Generative Models

Samuel Cheng

School of ECE
University of Oklahoma

Spring, 2018
(Slides credit to Goodfellow, Larochelle, Hinton)

Table of Contents

Review

- We talked about RNN previously. RNN can be treated as a kind of generative models. That is, able to generate samples from the model
- We will look into more generative models:
  - PixelCNN and PixelRNN
  - Generative adversarial networks (GANs)
  - Variational autoencoders

Supervised vs Unsupervised Learning

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc.

Fei-Fei Li & Justin Johnson & Serena Yeung   Lecture 13 - 4   May 18, 2017

Supervised vs Unsupervised Learning

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc.

→ Cat

Classification

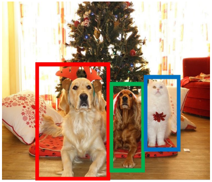
Supervised vs Unsupervised Learning

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc.



DOG, DOG, CAT

Object Detection

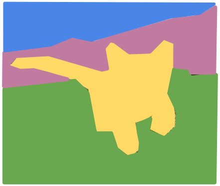
Supervised vs Unsupervised Learning

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc.



Semantic Segmentation

Fei-Fei Li & Justin Johnson & Serena Yeung Lecture 13 - 7 May 18, 2017

Supervised vs Unsupervised Learning

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc.

A cat sitting on a suitcase on the floor

Image captioning

Supervised vs Unsupervised Learning

Unsupervised Learning

Data: x
Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, feature learning, density estimation, etc.

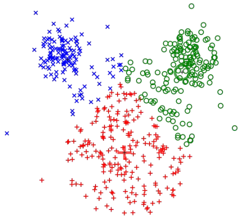
Supervised vs Unsupervised Learning

Unsupervised Learning

Data: x
Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, feature learning, density estimation, etc.



K-means clustering

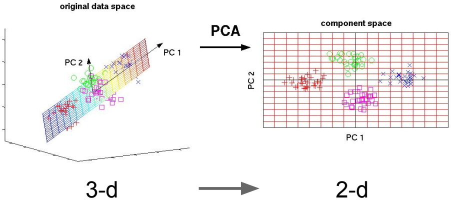
Supervised vs Unsupervised Learning

Unsupervised Learning

Data: x
Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, feature learning, density estimation, etc.



Principal Component Analysis
(Dimensionality reduction)

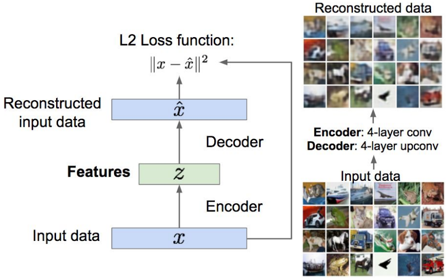
Supervised vs Unsupervised Learning

Unsupervised Learning

Data: x
Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, feature learning, density estimation, etc.



Autoencoders
(Feature learning)

Supervised vs Unsupervised Learning

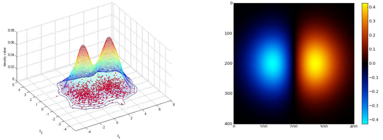
Unsupervised Learning

Data: x
Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, feature learning, density estimation, etc.

1-d density estimation



2-d density estimation

Supervised vs Unsupervised Learning

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc.

Unsupervised Learning

Data: x
Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, feature learning, density estimation, etc.

Fei-Fei Li & Justin Johnson & Serena Yeung   Lecture 13 - 14   May 18, 2017

Supervised vs Unsupervised Learning

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc.

Unsupervised Learning

Training data is cheap

Data: x
Just data, no labels!

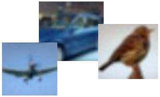
Holy grail: Solve unsupervised learning => understand structure of visual world

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, feature learning, density estimation, etc.

Generative Models

Given training data, generate new samples from same distribution



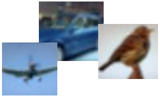
Training data ~ $p_{data}(x)$

Generated samples ~ $p_{model}(x)$

Want to learn $p_{model}(x)$ similar to $p_{data}(x)$

Generative Models

Given training data, generate new samples from same distribution



Training data ~ $p_{data}(x)$

Generated samples ~ $p_{model}(x)$

Want to learn $p_{model}(x)$ similar to $p_{data}(x)$

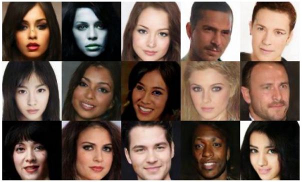
Addresses density estimation, a core problem in unsupervised learning

Several flavors:

- Explicit density estimation: explicitly define and solve for $p_{model}(x)$
- Implicit density estimation: learn model that can sample from $p_{model}(x)$ w/o explicitly defining it

Why Generative Models?

- Realistic samples for artwork, super-resolution, colorization, etc.



- Generative models of time-series data can be used for simulation and planning (reinforcement learning applications!)
- Training generative models can also enable inference of latent representations that can be useful as general features

Discriminative models vs generative models

- Discriminative models try to discriminate if one input is different from another. But it is not possible to generate samples from the models. Many classifiers are based on discriminative models, for example, support vector machines
- Generative models on the other hand can generate simulated data, for example, PixelCNN
- Many older machine learning problems are classification problems. Discriminative models provide a more direct solution and thus were more attractive
- Generative models have gained quite some attentions in recent years
  - Generate labeled simulation data for semi-supervised learning
  - Simulate data for planning and reinforcement learning

Discriminative models vs generative models

- Discriminative models try to discriminate if one input is different from another. But it is not possible to generate samples from the models. Many classifiers are based on discriminative models, for example, support vector machines
- Generative models on the other hand can generate simulated data, for example, PixelCNN
- Many older machine learning problems are classification problems. Discriminative models provide a more direct solution and thus were more attractive
- Generative models have gained quite some attentions in recent years
  - Generate labeled simulation data for semi-supervised learning
  - Simulate data for planning and reinforcement learning

Discriminative models vs generative models

- Discriminative models try to discriminate if one input is different from another. But it is not possible to generate samples from the models. Many classifiers are based on discriminative models, for example, support vector machines
- Generative models on the other hand can generate simulated data, for example, PixelCNN
- Many older machine learning problems are classification problems. Discriminative models provide a more direct solution and thus were more attractive
- Generative models have gained quite some attentions in recent years
  - Generate labeled simulation data for semi-supervised learning
  - Simulate data for planning and reinforcement learning

Discriminative models vs generative models

- Discriminative models try to discriminate if one input is different from another. But it is not possible to generate samples from the models. Many classifiers are based on discriminative models, for example, support vector machines
- Generative models on the other hand can generate simulated data, for example, PixelCNN
- Many older machine learning problems are classification problems. Discriminative models provide a more direct solution and thus were more attractive
- Generative models have gained quite some attentions in recent years
  - Generate labeled simulation data for semi-supervised learning
  - Simulate data for planning and reinforcement learning

Taxonomy of Generative Models

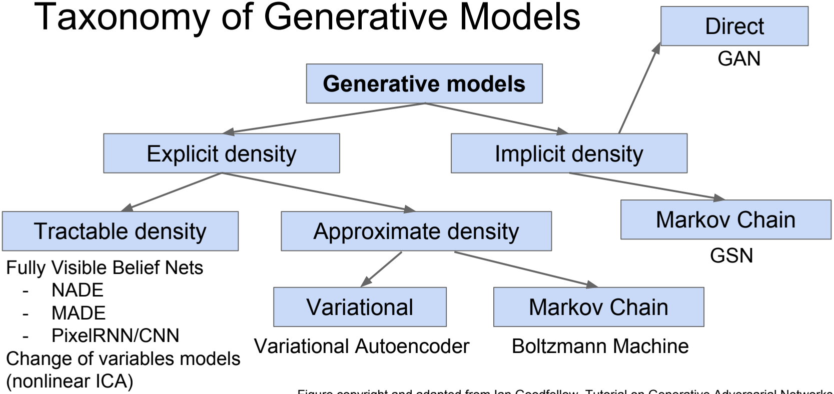
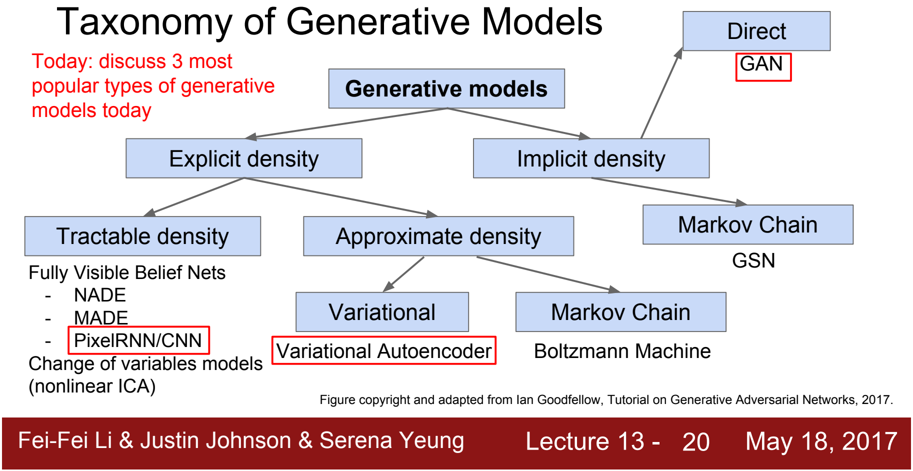


Figure copyright and adapted from Ian Goodfellow, Tutorial on Generative Adversarial Networks, 2017.

Fei-Fei Li & Justin Johnson & Serena Yeung Lecture 13 - 19 May 18, 2017


Taxonomy of Generative Models
Today: discuss 3 most popular types of generative models today
Generative models
Explicit density
Implicit density
Direct
GAN
Markov Chain
GSN
Tractable density
Approximate density
Fully Visible Belief Nets
- NADE
- MADE
- PixelRNN/CNN
Change of variables models (nonlinear ICA)
Variational
Variational Autoencoder
Markov Chain
Boltzmann Machine
Figure copyright and adapted from Ian Goodfellow, Tutorial on Generative Adversarial Networks, 2017.
Fei-Fei Li & Justin Johnson & Serena Yeung
Lecture 13 - 20
May 18, 2017

PixelRNN and PixelCNN

Fully visible belief network

Explicit density model

Use chain rule to decompose likelihood of an image x into product of 1-d distributions:

$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, ..., x_{i-1})$$

Likelihood of image x

Probability of i'th pixel value given all previous pixels

Then maximize likelihood of training data

Fully visible belief network

Explicit density model

Use chain rule to decompose likelihood of an image x into product of 1-d distributions:

$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, ..., x_{i-1})$$

Likelihood of image x

Probability of i'th pixel value given all previous pixels

Then maximize likelihood of training data

Complex distribution over pixel values => Express using a neural network!

Fei-Fei Li & Justin Johnson & Serena Yeung Lecture 13 - 23 May 18, 2017

Fully visible belief network

Explicit density model

Use chain rule to decompose likelihood of an image x into product of 1-d distributions:

$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, ..., x_{i-1})$$

Likelihood of image x

Probability of i'th pixel value given all previous pixels

Will need to define ordering of "previous pixels"

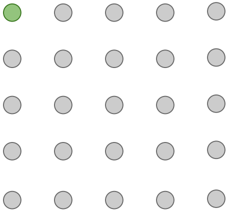
Complex distribution over pixel values => Express using a neural network!

Then maximize likelihood of training data

PixelRNN *[van der Oord et al. 2016]*

Generate image pixels starting from corner

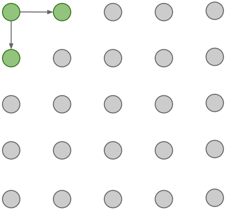
Dependency on previous pixels modeled using an RNN (LSTM)



PixelRNN *[van der Oord et al. 2016]*

Generate image pixels starting from corner

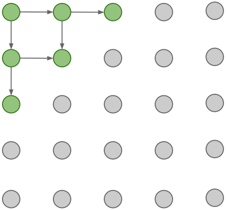
Dependency on previous pixels modeled using an RNN (LSTM)



PixelRNN *[van der Oord et al. 2016]*

Generate image pixels starting from corner

Dependency on previous pixels modeled using an RNN (LSTM)

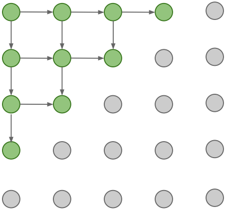


PixelRNN *[van der Oord et al. 2016]*

Generate image pixels starting from corner

Dependency on previous pixels modeled using an RNN (LSTM)

Drawback: sequential generation is slow!

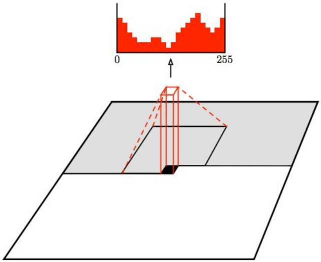


Fei-Fei Li & Justin Johnson & Serena Yeung Lecture 13 - 28 May 18, 2017

PixelCNN *[van der Oord et al. 2016]*

Still generate image pixels starting from corner

Dependency on previous pixels now modeled using a CNN over context region



PixelCNN *[van der Oord et al. 2016]*

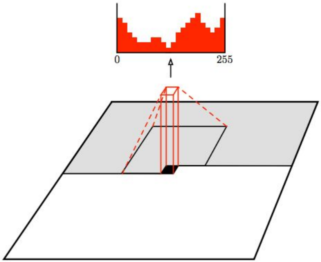
Still generate image pixels starting from corner

Dependency on previous pixels now modeled using a CNN over context region

Training: maximize likelihood of training images

$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, ..., x_{i-1})$$

Softmax loss at each pixel



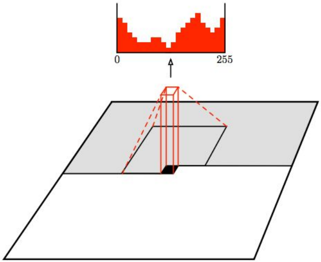
PixelCNN *[van der Oord et al. 2016]*

Still generate image pixels starting from corner

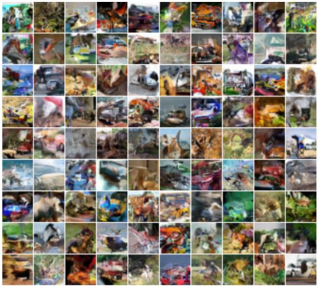
Dependency on previous pixels now modeled using a CNN over context region

Training is faster than PixelRNN
(can parallelize convolutions since context region values known from training images)

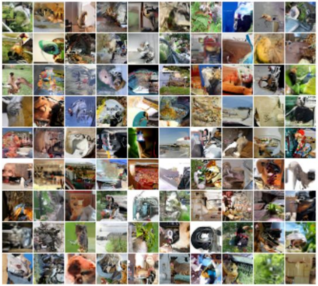
Generation must still proceed sequentially
=> still slow



Generation Samples



32x32 CIFAR-10



32x32 ImageNet

PixelRNN and PixelCNN

Pros:
- Can explicitly compute likelihood p(x)
- Explicit likelihood of training data gives good evaluation metric
- Good samples

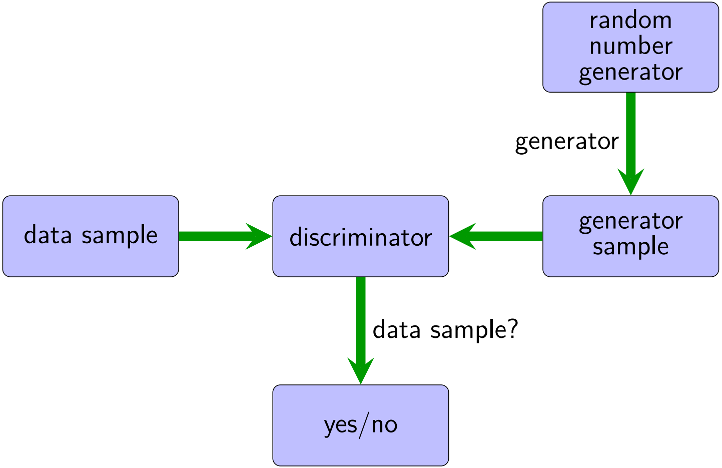
Con:
- Sequential generation => slow

Improving PixelCNN performance
- Gated convolutional layers
- Short-cut connections
- Discretized logistic loss
- Multi-scale
- Training tricks
- Etc…

See
- Van der Oord et al. NIPS 2016
- Salimans et al. 2017 (PixelCNN++)

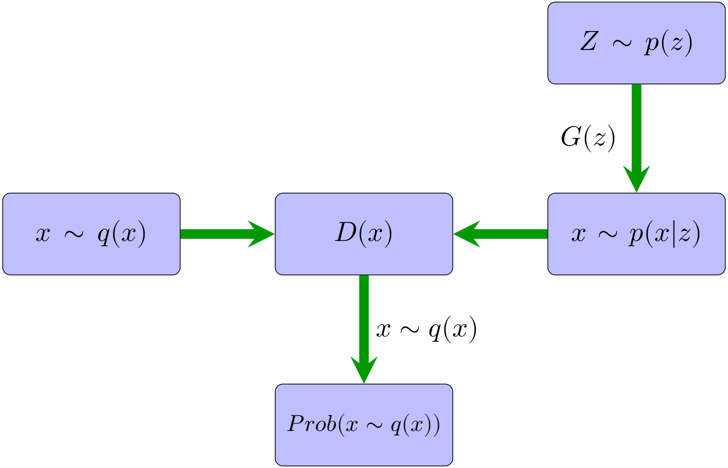
Generative adversarial networks (GANs)

Goodfellow *et al.* 2014



Generative adversarial networks (GANs)

Goodfellow *et al.* 2014



Minimax game of a GAN

- Probability of model data: $p_{model}(x) = \int_z p(z)p(x|z)dz$
- Probability of true data: $p_{data}(x) = q(x)$
- Discriminator wants to catch fake data

$$J^{(D)} = -E_{x \sim p_{data}} \log D(x) - E_z \log(1 - D(G(z)))$$
$$= -E_{x \sim p_{data}} \log D(x) - E_{x \sim p_{model}} \log(1 - D(x))$$

  - N.B. $J^{(D)}$ is just cross-entropy loss for correct classification
- Generator wants to fool the discriminator: $J^{(G)} = -J^{(D)}$
  - Since first term does not depend on $G(\cdot)$, we can simplify $J^{(G)}$ to

$$J^{(G)} = -E_z \log(1 - D(G(z)))$$

Minimax game of a GAN

- Probability of model data: $p_{model}(x) = \int_z p(z)p(x|z)dz$
- Probability of true data: $p_{data}(x) = q(x)$
- Discriminator wants to catch fake data

$$
\begin{aligned}
J^{(D)} &= -E_{x\sim p_{data}} \log D(x) - E_z \log(1 - D(G(z))) \\
&= -E_{x\sim p_{data}} \log D(x) - E_{x\sim p_{model}} \log(1 - D(x))
\end{aligned}
$$

  - N.B. $J^{(D)}$ is just cross-entropy loss for correct classification
- Generator wants to fool the discriminator: $J^{(G)} = -J^{(D)}$
  - Since first term does not depend on $G(\cdot)$, we can simplify $J^{(G)}$ to

$$J^{(G)} = -E_z \log(1 - D(G(z)))$$

Minimax game of a GAN

- Probability of model data: $p_{model}(x) = \int_z p(z)p(x|z)dz$
- Probability of true data: $p_{data}(x) = q(x)$
- Discriminator wants to catch fake data

$$\begin{aligned} J^{(D)} &= -E_{x\sim p_{data}} \log D(x) - E_z \log(1 - D(G(z))) \\ &= -E_{x\sim p_{data}} \log D(x) - E_{x\sim p_{model}} \log(1 - D(x)) \end{aligned}$$

