Mathematical
Institute

The Scattering Transform: a deterministic transform with depth; and into to auto-encoders

Theories of Deep Learning: C6.5,
Lecture / Video 12

Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford
Mathematics

Scattering Transform (Mallat 12')

Repeated application of deterministic transforms

UNIVERSITY OF OXFORD
Mathematical Institute

The Scattering Transform repeatedly applied a deterministic wavelet transform followed by $\sigma(x) = |x|$ as nonlinear activation

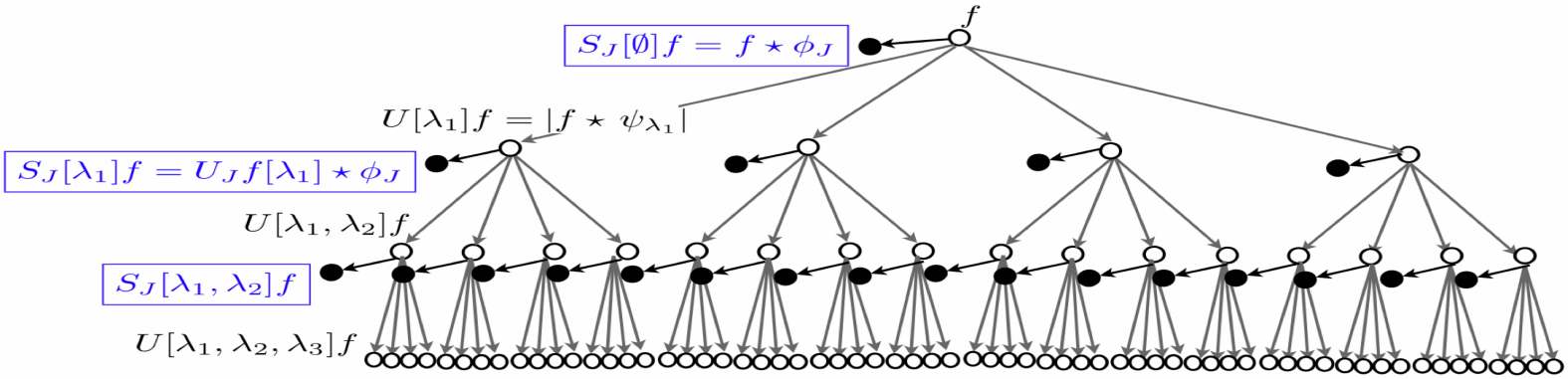


Figure 1: A scattering propagator U_J applied to f computes each $U[\lambda_1]f = |f \star \psi_{\lambda_1}|$ and outputs $S_J[\emptyset]f = f \star \phi_{2^J}$. Applying U_J to each $U[\lambda_1]f$ computes all $U[\lambda_1, \lambda_2]f$ and outputs $S_J[\lambda_1] = U[\lambda_1] \star \phi_{2^J}$. Applying iteratively U_J to each $U[p]f$ outputs $S_J[p]f = U[p]f \star \phi_{2^J}$ and computes the next path layer.

Depth allows the transform to become increasingly invariant to translation and small diffeomorphisms.

`https://arxiv.org/pdf/1101.2286.pdf`

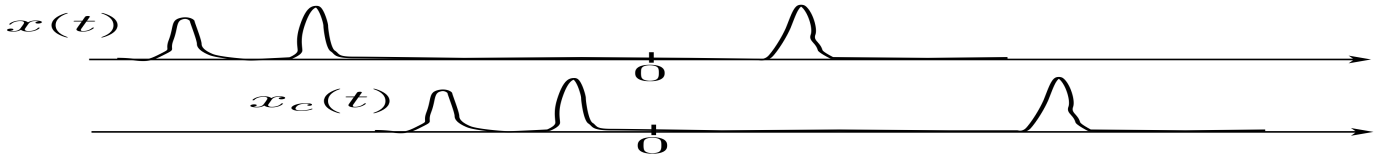
Classification as learning invariance (Mallat '13)

Projecting out invariants not needed for classification

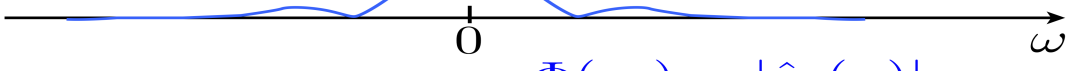
Mathematical Institute

Invariance to translations $x_c(t) = x(t - c)$

$$\forall c \in \mathbf{R} \quad , \quad \Phi(x_c) = \Phi(x) \ .$$



$\Phi(x) = |\hat{x}(\omega)|$: Fourier Modulus



$\Phi(x_c) = |\hat{x}_c(\omega)|$

Lipschitz stable to deformations $x_\tau(t) = x(t - \tau(t))$

small deformations of x $\Longrightarrow$ small modifications of $\Phi(x)$

$$\forall \tau \quad , \quad \|\Phi(x_\tau) - \Phi(x)\| \leq C \underbrace{\sup_t |\nabla \tau(t)|}_{\text{deformation size}} \|x\| \ .$$

