Autoencoders

Feed-forward neural network trained to reproduce its input to the output layer

Unsupervised learning: only use the input $X$ for learning

Encoder

$$h = f(x)$$
$$= \sigma(Wx + b)$$

Decoder

$$r = g(x)$$
$$= \sigma(W^*h + b)$$

$(Tied weights)$
Autoencoders

Loss function:

$$\text{Min } L(x, g(f(x)))$$

L is a loss function penalizing $g(f(x))$ for being dissimilar from $x$, e.g., mean squared error.

Train with backpropagation

When computing gradients with tied weights ($w^* = w^T$), $\nabla_W L(x, g(f(x)))$ is the sum of two gradients!

-- because $W$ is present in the encoder and in the decoder
Autoencoders

General structure:
- Encoder $f$: mapping $x$ to $h$
- Decoder $g$: mapping $h$ to $r$

Autoencoders may learn identity function precisely: $g(f(x)) = x$
⇒ Not useful!

Need to constrain complexity:
- By architectural constraint
- Penalty on internal representation
Autoencoders

Autoencoder types:

• Undercomplete Autoencoders
• Regularized Autoencoders
• Sparse Autoencoders
• Denoising Autoencoders
• Contractive Autoencoders
• ...

Penalty on internal representation (regularized autoencoders)

architectural constraint
Undercomplete Autoencoders

Constraint: Dimension of $h$ is smaller than $x$

$x \in \mathbb{R}^D, h \in \mathbb{R}^K$

Undercomplete autoencoders if $K < D$

Capture the most salient features
Undercomplete Autoencoders

Undercomplete autoencoders with:
✓ Decoder is linear transformation
✓ Loss $L$ is mean square error (MSE)

In this process, two tasks are accomplished:
1. Copy the input to output
2. Learn the principal subspace of training data as a **side-effect**
Undercomplete Autoencoders

If the encoder and decoder functions \((f, g)\) are nonlinear,

\[ \Rightarrow \text{A more powerful nonlinear generalization of PCA} \]

However,

Too large capacity of encoder and decoder

\[ \Rightarrow \text{can perform the copying task well, but fail to capture useful information on dataset} \]
Regularized Autoencoders

\[ x \in \mathbb{R}^D, h \in \mathbb{R}^K \]

What if \( K > D \) ? \quad => \quad \text{Overcomplete Autoencoders}

Regularized Autoencoders use a loss function that encourages the model to have some properties besides reproducing inputs:

- Sparsity representation (Sparse Autoencoders)
- Smallness of derivative of representation (Contractive Autoencoders)
- Robustness to noise or to missing inputs (Denoising Autoencoders)
Sparse Autoencoders

\[ L(x, g(f(x))) + \Omega(h) \]

- Loss for copying inputs
- Sparsity penalty
In general neural network, we are trying to find the **maximum likelihood**: $p(x|\theta)$

To do the maximum likelihood estimation (MLE), we often use the $\log(p(x|\theta))$ for simplification, from which we can get the loss function without regularization.

What about MAP (Maximum a posterior)?

$$p(\theta|x) \propto p(x|\theta) \ast p(\theta)$$

max $\log(p(\theta|x)) \Rightarrow$ max $\{ \log(p(x|\theta)) + \log(p(\theta)) \}$

Loss function  Regularization penalty
Sparse Autoencoders

\[
 \max \log(p(\theta|\mathbf{x})) \Rightarrow \max \{ \log(p(\mathbf{x}|\theta)) + \log(p(\theta)) \}
\]

What will happen if \( p(\theta) \) follows the **Gaussian Distribution**?

Consider the linear regression model, if

\[
 \omega \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda} \mathbf{I})
\]

\[
 p(w) = \frac{1}{\sqrt{|2\pi \frac{1}{\lambda} I|}} e^{-\frac{1}{2} w^T \frac{1}{\lambda} w}
\]

\[
 \Rightarrow \log p(w) = \log \left( \frac{1}{\sqrt{|2\pi \frac{1}{\lambda} I|}} \right) - \frac{\lambda}{2} w^T w \Rightarrow \text{L2 Norm}
\]

Gaussian Prior => L2 Norm

Similarly, Laplace Prior = > L1 Norm
Sparse Autoencoders

How to get the sparse penalty in sparse autoencoders?

Set the distribution over latent variable $h$

The joint distribution of $h$ and $x$ is given as:

$$p_{\text{model}}(x, h) = p_{\text{model}}(h)p_{\text{model}}(x|h)$$

$$\log p_{\text{model}}(x, h) = \log p_{\text{model}}(h) + \log p_{\text{model}}(x|h)$$
**Sparse Autoencoders**

\[ L(x, g(f(x))) + \Omega(h) \]

- Loss for copying inputs
- Sparsity penalty

Our target becomes:
Find a distribution of \( h \) which can have the characteristic of sparsity

Which distribution?

\[ \Rightarrow \text{Laplace distribution!} \]

\[
\log p_{model}(x, h) = \log p_{model}(h) + \log p_{model}(x|h)
\]

Sparse penalty
Sparse Autoencoders

Laplace distribution:

\[ p(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \]

\[ p_{\text{model}}(h_i) = \frac{\lambda}{2} e^{-\lambda|h_i|} \]

\[ -\log p_{\text{model}}(h) = \sum_i (\lambda|h_i|) - \log \frac{\lambda}{2} \]

L1 Norm \[ \Omega(h) \] Constant