  - N.B. $J^{(D)}$ is just cross-entropy loss for correct classification
- Generator wants to fool the discriminator: $J^{(G)} = -J^{(D)}$
  - Since first term does not depend on $G(\cdot)$, we can simplify $J^{(G)}$ to

$$J^{(G)} = -E_z \log(1 - D(G(z)))$$

Minimax game of a GAN

- Probability of model data: $p_{model}(x) = \int_z p(z)p(x|z)dz$
- Probability of true data: $p_{data}(x) = q(x)$
- Discriminator wants to catch fake data

$$
\begin{aligned}
J^{(D)} &= -E_{x\sim p_{data}} \log D(x) - E_z \log(1 - D(G(z))) \\
&= -E_{x\sim p_{data}} \log D(x) - E_{x\sim p_{model}} \log(1 - D(x))
\end{aligned}
$$

  - N.B. $J^{(D)}$ is just cross-entropy loss for correct classification
- Generator wants to fool the discriminator: $J^{(G)} = -J^{(D)}$
  - Since first term does not depend on $G(\cdot)$, we can simplify $J^{(G)}$ to

$$
J^{(G)} = -E_z \log(1 - D(G(z)))
$$

Nash equilibrium

- By game theory, Nash equilibriums exist
- One equilibrium is $G(\cdot)$ generate indifferentiable sample as the true data and $D(\cdot)$ will just make choices randomly (output 1 with probability 0.5)
  - This is the equilibrium that we are interested in

Optimal discriminator $D^*(x)$

By calculus of variations, for any $\Delta(x)$,

$$\left.\frac{\partial J^{(D)}(D^*(X)+\lambda\Delta(x))}{\partial\lambda}\right|_{\lambda=0}=0$$

$$\Rightarrow -\left.\frac{\partial E_{x\sim p_{data}}\log(D^*(x)+\lambda\Delta(x))}{\partial\lambda}-\frac{\partial E_{x\sim p_{model}}\log(1-D^*(x)-\lambda\Delta(x))}{\partial\lambda}\right|_{\lambda=0}=0$$

$$\Rightarrow -E_{x\sim p_{data}}\left[\frac{\Delta(x)}{D^*(x)+\lambda\Delta(x)}\right]+E_{x\sim p_{model}}\left.\left[\frac{\Delta(x)}{1-D^*(x)-\lambda\Delta(x)}\right]\right|_{\lambda=0}=0$$

$$\Rightarrow \int_x\left[\frac{p_{data}(x)}{D^*(x)}-\frac{p_{model}(x)}{1-D^*(x)}\right]\Delta(x)dx=0$$

$$\Rightarrow D^*(x)=\frac{p_{data}(x)}{p_{data}(x)+p_{model}(x)}$$

Optimal discriminator $D^*(x)$

By calculus of variations, for any $\Delta(x)$,

$$\left.\frac{\partial J^{(D)}(D^*(X)+\lambda\Delta(x))}{\partial\lambda}\right|_{\lambda=0}=0$$

$$\Rightarrow -\left.\frac{\partial E_{x\sim p_{data}}\log(D^*(x)+\lambda\Delta(x))}{\partial\lambda}-\frac{\partial E_{x\sim p_{model}}\log(1-D^*(x)-\lambda\Delta(x))}{\partial\lambda}\right|_{\lambda=0}=0$$

$$\Rightarrow -E_{x\sim p_{data}}\left[\frac{\Delta(x)}{D^*(x)+\lambda\Delta(x)}\right]+E_{x\sim p_{model}}\left.\left[\frac{\Delta(x)}{1-D^*(x)-\lambda\Delta(x)}\right]\right|_{\lambda=0}=0$$

$$\Rightarrow \int_x\left[\frac{p_{data}(x)}{D^*(x)}-\frac{p_{model}(x)}{1-D^*(x)}\right]\Delta(x)dx=0$$

$$\Rightarrow D^*(x)=\frac{p_{data}(x)}{p_{data}(x)+p_{model}(x)}$$

Optimal discriminator $D^*(x)$

By calculus of variations, for any $\Delta(x)$,

$$\left.\frac{\partial J^{(D)}(D^*(X)+\lambda\Delta(x))}{\partial\lambda}\right|_{\lambda=0}=0$$

$$\Rightarrow -\left.\frac{\partial E_{x\sim p_{data}}\log(D^*(x)+\lambda\Delta(x))}{\partial\lambda}-\frac{\partial E_{x\sim p_{model}}\log(1-D^*(x)-\lambda\Delta(x))}{\partial\lambda}\right|_{\lambda=0}=0$$

$$\Rightarrow -E_{x\sim p_{data}}\left[\frac{\Delta(x)}{D^*(x)+\lambda\Delta(x)}\right]+E_{x\sim p_{model}}\left.\left[\frac{\Delta(x)}{1-D^*(x)-\lambda\Delta(x)}\right]\right|_{\lambda=0}=0$$

$$\Rightarrow \int_x\left[\frac{p_{data}(x)}{D^*(x)}-\frac{p_{model}(x)}{1-D^*(x)}\right]\Delta(x)dx=0$$

$$\Rightarrow D^*(x)=\frac{p_{data}(x)}{p_{data}(x)+p_{model}(x)}$$

Optimal discriminator $D^*(x)$

By calculus of variations, for any $\Delta(x)$,

$$\left.\frac{\partial J^{(D)}(D^*(X)+\lambda\Delta(x))}{\partial\lambda}\right|_{\lambda=0}=0$$

$$\Rightarrow -\frac{\partial E_{x\sim p_{data}}\log(D^*(x)+\lambda\Delta(x))}{\partial\lambda}-\left.\frac{\partial E_{x\sim p_{model}}\log(1-D^*(x)-\lambda\Delta(x))}{\partial\lambda}\right|_{\lambda=0}=0$$

$$\Rightarrow -E_{x\sim p_{data}}\left[\frac{\Delta(x)}{D^*(x)+\lambda\Delta(x)}\right]+E_{x\sim p_{model}}\left.\left[\frac{\Delta(x)}{1-D^*(x)-\lambda\Delta(x)}\right]\right|_{\lambda=0}=0$$

$$\Rightarrow \int_x\left[\frac{p_{data}(x)}{D^*(x)}-\frac{p_{model}(x)}{1-D^*(x)}\right]\Delta(x)dx=0$$

$$\Rightarrow D^*(x)=\frac{p_{data}(x)}{p_{data}(x)+p_{model}(x)}$$

Optimal discriminator $D^*(x)$

By calculus of variations, for any $\Delta(x)$,

$$\left.\frac{\partial J^{(D)}(D^*(X)+\lambda\Delta(x))}{\partial\lambda}\right|_{\lambda=0}=0$$

$$\Rightarrow -\left.\frac{\partial E_{x\sim p_{data}}\log(D^*(x)+\lambda\Delta(x))}{\partial\lambda}-\frac{\partial E_{x\sim p_{model}}\log(1-D^*(x)-\lambda\Delta(x))}{\partial\lambda}\right|_{\lambda=0}=0$$

$$\Rightarrow -E_{x\sim p_{data}}\left[\frac{\Delta(x)}{D^*(x)+\lambda\Delta(x)}\right]+E_{x\sim p_{model}}\left.\left[\frac{\Delta(x)}{1-D^*(x)-\lambda\Delta(x)}\right]\right|_{\lambda=0}=0$$

$$\Rightarrow \int_x\left[\frac{p_{data}(x)}{D^*(x)}-\frac{p_{model}(x)}{1-D^*(x)}\right]\Delta(x)dx=0$$

$$\Rightarrow D^*(x)=\frac{p_{data}(x)}{p_{data}(x)+p_{model}(x)}$$

Non-saturating cost function

- The discriminator cost function $J^{(D)} = -E_{x \sim p_{data}} \log D(x) - E_z \log(1 - D(G(z)))$ is a very reasonable choice and usually will not be modified
- On the other hand, we have more freedom on choosing the generator cost
  - $E_z \log(1 - D(G(z)))$ is the intuitive choice for $J^{(G)}$ but it has a small gradient when $D(G(z))$ is small for all z
    - That is, generator is not able to fool the discriminator
    - Reasonable when we just started to train the generator
  - Instead, it is better to have $J^{(G)} = -E_z \log D(G(z))$
    - $-\log D(G(z)) \approx 0$ when $D(G(z)) \approx 1$: ignore samples that successfully fool the discriminator
    - $-\log D(G(z)) \gg 0$ when $D(G(z)) \approx 0$: emphasize samples that cannot fool the discriminator
    - When $D(G(z)) \approx 1$ for all z, we may need to switch back to the original cost function. But better yet, we should better train the discriminator

Non-saturating cost function

- The discriminator cost function $J^{(D)} = -E_{x\sim p_{data}} \log D(x) - E_z \log(1 - D(G(z)))$ is a very reasonable choice and usually will not be modified
- On the other hand, we have more freedom on choosing the generator cost
  - $E_z \log(1 - D(G(z)))$ is the intuitive choice for $J^{(G)}$ but it has a small gradient when $D(G(z))$ is small for all z
    - That is, generator is not able to fool the discriminator
    - Reasonable when we just started to train the generator
  - Instead, it is better to have $J^{(G)} = -E_z \log D(G(z))$
    - $-\log D(G(z)) \approx 0$ when $D(G(z)) \approx 1$: ignore samples that successfully fool the discriminator
    - $-\log D(G(z)) \gg 0$ when $D(G(z)) \approx 0$: emphasize samples that cannot fool the discriminator
    - When $D(G(z)) \approx 1$ for all z, we may need to switch back to the original cost function. But better yet, we should better train the discriminator

Non-saturating cost function

- The discriminator cost function $J^{(D)} = -E_{x\sim p_{data}} \log D(x) - E_z \log(1 - D(G(z)))$ is a very reasonable choice and usually will not be modified
- On the other hand, we have more freedom on choosing the generator cost
    - $E_z \log(1 - D(G(z)))$ is the intuitive choice for $J^{(G)}$ but it has a small gradient when $D(G(z))$ is small for all z
        - That is, generator is not able to fool the discriminator
        - Reasonable when we just started to train the generator
    - Instead, it is better to have $J^{(G)} = -E_z \log D(G(z))$
        - $-\log D(G(z)) \approx 0$ when $D(G(z)) \approx 1$: ignore samples that successfully fool the discriminator
        - $-\log D(G(z)) \gg 0$ when $D(G(z)) \approx 0$: emphasize samples that cannot fool the discriminator
        - When $D(G(z)) \approx 1$ for all z, we may need to switch back to the original cost function. But better yet, we should better train the discriminator

Non-saturating cost function

- The discriminator cost function $J^{(D)} = -E_{x \sim p_{data}} \log D(x) - E_z \log(1 - D(G(z)))$ is a very reasonable choice and usually will not be modified
- On the other hand, we have more freedom on choosing the generator cost
    - $E_z \log(1 - D(G(z)))$ is the intuitive choice for $J^{(G)}$ but it has a small gradient when $D(G(z))$ is small for all z
        - That is, generator is not able to fool the discriminator
        - Reasonable when we just started to train the generator
    - Instead, it is better to have $J^{(G)} = -E_z \log D(G(z))$
        - $-\log D(G(z)) \approx 0$ when $D(G(z)) \approx 1$: ignore samples that successfully fool the discriminator
        - $-\log D(G(z)) \gg 0$ when $D(G(z)) \approx 0$: emphasize samples that cannot fool the discriminator
        - When $D(G(z)) \approx 1$ for all z, we may need to switch back to the original cost function. But better yet, we should better train the discriminator

Training GANs: Two-player game

Ian Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

Minimax objective function:

$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

Alternate between:

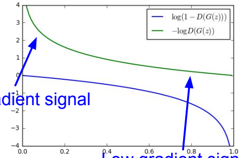
1. **Gradient ascent** on discriminator

$$\max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

2. **Instead: Gradient ascent** on generator, different objective

$$\max_{\theta_g} \mathbb{E}_{z \sim p(z)} \log(D_{\theta_d}(G_{\theta_g}(z)))$$

Instead of minimizing likelihood of discriminator being correct, now maximize likelihood of discriminator being wrong.
Same objective of fooling discriminator, but now higher gradient signal for bad samples => works much better! Standard in practice.



Some refinements

Training GAN is equivalent of finding the Nash equilibrium of a two-player non-cooperative game, which itself is a very hard problem. We will mention here a couple refinements to help find a better solution. You probably would like to check out Salimans' 16 also

- One-sided label smoothing
- Fixing batch-norm
- Mini-batch features
- Unrolled GAN

One-sided label smoothing
Salimans *et al.* 2016

- Default discriminator cost can also be written as

$$\begin{aligned}&\text{cross\_entropy}(\text{"}1\text{"},\text{discriminator}(\text{data}))\\+&\text{cross\_entropy}(\text{"}0\text{"},\text{discriminator}(\text{samples}))\end{aligned}$$

- Experiment shows that one-sided label smoothed cost enhance system stability

$$\begin{aligned}&\text{cross\_entropy}(\text{"}0.9\text{"},\text{discriminator}(\text{data}))\\+&\text{cross\_entropy}(\text{"}0\text{"},\text{discriminator}(\text{samples}))\end{aligned}$$

  - Essentially prevent extrapolating effect from extreme samples
  - Generally does not reduce classification accuracy, only confidence

One-sided label smoothing
Salimans *et al.* 2016

- Default discriminator cost can also be written as

$$\begin{aligned} &\text{cross\_entropy}(\text{"}1\text{"},\text{discriminator}(\text{data})) \\ +&\text{cross\_entropy}(\text{"}0\text{"}, \text{discriminator}(\text{samples})) \end{aligned}$$

- Experiment shows that one-sided label smoothed cost enhance system stability

$$\begin{aligned} &\text{cross\_entropy}(\text{"}0.9\text{"},\text{discriminator}(\text{data})) \\ +&\text{cross\_entropy}(\text{"}0\text{"}, \text{discriminator}(\text{samples})) \end{aligned}$$

  - Essentially prevent extrapolating effect from extreme samples
  - Generally does not reduce classification accuracy, only confidence

One-sided label smoothing
Salimans *et al.* 2016

- Default discriminator cost can also be written as

$$\begin{aligned} &\text{cross\_entropy}(\text{"}1\text{"}, \text{discriminator}(\text{data})) \\ +&\text{cross\_entropy}(\text{"}0\text{"}, \text{discriminator}(\text{samples})) \end{aligned}$$

- Experiment shows that one-sided label smoothed cost enhance system stability

$$\begin{aligned} &\text{cross\_entropy}(\text{"}0.9\text{"}, \text{discriminator}(\text{data})) \\ +&\text{cross\_entropy}(\text{"}0\text{"}, \text{discriminator}(\text{samples})) \end{aligned}$$

  - Essentially prevent extrapolating effect from extreme samples
  - Generally does not reduce classification accuracy, only confidence

One-sided label smoothing
Salimans *et al.* 2016

- It is important not to smooth the negative labels though, i.e., say

$$\text{cross\_entropy}(1-\alpha,\text{discriminator}(\text{data})) + \text{cross\_entropy}(\beta,\text{discriminator}(\text{samples}))$$

with $\beta > 0$

- Just follow the same derivation as before, we can get the optimum $D(x)$ as

$$D^*(x) = \frac{(1-\alpha)p_{data}(x) + \beta p_{model}(x)}{p_{data}(x) + p_{model}(x)}$$

- $\beta > 0$ tends to give undesirable bias of the discriminator to data generated by the model

Replacing positive classification targets with α and negative targets with β, the optimal discriminator becomes $D(\boldsymbol{x}) = \frac{\alpha p_{\text{data}}(\boldsymbol{x}) + \beta p_{\text{model}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_{\text{model}}(\boldsymbol{x})}$. The presence of p_{model} in the numerator is problematic because, in areas where p_{data} is approximately zero and p_{model} is large, erroneous samples from p_{model} have no incentive to move nearer to the data. We therefore smooth *only* the positive labels to α, leaving negative labels set to 0.

One-sided label smoothing
Salimans *et al.* 2016

- It is important not to smooth the negative labels though, i.e., say

$$\text{cross\_entropy}(1-\alpha, \text{discriminator}(\text{data}))$$
$$+\text{cross\_entropy}(\beta, \text{discriminator}(\text{samples}))$$

with $\beta > 0$

- Just follow the same derivation as before, we can get the optimum $D(x)$ as

$$D^*(x) = \frac{(1-\alpha)p_{data}(x) + \beta p_{model}(x)}{p_{data}(x) + p_{model}(x)}$$

- $\beta > 0$ tends to give undesirable bias of the discriminator to data generated by the model

Replacing positive classification targets with α and negative targets with β, the optimal discriminator becomes $D(\boldsymbol{x}) = \frac{\alpha p_{\text{data}}(\boldsymbol{x}) + \beta p_{\text{model}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_{\text{model}}(\boldsymbol{x})}$. The presence of p_{model} in the numerator is problematic because, in areas where p_{data} is approximately zero and p_{model} is large, erroneous samples from p_{model} have no incentive to move nearer to the data. We therefore smooth *only* the positive labels to α, leaving negative labels set to 0.

One-sided label smoothing

Salimans *et al.* 2016

- It is important not to smooth the negative labels though, i.e., say

$$\text{cross\_entropy}(1-\alpha, \text{discriminator}(\text{data})) + \text{cross\_entropy}(\beta, \text{discriminator}(\text{samples}))$$

  with $\beta > 0$
- Just follow the same derivation as before, we can get the optimum $D(x)$ as

$$D^*(x) = \frac{(1-\alpha)p_{data}(x) + \beta p_{model}(x)}{p_{data}(x) + p_{model}(x)}$$

- $\beta > 0$ tends to give undesirable bias of the discriminator to data generated by the model

Replacing positive classification targets with α and negative targets with β, the optimal discriminator becomes $D(\boldsymbol{x}) = \frac{\alpha p_{\text{data}}(\boldsymbol{x}) + \beta p_{\text{model}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_{\text{model}}(\boldsymbol{x})}$. The presence of p_{model} in the numerator is problematic because, in areas where p_{data} is approximately zero and p_{model} is large, erroneous samples from p_{model} have no incentive to move nearer to the data. We therefore smooth *only* the positive labels to α, leaving negative labels set to 0.

Issue on batch normalization
Goodfellow 2016

Batch normalization is preferred and highly recommended. But it can cause strong intra-batch correlation

Fixing batch norm

- Reference batch norm: one possible approach is keep one reference batch and always normalized based on that batch. That is, always subtract mean from that of the reference batch and adjust variance to that of the reference batch
  - Can easily overfit to the particular reference batch
- Virtual batch norm: a partial solution by combining the reference batch norm and conventional batch norm. Fix a reference batch, but every time inputs are normalize to the net mean and variance of the virtual batch containing both inputs and all elements of the reference batch

Fixing batch norm

- Reference batch norm: one possible approach is keep one reference batch and always normalized based on that batch. That is, always subtract mean from that of the reference batch and adjust variance to that of the reference batch
  - Can easily overfit to the particular reference batch
- Virtual batch norm: a partial solution by combining the reference batch norm and conventional batch norm. Fix a reference batch, but every time inputs are normalize to the net mean and variance of the virtual batch containing both inputs and all elements of the reference batch

Balancing G and D

- Usually it is more preferable to have a bigger and deeper D
- Some researchers also run more D steps than G steps. The results are mixed though
- Do not try to limit D from being "too smart"
  - The original theoretical justification is that D is supposed to be perfect

Balancing G and D

- Usually it is more preferable to have a bigger and deeper D
- Some researchers also run more D steps than G steps. The results are mixed though
- Do not try to limit D from being "too smart"
  - The original theoretical justification is that D is supposed to be perfect

Balancing G and D

- Usually it is more preferable to have a bigger and deeper D
- Some researchers also run more D steps than G steps. The results are mixed though
- Do not try to limit D from being "too smart"
    - The original theoretical justification is that D is supposed to be perfect

Mode collapse

Metz *et al.* 2016

Below demonstrates why D should be smart.

