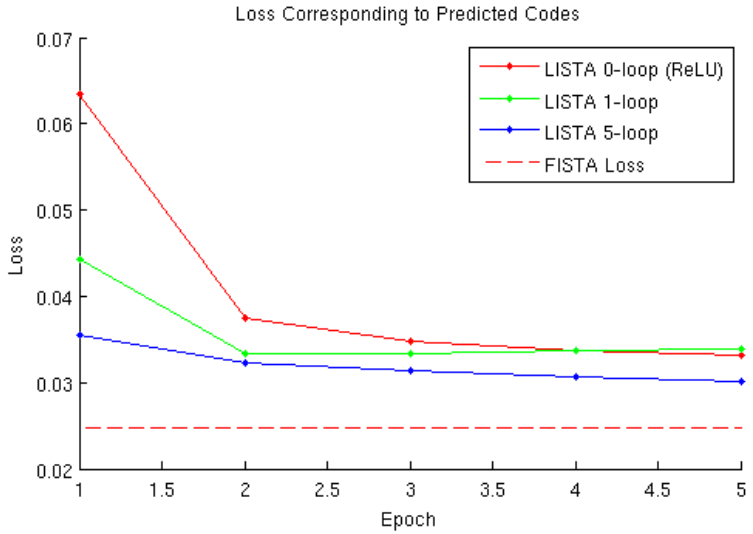


$$S = I - \frac{1}{L} W_d^T W_d \text{ and } W_e = \frac{1}{L} W_d^T$$

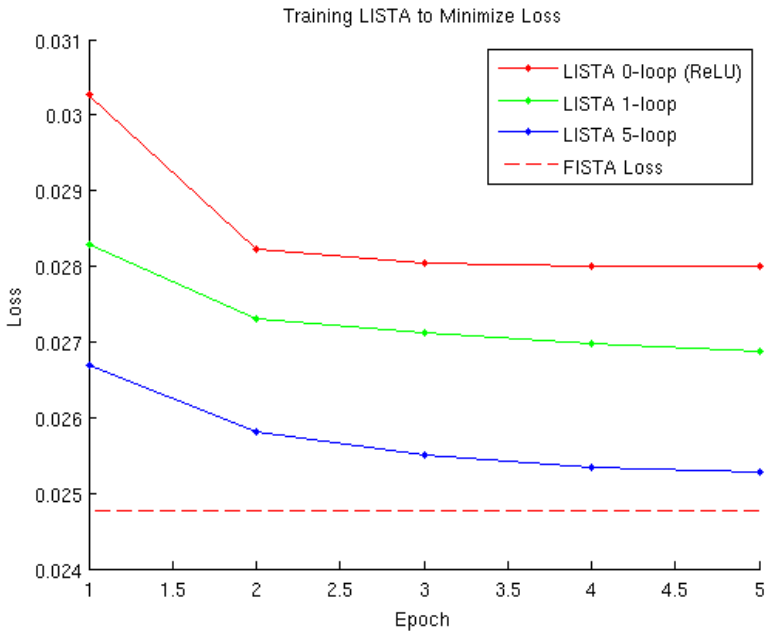
- Learned ISTA is a method for approximating the inference step by a feed forward network
- Originally designed/tested to perform inference in a *fixed* dictionary. Does it also work when jointly learning the dictionary, i.e. as an encoder?

LISTA Inference





LISTA Inference



- LISTA works best when minimizing the loss (not code prediction error) directly with a *fixed* dictionary
- It it is not beneficial when learning the dictionary

Contractive/Saturating Regularization

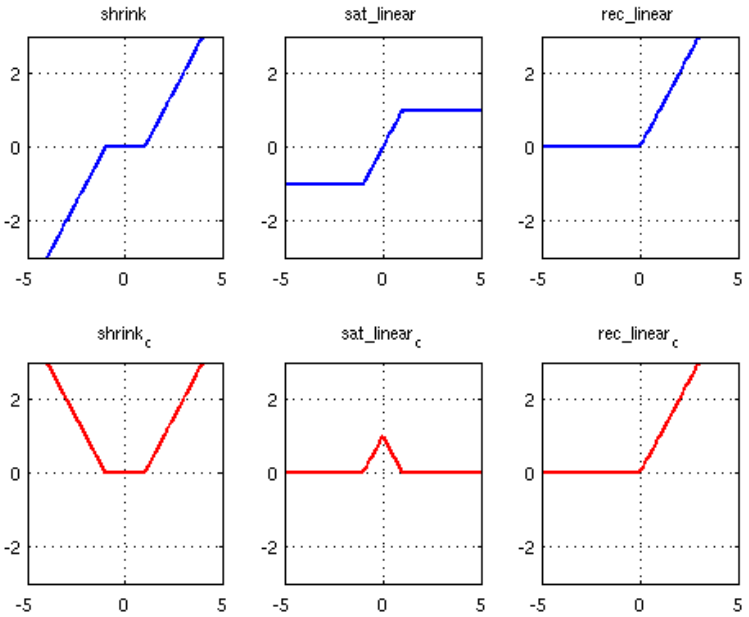
Contractive regularization

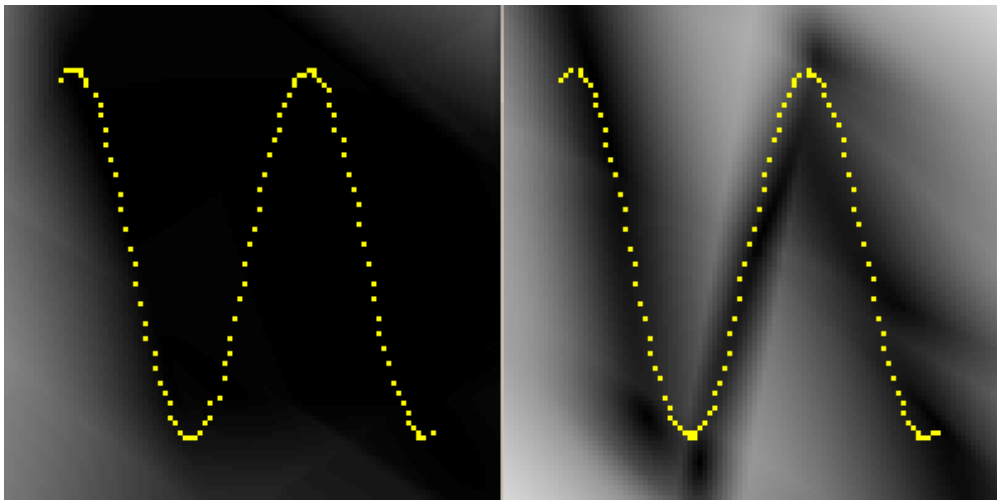
- Penalize the norm of the Jacobian of the representation, w.r.t. the inputs at each sample
- This promotes stability (invariance) at the samples in all directions except in the tangent space of the manifold
- Let $f(\cdot)$ be a point-wise nonlinear activation function
- $L_{CAE} = \sum_{x \in D_n} \|x - W_d f(W_e x)\|_2^2 + \lambda \sum_{ij} \left(\frac{\partial f_j(W_e x)}{\partial x_i} \right)^2$