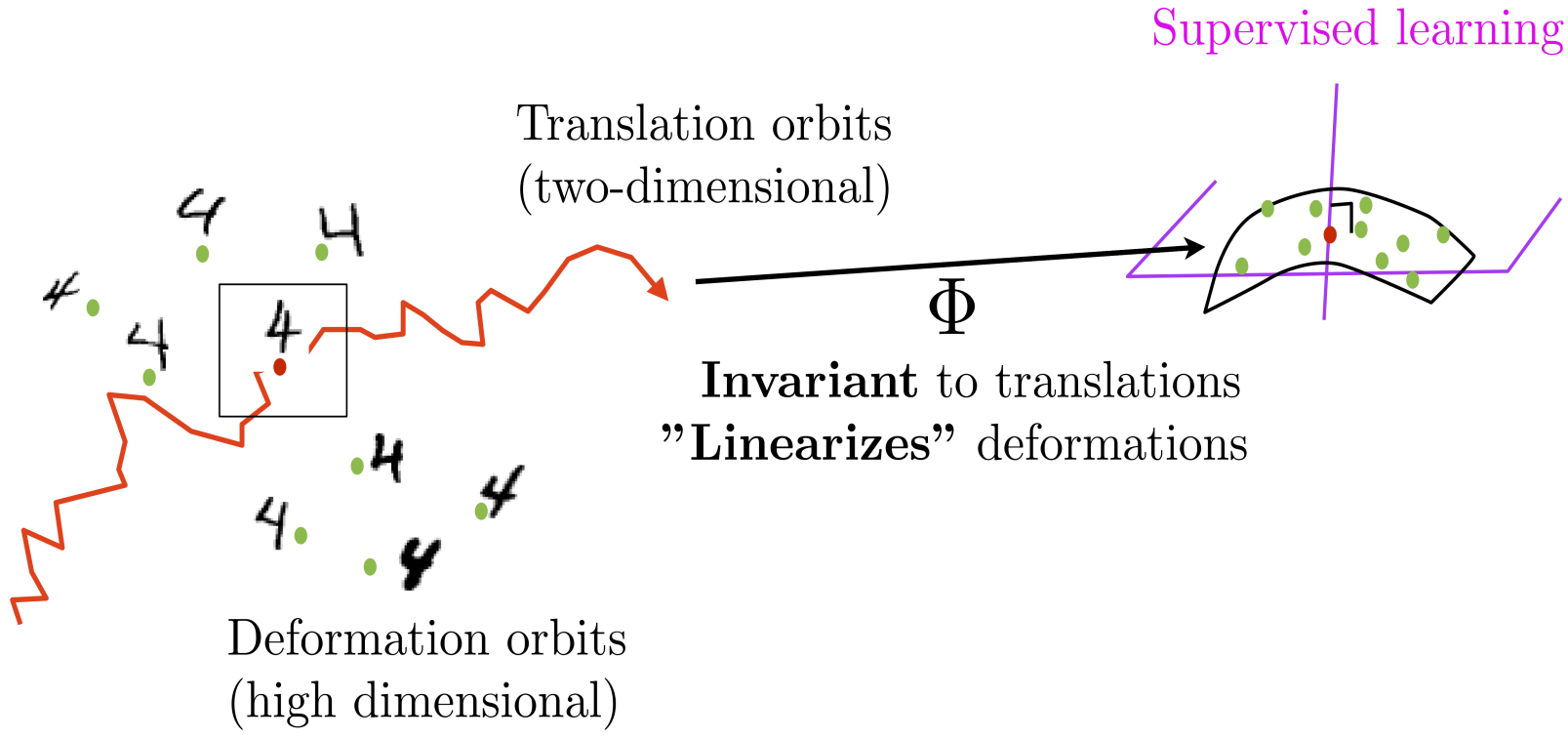
http://lcsl.mit.edu/ldr-workshop/Home.html

Linearising deformations (Mallat '13)

Projecting out invariants not needed for classification

University of Oxford
Mathematical Institute

- Specific deformation invariance must be learned.

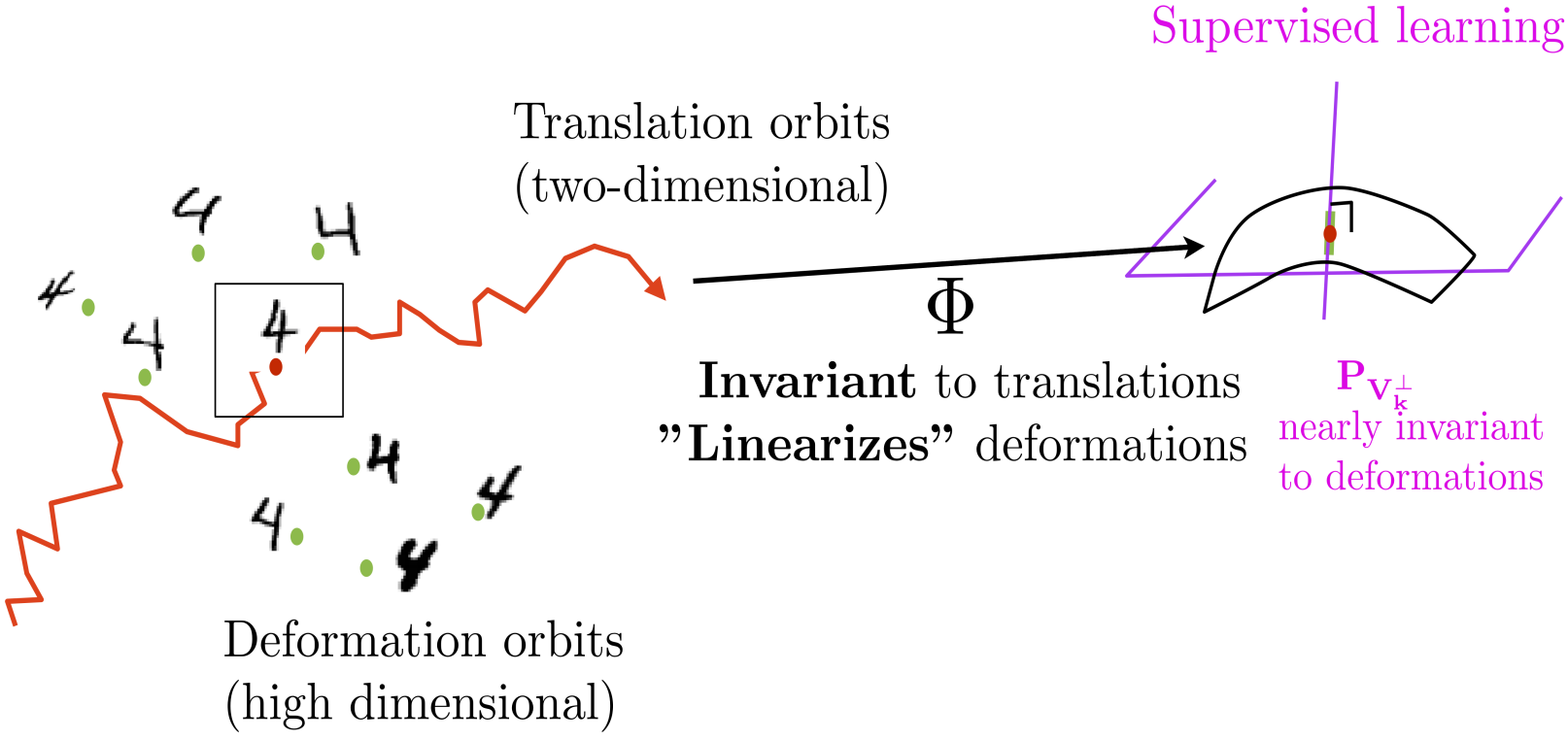


`http://lcsl.mit.edu/ldr-workshop/Home.html`

Linearising deformations (Mallat '13)

Projecting out invariants not needed for classification

- Specific deformation invariance must be learned.