- Basically the minmax and the minmax problem is not the same and can lead to drastically different solutions

$$\min_G \max_D V(G, D) \neq \max_D \min_G V(G, D)$$

- D in the inner loop: converge to the correct distribution
- G in the inner loop: place all mass on most likely point

Mode collapse
Metz *et al.* 2016

Below demonstrates why D should be smart.

- Basically the minmax and the minmax problem is not the same and can lead to drastically different solutions

$$\min_G \max_D V(G, D) \neq \max_D \min_G V(G, D)$$

- D in the inner loop: converge to the correct distribution
- G in the inner loop: place all mass on most likely point

Minibatch features
Salimans *et al.* 2016

- Mode collapse can lead to low diversity of generated data
- One attempt to mitigate this problem is to introduce the so-called minibatch features
  - Basically classify each example by comparing the features to other members in the minibatch
  - Reject a sample if the feature to close to existing ones

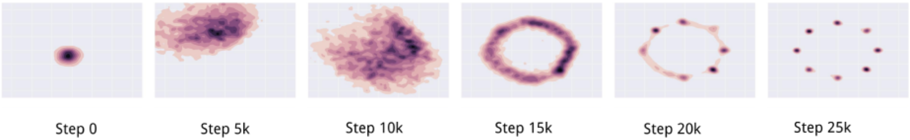
Minibatch features
Salimans *et al.* 2016

- Mode collapse can lead to low diversity of generated data
- One attempt to mitigate this problem is to introduce the so-called minibatch features
  - Basically classify each example by comparing the features to other members in the minibatch
  - Reject a sample if the feature to close to existing ones

Unrolled Gans

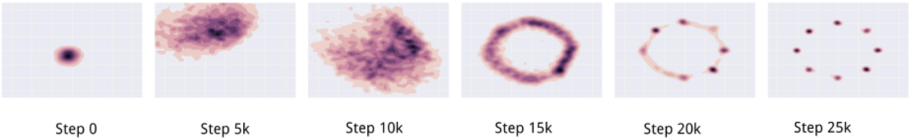
Metz *et al.* 2016

- A more direct approach was proposed by Google brain
- Trying to ensure that the generated sample is a solution of the minmax rather than the maxmin problem
- Have the generator to unroll k future steps and predict what discriminator will think of the current sample
  - Since generator is the one who unrolls, generator is in the outer loop and discriminator is in the inner loop
  - We ensure that we have solution approximating a minmax rather than maxmin problem



Unrolled Gans
Metz *et al.* 2016

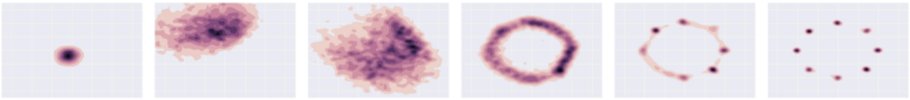
- A more direct approach was proposed by Google brain
- Trying to ensure that the generated sample is a solution of the minmax rather than the maxmin problem
- Have the generator to unroll k future steps and predict what discriminator will think of the current sample
  - Since generator is the one who unrolls, generator is in the outer loop and discriminator is in the inner loop
  - We ensure that we have solution approximating a minmax rather than maxmin problem



Unrolled Gans

Metz *et al.* 2016

- A more direct approach was proposed by Google brain
- Trying to ensure that the generated sample is a solution of the minmax rather than the maxmin problem
- Have the generator to unroll k future steps and predict what discriminator will think of the current sample
  - Since generator is the one who unrolls, generator is in the outer loop and discriminator is in the inner loop
  - We ensure that we have solution approximating a minmax rather than maxmin problem



Step 0 Step 5k Step 10k Step 15k Step 20k Step 25k

Deep convolutional GAN (DCGAN)

Generative Adversarial Nets: Convolutional Architectures

Generator is an upsampling network with fractionally-strided convolutions
Discriminator is a convolutional network

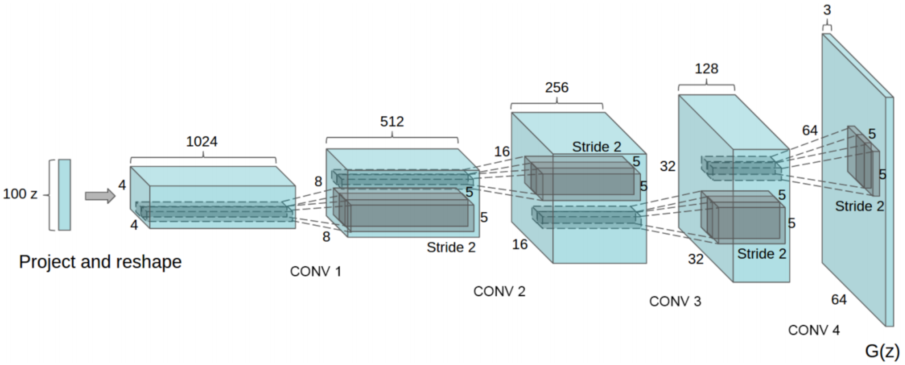
Architecture guidelines for stable Deep Convolutional GANs

- Replace any pooling layers with strided convolutions (discriminator) and fractional-strided convolutions (generator).
- Use batchnorm in both the generator and the discriminator.
- Remove fully connected hidden layers for deeper architectures.
- Use ReLU activation in generator for all layers except for the output, which uses Tanh.
- Use LeakyReLU activation in the discriminator for all layers.

Radford et al, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", ICLR 2016

Deep convolutional GAN (DCGAN)

Radford *et al.* 2016

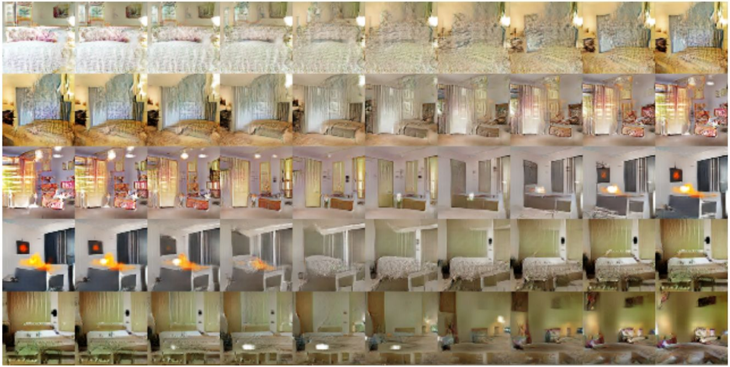


Generated bedroom after 5 epochs (LSUN dataset)

Radford *et al.* 2016

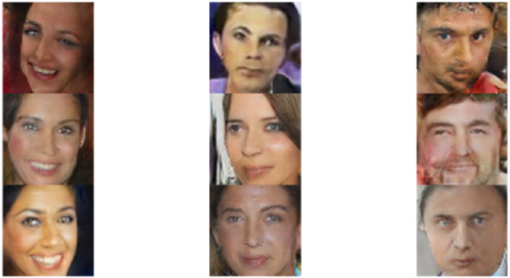
Generative Adversarial Nets: Convolutional Architectures

Interpolating between random points in laten space

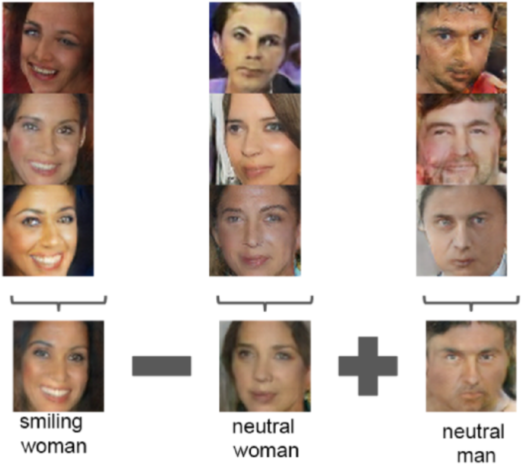


Radford et al, ICLR 2016

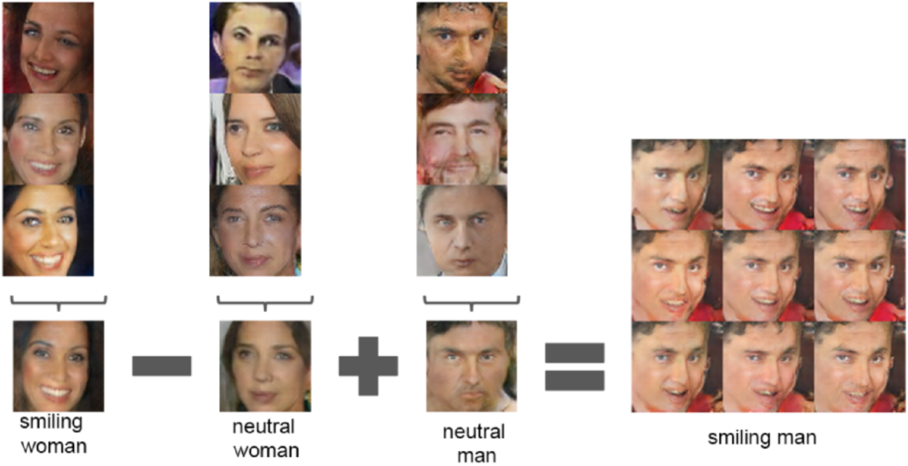
Vector arithmetics
Radford *et al.* 2016



Vector arithmetics
Radford *et al.* 2016



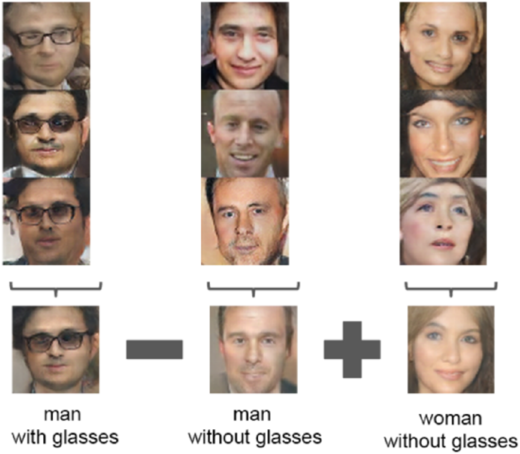
Vector arithmetics
Radford *et al.* 2016



Vector arithmetics

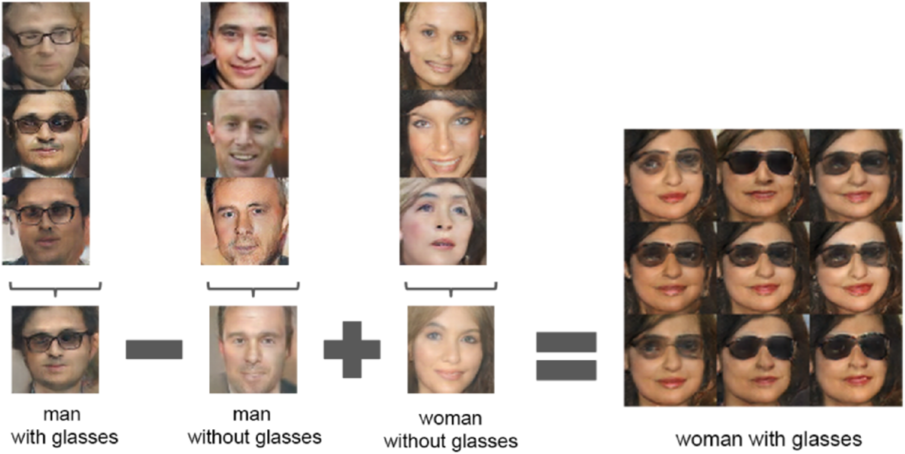
Radford *et al.* 2016

Vector arithmetics
Radford *et al.* 2016



Vector arithmetics

Radford *et al.* 2016

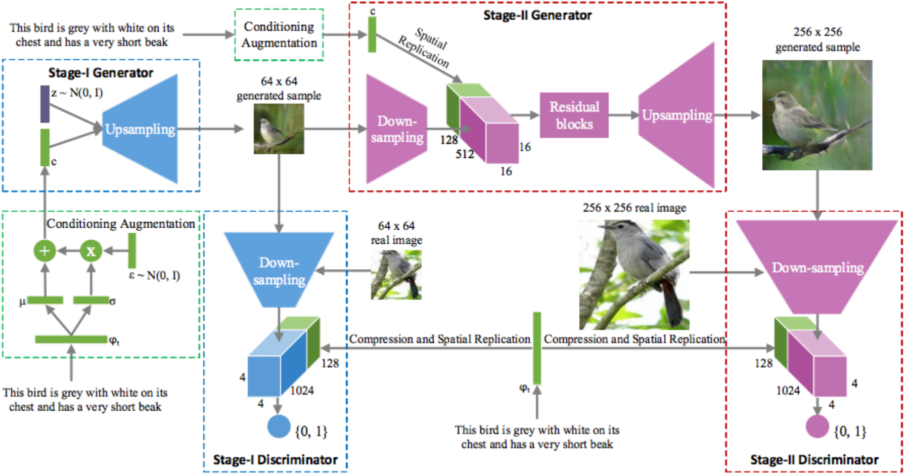


Some failure cases

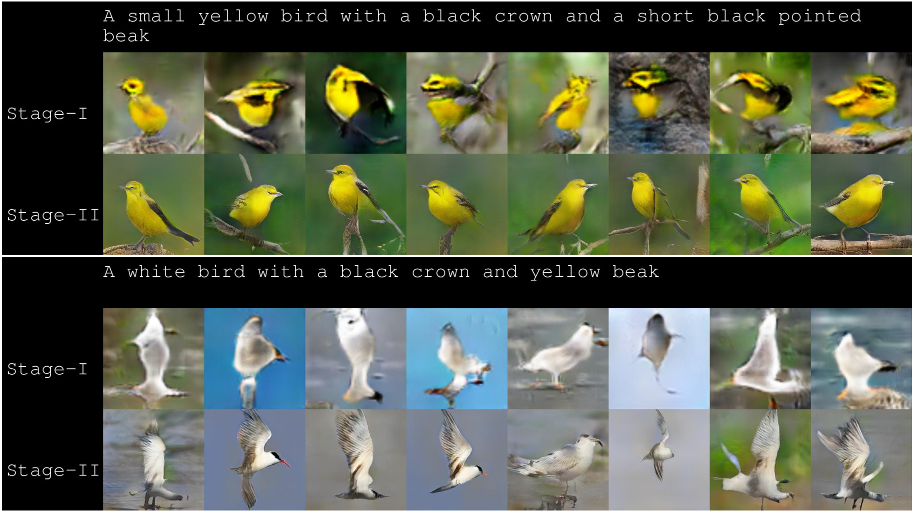
(Goodfellow 2016)

StackGAN

Zhang et al. 2016



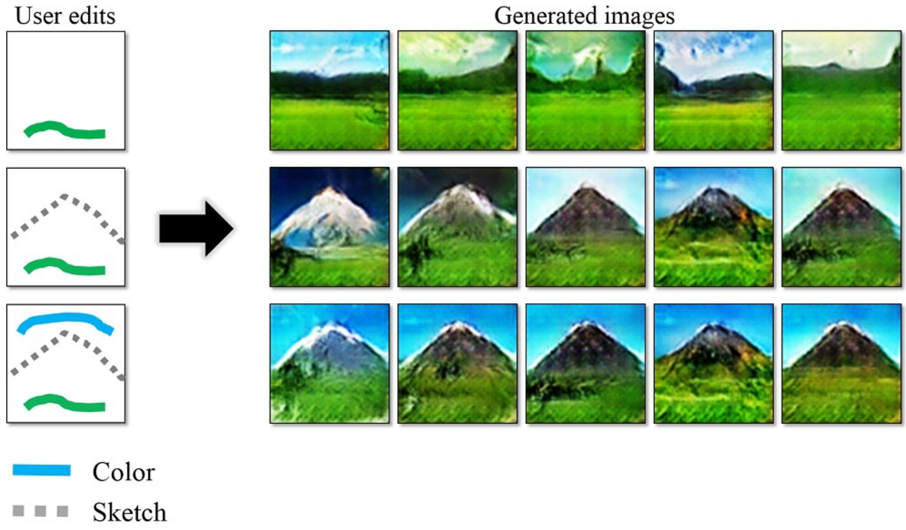
StackGAN


A small yellow bird with a black crown and a short black pointed beak
Stage-I
Stage-II
A white bird with a black crown and yellow beak
Stage-I
Stage-II

StackGAN

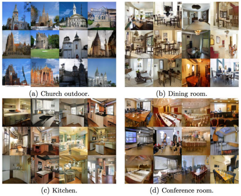

This flower has long thin yellow petals and a lot of yellow anthers in the center
Stage-I
Stage-II
This flower is white, pink, and yellow in color, and has petals that are multi colored
Stage-I
Stage-II

iGAN
Zhu *et al.* 2016

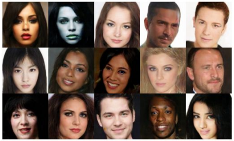


2017: Year of the GAN

Better training and generation

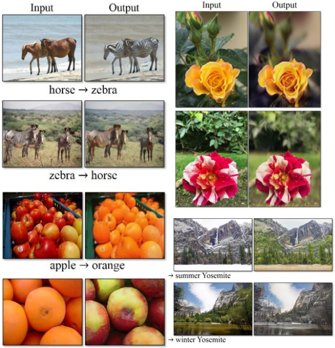


LSGAN. Mao et al. 2017.



BEGAN. Bertholet et al. 2017.

Source->Target domain transfer



CycleGAN. Zhu et al. 2017.

Text -> Image Synthesis

Reed et al. 2017.