Saturating regularization

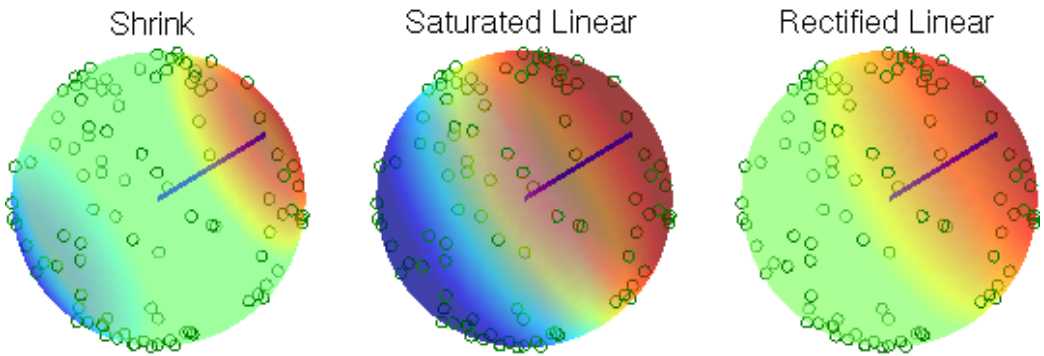
- Assuming $f(\cdot)$ has zero gradient regions, introduce a complimentary function $f_c(\cdot)$ to $f(\cdot)$
- $f_c(\cdot)$ is the distance transform to the nearest gradient region
- The motivation is to obtain an implicit parameterization of the data manifold. The data manifold is implicitly embedded in the reconstruction error function
- $L_{SAT} = \sum_{x \in D_n} \|x - W_d f(W_e x)\|_2^2 + \lambda f_c(W_e x)$

Saturating Auto-Encoders



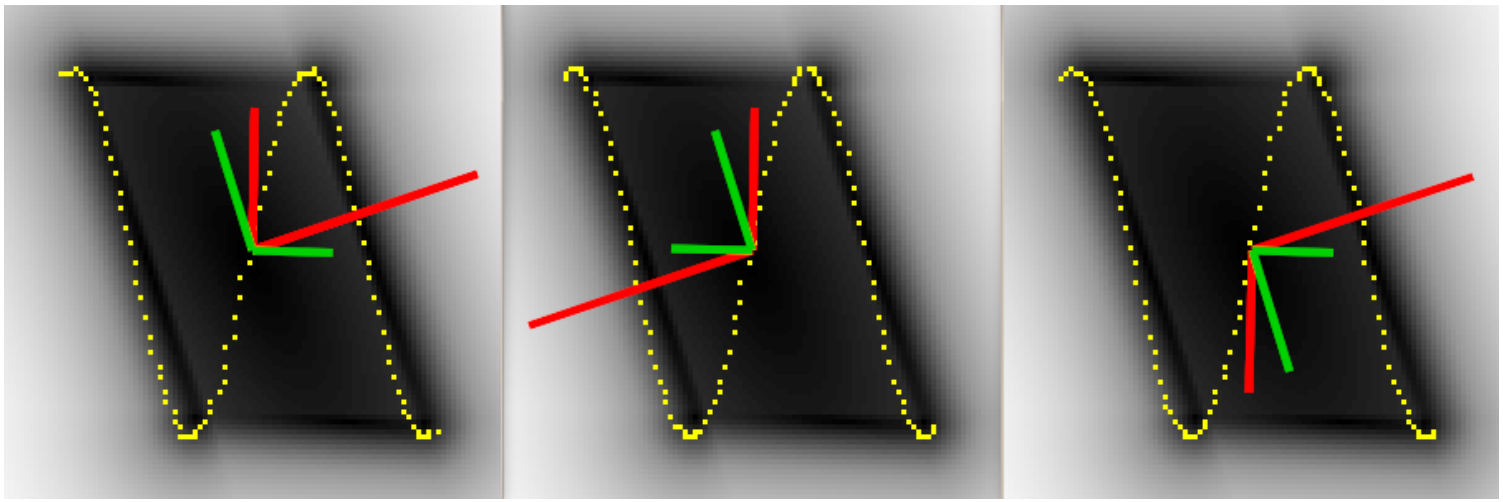


Saturating Auto-Encoders

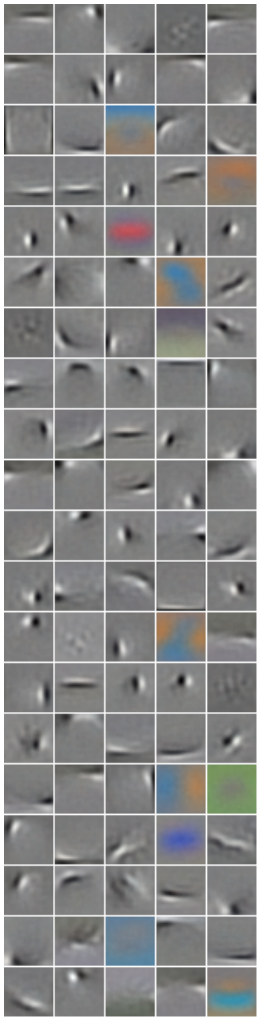


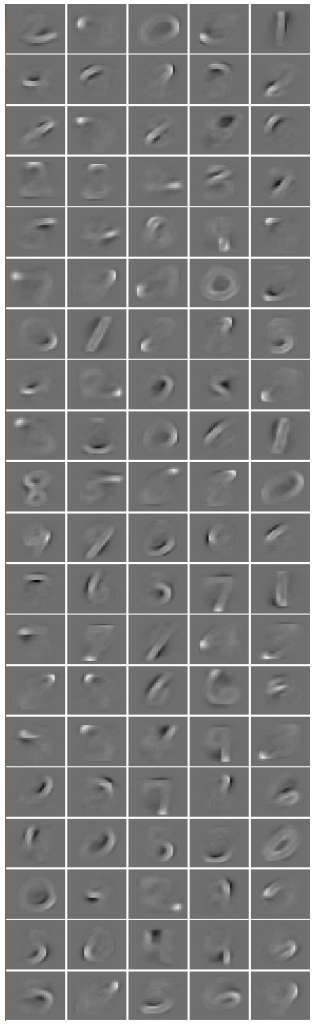
- Shrink and ReLU saturating auto-encoders are equivalent to L_1 regularized auto-encoders with corresponding nonlinearities
- Saturating auto-encoders with sigmoid-like activation functions regularize the representation by promoting binary codes

Saturated Linear Activations

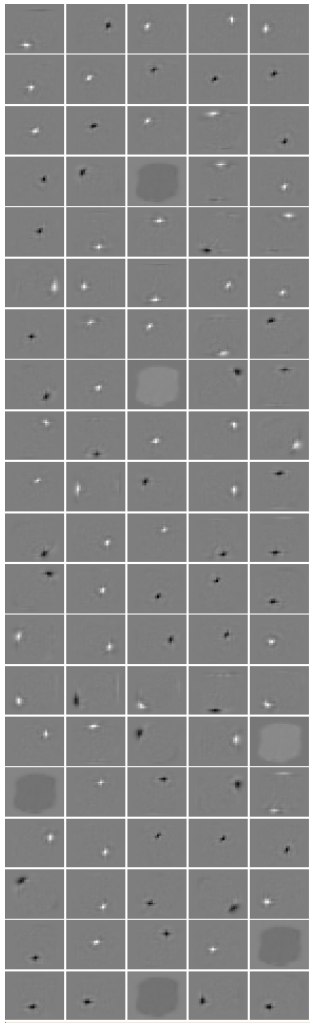


Shrink/ReLU Activations





Saturated Linear Activations



Lessons

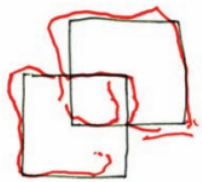
- Contractive regularization is effective only within a small ball around each sample (same with denoising auto-encoder)
- Saturated-linear auto-encoder does not work for certain datasets (e.g. MNIST)
- Parameterization is important. Which portions of the input space are assigned a higher energy? Do we have any reasonable hope of seeing any samples from the test set there?