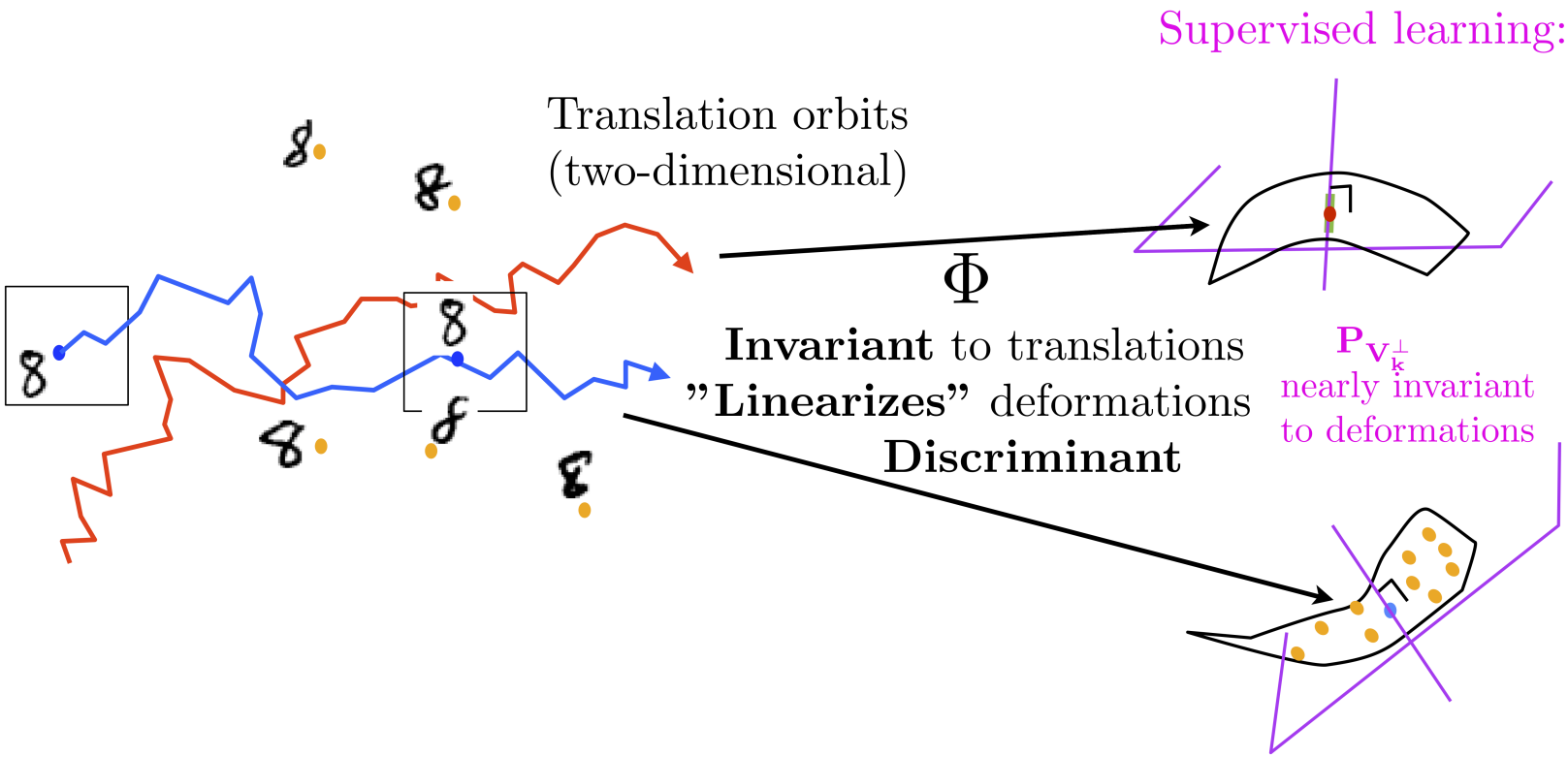


`http://lcsl.mit.edu/ldr-workshop/Home.html`

Linearising deformations (Mallat '13)

Projecting out invariants not needed for classification

- Specific deformation invariance must be learned.



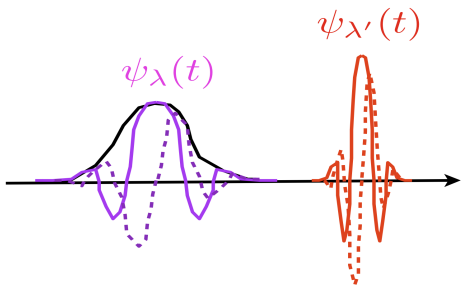
`http://lcsl.mit.edu/ldr-workshop/Home.html`

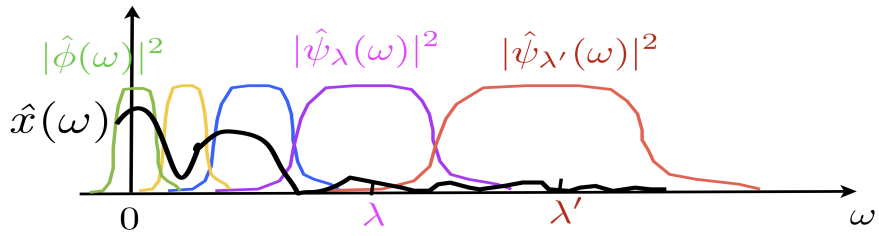
Wavelet Transform as frequency tiling (Mallat '13)

Wavelets decompose function into local frequency information

University of Oxford
Mathematical Institute

- Complex wavelet: $\psi(t) = \psi^a(t) + i\,\psi^b(t)$
- Dilated: $\psi_\lambda(t) = 2^{-j}\,\psi(2^{-j}t)$ with $\lambda = 2^{-j}$.