Many GAN applications

Pix2pix. Isola 2017. Many examples at https://phillipi.github.io/pix2pix/

"The GAN Zoo"

See also: https://github.com/soumith/ganhacks for tips and tricks for trainings GANs

- GAN - Generative Adversarial Networks
- 3D-GAN - Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling
- acGAN - Face Aging With Conditional Generative Adversarial Networks
- AC-GAN - Conditional Image Synthesis With Auxiliary Classifier GANs
- AdaGAN - AdaGAN: Boosting Generative Models
- AEGAN - Learning Inverse Mapping by Autoencoder based Generative Adversarial Nets
- AffGAN - Amortised MAP Inference for Image Super-resolution
- AL-CGAN - Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts
- ALI - Adversarially Learned Inference
- AM-GAN - Generative Adversarial Nets with Labeled Data by Activation Maximization
- AnoGAN - Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery
- ArtGAN - ArtGAN: Artwork Synthesis with Conditional Categorial GANs
- b-GAN - b-GAN: Unified Framework of Generative Adversarial Networks
- Bayesian GAN - Deep and Hierarchical Implicit Models
- BEGAN - BEGAN: Boundary Equilibrium Generative Adversarial Networks
- BiGAN - Adversarial Feature Learning
- BS-GAN - Boundary-Seeking Generative Adversarial Networks
- CGAN - Conditional Generative Adversarial Nets
- CaloGAN - CaloGAN: Simulating 3D High Energy Particle Showers in Multi-Layer Electromagnetic Calorimeters with Generative Adversarial Networks
- CCGAN - Semi-Supervised Learning with Context-Conditional Generative Adversarial Networks
- CatGAN - Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks
- CoGAN - Coupled Generative Adversarial Networks
- Context-RNN-GAN - Contextual RNN-GANs for Abstract Reasoning Diagram Generation
- C-RNN-GAN - C-RNN-GAN: Continuous recurrent neural networks with adversarial training
- CS-GAN - Improving Neural Machine Translation with Conditional Sequence Generative Adversarial Nets
- CVAE-GAN - CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training
- CycleGAN - Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks
- DTN - Unsupervised Cross-Domain Image Generation
- DCGAN - Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
- DiscoGAN - Learning to Discover Cross-Domain Relations with Generative Adversarial Networks
- DR-GAN - Disentangled Representation Learning GAN for Pose-Invariant Face Recognition
- DualGAN - DualGAN: Unsupervised Dual Learning for Image-to-Image Translation
- EBGAN - Energy-based Generative Adversarial Network
- f-GAN - f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization
- FF-GAN - Towards Large-Pose Face Frontalization in the Wild
- GAWWN - Learning What and Where to Draw
- GeneGAN - GeneGAN: Learning Object Transfiguration and Attribute Subspace from Unpaired Data
- Geometric GAN - Geometric GAN
- GoGAN - Gang of GANs: Generative Adversarial Networks with Maximum Margin Ranking
- GP-GAN - GP-GAN: Towards Realistic High-Resolution Image Blending
- IAN - Neural Photo Editing with Introspective Adversarial Networks
- iGAN - Generative Visual Manipulation on the Natural Image Manifold
- IcGAN - Invertible Conditional GANs for image editing
- ID-CGAN - Image De-raining Using a Conditional Generative Adversarial Network
- Improved GAN - Improved Techniques for Training GANs
- InfoGAN - InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets
- LAGAN - Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis
- LAPGAN - Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks

https://github.com/hindupuravinash/the-gan-zoo

Fei-Fei Li & Justin Johnson & Serena Yeung | Lecture 13 - 129 | May 18, 2017

GANs

Don't work with an explicit density function
Take game-theoretic approach: learn to generate from training distribution through 2-player game

Pros:

- Beautiful, state-of-the-art samples!

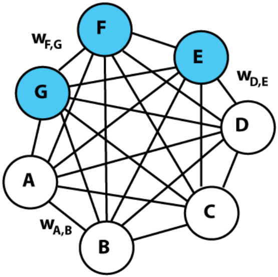
Cons:

- Trickier / more unstable to train
- Can't solve inference queries such as p(x), p(z|x)

Active areas of research:

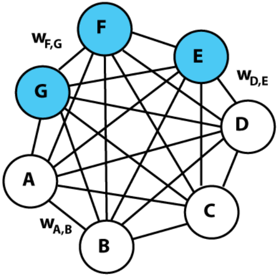
- Better loss functions, more stable training (Wasserstein GAN, LSGAN, many others)
- Conditional GANs, GANs for all kinds of applications

Boltzmann machines



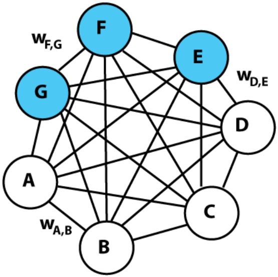
- Boltzmann machines were invented by Geoffrey Hinton and Terry Sejnowski in 1985
- It is a binary generative model
- Probability of a "configuration" is government by the Boltzmann distribution $\frac{\exp(-E(x))}{Z}$, where Z is a normalization factor and called the partition function (a name originated from statistical physics)
- The energy function $E(x)$ has a very simple form $E(x) = -x^T W x - c^T x$
- Typically some variables are hidden whereas others are visible

Boltzmann machines



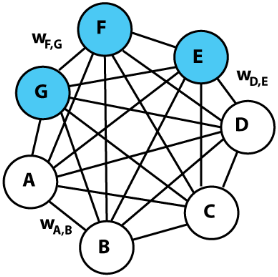
- Boltzmann machines were invented by Geoffrey Hinton and Terry Sejnowski in 1985
- It is a binary generative model
- Probability of a "configuration" is government by the Boltzmann distribution $\frac{\exp(-E(x))}{Z}$, where Z is a normalization factor and called the partition function (a name originated from statistical physics)
- The energy function $E(x)$ has a very simple form $E(x) = -x^T W x - c^T x$
- Typically some variables are hidden whereas others are visible

Boltzmann machines



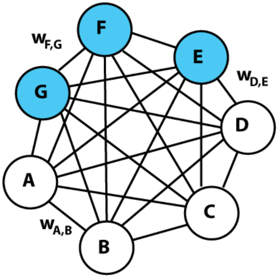
- Boltzmann machines were invented by Geoffrey Hinton and Terry Sejnowski in 1985
- It is a binary generative model
- Probability of a "configuration" is government by the Boltzmann distribution $\frac{\exp(-E(x))}{Z}$, where Z is a normalization factor and called the partition function (a name originated from statistical physics)
- The energy function $E(x)$ has a very simple form $E(x) = -x^T W x - c^T x$
- Typically some variables are hidden whereas others are visible

Boltzmann machines



- Boltzmann machines were invented by Geoffrey Hinton and Terry Sejnowski in 1985
- It is a binary generative model
- Probability of a "configuration" is government by the Boltzmann distribution $\frac{\exp(-E(x))}{Z}$, where Z is a normalization factor and called the partition function (a name originated from statistical physics)
- The energy function $E(x)$ has a very simple form $E(x) = -x^T W x - c^T x$
- Typically some variables are hidden whereas others are visible

Boltzmann machines



- Boltzmann machines were invented by Geoffrey Hinton and Terry Sejnowski in 1985
- It is a binary generative model
- Probability of a "configuration" is government by the Boltzmann distribution $\frac{\exp(-E(x))}{Z}$, where Z is a normalization factor and called the partition function (a name originated from statistical physics)
- The energy function $E(x)$ has a very simple form $E(x) = -x^T W x - c^T x$
- Typically some variables are hidden whereas others are visible

Restricted Boltzmann machines

- Boltzmann machine is a very powerful model. But with unconstrained connectivity, there are not known *efficient* methods to learn data and conduct inference for practical problems
- Consequently, restricted Boltzmann machine (RBM) (originally called Harmonium) was introduced by Paul Smolensky in 1986. It restricted the hidden units and the visible units from connecting to themselves
- The model rose to prominence after fast learning algorithm was invented by Hinton and his collaborators in mid-2000s

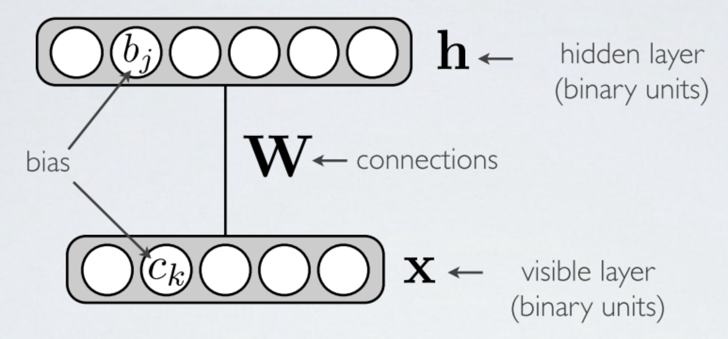
Restricted Boltzmann machines

- Boltzmann machine is a very powerful model. But with unconstrained connectivity, there are not known *efficient* methods to learn data and conduct inference for practical problems
- Consequently, restricted Boltzmann machine (RBM) (originally called Harmonium) was introduced by Paul Smolensky in 1986. It restricted the hidden units and the visible units from connecting to themselves
- The model rose to prominence after fast learning algorithm was invented by Hinton and his collaborators in mid-2000s

Restricted Boltzmann machines

- Boltzmann machine is a very powerful model. But with unconstrained connectivity, there are not known *efficient* methods to learn data and conduct inference for practical problems
- Consequently, restricted Boltzmann machine (RBM) (originally called Harmonium) was introduced by Paul Smolensky in 1986. It restricted the hidden units and the visible units from connecting to themselves
- The model rose to prominence after fast learning algorithm was invented by Hinton and his collaborators in mid-2000s

Restricted Boltzmann machines

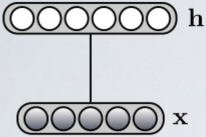


Energy function: $E(x,h) = -h^TWx - c^Tx - b^Th$

Distribution:

$$p(x,h) = \frac{\exp(-E(x,h))}{Z} = \frac{\exp(h^TWx)\exp(c^Tx)\exp(b^Th)}{Z}$$

Conditional probabilities

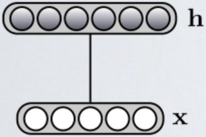


$$p(\mathbf{h}|\mathbf{x}) = \prod_j p(h_j|\mathbf{x})$$

$$p(h_j = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(b_j + \mathbf{W}_{j\cdot}\mathbf{x}))}$$

$$= \mathrm{sigm}(b_j + \mathbf{W}_{j\cdot}\mathbf{x})$$

j th row of $\mathbf{W}$



$$p(\mathbf{x}|\mathbf{h}) = \prod_k p(x_k|\mathbf{h})$$

$$p(x_k = 1|\mathbf{h}) = \frac{1}{1 + \exp(-(c_k + \mathbf{h}^\top \mathbf{W}_{\cdot k}))}$$

$$= \mathrm{sigm}(c_k + \mathbf{h}^\top \mathbf{W}_{\cdot k})$$

k th column of $\mathbf{W}$

Derivation of conditional probabilities

$$p(h|x) = \frac{p(x,h)}{\sum_{h'} p(x,h')} = \frac{\exp(h^T W x + c^T x + b^T h)/Z}{\sum_{h' \in \{0,1\}^M} \exp(h'^T W x + c^T x + b^T h')/Z}$$

$$= \frac{\exp\left(\sum_i h_i W_i x + b_i h_i\right)}{\sum_{h_1' \in \{0,1\}} \cdots \sum_{h_M' \in \{0,1\}} \exp(\sum_i h_i' W_i x + b_i h_i')} \qquad \left(W = \begin{pmatrix} W_1 \\ \cdots \\ W_M \end{pmatrix} \right)$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\sum_{h_1' \in \{0,1\}} \cdots \sum_{h_M' \in \{0,1\}} \prod_i \exp(h_i' W_i x + b_i h_i')}$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\left(\sum_{h_1' \in \{0,1\}} \exp(h_1' W_1 x + b_1 h_1')\right) \cdots \left(\sum_{h_M' \in \{0,1\}} \exp(h_M' W_M x + b_M h_M')\right)}$$

$$= \prod_i \frac{\exp\left(h_i W_i x + c^T x + b_i h_i\right)/Z}{\left(\sum_{h_i' \in \{0,1\}} \exp(h_i' W_i x + c^T x + b_i h_i')\right)/Z} = \prod_i p(h_i|x)$$

N.B. Can also be obtained immediately since $h_1, h_2, \cdots, h_M$ are conditionally independent given x

Derivation of conditional probabilities

$$p(h|x) = \frac{p(x,h)}{\sum_{h'} p(x,h')} = \frac{\exp(h^T W x + c^T x + b^T h)/Z}{\sum_{h' \in \{0,1\}^M} \exp(h'^T W x + c^T x + b^T h')/Z}$$

$$= \frac{\exp\left(\sum_i h_i W_i x + b_i h_i\right)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_M \in \{0,1\}} \exp(\sum_i h'_i W_i x + b_i h'_i)} \qquad \left(W = \begin{pmatrix} W_1 \\ \cdots \\ W_M \end{pmatrix} \right)$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_M \in \{0,1\}} \prod_i \exp(h'_i W_i x + b_i h'_i)}$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 W_1 x + b_1 h'_1)\right) \cdots \left(\sum_{h'_M \in \{0,1\}} \exp(h'_M W_M x + b_M h'_M)\right)}$$

$$= \prod_i \frac{\exp\left(h_i W_i x + c^T x + b_i h_i\right)/Z}{\left(\sum_{h'_i \in \{0,1\}} \exp(h'_i W_i x + c^T x + b_i h'_i)\right)/Z} = \prod_i p(h_i|x)$$

N.B. Can also be obtained immediately since $h_1, h_2, \cdots, h_M$ are conditionally independent given x

Derivation of conditional probabilities

$$p(h|x) = \frac{p(x,h)}{\sum_{h'} p(x,h')} = \frac{\exp(h^T W x + c^T x + b^T h)/Z}{\sum_{h' \in \{0,1\}^M} \exp(h'^T W x + c^T x + b^T h')/Z}$$

$$= \frac{\exp\left(\sum_i h_i W_i x + b_i h_i\right)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_M \in \{0,1\}} \exp(\sum_i h'_i W_i x + b_i h'_i)} \qquad \left(W = \begin{pmatrix} W_1 \\ \cdots \\ W_M \end{pmatrix} \right)$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_M \in \{0,1\}} \prod_i \exp(h'_i W_i x + b_i h'_i)}$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 W_1 x + b_1 h'_1)\right) \cdots \left(\sum_{h'_M \in \{0,1\}} \exp(h'_M W_M x + b_M h'_M)\right)}$$

$$= \prod_i \frac{\exp\left(h_i W_i x + c^T x + b_i h_i\right)/Z}{\left(\sum_{h'_i \in \{0,1\}} \exp(h'_i W_i x + c^T x + b_i h'_i)\right)/Z} = \prod_i p(h_i|x)$$

N.B. Can also be obtained immediately since $h_1, h_2, \cdots, h_M$ are conditionally independent given x

Derivation of conditional probabilities

$$p(h|x) = \frac{p(x,h)}{\sum_{h'} p(x,h')} = \frac{\exp(h^T W x + c^T x + b^T h)/Z}{\sum_{h' \in \{0,1\}^M} \exp(h'^T W x + c^T x + b^T h')/Z}$$

$$= \frac{\exp\left(\sum_i h_i W_i x + b_i h_i\right)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_M \in \{0,1\}} \exp(\sum_i h'_i W_i x + b_i h'_i)} \qquad \left(W = \begin{pmatrix} W_1 \\ \cdots \\ W_M \end{pmatrix} \right)$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_M \in \{0,1\}} \prod_i \exp(h'_i W_i x + b_i h'_i)}$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 W_1 x + b_1 h'_1)\right) \cdots \left(\sum_{h'_M \in \{0,1\}} \exp(h'_M W_M x + b_M h'_M)\right)}$$

$$= \prod_i \frac{\exp\left(h_i W_i x + c^T x + b_i h_i\right)/Z}{\left(\sum_{h'_i \in \{0,1\}} \exp(h'_i W_i x + c^T x + b_i h'_i)\right)/Z} = \prod_i p(h_i|x)$$

N.B. Can also be obtained immediately since $h_1, h_2, \cdots, h_M$ are conditionally independent given x

Derivation of conditional probabilities

$$p(h|x) = \frac{p(x,h)}{\sum_{h'} p(x,h')} = \frac{\exp(h^T W x + c^T x + b^T h)/Z}{\sum_{h' \in \{0,1\}^M} \exp(h'^T W x + c^T x + b^T h')/Z}$$

$$= \frac{\exp\left(\sum_i h_i W_i x + b_i h_i\right)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_M \in \{0,1\}} \exp(\sum_i h'_i W_i x + b_i h'_i)} \qquad \left(W = \begin{pmatrix} W_1 \\ \cdots \\ W_M \end{pmatrix} \right)$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_M \in \{0,1\}} \prod_i \exp(h'_i W_i x + b_i h'_i)}$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 W_1 x + b_1 h'_1)\right) \cdots \left(\sum_{h'_M \in \{0,1\}} \exp(h'_M W_M x + b_M h'_M)\right)}$$

$$= \prod_i \frac{\exp\left(h_i W_i x + c^T x + b_i h_i\right)/Z}{\left(\sum_{h'_i \in \{0,1\}} \exp(h'_i W_i x + c^T x + b_i h'_i)\right)/Z} = \prod_i p(h_i|x)$$

N.B. Can also be obtained immediately since $h_1, h_2, \cdots, h_M$ are conditionally independent given x

Derivation of conditional probabilities

$$p(h|x) = \frac{p(x,h)}{\sum_{h'} p(x,h')} = \frac{\exp(h^T W x + c^T x + b^T h)/Z}{\sum_{h' \in \{0,1\}^M} \exp(h'^T W x + c^T x + b^T h')/Z}$$

$$= \frac{\exp\left(\sum_i h_i W_i x + b_i h_i\right)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_M \in \{0,1\}} \exp(\sum_i h'_i W_i x + b_i h'_i)} \qquad \left(W = \begin{pmatrix} W_1 \\ \cdots \\ W_M \end{pmatrix} \right)$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_M \in \{0,1\}} \prod_i \exp(h'_i W_i x + b_i h'_i)}$$

$$= \frac{\prod_i \exp\left(h_i W_i x + b_i h_i\right)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 W_1 x + b_1 h'_1)\right) \cdots \left(\sum_{h'_M \in \{0,1\}} \exp(h'_M W_M x + b_M h'_M)\right)}$$

$$= \prod_i \frac{\exp\left(h_i W_i x + c^T x + b_i h_i\right)/Z}{\left(\sum_{h'_i \in \{0,1\}} \exp(h'_i W_i x + c^T x + b_i h'_i)\right)/Z} = \prod_i p(h_i|x)$$

N.B. Can also be obtained immediately since $h_1, h_2, \cdots, h_M$ are conditionally independent given x

Derivation of conditional probabilities

$$p(h_i = 1|x) = \frac{\exp\left(W_i x + b_i\right)}{\left(\sum_{h_i' \in \{0,1\}} \exp(h_i' W_i x + b_i h_i')\right)}$$

$$= \frac{\exp\left(W_i x + b_i\right)}{(1 + \exp(\ W_i x + b_i))}$$

$$= \text{sigm}(b_i + W_i x)$$

Derivation of conditional probabilities

$$p(h_i = 1|x) = \frac{\exp\left(W_i x + b_i\right)}{\left(\sum_{h_i' \in \{0,1\}} \exp(h_i' W_i x + b_i h_i')\right)}$$

$$= \frac{\exp\left(W_i x + b_i\right)}{(1 + \exp(\ W_i x + b_i))}$$

$$= \text{sigm}(b_i + W_i x)$$

Derivation of conditional probabilities

$$
\begin{aligned}
p(h_i = 1|x) &= \frac{\exp\left(W_i x + b_i\right)}{\left(\sum_{h_i' \in \{0,1\}} \exp(h_i' W_i x + b_i h_i')\right)} \\
&= \frac{\exp\left(W_i x + b_i\right)}{(1 + \exp(\ W_i x + b_i))} \\
&= \mathsf{sigm}(b_i + W_i x)
\end{aligned}
$$

Data generation

Equipped with the conditional probabilities $p(x|h)$ and $p(h|x)$, we can generate simulated data given some hidden variables h' using Gibbs sampling

- Sample x' from $p(x|h')$
- Sample h'' from $p(h|x')$
- Sample x'' from $p(x|h'')$
- ⋯