The Role of Time

FIRST SIGHT

Only the Parts, Not a Whole

A newly sighted child's tracings reveal a piecemeal view of even two-dimensional figures. Each area of overlapping boxes was perceived separately, as indicated by red squiggles. The boy also saw segments of a cow and of a ball and its shadow—all delineated in green—to be distinct objects. Consequently, he was unable to identify any of these images.



we found a recapitulation of his experience. Many months after encountering difficulties on the image-parsing task, they begin to succeed in organizing their perceptions into coherent objects. The time needed to acquire this skill appears to depend on the age at which the child received treatment, with younger children learning more rapidly than older ones.

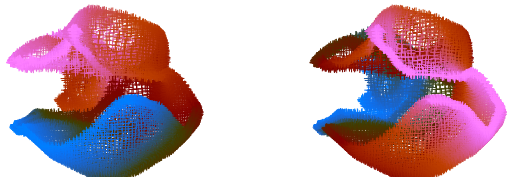
en

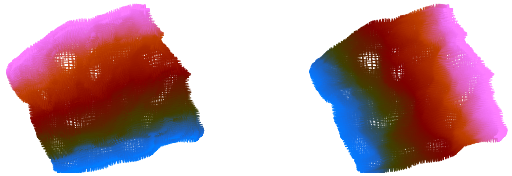
What underlies this improvement? Theory suggests that motion may play the part of a "teacher," training the visual system to parse images even when they are static. With the rule "things that move together belong together," a person's visual system can eventually learn to group images via static attributes such as color and orientation.

The brain, of course, does more than pick out the elements of a visual scene. It also connects to the realms of sound, touch, smell and taste—creating a sensory panorama via intermodal

Time is a one dimensional parametrization of the data-manifold, neighbors in time should be neighbors in feature space.

Learning a Temporally Coherent Metric





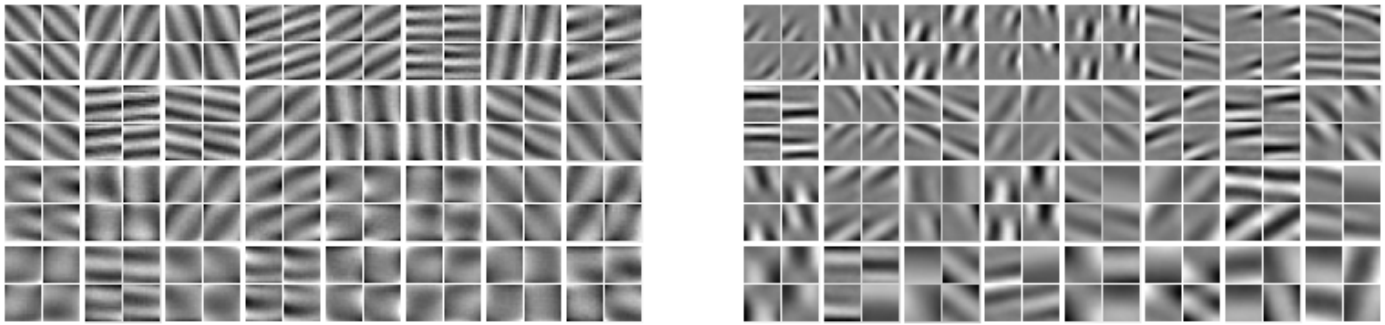
$$L(x_t, x_{t'}, W) = \begin{cases} \|G_W(x_t) - G_W(x_{t'})\|_p, & \text{if } |t - t'| = 1 \\ \max(0, m - \|G_W(x_t) - G_W(x_{t'})\|_p) & \text{if } |t - t'| > 1 \end{cases}$$

Learning Sparse and Slow Features

$$L(x_t, x_{t'}, W) = \sum_{\tau=\{t,t'\}} (\|W_d h_\tau - x_\tau\| + \alpha|h_t|) + \beta \sum_{i=1}^{K} \left| \|h_t\|^{P_i} - \|h_{t'}\|^{P_i} \right|$$

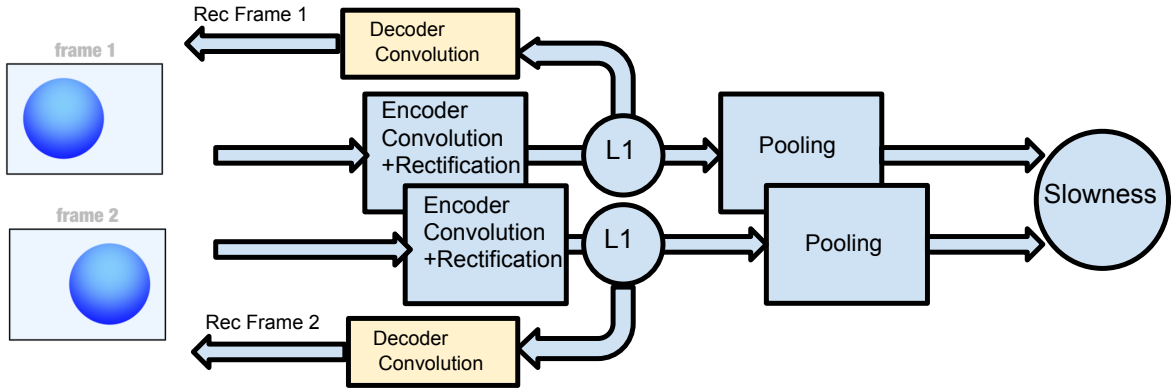
Where the vector of hidden activations for the sample at time t is denoted by $h_t = f(W_e x_t)$, and $\|h_t\|^{P_i}$ denotes a local L_2 pooling over the neighborhood P_i

Learning Sparse and Slow Features



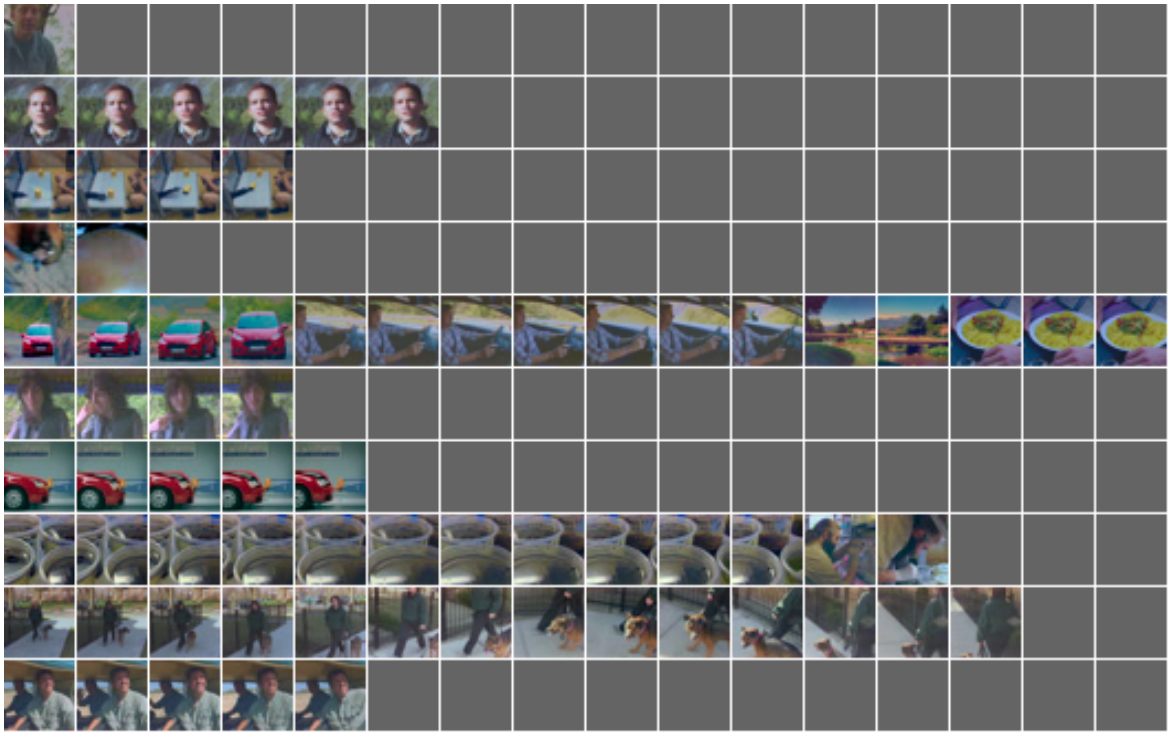
Learned bases with (right) and without (left) L_1 penalty