- Wavelet transform: $x \star \psi_\lambda(t) = \int x(u)\,\psi_\lambda(t-u)\,du$

$$Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$$

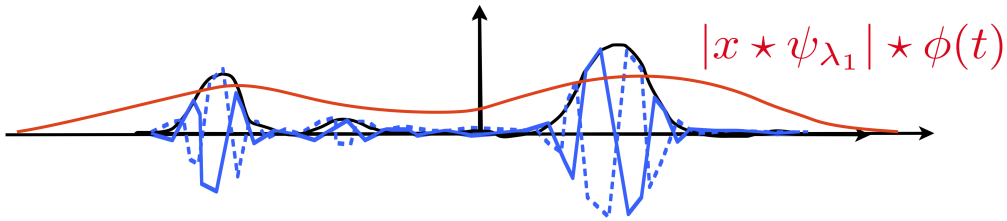
Unitary: $\|Wx\|^2 = \|x\|^2$.

`http://lcsl.mit.edu/ldr-workshop/Home.html`

Modulus and averaging in wavelet domain (Mallat '13)

Smoothing to identify discontinuities and have energy decay

University of Oxford
Mathematical Institute



- The modulus $|x \star \psi_{\lambda_1}|$ is a regular envelop
- The average $|x \star \psi_{\lambda_1}| \star \phi(t)$ is invariant to small translations relatively to the support of ϕ.
- Full translation invariance at the limit:

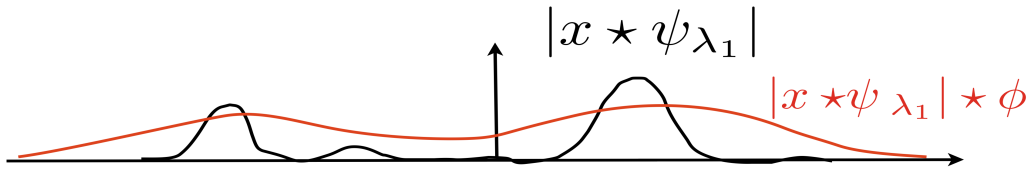
$$\lim_{\phi \to 1} |x \star \psi_{\lambda_1}| \star \phi(t) = \int |x \star \psi_{\lambda_1}(u)| \, du = \|x \star \psi_{\lambda_1}\|_1$$

`http://lcsl.mit.edu/ldr-workshop/Home.html`

Second layer of the scattering transform (Mallat '13)

Increased smoothness with depth

Mathematical Institute



- The high frequencies of $|x \star \psi_{\lambda_1}|$ are in wavelet coefficients:

$$W|x \star \psi_{\lambda_1}| = \begin{pmatrix} |x \star \psi_{\lambda_1}| \star \phi(t) \\ |x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t) \end{pmatrix}_{t,\lambda_2}$$

- Translation invariance by time averaging the amplitude:

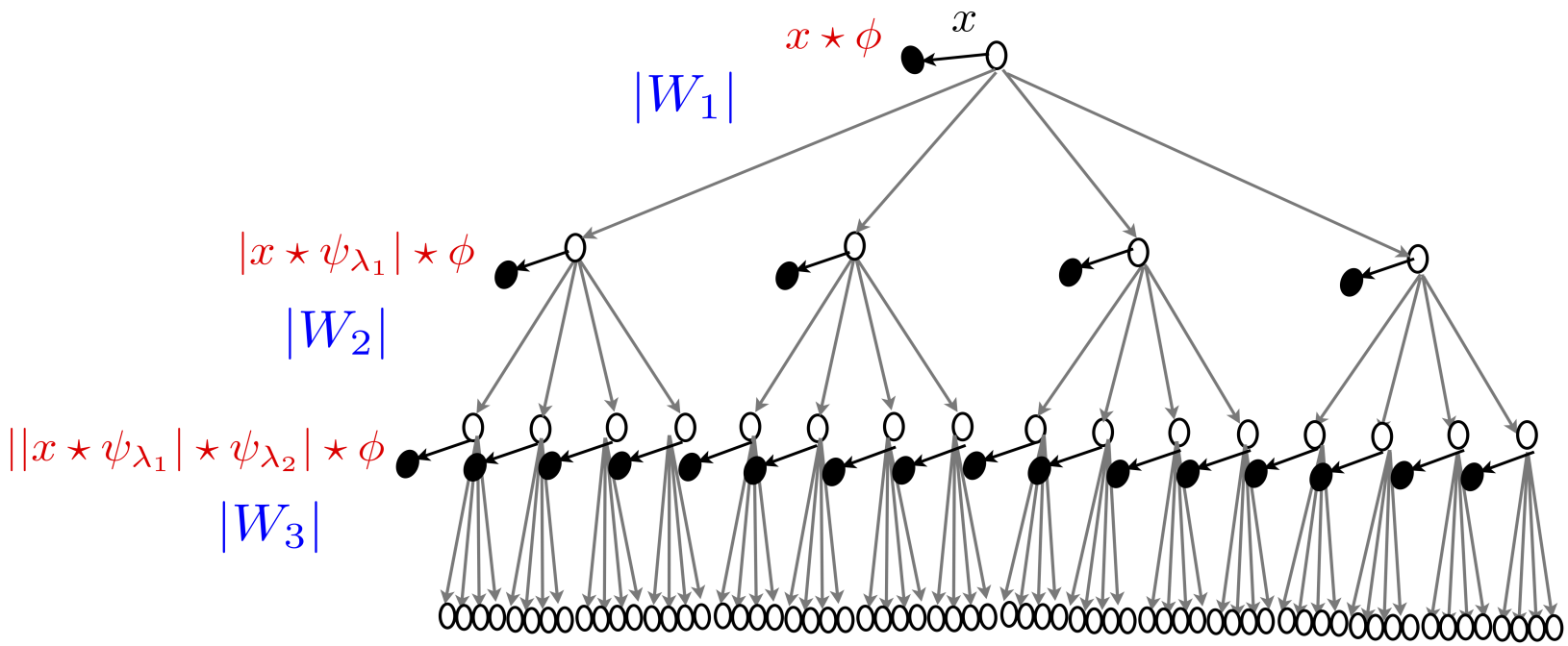
$$\forall \lambda_1 , \lambda_2 , \quad ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t)$$