Marginal probability $p(x)$

$$p(x) = \sum_{h\in\{0,1\}^M} \exp(h^T W x + c^T x + b^T h)/Z$$

$$= \frac{exp(c^T x)}{Z} \sum_{h_1\in\{0,1\}} \cdots \sum_{h_M\in\{0,1\}} \exp\left(\sum_i h_i W_i x + b_i h_i\right)$$

$$= \frac{\exp(c^T x)}{Z} \left(\sum_{h_1\in\{0,1\}} e^{(h_1 W_1 x + b_1 h_1)}\right) \cdots \left(\sum_{h_M\in\{0,1\}} e^{(h_M W_M x + b_M h_M)}\right)$$

$$= \frac{\exp(c^T x)}{Z} (1 + e^{(W_1 x + b_1)}) \cdots (1 + e^{(W_M x + b_M)})$$

$$= \frac{\exp(c^T x)}{Z} \exp\left(\log(1 + e^{(W_1 x + b_1)}) + \cdots + \log(1 + e^{(W_M x + b_M)})\right)$$

$$= \exp\left(c^T x + \sum_i \log(1 + e^{(W_i x + b_i)})\right)/Z$$

Marginal probability $p(x)$

$$
\begin{aligned}
p(x) &= \sum_{h\in\{0,1\}^M} \exp(h^T W x + c^T x + b^T h)/Z \\
&= \frac{exp(c^T x)}{Z} \sum_{h_1\in\{0,1\}} \cdots \sum_{h_M\in\{0,1\}} \exp\left(\sum_i h_i W_i x + b_i h_i\right) \\
&= \frac{\exp(c^T x)}{Z} \left(\sum_{h_1\in\{0,1\}} e^{(h_1 W_1 x + b_1 h_1)}\right) \cdots \left(\sum_{h_M\in\{0,1\}} e^{(h_M W_M x + b_M h_M)}\right) \\
&= \frac{\exp(c^T x)}{Z} \left(1 + e^{(W_1 x + b_1)}\right) \cdots \left(1 + e^{(W_M x + b_M)}\right) \\
&= \frac{\exp(c^T x)}{Z} \exp\left(\log(1 + e^{(W_1 x + b_1)}) + \cdots + \log(1 + e^{(W_M x + b_M)})\right) \\
&= \exp\left(c^T x + \sum_i \log(1 + e^{(W_i x + b_i)})\right)/Z
\end{aligned}
$$

Marginal probability $p(x)$

$$
\begin{aligned}
p(x) &= \sum_{h\in\{0,1\}^M} \exp(h^T W x + c^T x + b^T h)/Z \\
&= \frac{exp(c^T x)}{Z} \sum_{h_1\in\{0,1\}} \cdots \sum_{h_M\in\{0,1\}} \exp\left(\sum_i h_i W_i x + b_i h_i\right) \\
&= \frac{\exp(c^T x)}{Z} \left(\sum_{h_1\in\{0,1\}} e^{(h_1 W_1 x + b_1 h_1)}\right) \cdots \left(\sum_{h_M\in\{0,1\}} e^{(h_M W_M x + b_M h_M)}\right) \\
&= \frac{\exp(c^T x)}{Z} (1 + e^{(W_1 x + b_1)}) \cdots (1 + e^{(W_M x + b_M)}) \\
&= \frac{\exp(c^T x)}{Z} \exp\left(\log(1 + e^{(W_1 x + b_1)}) + \cdots + \log(1 + e^{(W_M x + b_M)})\right) \\
&= \exp\left(c^T x + \sum_i \log(1 + e^{(W_i x + b_i)})\right)/Z
\end{aligned}
$$

Marginal probability $p(x)$

$$
\begin{aligned}
p(x) &= \sum_{h\in\{0,1\}^M} \exp(h^T W x + c^T x + b^T h)/Z \\
&= \frac{exp(c^T x)}{Z} \sum_{h_1\in\{0,1\}} \cdots \sum_{h_M\in\{0,1\}} \exp\left(\sum_i h_i W_i x + b_i h_i\right) \\
&= \frac{\exp(c^T x)}{Z} \left(\sum_{h_1\in\{0,1\}} e^{(h_1 W_1 x + b_1 h_1)}\right) \cdots \left(\sum_{h_M\in\{0,1\}} e^{(h_M W_M x + b_M h_M)}\right) \\
&= \frac{\exp(c^T x)}{Z} \left(1 + e^{(W_1 x + b_1)}\right) \cdots \left(1 + e^{(W_M x + b_M)}\right) \\
&= \frac{\exp(c^T x)}{Z} \exp\left(\log(1 + e^{(W_1 x + b_1)}) + \cdots + \log(1 + e^{(W_M x + b_M)})\right) \\
&= \exp\left(c^T x + \sum_i \log(1 + e^{(W_i x + b_i)})\right)/Z
\end{aligned}
$$

Marginal probability $p(x)$

$$
\begin{aligned}
p(x) &= \sum_{h\in\{0,1\}^M} \exp(h^T W x + c^T x + b^T h)/Z \\
&= \frac{exp(c^T x)}{Z} \sum_{h_1\in\{0,1\}} \cdots \sum_{h_M\in\{0,1\}} \exp\left(\sum_i h_i W_i x + b_i h_i\right) \\
&= \frac{\exp(c^T x)}{Z} \left(\sum_{h_1\in\{0,1\}} e^{(h_1 W_1 x + b_1 h_1)}\right) \cdots \left(\sum_{h_M\in\{0,1\}} e^{(h_M W_M x + b_M h_M)}\right) \\
&= \frac{\exp(c^T x)}{Z} \left(1 + e^{(W_1 x + b_1)}\right) \cdots \left(1 + e^{(W_M x + b_M)}\right) \\
&= \frac{\exp(c^T x)}{Z} \exp\left(\log(1 + e^{(W_1 x + b_1)}) + \cdots + \log(1 + e^{(W_M x + b_M)})\right) \\
&= \exp\left(c^T x + \sum_i \log(1 + e^{(W_i x + b_i)})\right) / Z
\end{aligned}
$$

Marginal probability $p(x)$

$$
\begin{aligned}
p(x) &= \sum_{h\in\{0,1\}^M} \exp(h^T W x + c^T x + b^T h)/Z \\
&= \frac{exp(c^T x)}{Z} \sum_{h_1\in\{0,1\}} \cdots \sum_{h_M\in\{0,1\}} \exp\left(\sum_i h_i W_i x + b_i h_i\right) \\
&= \frac{\exp(c^T x)}{Z} \left(\sum_{h_1\in\{0,1\}} e^{(h_1 W_1 x + b_1 h_1)}\right) \cdots \left(\sum_{h_M\in\{0,1\}} e^{(h_M W_M x + b_M h_M)}\right) \\
&= \frac{\exp(c^T x)}{Z} \left(1 + e^{(W_1 x + b_1)}\right) \cdots \left(1 + e^{(W_M x + b_M)}\right) \\
&= \frac{\exp(c^T x)}{Z} \exp\left(\log(1 + e^{(W_1 x + b_1)}) + \cdots + \log(1 + e^{(W_M x + b_M)})\right) \\
&= \exp\left(c^T x + \sum_i \log(1 + e^{(W_i x + b_i)})\right)/Z
\end{aligned}
$$

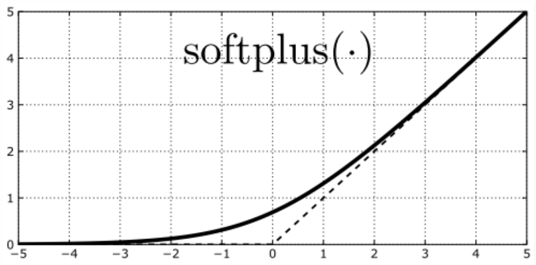
$$p(x) = \exp\left(c^T x + \sum_i \log(1 + e^{(W_i x + b_i)})\right)/Z$$

$$= \exp\left(c^T x + \sum_i \text{softplus}(W_i x + b_i)\right)/Z \triangleq \exp(-F(x))/Z,$$

where $F(x)$ is known to be free energy, a term borrowed from statistical physics. Note that $\frac{\partial \text{softplus(t)}}{\partial t} = \text{sigmod}(t)$

$$p(x) = \exp\left(c^T x + \sum_i \log(1 + e^{(W_i x + b_i)})\right) / Z$$
$$= \exp\left(c^T x + \sum_i \text{softplus}(W_i x + b_i)\right) / Z \triangleq \exp(-F(x))/Z,$$

where $F(x)$ is known to be free energy, a term borrowed from statistical physics. Note that $\frac{\partial \text{softplus(t)}}{\partial t} = \text{sigmod}(t)$



Training RBM

Use the cross entropy loss,

$$l(\theta) = \frac{1}{T}\sum_{t=1}^{T} -\log p(x^{(t)}) = \frac{1}{T}\sum_{t=1}^{T} F(x^{(t)}) - \log Z,$$

where $Z = \sum_x \exp(-F(x))$. And

$$\frac{\partial - \log p(x^{(t)})}{\partial \theta} = \frac{\partial F(x^{(t)})}{\partial \theta} - \sum_x \frac{\exp(-F(x))}{Z}\frac{\partial F(x)}{\partial \theta}$$

$$= \underbrace{\frac{\partial F(x^{(t)})}{\partial \theta}}_{\text{positive phase}} - \underbrace{E\left[\frac{\partial F(x)}{\partial \theta}\right]}_{\text{negative phase}}$$

N.B. The naming of the terms is not related to the sign in the equation. It refers to the fact that adjusting the +ve phase terms to increase the probability of the training data and the -ve terms to decrease the probability of the rest of x

Training RBM

Use the cross entropy loss,

$$l(\theta) = \frac{1}{T}\sum_{t=1}^{T} -\log p(x^{(t)}) = \frac{1}{T}\sum_{t=1}^{T} F(x^{(t)}) - \log Z,$$

where $Z = \sum_x \exp(-F(x))$. And

$$\frac{\partial - \log p(x^{(t)})}{\partial \theta} = \frac{\partial F(x^{(t)})}{\partial \theta} - \sum_x \frac{\exp(-F(x))}{Z} \frac{\partial F(x)}{\partial \theta}$$

$$= \underbrace{\frac{\partial F(x^{(t)})}{\partial \theta}}_{\text{positive phase}} - \underbrace{E\left[\frac{\partial F(x)}{\partial \theta}\right]}_{\text{negative phase}}$$

N.B. The naming of the terms is not related to the sign in the equation. It refers to the fact that adjusting the +ve phase terms to increase the probability of the training data and the -ve terms to decrease the probability of the rest of x

Training RBM

Use the cross entropy loss,

$$l(\theta) = \frac{1}{T}\sum_{t=1}^{T} -\log p(x^{(t)}) = \frac{1}{T}\sum_{t=1}^{T} F(x^{(t)}) - \log Z,$$

where $Z = \sum_x \exp(-F(x))$. And

$$\frac{\partial - \log p(x^{(t)})}{\partial \theta} = \frac{\partial F(x^{(t)})}{\partial \theta} - \sum_x \frac{\exp(-F(x))}{Z} \frac{\partial F(x)}{\partial \theta}$$

$$= \underbrace{\frac{\partial F(x^{(t)})}{\partial \theta}}_{\text{positive phase}} - \underbrace{E\left[\frac{\partial F(x)}{\partial \theta}\right]}_{\text{negative phase}}$$

N.B. The naming of the terms is not related to the sign in the equation. It refers to the fact that adjusting the +ve phase terms to increase the probability of the training data and the -ve terms to decrease the probability of the rest of x

Training RBM

Use the cross entropy loss,

$$l(\theta) = \frac{1}{T}\sum_{t=1}^{T} -\log p(x^{(t)}) = \frac{1}{T}\sum_{t=1}^{T} F(x^{(t)}) - \log Z,$$

where $Z = \sum_x \exp(-F(x))$. And

$$\frac{\partial - \log p(x^{(t)})}{\partial \theta} = \frac{\partial F(x^{(t)})}{\partial \theta} - \sum_x \frac{\exp(-F(x))}{Z}\frac{\partial F(x)}{\partial \theta}$$

$$= \underbrace{\frac{\partial F(x^{(t)})}{\partial \theta}}_{\text{positive phase}} - \underbrace{E\left[\frac{\partial F(x)}{\partial \theta}\right]}_{\text{negative phase}}$$

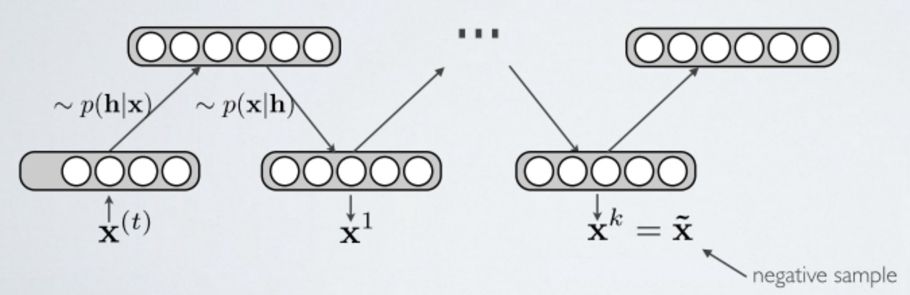
N.B. The naming of the terms is not related to the sign in the equation. It refers to the fact that adjusting the $+$ve phase terms to increase the probability of the training data and the -ve terms to decrease the probability of the rest of x

Contrastive divergence (CD-k)

The negative phase term is very hard to compute exactly as we need to sum over all x. The natural way out is to approximate using sampling $\Rightarrow$ contrastive divergence (CD-k) training

Key idea:

1. Start sampling chain at $x^{(t)}$
2. Obtain the point $\tilde{x}$ with k Gibbs sampling steps
3. Replace the expectation by a point estimate at $\tilde{x}$



N.B. CD-1 works surprisingly well in practice

Parameters update

So we have $\frac{\partial l(\theta)}{\partial \theta} = \frac{\partial F(x^{(t)})}{\partial \theta} - \frac{\partial F(\tilde{x})}{\partial \theta}$. Recall that

$$F(x) = -c^T x - \sum_i \text{softplus}(W_i x + b_i)$$

$$\frac{\partial F(x)}{\partial c_i} = -x_i$$

$$\frac{\partial F(x)}{\partial b_i} = -\text{sigmoid}(W_i x + b_i)$$

$$\frac{\partial F(x)}{\partial W_{ij}} = -\text{sigmoid}(W_i x + b_i) x_j$$

This gives us

$$c \Leftarrow c + \alpha(x^{(t)} - \tilde{x})$$

$$b \Leftarrow b + \alpha(\text{sigmoid}(W x^{(t)} + b) - \text{sigmoid}(W\tilde{x} + b))$$

$$W \Leftarrow W + \alpha(\text{sigmoid}(W x^{(t)} + b) {x^{(t)}}^T - \text{sigmoid}(W\tilde{x} + b)\tilde{x}^T)$$

Parameters update

So we have $\frac{\partial l(\theta)}{\partial \theta} = \frac{\partial F(x^{(t)})}{\partial \theta} - \frac{\partial F(\tilde{x})}{\partial \theta}$. Recall that

$$F(x) = -c^T x - \sum_i \mathsf{softplus}(W_i x + b_i)$$

$$\frac{\partial F(x)}{\partial c_i} = -x_i$$

$$\frac{\partial F(x)}{\partial b_i} = -\mathsf{sigmoid}(W_i x + b_i)$$

$$\frac{\partial F(x)}{\partial W_{ij}} = -\mathsf{sigmoid}(W_i x + b_i) x_j$$

This gives us

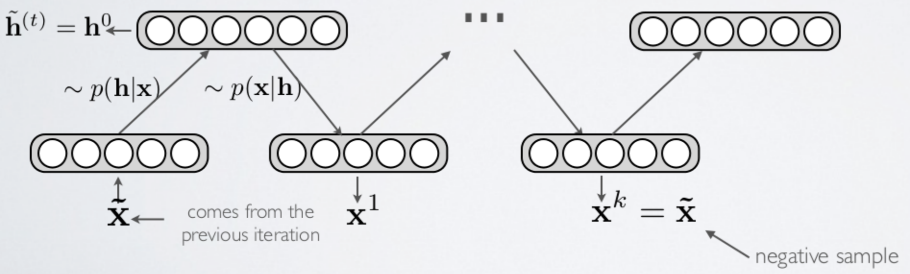
$$c \Leftarrow c + \alpha(x^{(t)} - \tilde{x})$$

$$b \Leftarrow b + \alpha(\mathsf{sigmoid}(W x^{(t)} + b) - \mathsf{sigmoid}(W\tilde{x} + b))$$

$$W \Leftarrow W + \alpha(\mathsf{sigmoid}(W x^{(t)} + b) {x^{(t)}}^T - \mathsf{sigmoid}(W\tilde{x} + b)\tilde{x}^T)$$

Persistent CD
Tieleman, ICML 2008

- Idea: Instead of initializing the chain to $x^{(t)}$, initialize the chain to the negative sample of the last iteration
- This has a similar effect of CD-k with a large k and yet can have much lower complexity



Gaussian-Bernoulli RBM
Extension to continuous variables

- RBM is a binary model and thus is not suitable for continuous data
- One simple extension to allow the visible variables x to be continuous while keeping the hidden variables h to be binary
- In particular, we can simply add a quadratic term $\frac{1}{2}x^T x$ to the energy function, i.e.,

$$E(x,h) = -h^T W x - c^T x - b^T h + \frac{1}{2}x^T x$$

  to get Gaussian distributed $p(x|h)$
- For efficient training, the input data are typically preprocessed with zero-mean and unit variance
- A smaller learning rate is needed compared to a regular RBM

Gaussian-Bernoulli RBM

Extension to continuous variables

- RBM is a binary model and thus is not suitable for continuous data
- One simple extension to allow the visible variables x to be continuous while keeping the hidden variables h to be binary
- In particular, we can simply add a quadratic term $\frac{1}{2}x^Tx$ to the energy function, i.e.,

$$E(x,h) = -h^TWx - c^Tx - b^Th + \frac{1}{2}x^Tx$$

  to get Gaussian distributed $p(x|h)$
- For efficient training, the input data are typically preprocessed with zero-mean and unit variance
- A smaller learning rate is needed compared to a regular RBM

Gaussian-Bernoulli RBM

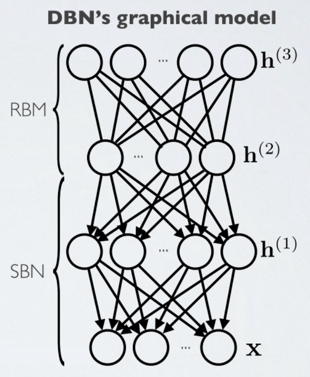
Extension to continuous variables

- RBM is a binary model and thus is not suitable for continuous data
- One simple extension to allow the visible variables x to be continuous while keeping the hidden variables h to be binary
- In particular, we can simply add a quadratic term $\frac{1}{2}x^T x$ to the energy function, i.e.,

$$E(x,h) = -h^T W x - c^T x - b^T h + \frac{1}{2}x^T x$$

  to get Gaussian distributed $p(x|h)$
- For efficient training, the input data are typically preprocessed with zero-mean and unit variance
- A smaller learning rate is needed compared to a regular RBM

Deep belief networks (DBN)