Spatially Coherent Features

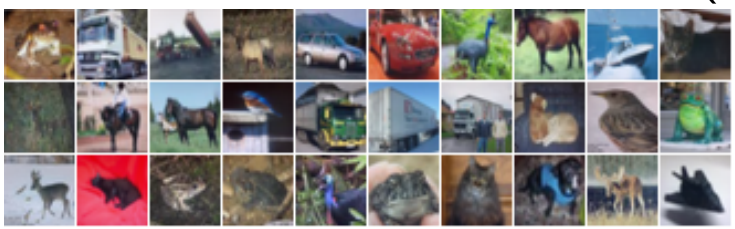


- What criterion should determine the values of α and β?
- What is the correct trade-off between discriminability and invariance?

Datasets

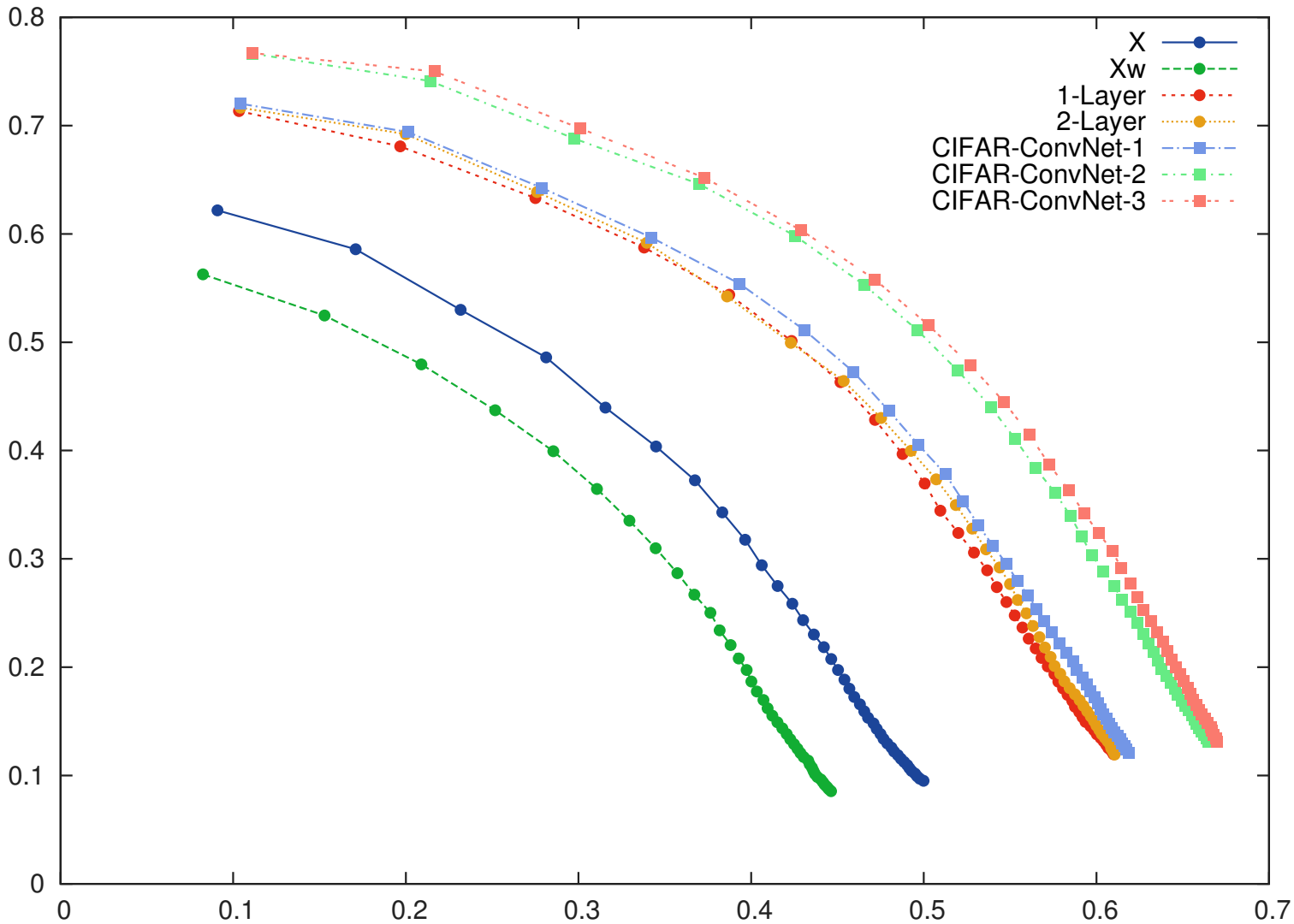


Train frames = 125,000, Test frames = 23,337 (1,626 unique scenes)