`http://lcsl.mit.edu/ldr-workshop/Home.html`

Scattering transform (Mallat '13)

Lipshitz continuous, inputs contract to one another

Mathematical Institute



- Cascade of contractive operators

$$\| |W_k|x - |W_k|x' \| \leq \|x - x'\| \quad \text{with} \quad \| |W_k|x \| = \|x\| \ .$$

`http://lcsl.mit.edu/ldr-workshop/Home.html`

Scattering transform properties(Mallat '13)

Stability to deformormations

$$Sx = \begin{pmatrix} x \star \phi\,(u) \\ |x \star \psi_{\lambda_1}| \star \phi(u) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(u) \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(u) \\ \cdots \end{pmatrix}_{u,\lambda_1,\lambda_2,\lambda_3,\ldots}$$

Theorem: *For appropriate wavelets, a scattering is*

contractive $\|Sx - Sy\| \leq \|x - y\|$

preserves norms $\|Sx\| = \|x\|$

stable to deformations $x_\tau(t) = x(t - \tau(t))$

$$\|Sx - Sx_\tau\| \leq C \sup_t |\nabla \tau(t)|\, \|x\|$$

$\Rightarrow$ linear discriminative classification from $\Phi x = Sx$

`http://lcsl.mit.edu/ldr-workshop/Home.html`

Scattering Transform: energy decay (Mallat 12')

The transform can be truncated stably

UNIVERSITY OF OXFORD
Mathematical Institute

Lemma

For suitably chosen wavelet transforms (see Theorem 2.6 in footnote) then for all $f \in L^2(\mathbb{R}^d)$

$$\lim_{m\to\infty} \|U[\Lambda_J^m]f\|^2 = \lim_{m\to\infty} \sum_{n=m}^{\infty} \|S_J[\Lambda_J^n]f\|^2 = 0$$

where $U[\lambda]f = |f \star \psi_\lambda|$ and $S_J[\lambda]f = \phi_j \star U[\lambda]f$ and $\|S_J[P_J]f\| = \|f\|$. Morevover, for all $c \in \mathbb{R}^d$

$$\lim_{J\to\infty} \|S_J[P_J]f - S_J[P_J]L_c f\| = 0$$

where $L_c f = f(x - c)$ is the translation operator.

`https://arxiv.org/pdf/1101.2286.pdf`

Scattering Transform: energy decay (Mallat 13')

Energy decay on CalTech-101

University of Oxford
Mathematical Institute

TABLE 1
Percentage of Energy $\sum_{p \in \mathcal{P}^m} \|S[p]x\|^2 / \|x\|^2$ of Scattering Coefficients on Frequency-Decreasing Paths of Length m, Depending upon J

J	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m \leq 3$
1	95.1	4.86	-	-	-	99.96
2	87.56	11.97	0.35	-	-	99.89
3	76.29	21.92	1.54	0.02	-	99.78
4	61.52	33.87	4.05	0.16	0	99.61
5	44.6	45.26	8.9	0.61	0.01	99.37
6	26.15	57.02	14.4	1.54	0.07	99.1
7	0	73.37	21.98	3.56	0.25	98.91

These average values are computed on the Caltech-101 database, with zero mean and unit variance images.

https://www.di.ens.fr/data/publications/papers/pami-final.pdf

Scattering Transform: MNIST classification (Mallat 13')

Accuracy on MNIST based on training size

UNIVERSITY OF OXFORD
Mathematical Institute

TABLE 4
Percentage of Errors of MNIST Classifiers, Depending on the Training Size

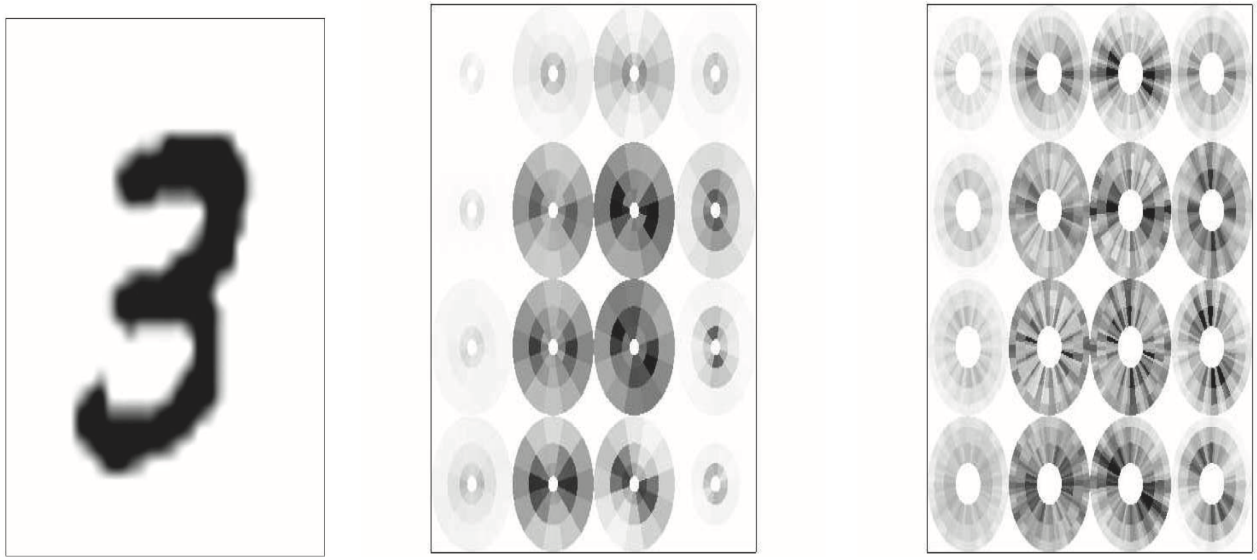
Training size	x PCA	x SVM	Wind. Four. PCA	Wind. Four. SVM	Scat. $\overline{m}=1$ PCA	Scat. $\overline{m}=1$ SVM	Scat. $\overline{m}=2$ PCA	Scat. $\overline{m}=2$ SVM	Conv. Net.
300	14.5	15.4	7.35	7.4	5.7	8	**4.7**	5.6	7.18
1000	7.2	8.2	3.74	3.74	2.35	4	**2.3**	2.6	3.21
2000	5.8	6.5	2.99	2.9	1.7	2.6	**1.3**	1.8	2.53
5000	4.9	4	2.34	2.2	1.6	1.6	**1.03**	1.4	1.52
10000	4.55	3.11	2.24	1.65	1.5	1.23	0.88	1	**0.85**
20000	4.25	2.2	1.92	1.15	1.4	0.96	0.79	**0.58**	0.76
40000	4.1	1.7	1.85	0.9	1.36	0.75	0.74	**0.53**	0.65
60000	4.3	1.4	1.80	0.8	1.34	0.62	0.7	**0.43**	0.53