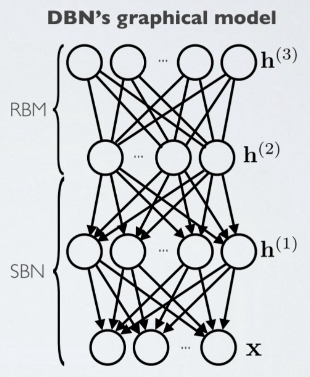


- DBN is a generative model that mixes undirected and directed connections
- Top 2 layers' distribution $p(h^{(2)}, h^{(3)})$ is an RBN
- Other layers form a Bayesian network:
  - The conditional distributions of layers given the one above it are

$$p(h_i^{(1)} = 1 | h^{(2)}) = \text{sigm}(b_i^{(1)} + W^{(2)}{}_{;i} h^{(2)})$$
$$p(h_i^{(1)} = 1 | h^{(1)}) = \text{sigm}(b_i^{(0)} + W^{(1)}{}_{;i} h^{(1)})$$

  - This is referred to as a sigmoid belief network (SBN)
- Note that DBN is not a feed-forward network

Deep belief networks (DBN)

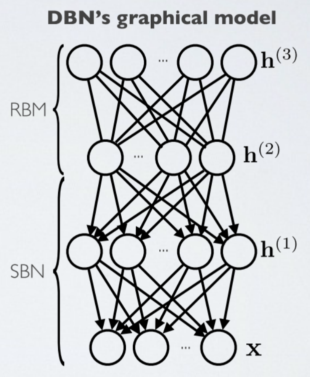


- DBN is a generative model that mixes undirected and directed connections
- Top 2 layers' distribution $p(h^{(2)}, h^{(3)})$ is an RBN
- Other layers form a Bayesian network:
  - The conditional distributions of layers given the one above it are

$$p(h_i^{(1)} = 1 | h^{(2)}) = \text{sigm}(b_i^{(1)} + W^{(2)}{}_i h^{(2)})$$
$$p(h_i^{(1)} = 1 | h^{(1)}) = \text{sigm}(b_i^{(0)} + W^{(1)}{}_i h^{(1)})$$

  - This is referred to as a sigmoid belief network (SBN)
- Note that DBN is not a feed-forward network

Deep belief networks (DBN)



- DBN is a generative model that mixes undirected and directed connections
- Top 2 layers' distribution $p(h^{(2)}, h^{(3)})$ is an RBN
- Other layers form a Bayesian network:
  - The conditional distributions of layers given the one above it are

$$p(h_i^{(1)} = 1 | h^{(2)}) = \text{sigm}(b_i^{(1)} + W^{(2)}{}_i h^{(2)})$$
$$p(h_i^{(1)} = 1 | h^{(1)}) = \text{sigm}(b_i^{(0)} + W^{(1)}{}_i h^{(1)})$$

  - This is referred to as a sigmoid belief network (SBN)
- Note that DBN is not a feed-forward network

History of DBNs

According to HInton's coursera's course

- Professor Hinton was working on algorithms to train Sigmoid belief network but gave up after many different ideas
- He moved on to work with RBMs and invented the CD-k algorithm for training RBMs
- Since CD-k is very effective, it is very tempting to think if one can train a Sigmoid belief network one layer at a time by treating each layer as a RBM
  - The procedure is working great. But it actually trains a different model, the DBN instead of SBN (with some complicated math behind), pointed out by Yee-Whye Teh
- DBN is actually the first successful deep neural network model and revived the entire neural network field
- Try not to get confused of DBN with deep Boltzmann machines (DBMs), where each layer is composed of an RBM

History of DBNs
According to HInton's coursera's course

- Professor Hinton was working on algorithms to train Sigmoid belief network but gave up after many different ideas
- He moved on to work with RBMs and invented the CD-k algorithm for training RBMs
- Since CD-k is very effective, it is very tempting to think if one can train a Sigmoid belief network one layer at a time by treating each layer as a RBM
    - The procedure is working great. But it actually trains a different model, the DBN instead of SBN (with some complicated math behind), pointed out by Yee-Whye Teh
- DBN is actually the first successful deep neural network model and revived the entire neural network field
- Try not to get confused of DBN with deep Boltzmann machines (DBMs), where each layer is composed of an RBM

History of DBNs

According to HInton's coursera's course

- Professor Hinton was working on algorithms to train Sigmoid belief network but gave up after many different ideas
- He moved on to work with RBMs and invented the CD-k algorithm for training RBMs
- Since CD-k is very effective, it is very tempting to think if one can train a Sigmoid belief network one layer at a time by treating each layer as a RBM
    - The procedure is working great. But it actually trains a different model, the DBN instead of SBN (with some complicated math behind), pointed out by Yee-Whye Teh
- DBN is actually the first successful deep neural network model and revived the entire neural network field
- Try not to get confused of DBN with deep Boltzmann machines (DBMs), where each layer is composed of an RBM

History of DBNs

According to HInton's coursera's course

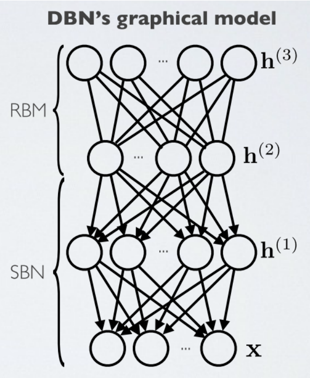
- Professor Hinton was working on algorithms to train Sigmoid belief network but gave up after many different ideas
- He moved on to work with RBMs and invented the CD-k algorithm for training RBMs
- Since CD-k is very effective, it is very tempting to think if one can train a Sigmoid belief network one layer at a time by treating each layer as a RBM
    - The procedure is working great. But it actually trains a different model, the DBN instead of SBN (with some complicated math behind), pointed out by Yee-Whye Teh
- DBN is actually the first successful deep neural network model and revived the entire neural network field
- Try not to get confused of DBN with deep Boltzmann machines (DBMs), where each layer is composed of an RBM

History of DBNs

According to HInton's coursera's course

- Professor Hinton was working on algorithms to train Sigmoid belief network but gave up after many different ideas
- He moved on to work with RBMs and invented the CD-k algorithm for training RBMs
- Since CD-k is very effective, it is very tempting to think if one can train a Sigmoid belief network one layer at a time by treating each layer as a RBM
    - The procedure is working great. But it actually trains a different model, the DBN instead of SBN (with some complicated math behind), pointed out by Yee-Whye Teh
- DBN is actually the first successful deep neural network model and revived the entire neural network field
- Try not to get confused of DBN with deep Boltzmann machines (DBMs), where each layer is composed of an RBM

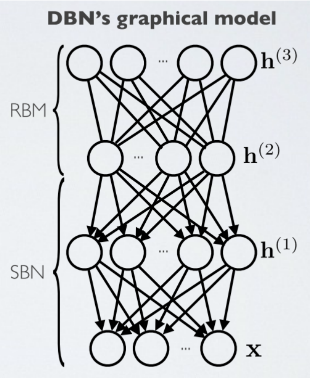
Pretraining of DBNs



As mentioned in the previous slide

- Treat the bottom two layers as an RBM and train it with the input data x
- Treat the next two layers as an RBM and train it with the $h^{(1)}$ obtained in the last step
- Keep continuing while keeping the trained weights

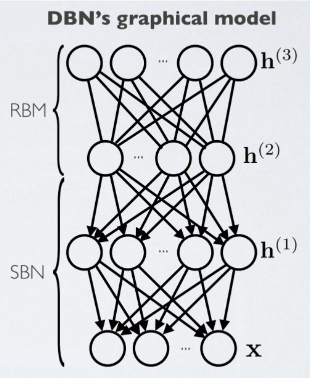
Pretraining of DBNs



As mentioned in the previous slide

- Treat the bottom two layers as an RBM and train it with the input data x
- Treat the next two layers as an RBM and train it with the $h^{(1)}$ obtained in the last step
- Keep continuing while keeping the trained weights

Pretraining of DBNs



As mentioned in the previous slide

- Treat the bottom two layers as an RBM and train it with the input data x
- Treat the next two layers as an RBM and train it with the $h^{(1)}$ obtained in the last step
- Keep continuing while keeping the trained weights

Fine-tuning of DBN

Up-down algorithm (aka contrastive wake-sleep algorithm)

After learning many layers of features, we can fine-tune the features to improve generation

1. Do a stochastic bottom-up pass
   - Construct hidden variables with reconstruction weight R (initialized as the transpose of W)
   - Use the approximated hidden variables to fine tune W
2. Do a few iterations of sampling in the top level RBM
   - Adjust top-level RBM weights using CD-k
3. Do a stochastic top-down pass
   - Generate simulation data and use that to fine-tune the reconstruction weights R

Fine-tuning of DBN

Up-down algorithm (aka contrastive wake-sleep algorithm)

After learning many layers of features, we can fine-tune the features to improve generation

1. Do a stochastic bottom-up pass
   - Construct hidden variables with reconstruction weight R (initialized as the transpose of W)
   - Use the approximated hidden variables to fine tune W
2. Do a few iterations of sampling in the top level RBM
   - Adjust top-level RBM weights using CD-k
3. Do a stochastic top-down pass
   - Generate simulation data and use that to fine-tune the reconstruction weights R

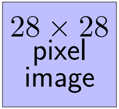
Fine-tuning of DBN

Up-down algorithm (aka contrastive wake-sleep algorithm)

After learning many layers of features, we can fine-tune the features to improve generation

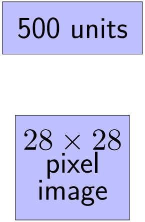
1. Do a stochastic bottom-up pass
   - Construct hidden variables with reconstruction weight R (initialized as the transpose of W)
   - Use the approximated hidden variables to fine tune W
2. Do a few iterations of sampling in the top level RBM
   - Adjust top-level RBM weights using CD-k
3. Do a stochastic top-down pass
   - Generate simulation data and use that to fine-tune the reconstruction weights R

MNIST example



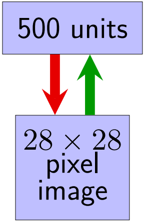
- Test on MNIST dataset
- Train 500 hidden units with the image block as input
- Train another 500 hidden units with the trained 500 hidden units as input
- Prepare another 2000 hidden units
- Train the 2000 hidden units with the previously trained 500 hidden units and target labels as input
- Error rate is about 1%

MNIST example



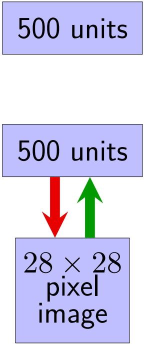
- Test on MNIST dataset
- Train 500 hidden units with the image block as input
- Train another 500 hidden units with the trained 500 hidden units as input
- Prepare another 2000 hidden units
- Train the 2000 hidden units with the previously trained 500 hidden units and target labels as input
- Error rate is about 1%

MNIST example



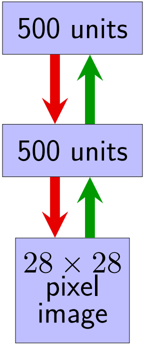
- Test on MNIST dataset
- Train 500 hidden units with the image block as input
- Train another 500 hidden units with the trained 500 hidden units as input
- Prepare another 2000 hidden units
- Train the 2000 hidden units with the previously trained 500 hidden units and target labels as input
- Error rate is about 1%

MNIST example



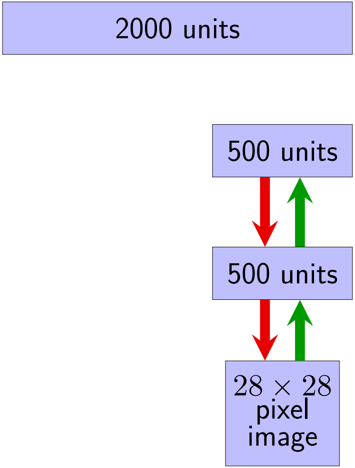
- Test on MNIST dataset
- Train 500 hidden units with the image block as input
- Train another 500 hidden units with the trained 500 hidden units as input
- Prepare another 2000 hidden units
- Train the 2000 hidden units with the previously trained 500 hidden units and target labels as input
- Error rate is about 1%

MNIST example



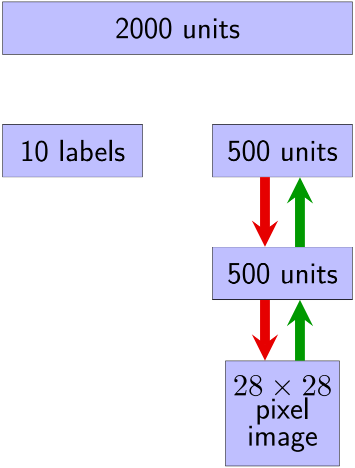
- Test on MNIST dataset
- Train 500 hidden units with the image block as input
- Train another 500 hidden units with the trained 500 hidden units as input
- Prepare another 2000 hidden units
- Train the 2000 hidden units with the previously trained 500 hidden units and target labels as input
- Error rate is about 1%

MNIST example



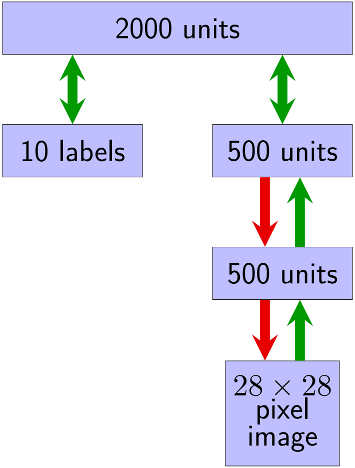
- Test on MNIST dataset
- Train 500 hidden units with the image block as input
- Train another 500 hidden units with the trained 500 hidden units as input
- Prepare another 2000 hidden units
- Train the 2000 hidden units with the previously trained 500 hidden units and target labels as input
- Error rate is about 1%

MNIST example



- Test on MNIST dataset
- Train 500 hidden units with the image block as input
- Train another 500 hidden units with the trained 500 hidden units as input
- Prepare another 2000 hidden units
- Train the 2000 hidden units with the previously trained 500 hidden units and target labels as input
- Error rate is about 1%

MNIST example



- Test on MNIST dataset
- Train 500 hidden units with the image block as input
- Train another 500 hidden units with the trained 500 hidden units as input
- Prepare another 2000 hidden units
- Train the 2000 hidden units with the previously trained 500 hidden units and target labels as input
- Error rate is about 1%

Demo

http://www.cs.toronto.edu/~hinton/adi/index.htm

Summary of Boltzmann machines and DBN

- Restricted Boltzmann machines (RBMs) and deep belief networks (DBNs) are both generative models
- RBMs can be trained efficiently with contrastive divergence (CD-k) algorithm
- DBNs can be trained by first pre-trained each pair of layers as an RBM and then fine-tune with up-down algorithm
- DBNs are the earliest deep neural network model and essential the starting point of "deep learning" research

Why autoencoders? Dimension reduction

- As name suggests, the objective of dimension of reduction is to decrease the dimension of input signals to ease later processing
  - It is often a preprocessing step
  - Was commonly used to compress features
- It is a very old problem. The most representative algorithm is the principal component analysis (PCA)

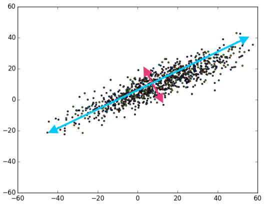
Why autoencoders? Dimension reduction

- As name suggests, the objective of dimension of reduction is to decrease the dimension of input signals to ease later processing
  - It is often a preprocessing step
  - Was commonly used to compress features
- It is a very old problem. The most representative algorithm is the principal component analysis (PCA)

Why autoencoders? Dimension reduction

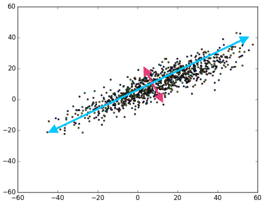
- As name suggests, the objective of dimension of reduction is to decrease the dimension of input signals to ease later processing
    - It is often a preprocessing step
    - Was commonly used to compress features
- It is a very old problem. The most representative algorithm is the principal component analysis (PCA)

Principal component analysis (PCA)



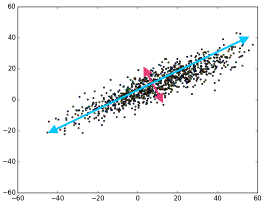
- Take N-dimensional data and find the M orthogonal directions in which the data have the most variance
  - We can represent an N-dimensional datapoint by its projections onto the M principal directions (i.e., with highest variances)
  - This loses all information about where the datapoint is located in the remaining orthogonal directions

Principal component analysis (PCA)



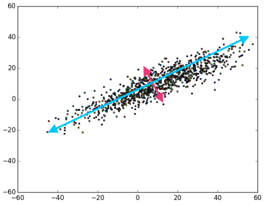
- Take N-dimensional data and find the M orthogonal directions in which the data have the most variance
  - We can represent an N-dimensional datapoint by its projections onto the M principal directions (i.e., with highest variances)
  - This loses all information about where the datapoint is located in the remaining orthogonal directions

Principal component analysis (PCA)



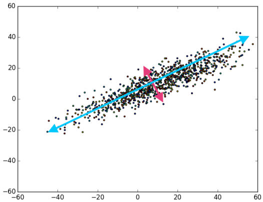
- Take N-dimensional data and find the M orthogonal directions in which the data have the most variance
  - We can represent an N-dimensional datapoint by its projections onto the M principal directions (i.e., with highest variances)
  - This loses all information about where the datapoint is located in the remaining orthogonal directions

PCA reconstruction



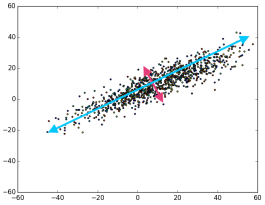
- We reconstruct by using the mean value (over all the data) on the $N - M$ directions that are not represented.
  - The reconstruction error is the sum over the variances over all these unrepresented directions
    - The variances are just eigenvalues of covariance matrix of the data
- PCA is "optimum"
  - Since we keep the largest variance components, on average the distortion is minimum among all linear dimension reduction methods

PCA reconstruction



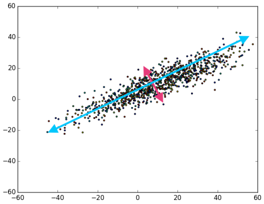
- We reconstruct by using the mean value (over all the data) on the $N - M$ directions that are not represented.
  - The reconstruction error is the sum over the variances over all these unrepresented directions
    - The variances are just eigenvalues of covariance matrix of the data
- PCA is "optimum"
  - Since we keep the largest variance components, on average the distortion is minimum among all linear dimension reduction methods

PCA reconstruction



- We reconstruct by using the mean value (over all the data) on the $N - M$ directions that are not represented.
  - The reconstruction error is the sum over the variances over all these unrepresented directions
    - The variances are just eigenvalues of covariance matrix of the data
- PCA is "optimum"
  - Since we keep the largest variance components, on average the distortion is minimum among all linear dimension reduction methods

PCA reconstruction



- We reconstruct by using the mean value (over all the data) on the $N - M$ directions that are not represented.
    - The reconstruction error is the sum over the variances over all these unrepresented directions
        - The variances are just eigenvalues of covariance matrix of the data
- PCA is "optimum"
    - Since we keep the largest variance components, on average the distortion is minimum among all linear dimension reduction methods

Math review: Singular value decomposition (SVD)

For any $N \times K$ matrix A (assume $K \leq N$), we can decompose it into product of three matrices

$$\begin{pmatrix} \\ A \\ \\ \end{pmatrix} = \begin{pmatrix} \\ U \\ \\ \end{pmatrix} \begin{pmatrix} \\ D \\ \\ \end{pmatrix} \begin{pmatrix} V \end{pmatrix}^T,$$

where U is $N \times N$, D is $N \times K$, and V is $K \times K$. Moreover,

- U is orthonormal, i.e., $U^T U = I$
- D is rectangular diagonal
- V is orthonormal, i.e., $V^T V = I$