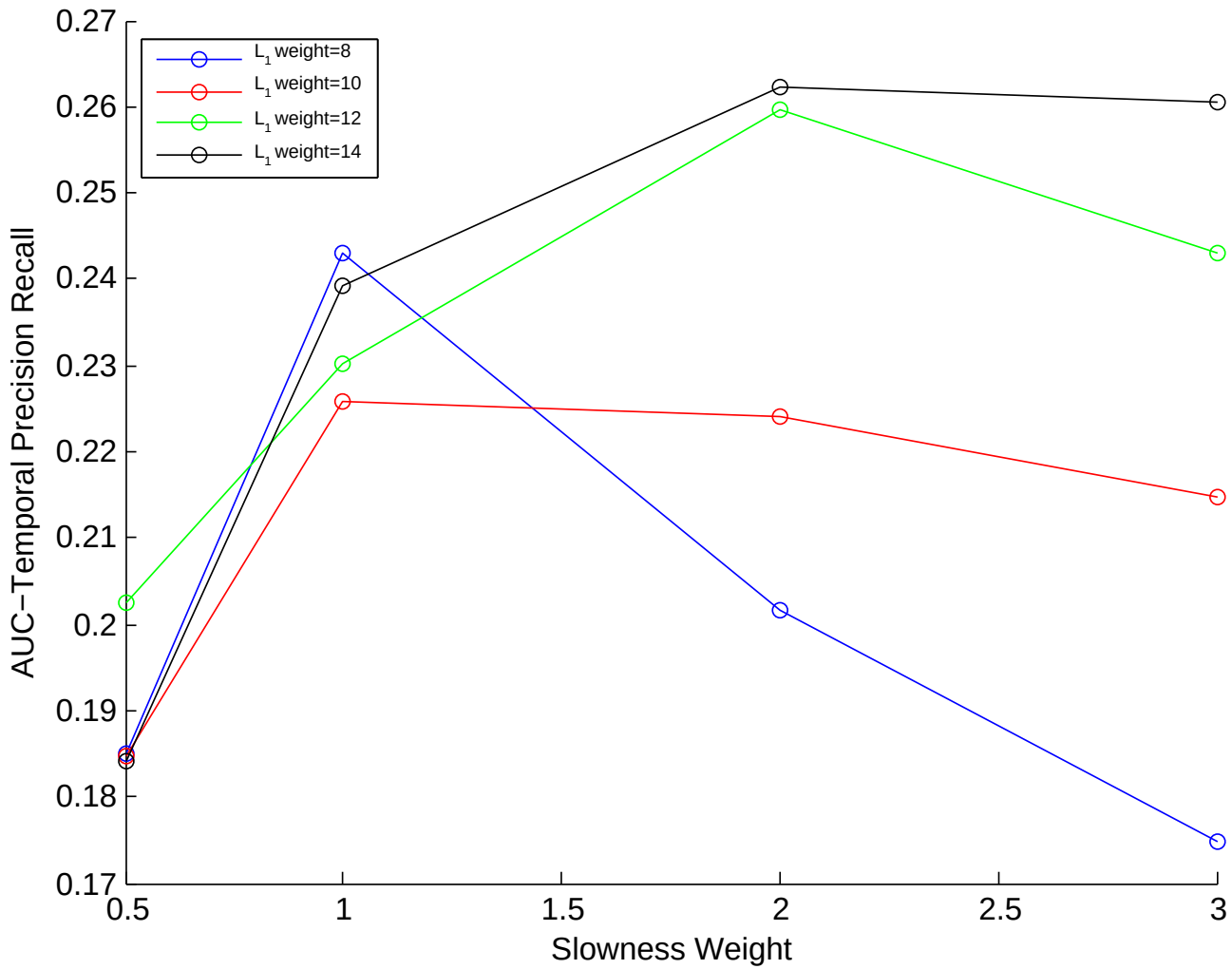


Train images = 50,000, Test images 10,000

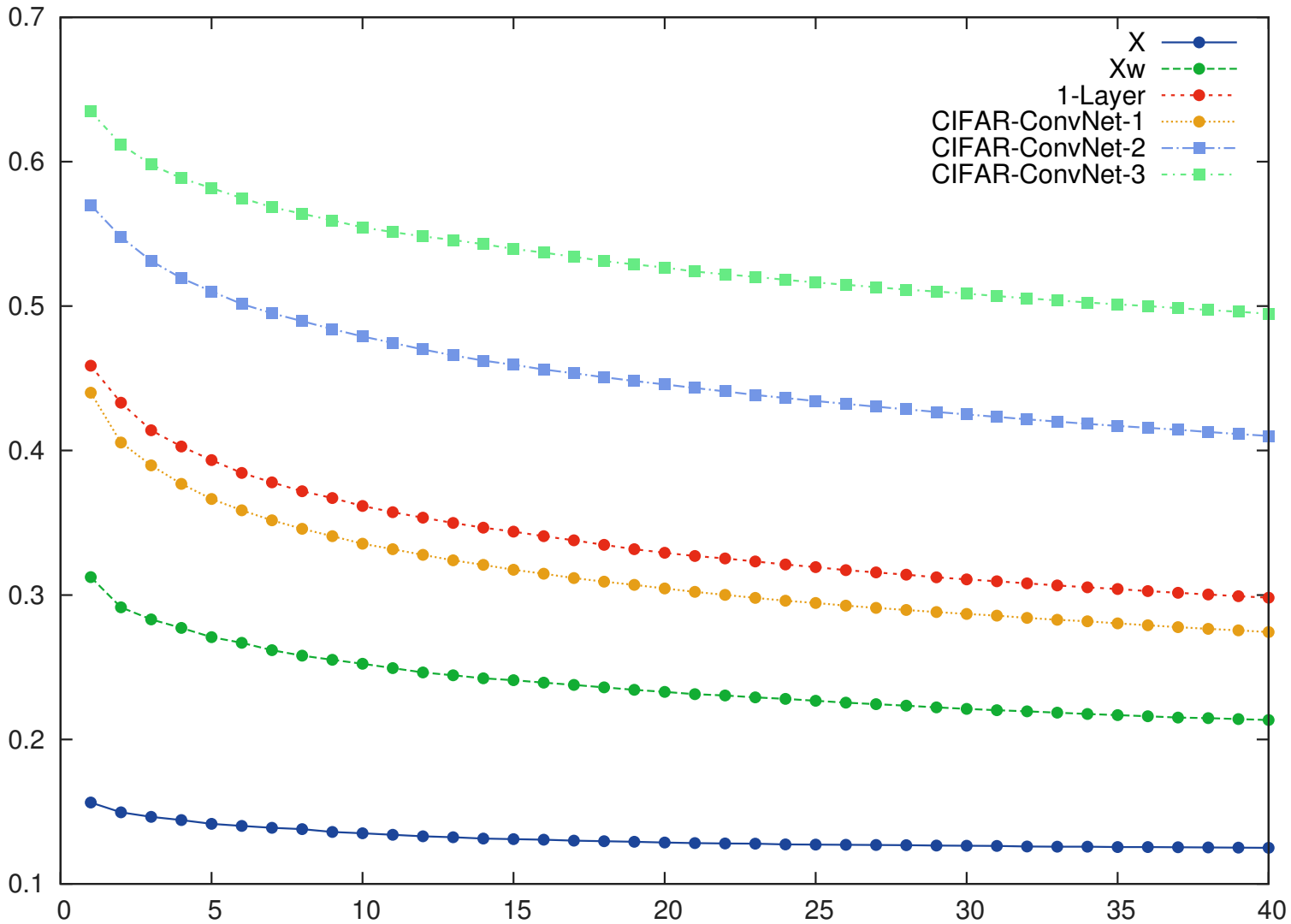
Scene Precision-Recall



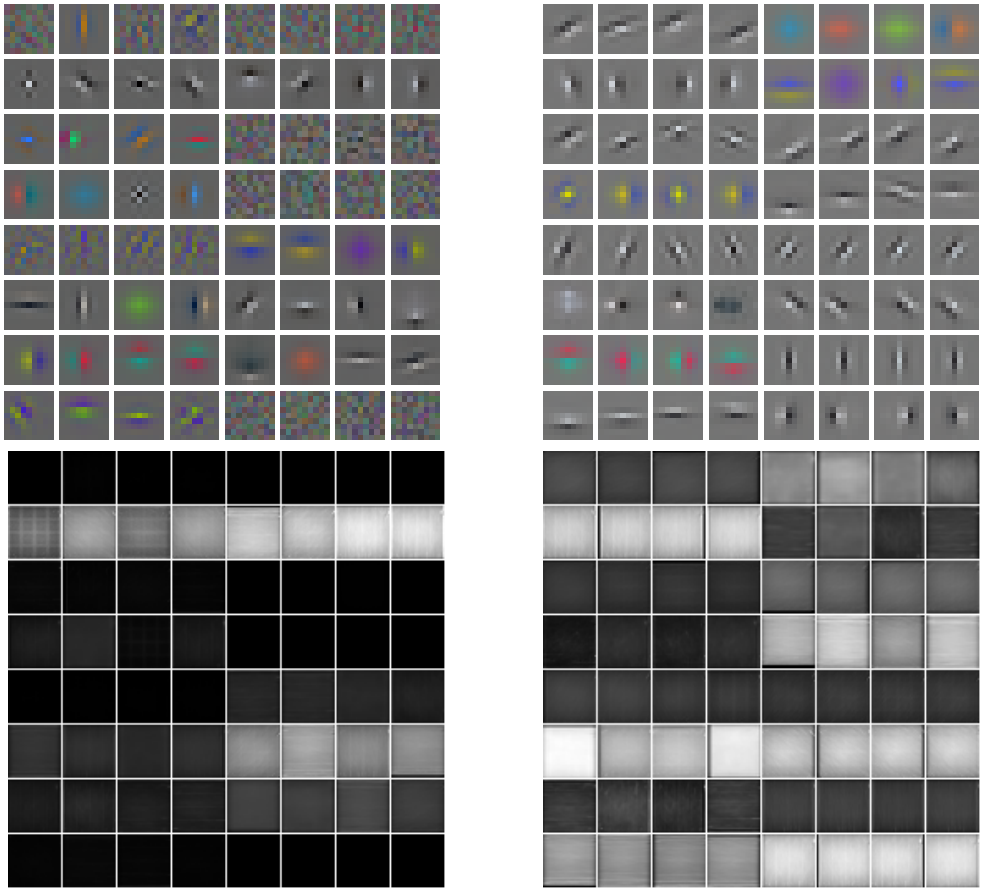