`https://www.di.ens.fr/data/publications/papers/pami-final.pdf`

Scattering Transform: MNIST digit 3 (Mallat 13')

Example of energy in a scattering transform

Mathematical Institute



(a) (b) (c)

Fig. 7. (a) Image $X(u)$ of a digit "3." (b) Arrays of windowed scattering coefficients $S[p]X(u)$ of order $m = 1$, with u sampled at intervals of $2^J = 8$ pixels. (c) Windowed scattering coefficients $S[p]X(u)$ of order $m = 2$.

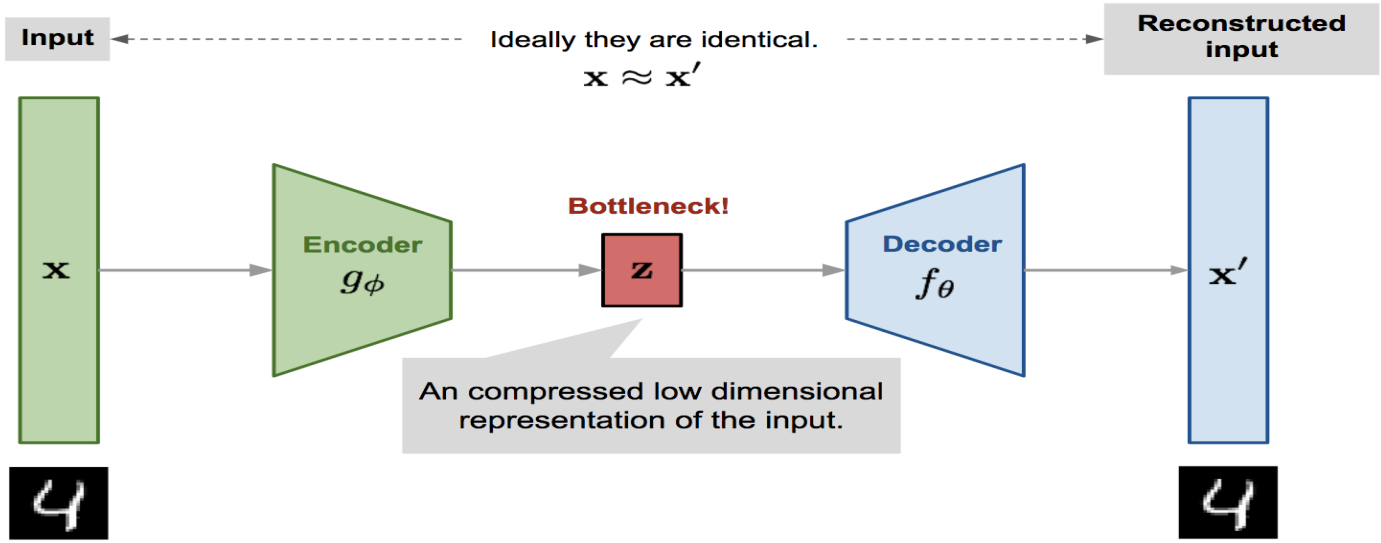
https://www.di.ens.fr/data/publications/papers/pami-final.pdf

Introduction to Variational Autoencoders.

Autoencoder (AE) Illustration

Restricting the number of data parameters

University of Oxford
Mathematical Institute



The parameters, (θ, ϕ), of the autoencoder are then learned:

$$\mathcal{L}(\theta, \phi) = m^{-1} \sum_{\mu=1}^{m} l(x_\mu, f_\theta(g_\phi(x_\mu)))$$

https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

The parameters, (θ, ϕ), of the autoencoder are then learned:

$$\mathcal{L}(\theta, \phi) = n^{-1} \sum_{\mu=1}^{n} l(x_\mu, f_\theta(g_\phi(x_\mu)))$$

Consider a simple model where the encoder and decoder are linear, that is $g_\phi(x) = \Phi x$ where $\Phi \in \mathbb{R}^{r \times p}$ with $r < p$, and the linear decoder $f_\theta(z) = \Theta z$ with $\Theta \in \mathbb{R}^{p \times r}$.
Moreover, consider an entrywise ℓ_2^2 error for $l(x_\mu, f_\theta(g_\phi(x_\mu)))$, then