Has nice geometric interpretation. Roughly speaking, any linear transform can be decompose into rotation, scaling, and rotation again

Math review: Singular value decomposition (SVD)

For any $N \times K$ matrix A (assume $K \leq N$), we can decompose it into product of three matrices

$$\begin{pmatrix} \\ & A & \\ \\ \end{pmatrix} = \begin{pmatrix} \\ & U & \\ \\ \end{pmatrix} \begin{pmatrix} \\ & D & \\ \\ \end{pmatrix} \begin{pmatrix} & V & \end{pmatrix}^T,$$

where U is $N \times N$, D is $N \times K$, and V is $K \times K$. Moreover,

- U is orthonormal, i.e., $U^TU = I$
- D is rectangular diagonal
- V is orthonormal, i.e., $V^TV = I$

Has nice geometric interpretation. Roughly speaking, any linear transform can be decompose into rotation, scaling, and rotation again

Math review: Singular value decomposition (SVD)

For any $N \times K$ matrix A (assume $K \leq N$), we can decompose it into product of three matrices

$$\left(\quad A \quad \right) = \left(\quad U \quad \right) \left(\quad D \quad \right) \left(\; V \; \right)^T,$$

where U is $N \times N$, D is $N \times K$, and V is $K \times K$. Moreover,

- U is orthonormal, i.e., $U^T U = I$
- D is rectangular diagonal
- V is orthonormal, i.e., $V^T V = I$

Has nice geometric interpretation. Roughly speaking, any linear transform can be decompose into rotation, scaling, and rotation again

SVD and PCA

- Let $X = [x_1, x_2, \cdots, x_K]$ be the matrix with columns as data vectors. We can decompose $X = U\Sigma V^T$ using SVD
- Assume X is zero-mean, the covariance matrix C is just $C \approx \frac{XX^T}{k}$
- Note that $C \sim U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T$, thus singular values are just square root of eigenvalues
  - Since PCA is in effect keeping the M largest eigenvalues of the covariance matrix, it is the same as keeping the M largest singular values of X
- One can easily verify that. Let $\hat{X} = U\hat{\Sigma}V^T$, where $\hat{\Sigma}$ only keeps the M largest singular values, then

$$
\begin{aligned}
Error = \sum_i (x-\hat{x})^T(x-\hat{x}) &= tr((X-\hat{X})^T(X-\hat{X})) \\
&= tr(V(\Sigma-\hat{\Sigma})U^T U(\Sigma-\hat{\Sigma})V^T) = tr(V(\Sigma-\hat{\Sigma})(\Sigma-\hat{\Sigma})V^T) \\
&= tr(((\Sigma-\hat{\Sigma})V^T)V(\Sigma-\hat{\Sigma})) = tr((\Sigma-\hat{\Sigma})^2) \\
&= \text{Sum of eigenvalues excluding the } M \text{ largest ones}
\end{aligned}
$$

SVD and PCA

- Let $X = [x_1, x_2, \cdots, x_K]$ be the matrix with columns as data vectors. We can decompose $X = U\Sigma V^T$ using SVD
- Assume X is zero-mean, the covariance matrix C is just $C \approx \frac{XX^T}{k}$
- Note that $C \sim U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T$, thus singular values are just square root of eigenvalues
  - Since PCA is in effect keeping the M largest eigenvalues of the covariance matrix, it is the same as keeping the M largest singular values of X
- One can easily verify that. Let $\hat{X} = U\hat{\Sigma}V^T$, where $\hat{\Sigma}$ only keeps the M largest singular values, then

$$
\begin{aligned}
Error = & \sum_i (x-\hat{x})^T(x-\hat{x}) = tr((X-\hat{X})^T(X-\hat{X})) \\
= & tr(V(\Sigma-\hat{\Sigma})U^TU(\Sigma-\hat{\Sigma})V^T) = tr(V(\Sigma-\hat{\Sigma})(\Sigma-\hat{\Sigma})V^T) \\
= & tr(((\Sigma-\hat{\Sigma})V^T)V(\Sigma-\hat{\Sigma})) = tr((\Sigma-\hat{\Sigma})^2) \\
= & \text{Sum of eigenvalues excluding the } M \text{ largest ones}
\end{aligned}
$$

SVD and PCA

- Let $X = [x_1, x_2, \cdots, x_K]$ be the matrix with columns as data vectors. We can decompose $X = U\Sigma V^T$ using SVD
- Assume X is zero-mean, the covariance matrix C is just $C \approx \frac{XX^T}{k}$
- Note that $C \sim U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T$, thus singular values are just square root of eigenvalues
  - Since PCA is in effect keeping the M largest eigenvalues of the covariance matrix, it is the same as keeping the M largest singular values of X
- One can easily verify that. Let $\hat{X} = U\hat{\Sigma}V^T$, where $\hat{\Sigma}$ only keeps the M largest singular values, then

$$
\begin{aligned}
Error = \sum_i (x - \hat{x})^T (x - \hat{x}) &= tr((X - \hat{X})^T (X - \hat{X})) \\
&= tr(V(\Sigma - \hat{\Sigma})U^T U(\Sigma - \hat{\Sigma})V^T) = tr(V(\Sigma - \hat{\Sigma})(\Sigma - \hat{\Sigma})V^T) \\
&= tr(((\Sigma - \hat{\Sigma})V^T)V(\Sigma - \hat{\Sigma})) = tr((\Sigma - \hat{\Sigma})^2) \\
&= \text{Sum of eigenvalues excluding the } M \text{ largest ones}
\end{aligned}
$$

SVD and PCA

- Let $X = [x_1, x_2, \cdots, x_K]$ be the matrix with columns as data vectors. We can decompose $X = U\Sigma V^T$ using SVD
- Assume X is zero-mean, the covariance matrix C is just $C \approx \frac{XX^T}{k}$
- Note that $C \sim U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T$, thus singular values are just square root of eigenvalues
  - Since PCA is in effect keeping the M largest eigenvalues of the covariance matrix, it is the same as keeping the M largest singular values of X
- One can easily verify that. Let $\hat{X} = U\hat{\Sigma}V^T$, where $\hat{\Sigma}$ only keeps the M largest singular values, then

$$
\begin{aligned}
Error = \sum_i (x-\hat{x})^T(x-\hat{x}) &= tr((X-\hat{X})^T(X-\hat{X})) \\
&= tr(V(\Sigma-\hat{\Sigma})U^T U(\Sigma-\hat{\Sigma})V^T) = tr(V(\Sigma-\hat{\Sigma})(\Sigma-\hat{\Sigma})V^T) \\
&= tr(((\Sigma-\hat{\Sigma})V^T)V(\Sigma-\hat{\Sigma})) = tr((\Sigma-\hat{\Sigma})^2) \\
&= \text{Sum of eigenvalues excluding the } M \text{ largest ones}
\end{aligned}
$$

SVD and PCA

- Let $X = [x_1, x_2, \cdots, x_K]$ be the matrix with columns as data vectors. We can decompose $X = U\Sigma V^T$ using SVD
- Assume X is zero-mean, the covariance matrix C is just $C \approx \frac{XX^T}{k}$
- Note that $C \sim U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T$, thus singular values are just square root of eigenvalues
  - Since PCA is in effect keeping the M largest eigenvalues of the covariance matrix, it is the same as keeping the M largest singular values of X
- One can easily verify that. Let $\hat{X} = U\hat{\Sigma}V^T$, where $\hat{\Sigma}$ only keeps the M largest singular values, then

$$Error = \sum_i (x - \hat{x})^T (x - \hat{x}) = tr((X - \hat{X})^T (X - \hat{X}))$$

$$= tr(V(\Sigma - \hat{\Sigma})U^T U(\Sigma - \hat{\Sigma})V^T) = tr(V(\Sigma - \hat{\Sigma})(\Sigma - \hat{\Sigma})V^T)$$

$$= tr(((\Sigma - \hat{\Sigma})V^T)V(\Sigma - \hat{\Sigma})) = tr((\Sigma - \hat{\Sigma})^2)$$

$=$Sum of eigenvalues excluding the M largest ones

SVD and PCA

- Let $X = [x_1, x_2, \cdots, x_K]$ be the matrix with columns as data vectors. We can decompose $X = U\Sigma V^T$ using SVD
- Assume X is zero-mean, the covariance matrix C is just $C \approx \frac{XX^T}{k}$
- Note that $C \sim U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T$, thus singular values are just square root of eigenvalues
  - Since PCA is in effect keeping the M largest eigenvalues of the covariance matrix, it is the same as keeping the M largest singular values of X
- One can easily verify that. Let $\hat{X} = U\hat{\Sigma}V^T$, where $\hat{\Sigma}$ only keeps the M largest singular values, then

$$Error = \sum_i (x - \hat{x})^T (x - \hat{x}) = tr((X - \hat{X})^T (X - \hat{X}))$$

$$= tr(V(\Sigma - \hat{\Sigma})U^T U(\Sigma - \hat{\Sigma})V^T) = tr(V(\Sigma - \hat{\Sigma})(\Sigma - \hat{\Sigma})V^T)$$

$$= tr(((\Sigma - \hat{\Sigma})V^T)V(\Sigma - \hat{\Sigma})) = tr((\Sigma - \hat{\Sigma})^2)$$

$$= \text{Sum of eigenvalues excluding the } M \text{ largest ones}$$

SVD and PCA

- Let $X = [x_1, x_2, \cdots, x_K]$ be the matrix with columns as data vectors. We can decompose $X = U\Sigma V^T$ using SVD
- Assume X is zero-mean, the covariance matrix C is just $C \approx \frac{XX^T}{k}$
- Note that $C \sim U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T$, thus singular values are just square root of eigenvalues
  - Since PCA is in effect keeping the M largest eigenvalues of the covariance matrix, it is the same as keeping the M largest singular values of X
- One can easily verify that. Let $\hat{X} = U\hat{\Sigma}V^T$, where $\hat{\Sigma}$ only keeps the M largest singular values, then

$$Error = \sum_i (x - \hat{x})^T (x - \hat{x}) = tr((X - \hat{X})^T (X - \hat{X}))$$

$$= tr(V(\Sigma - \hat{\Sigma})U^T U(\Sigma - \hat{\Sigma})V^T) = tr(V(\Sigma - \hat{\Sigma})(\Sigma - \hat{\Sigma})V^T)$$

$$= tr(((\Sigma - \hat{\Sigma})V^T)V(\Sigma - \hat{\Sigma})) = tr((\Sigma - \hat{\Sigma})^2)$$

$=$Sum of eigenvalues excluding the M largest ones

SVD and PCA

- Let $X = [x_1, x_2, \cdots, x_K]$ be the matrix with columns as data vectors. We can decompose $X = U\Sigma V^T$ using SVD
- Assume X is zero-mean, the covariance matrix C is just $C \approx \frac{XX^T}{k}$
- Note that $C \sim U\Sigma V^T(U\Sigma V^T)^T = U\Sigma^2 U^T$, thus singular values are just square root of eigenvalues
  - Since PCA is in effect keeping the M largest eigenvalues of the covariance matrix, it is the same as keeping the M largest singular values of X
- One can easily verify that. Let $\hat{X} = U\hat{\Sigma}V^T$, where $\hat{\Sigma}$ only keeps the M largest singular values, then

$$Error = \sum_i (x - \hat{x})^T (x - \hat{x}) = tr((X - \hat{X})^T (X - \hat{X}))$$

$$= tr(V(\Sigma - \hat{\Sigma})U^T U(\Sigma - \hat{\Sigma})V^T) = tr(V(\Sigma - \hat{\Sigma})(\Sigma - \hat{\Sigma})V^T)$$

$$= tr(((\Sigma - \hat{\Sigma})V^T)V(\Sigma - \hat{\Sigma})) = tr((\Sigma - \hat{\Sigma})^2)$$

$=$Sum of eigenvalues excluding the M largest ones

SVD and PCA

- Let $X = [x_1, x_2, \cdots, x_K]$ be the matrix with columns as data vectors. We can decompose $X = U\Sigma V^T$ using SVD
- Assume X is zero-mean, the covariance matrix C is just $C \approx \frac{XX^T}{k}$
- Note that $C \sim U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T$, thus singular values are just square root of eigenvalues
    - Since PCA is in effect keeping the M largest eigenvalues of the covariance matrix, it is the same as keeping the M largest singular values of X
- One can easily verify that. Let $\hat{X} = U\hat{\Sigma}V^T$, where $\hat{\Sigma}$ only keeps the M largest singular values, then

$$
\begin{aligned}
Error = &\sum_i (x - \hat{x})^T (x - \hat{x}) = tr((X - \hat{X})^T (X - \hat{X})) \\
=&tr(V(\Sigma - \hat{\Sigma})U^T U(\Sigma - \hat{\Sigma})V^T) = tr(V(\Sigma - \hat{\Sigma})(\Sigma - \hat{\Sigma})V^T) \\
=&tr(((\Sigma - \hat{\Sigma})V^T)V(\Sigma - \hat{\Sigma})) = tr((\Sigma - \hat{\Sigma})^2) \\
=&\text{Sum of eigenvalues excluding the } M \text{ largest ones}
\end{aligned}
$$

SVD and PCA

- Let $X = [x_1, x_2, \cdots, x_K]$ be the matrix with columns as data vectors. We can decompose $X = U\Sigma V^T$ using SVD
- Assume X is zero-mean, the covariance matrix C is just $C \approx \frac{XX^T}{k}$
- Note that $C \sim U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T$, thus singular values are just square root of eigenvalues
  - Since PCA is in effect keeping the M largest eigenvalues of the covariance matrix, it is the same as keeping the M largest singular values of X
- One can easily verify that. Let $\hat{X} = U\hat{\Sigma}V^T$, where $\hat{\Sigma}$ only keeps the M largest singular values, then

$$
\begin{aligned}
Error =& \sum_i (x - \hat{x})^T (x - \hat{x}) = tr((X - \hat{X})^T (X - \hat{X})) \\
=& tr(V(\Sigma - \hat{\Sigma})U^T U(\Sigma - \hat{\Sigma})V^T) = tr(V(\Sigma - \hat{\Sigma})(\Sigma - \hat{\Sigma})V^T) \\
=& tr(((\Sigma - \hat{\Sigma})V^T)V(\Sigma - \hat{\Sigma})) = tr((\Sigma - \hat{\Sigma})^2) \\
=& \text{Sum of eigenvalues excluding the } M \text{ largest ones}
\end{aligned}
$$

Optimal linear decoder ⇒ optimal linear encoder

- PCA is optimum when things are "linear"
- Interesting to know that as far as decoding is linear, the optimal encoding is linear (PCA) as well
    - That is, if $\hat{X} = Wh(X)$ for some optimal W
    - $\Rightarrow h(X) = TX$ for some optimal T

Optimal linear decoder ⇒ optimal linear encoder

- PCA is optimum when things are "linear"
- Interesting to know that as far as decoding is linear, the optimal encoding is linear (PCA) as well
  - That is, if $\hat{X} = Wh(X)$ for some optimal W
  - $\Rightarrow h(X) = TX$ for some optimal T

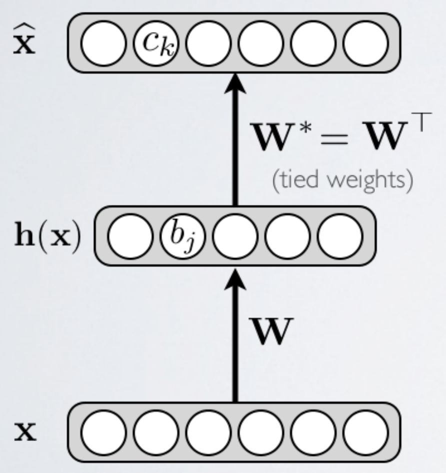
Optimal linear decoder ⇒ optimal linear encoder

- PCA is optimum when things are "linear"
- Interesting to know that as far as decoding is linear, the optimal encoding is linear (PCA) as well
  - That is, if $\hat{X} = Wh(X)$ for some optimal W
  - $\Rightarrow h(X) = TX$ for some optimal T

Optimal linear decoder $\Rightarrow$ optimal linear encoder

- PCA is optimum when things are “linear”
- Interesting to know that as far as decoding is linear, the optimal encoding is linear (PCA) as well
  - That is, if $\hat{X} = Wh(X)$ for some optimal W
  - $\Rightarrow h(X) = TX$ for some optimal T

Autoencoders

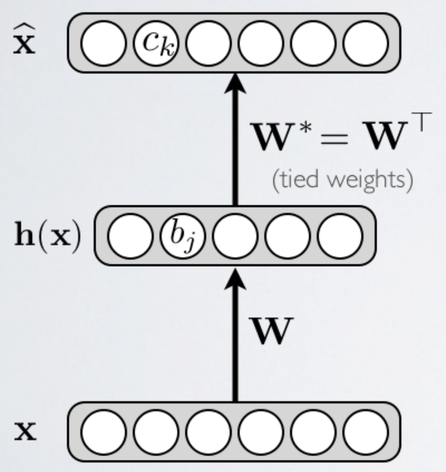


- Autoencoder is a way to perform dimension reduction with neural networks

$$h(x) = \text{sigm}(b + Wx)$$
$$\hat{x} = c + W^* h(x)$$

- loss $= \|x - \hat{x}\|$
- N.B., as the decoder is linear, the optimum autoencoder is just equivalent to PCA

Autoencoders

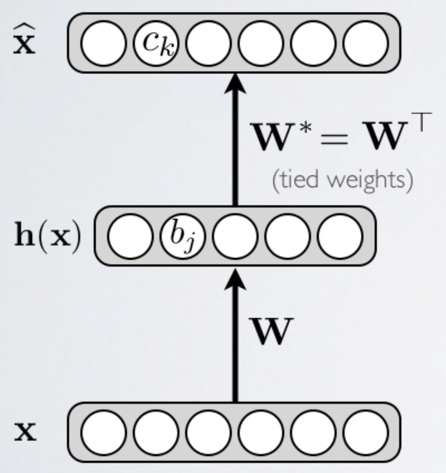


- Autoencoder is a way to perform dimension reduction with neural networks

$$h(x) = \text{sigm}(b + Wx)$$
$$\hat{x} = c + W^* h(x)$$

- loss $= \|x - \hat{x}\|$
- N.B., as the decoder is linear, the optimum autoencoder is just equivalent to PCA

Autoencoders

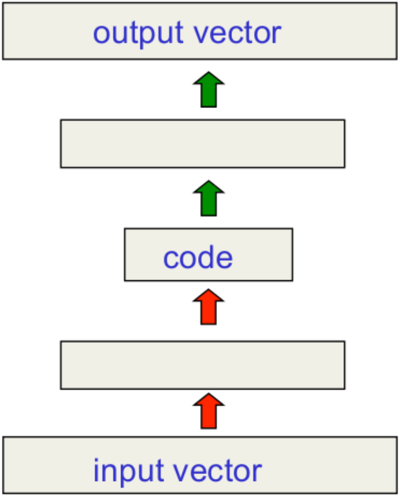


- Autoencoder is a way to perform dimension reduction with neural networks

$$h(x) = \text{sigm}(b + Wx)$$
$$\hat{x} = c + W^* h(x)$$

- loss $= \|x - \hat{x}\|$
- N.B., as the decoder is linear, the optimum autoencoder is just equivalent to PCA