Scene Precision-Recall



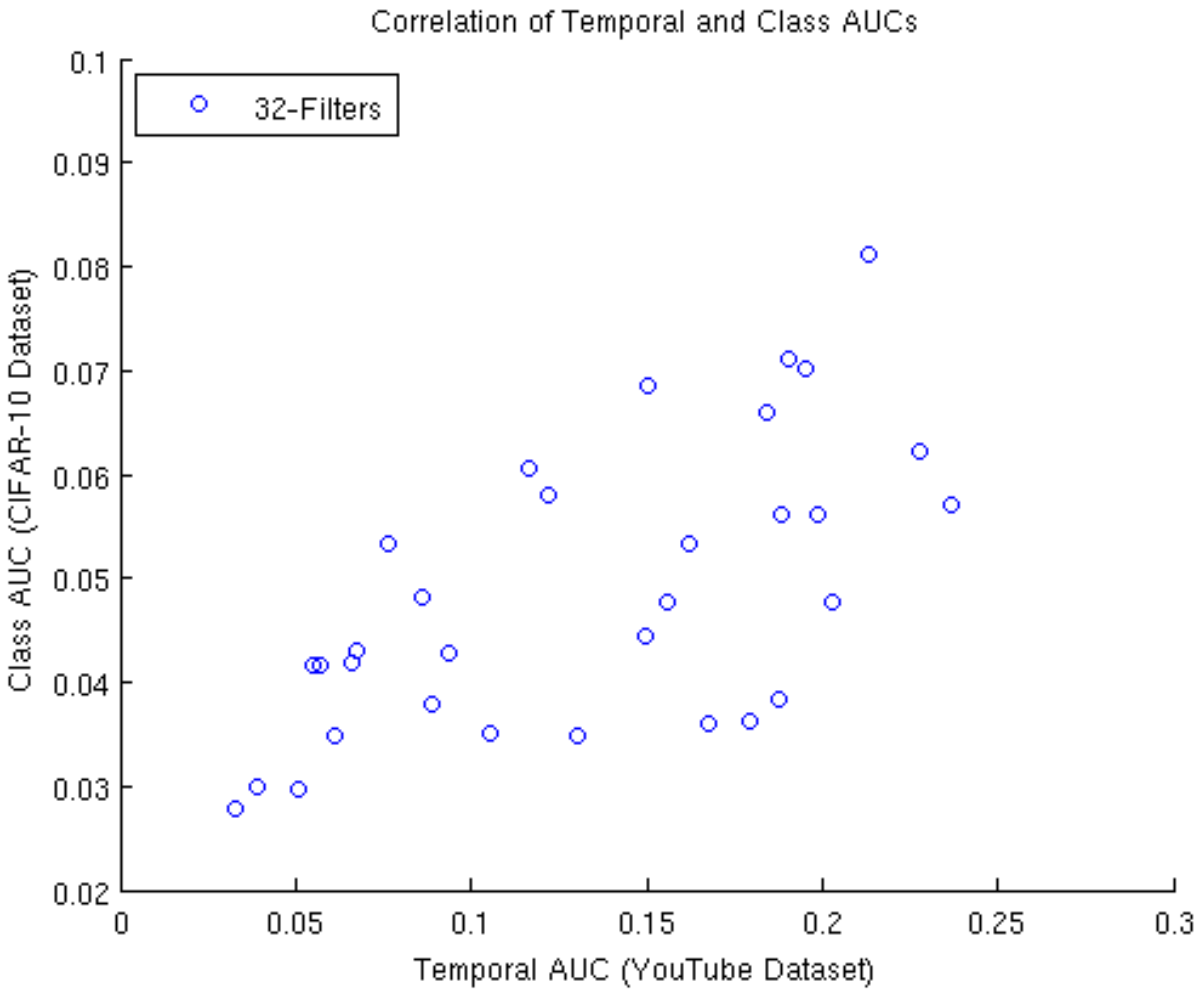
Class Precision-Recall



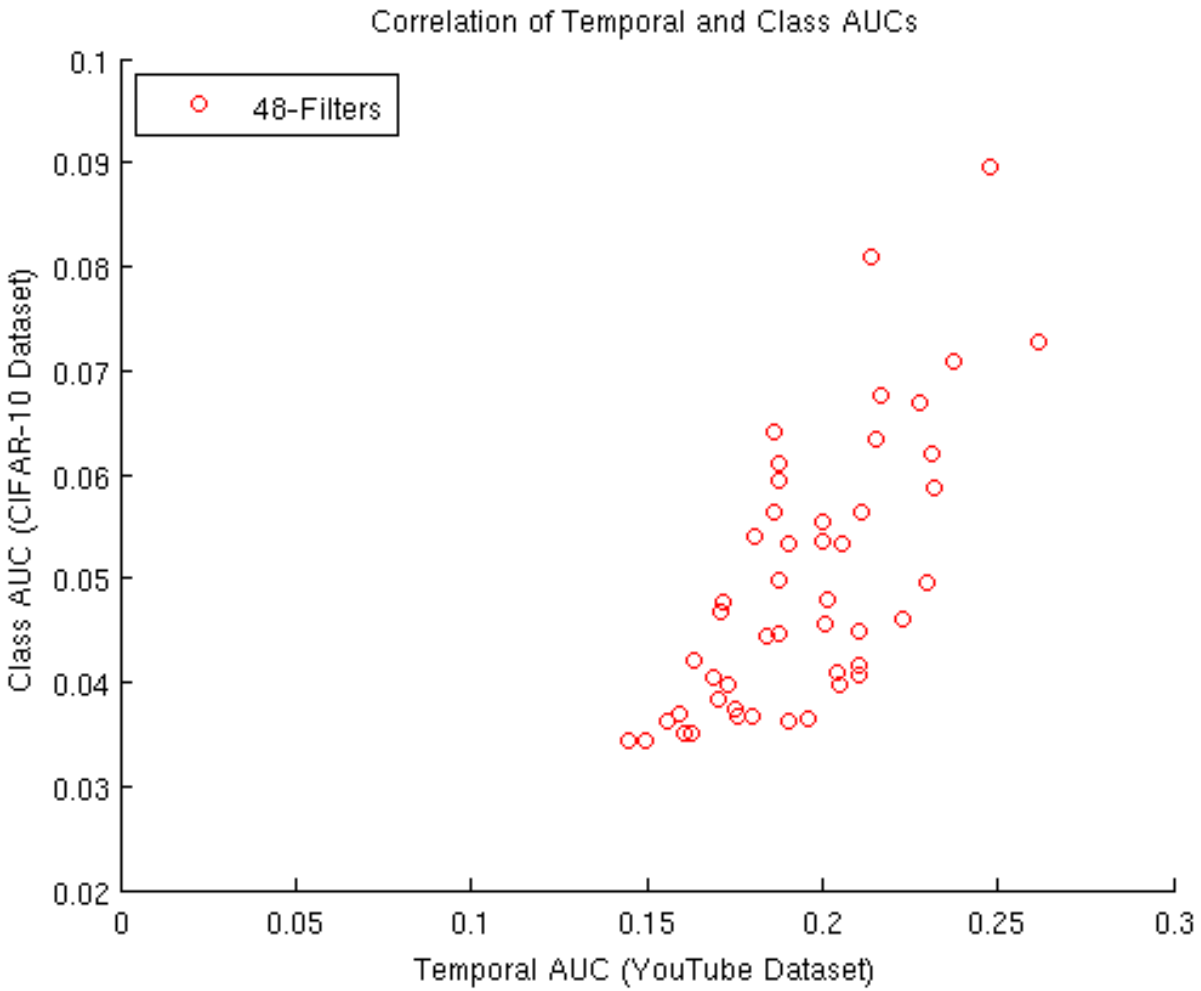
Learned Filter Banks



