CS-667 Advanced Machine Learning

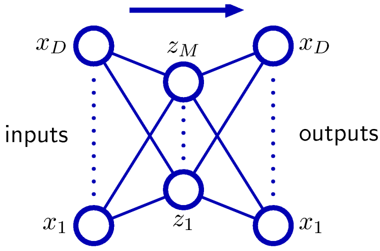
Nazar Khan

PUCIT

Autoassociative Neural Networks (Autoencoders)

Autoassociative Neural Networks

- Neural nets learn the mapping from inputs $\mathbf{x}_n$ to targets $\mathbf{t}_n$.
- If target is set to the input vector itself ($\mathbf{t}_n = \mathbf{x}_n$), the network learns to associate each input vector with itself.
- This is called an *autoassociative mapping* and the network is called an *autoassociative network*.



- Autoassociative nets perform unsupervised learning.

Autoassociative Neural Networks

- For $M < D$, hidden layer output $\mathbf{z} \in \mathbb{R}^M$ represents *dimensionality reduction*.
- Also called *autoencoders*.
- Serve as building blocks of deep learning. They enable deep architectures to be trained *properly*.

Autoencoders and Deep Learning

- Before deep learning, architectures with many layers suffered from the *vanishing gradient problem*.
- Gradients backpropagated to early layers had very small magnitudes.
- So the learning effectively took place in the later layers only.
- This meant that the architecture was effectively reduced from deep to shallow.
- Deep learning uses autoencoders to *pre-train* the weights of the early layers in an unsupervised fashion.
- When these weights are initial weights, standard backpropagation successfully trains all layers.

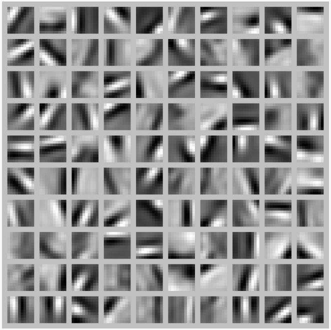
Two Layer Autoassociative Nets

Equivalence with PCA

- It can be proven that for two layer autoassociative nets, outputs of the M hidden neurons correspond to projection of $\mathbf{x}$ onto the M-dimensional subspace spanned by the first M principal components of the data.
- This is true when activation functions of the hidden neurons are linear as well as when they are non-linear.
- Weights of hidden neurons form the basis set that spans the principal subspace.
- However, they need not be orthogonal or normalized.
- There is *no advantage* over standard PCA methods that guarantee
    - correct solution
    - in finite time
    - ordered eigenvalues
    - orthonormal eigenvectors.

Autoassociative Nets

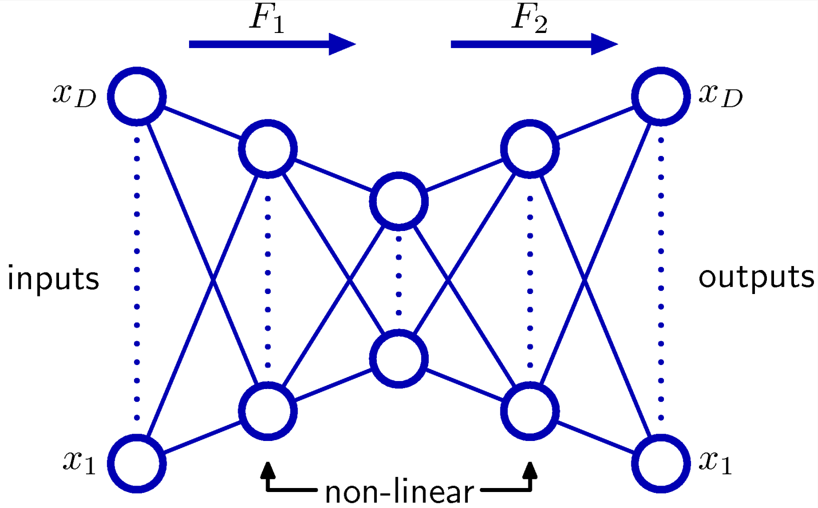
Visualization of weights

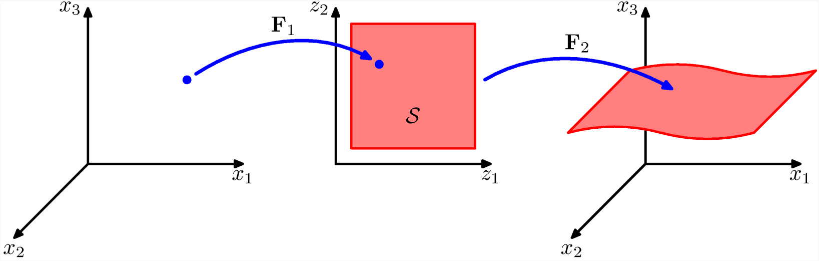


From `http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/`

The hidden neurons learned to detect edges of different orientations at different positions. Simple biological neurons ($V1$ layer) also respond to such edge-like inputs.

Multilayer Autoassociative Nets
Nonlinear PCA




x_3
x_1
x_2
$\mathbf{F}_1$
z_2
z_1
$\mathcal{S}$
$\mathbf{F}_2$
x_3
x_1
x_2