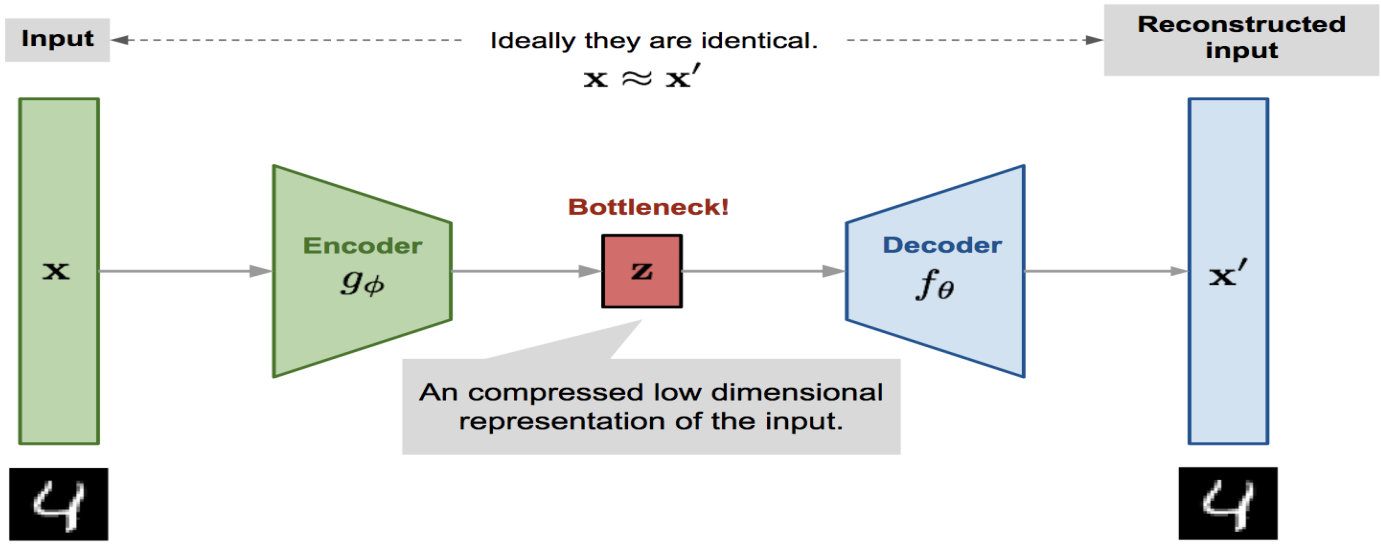
$$\mathcal{L}(\theta, \phi) = n^{-1} \|X - \Theta\Phi X\|_F^2$$

where $\Theta\Phi$ is a learned rank r matrix, whose optimal solution is the projector of X to its leading r singular space.

Autoencoder extend PCA

More complex maps to low parameter space

Mathematical Institute



The autoencoder framework allows $g_\phi(\cdot)$ and $f_\theta(\cdot)$ to be more general than linear, and in particular to benefit from the expressively of depth and introduce variation.
`https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html`

PCA vs 3 layer Autoencoder: MNIST (Hinton et al. 06')

Improved separation of data classes

Mathematical Institute

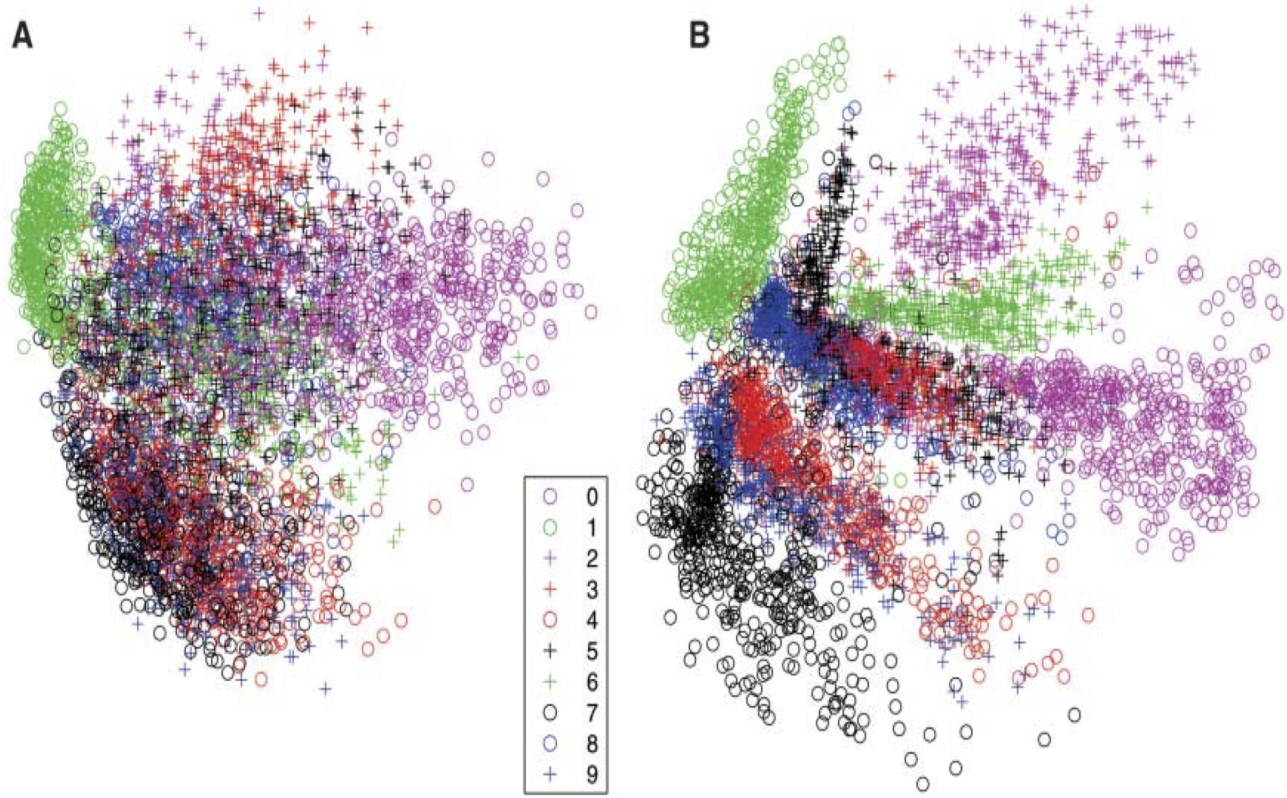


Fig. 3. (**A**) The two-dimensional codes for 500 digits of each class produced by taking the first two principal components of all 60,000 training images. (**B**) The two-dimensional codes found by a 784-1000-500-250-2 autoencoder. For an alternative visualization, see (*8*).

http://science.sciencemag.org/content/313/5786/504

k-sparse autoencoders (Makhzani et al. 13')

Low dimensionality through sparsity

UNIVERSITY OF OXFORD
Mathematical Institute

k-Sparse Autoencoders:

Training:

1) Perform the feedforward phase and compute

$$\boldsymbol{z} = W^\top \boldsymbol{x} + \boldsymbol{b}$$

2) Find the k largest activations of $\boldsymbol{z}$ and set the rest to zero.

$$\boldsymbol{z}_{(\Gamma)^c} = 0 \quad \text{where} \quad \Gamma = \text{supp}_k(\boldsymbol{z})$$

3) Compute the output and the error using the sparsified $\boldsymbol{z}$.

$$\hat{\boldsymbol{x}} = W\boldsymbol{z} + \boldsymbol{b}'$$
$$E = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2$$

3) Backpropagate the error through the k largest activations defined by Γ and iterate.