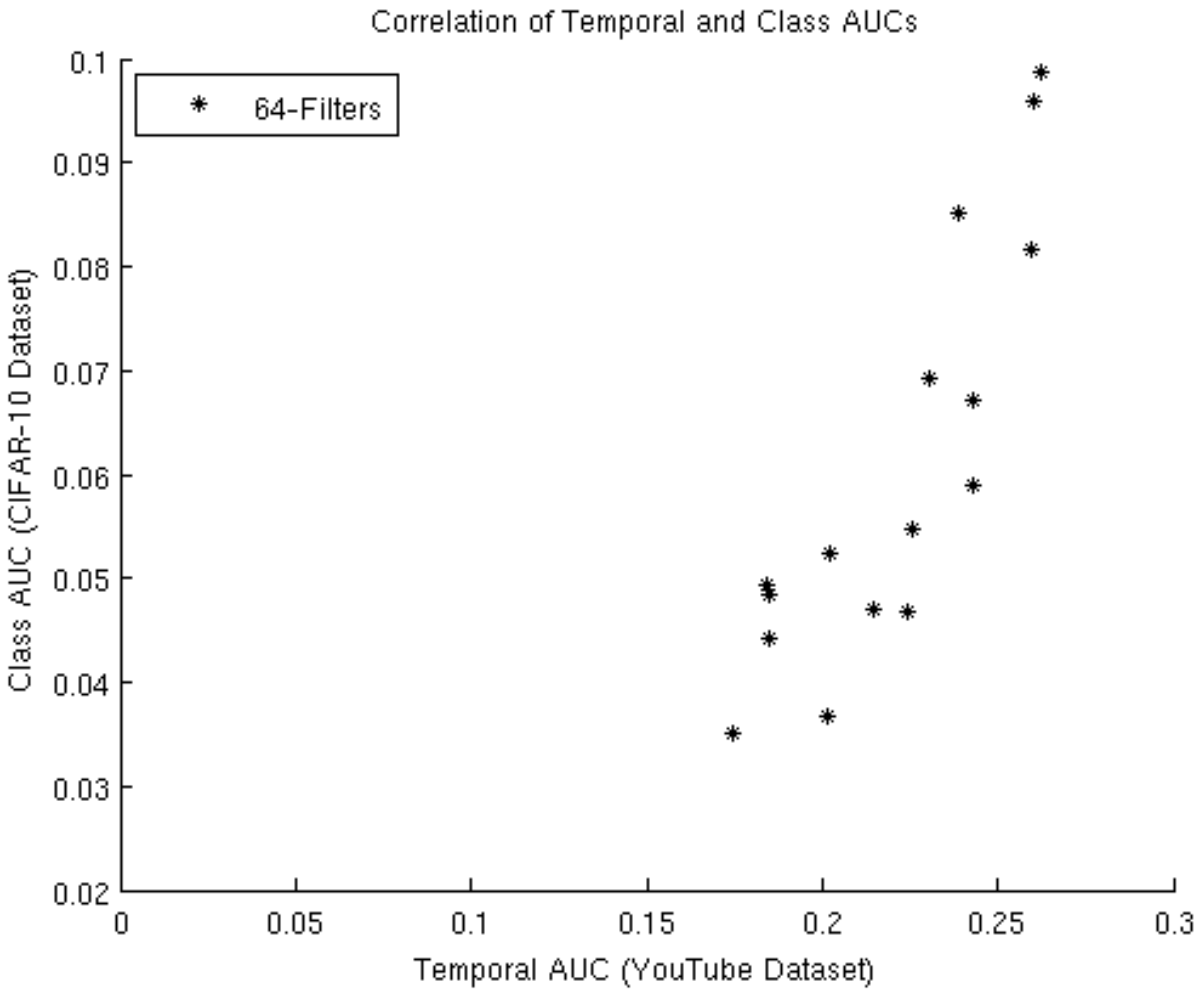
Relationship between Performance on Two Datasets



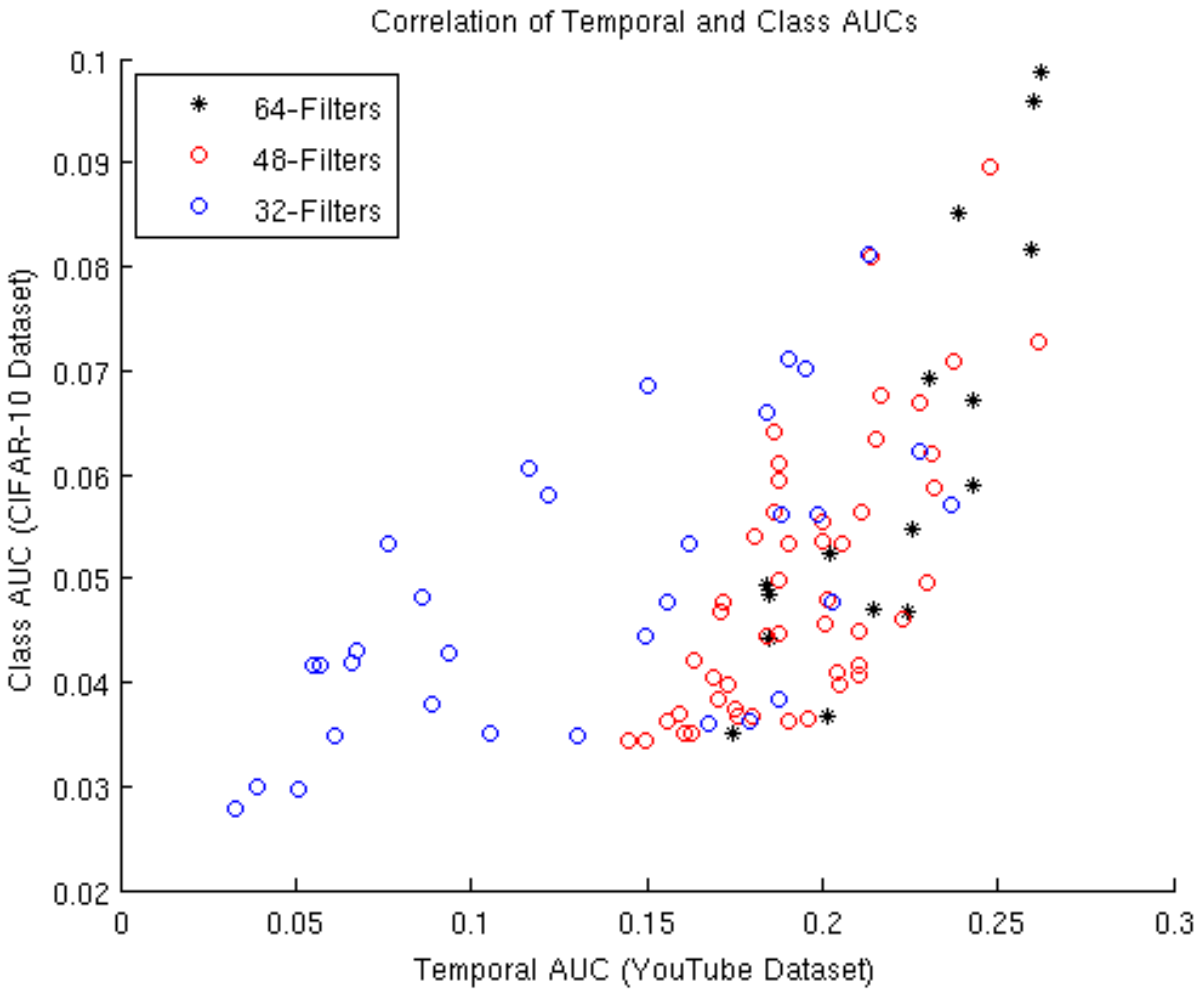
Relationship between Performance on Two Datasets



Relationship between Performance on Two Datasets



Relationship between Performance on Two Datasets

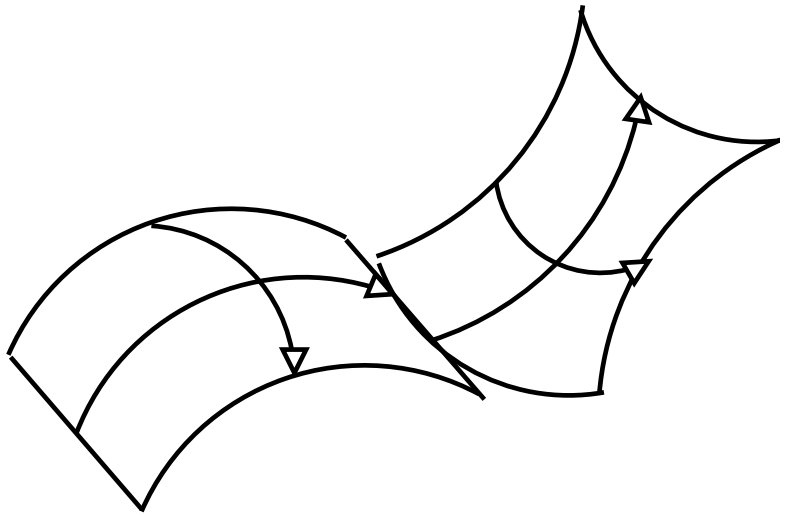


Shortcomings of the Model

- Does not regularize/output where (phase), only the what (magnitude)
- Indirect control on the informativeness of the representation because reconstruction is enforced *before* pooling
- Equivalent to reconstructing from a pooling operator which outputs all the moments
- Phase is unregularized
- Difficult to train multiple layers greedily

Other Temporal Objectives

Learn a representation that is information preserving and locally linearizes temporal trajectories



Let $\phi(x_t)$ be the feature-space representation of the sample at t

$$\|2\phi(x_t) - \phi(x_{t-1}) - \phi(x_{t+1})\|$$

Pooling with Phase

Magnitude: $m_i = \max_{z_j \in N_i(z_j)} z_j$ or $m_i = \frac{1}{\beta} log \left(\sum_{z_j \in N_i} e^{\beta z_j} \right)$

Phase: $p_k = \frac{e^{\beta z_k}}{\sum_{z_j \in N_i} e^{\beta z_j}}$

First moment of the phase: $\mu_i = \sum_{p_j \in N_i} x_j p_j$

- Assign a topology on the support N_i, and let x_j be a monotonic function of j
- μ_i correspond to the offset within N_i
- Constant speed translations of the activations along their support correspond to *linear variations* in μ
- Pooling across feature maps parameterizes arbitrary transformations

Applications

Supervised

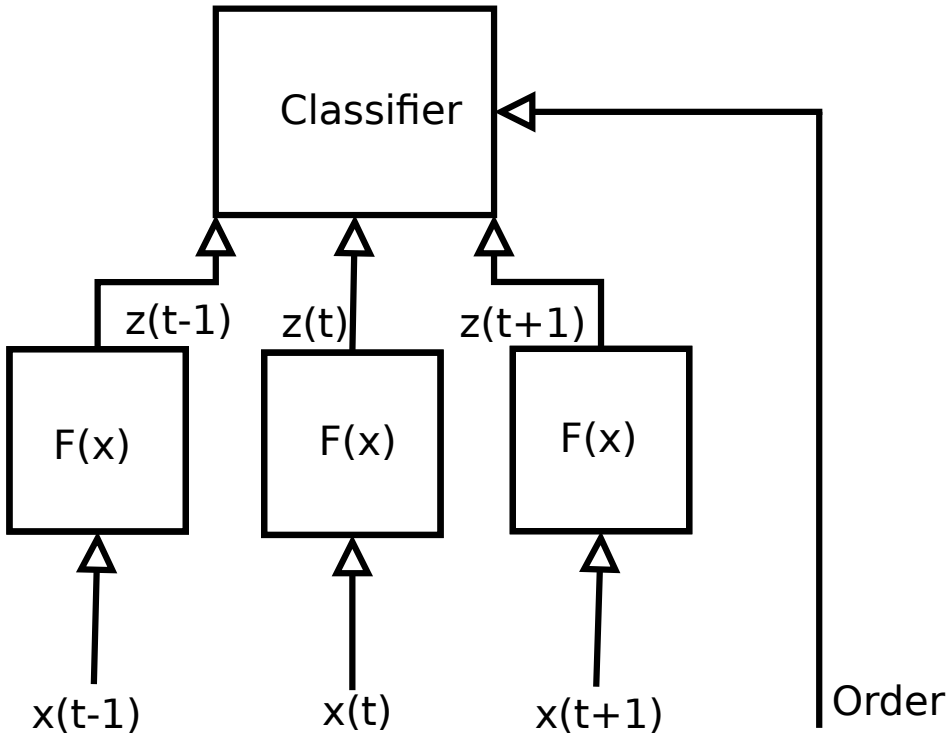
- With spatial pooling $\|\mu_i - \mu_j\|$ is the distance between the activations of features i and j
- End-to-end training of parts models

Unsupervised

- Minimize $\|2\mu_t - \mu_{t-1} - \mu_{t+1}\|$, provided that $m_{t-1} > 0$, $m_t > 0$, and $m_{t+1} > 0$, and minimize $|m_t - m_{t-1}|$
- This induces the pool group activations to be piecewise constant, while linearizing the phase within each group
- Phase moments and magnitude can be used to generate a reconstruction of the input

Supervised-Like Training with Time

Is it possible to conceive of an objective that implicitly forces temporal linearization?



Thank you

(Questions?)