Sparse Encoding:

Compute the features $\boldsymbol{h} = W^\top \boldsymbol{x} + \boldsymbol{b}$. Find its αk largest activations and set the rest to zero.

$$\boldsymbol{h}_{(\Gamma)^c} = 0 \quad \text{where} \quad \Gamma = \text{supp}_{\alpha k}(\boldsymbol{h})$$

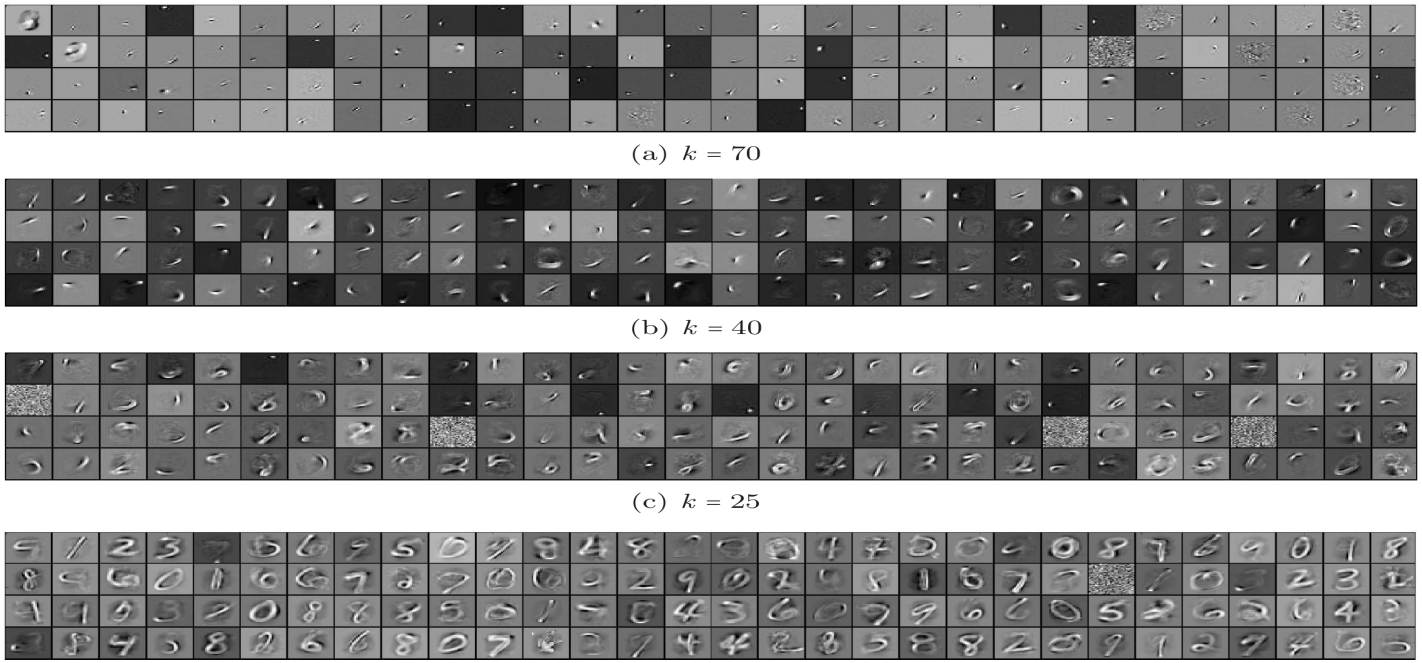
This framework includes nonlinearity and can be rigorously analysed using techniques from sparse approximation, but it lacks depth.

`https://arxiv.org/pdf/1312.5663.pdf`

k-sparse autoencoders (Makhzani et al. 13')

Learned elements: MNIST

UNIVERSITY OF OXFORD
Mathematical Institute



(a) $k = 70$

(b) $k = 40$

(c) $k = 25$

(d) $k = 10$

Figure 1. Filters of the k-sparse autoencoder for different sparsity levels k, learnt from MNIST with 1000 hidden units.

Elements learned depend on number of components, sparsity, allowed; k small are class elements, k large are basis elements.

`https://arxiv.org/pdf/1312.5663.pdf`

k-sparse autoencoders (Makhzani et al. 13')

Performance vs other autoencoders

UNIVERSITY OF OXFORD
Mathematical Institute

	Error Rate
Raw Pixels	7.20%
RBM	1.81%
Dropout Autoencoder (50% hidden)	1.80%
Denoising Autoencoder (20% input dropout)	1.95%
Dropout + Denoising Autoencoder (20% input and 50% hidden)	1.60%
k-Sparse Autoencoder, $k = 40$	1.54%
k-Sparse Autoencoder, $k = 25$	1.35%
k-Sparse Autoencoder, $k = 10$	2.10%

Table 1. Performance of unsupervised learning methods (without fine-tuning) with 1000 hidden units on **MNIST**.

	Error
Without Pre-Training	1.60%
RBM + F.T.	1.24%
Shallow Dropout AE + F.T. (%50 hidden)	1.05%
Denoising AE + F.T. (%20 input dropout)	1.20%
Deep Dropout AE + F.T. (Layer-wise pre-training, %50 hidden)	0.85%
k-Sparse AE + F.T. (k=25)	1.08%
Deep k-Sparse AE + F.T. (Layer-wise pre-training)	0.97%

Table 3. Performance of supervised learning methods on **MNIST**. Pre-training was performed using the corresponding unsupervised learning algorithm with 1000 hidden units, and then the model was fine-tuned.

`https://arxiv.org/pdf/1312.5663.pdf`