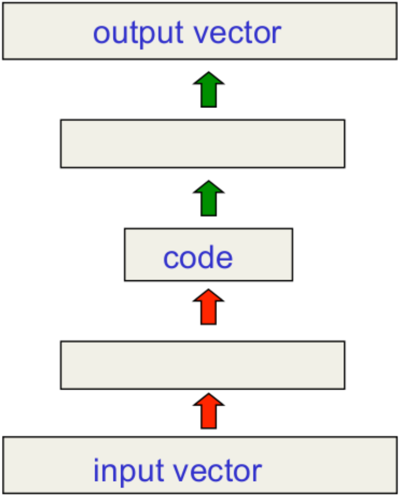
Deep autoencoders
Hinton & Salakhutdinov, Science 2006



- When using multiple layers, PCA is no longer optimal for continuous input
- The introduced nonlinearity can efficiently represent data that lies on a non-linear manifold
- It was an old idea (dated back to 80's) but it was considered to be very hard to train

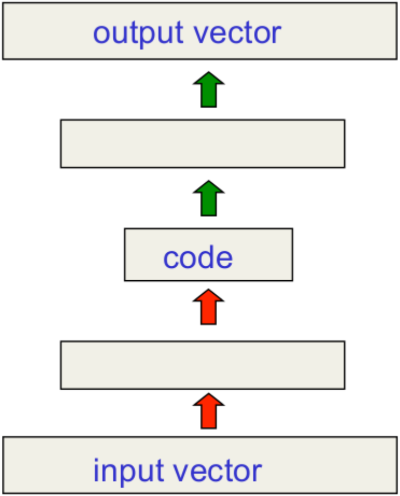
Deep autoencoders

Hinton & Salakhutdinov, Science 2006



- When using multiple layers, PCA is no longer optimal for continuous input
- The introduced nonlinearity can efficiently represent data that lies on a non-linear manifold
- It was an old idea (dated back to 80's) but it was considered to be very hard to train

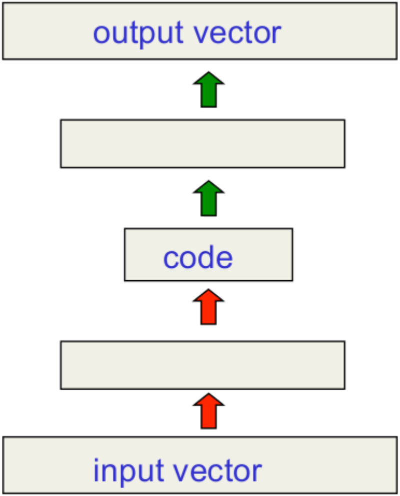
Deep autoencoders
Hinton & Salakhutdinov, Science 2006



- When using multiple layers, PCA is no longer optimal for continuous input
- The introduced nonlinearity can efficiently represent data that lies on a non-linear manifold
- It was an old idea (dated back to 80's) but it was considered to be very hard to train

Deep autoencoders

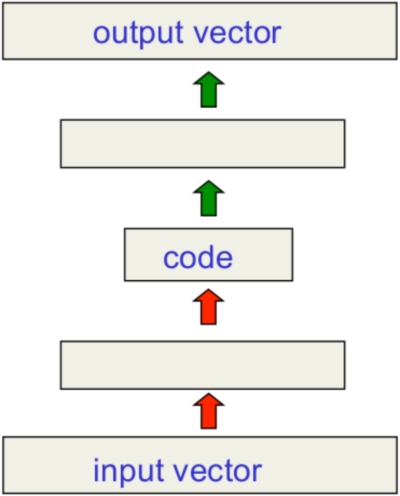
Hinton & Salakhutdinov, Science 2006



- First really successful deep autoencoder was trained in 2006 by Hinton's group
- It uses layer-by-layer RBM pre-training as described earlier
- Just use regular backprob for fine-tuning

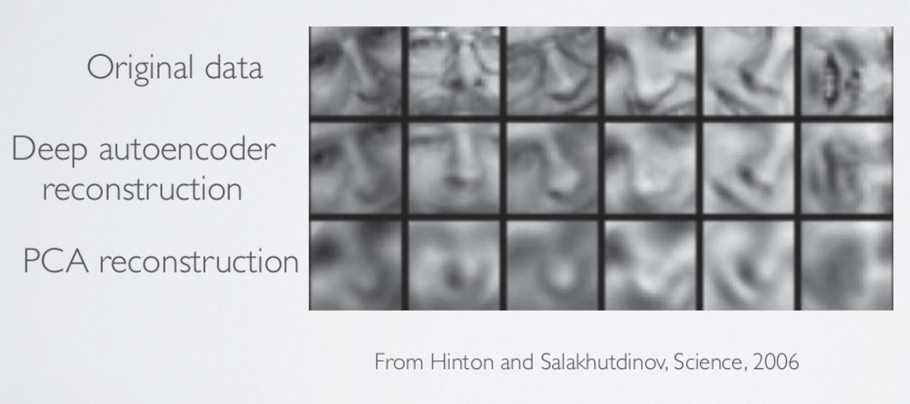
Deep autoencoders

Hinton & Salakhutdinov, Science 2006



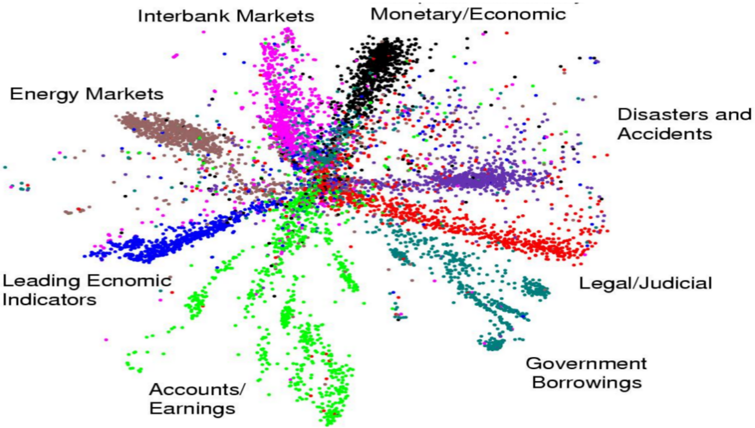
- First really successful deep autoencoder was trained in 2006 by Hinton's group
- It uses layer-by-layer RBM pre-training as described earlier
- Just use regular backprob for fine-tuning

Deep autoencoder vs PCA



Deep autoencoder for 400,000 business documents
Hinton 2006

First compress all documents to 2 numbers using deep auto.
Then use different colors for different document categories



Deep autoencoder for 400,000 image retrieval

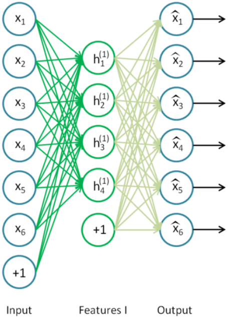
Hinton 2006

Leftmost column is the search image.

Other columns are the images that have the most similar feature activities in the last hidden layer.

Stacked autoencoders

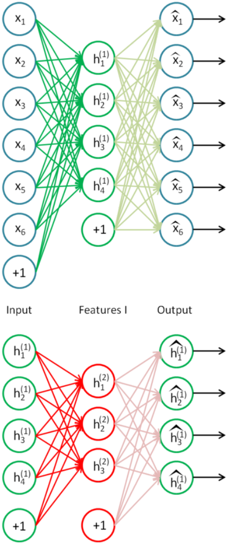
Alternative pretraining approach



- Besides pre-training using RBMs, we may also "expand" a deep autoencoders as a stack of shallow autoecoders
- Shallow autoencoders are easier to train than RBM

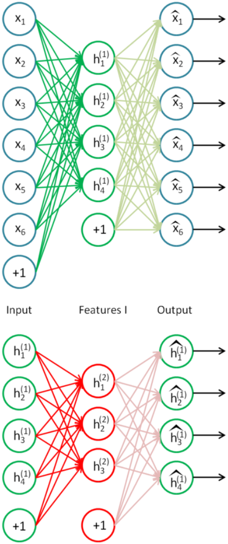
Stacked autoencoders

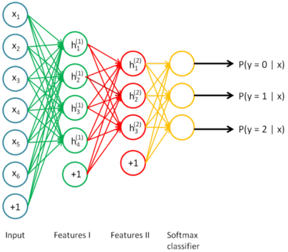
Alternative pretraining approach



- Besides pre-training using RBMs, we may also "expand" a deep autoencoders as a stack of shallow autoecoders
- Shallow autoencoders are easier to train than RBM

Stacked autoencoders
Alternative pretraining approach

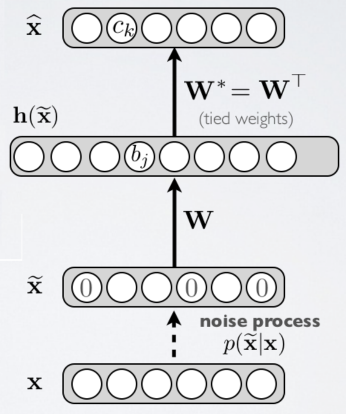




- Besides pre-training using RBMs, we may also "expand" a deep autoencoders as a stack of shallow autoecoders
- Shallow autoencoders are easier to train than RBM

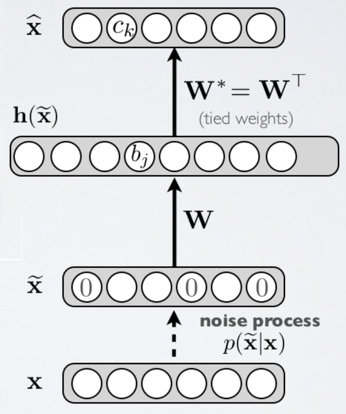
Denoising autoencoders

Vincent *et al.* 2008



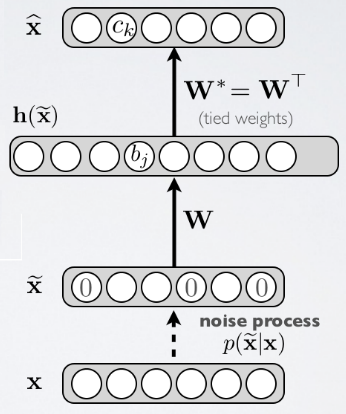
- Idea: representation should be robust to introduction of noise
  - Randomly assign bits to zero for binary case
    - Similar to dropout but for inputs only
  - Gaussian additive noise for continuous case
- Loss function compares $\hat{x}$ with noiseless input x

Denoising autoencoders
Vincent *et al.* 2008



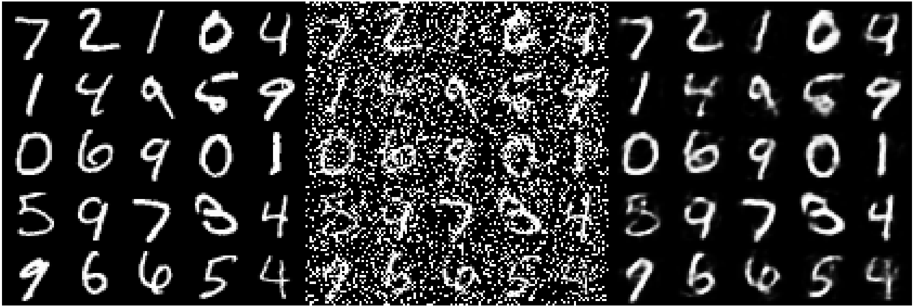
- Idea: representation should be robust to introduction of noise
  - Randomly assign bits to zero for binary case
    - Similar to dropout but for inputs only
  - Gaussian additive noise for continuous case
- Loss function compares $\hat{x}$ with noiseless input x

Denoising autoencoders
Vincent *et al.* 2008



- Idea: representation should be robust to introduction of noise
  - Randomly assign bits to zero for binary case
    - Similar to dropout but for inputs only
  - Gaussian additive noise for continuous case
- Loss function compares $\hat{x}$ with noiseless input x

Denoising autoencoders



Contractive autoencoders
Rifai *et al.* 2011

- Idea: encourage robustness of the model by forcing the hidden units to be insensitive to slight change of inputs
- Achieve this by penalizing the squared gradient of each hidden activity w.r.t. the inputs

$$L(x) \to L(x) + \lambda \|\nabla_x h(x)\|_F^2$$

- Pros and cons
  - + deterministic gradient ⇒ can use second order optimizers
  - + could be more stable than denoising autoencoder, which needs to use a sampled gradient
  - - Need to compute Jacobian of hidden layer
  - - More complex than denoising autoencoder, which just needs to add one two lines of code

Contractive autoencoders
Rifai *et al.* 2011

- Idea: encourage robustness of the model by forcing the hidden units to be insensitive to slight change of inputs
- Achieve this by penalizing the squared gradient of each hidden activity w.r.t. the inputs

$$L(x) \to L(x) + \lambda \|\nabla_x h(x)\|_F^2$$

- Pros and cons
  - + deterministic gradient ⇒ can use second order optimizers
  - + could be more stable than denoising autoencoder, which needs to use a sampled gradient
  - - Need to compute Jacobian of hidden layer
  - - More complex than denoising autoencoder, which just needs to add one two lines of code

Contractive autoencoders
Rifai *et al.* 2011

- Idea: encourage robustness of the model by forcing the hidden units to be insensitive to slight change of inputs
- Achieve this by penalizing the squared gradient of each hidden activity w.r.t. the inputs

$$L(x) \to L(x) + \lambda \|\nabla_x h(x)\|_F^2$$

- Pros and cons
  - + deterministic gradient ⇒ can use second order optimizers
  - + could be more stable than denoising autoencoder, which needs to use a sampled gradient
  - - Need to compute Jacobian of hidden layer
  - - More complex than denoising autoencoder, which just needs to add one two lines of code

Contractive autoencoders
Rifai *et al.* 2011

- Idea: encourage robustness of the model by forcing the hidden units to be insensitive to slight change of inputs
- Achieve this by penalizing the squared gradient of each hidden activity w.r.t. the inputs

$$L(x) \to L(x) + \lambda \|\nabla_x h(x)\|_F^2$$

- Pros and cons
  - + deterministic gradient ⇒ can use second order optimizers
  - + could be more stable than denoising autoencoder, which needs to use a sampled gradient
  - - Need to compute Jacobian of hidden layer
  - - More complex than denoising autoencoder, which just needs to add one two lines of code

Contractive autoencoders
Rifai *et al.* 2011

- Idea: encourage robustness of the model by forcing the hidden units to be insensitive to slight change of inputs
- Achieve this by penalizing the squared gradient of each hidden activity w.r.t. the inputs

$$L(x) \to L(x) + \lambda \|\nabla_x h(x)\|_F^2$$

- Pros and cons
  - $+$ deterministic gradient $\Rightarrow$ can use second order optimizers
  - $+$ could be more stable than denoising autoencoder, which needs to use a sampled gradient
  - - Need to compute Jacobian of hidden layer
  - - More complex than denoising autoencoder, which just needs to add one two lines of code

Contractive autoencoders
Rifai *et al.* 2011

- Idea: encourage robustness of the model by forcing the hidden units to be insensitive to slight change of inputs
- Achieve this by penalizing the squared gradient of each hidden activity w.r.t. the inputs

$$L(x) \to L(x) + \lambda \|\nabla_x h(x)\|_F^2$$

- Pros and cons
  - $+$ deterministic gradient $\Rightarrow$ can use second order optimizers
  - $+$ could be more stable than denoising autoencoder, which needs to use a sampled gradient
  - - Need to compute Jacobian of hidden layer
  - - More complex than denoising autoencoder, which just needs to add one two lines of code

Remark on pretraining

What are the disadvantages of pretraining deep neural networks by stacking autoencoders?

Answer Request Follow 55 Comment Downvote

1 Answer

Yoshua Bengio, My lab has been one of the three that started the deep learning approach, back in 2006, along with Hinton's...

Answered Aug 14, 2014 · Upvoted by Zeeshan Zia, PhD in Computer Vision and Machine Learning and Jason Li, AI researcher.

The same disadvantage as other layer-wise pre-training techniques: it is greedy, i.e., it does not try to tune the lower layers in a way that will make the work of higher layers easier. But that will change soon with a new approach I am working on!

Remark on pretraining

Ian Goodfellow, Lead author of the Deep Learning textbook: http://www.deeplearningbook.org

Answered Sep 28, 2016 · Upvoted by Aaditya Prakash, Graduate student in Computer Vision and Deep Learning and Abhinav Maurya, PhD Student (Machine Learning, Public Policy) at CMU

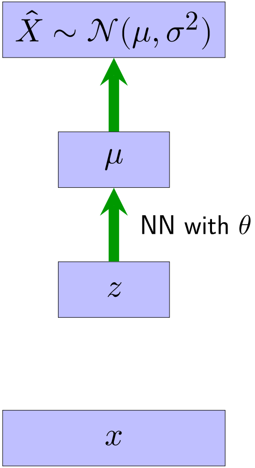
Autoencoders are useful for some things, but turned out not to be nearly as necessary as we once thought. Around 10 years ago, we thought that deep nets would not learn correctly if trained with only backprop of the supervised cost. We thought that deep nets would also need an unsupervised cost, like the autoencoder cost, to regularize them. When Google Brain built their first very large neural network to recognize objects in images, it was an autoencoder (and it didn't work very well at recognizing objects compared to later approaches). Today, we know we are able to recognize images just by using backprop on the supervised cost as long as there is enough labeled data. There are other tasks where we do still use autoencoders, but they're not the fundamental solution to training deep nets that people once thought they were going to be.

Variational autoencoders

"Generative autoencoders" $\Rightarrow$ variational autoencoders

- Instead of spitting out an approximate for the input
- The network spits out parameters of a distribution

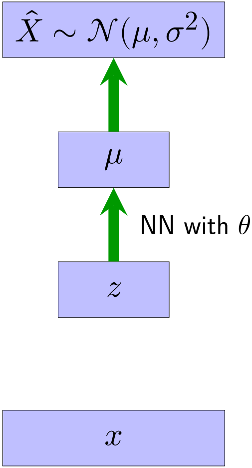
Variational autoencoder
Kingma and Willing 2014



- $p(z|x) = \frac{p(z)p_\theta(x|z)}{p(x)} = \frac{p(z)p_\theta(x|z)}{\int p(z)p_\theta(x|z)dz}$
- For simplicity, pick $p(z) = \mathcal{N}(z; 0, 1)$ and $p_\theta(x|z) = \mathcal{N}(\mu, \sigma^2)$, the posterior $p(z|x)$ is still intractable since computing $p(x)$ needs to integrate over all possible z
- We might use MAP or Monte Carlo sampling (MCMC) to estimate $p(z|x)$ but
  - MAP: - too biased
  - MCMC: - too expensive
  - $\Rightarrow$ Variational inference

Variational autoencoder

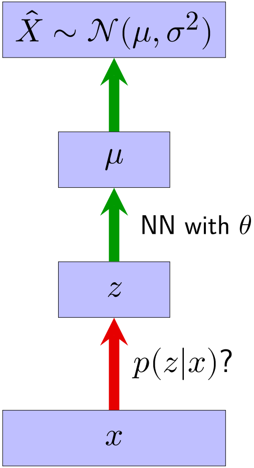
Kingma and Willing 2014



- $p(z|x) = \frac{p(z)p_\theta(x|z)}{p(x)} = \frac{p(z)p_\theta(x|z)}{\int p(z)p_\theta(x|z)dz}$
- For simplicity, pick $p(z) = \mathcal{N}(z; 0, 1)$ and $p_\theta(x|z) = \mathcal{N}(\mu, \sigma^2)$, the posterior $p(z|x)$ is still intractable since computing $p(x)$ needs to integrate over all possible z
- We might use MAP or Monte Carlo sampling (MCMC) to estimate $p(z|x)$ but
  - MAP: - too biased
  - MCMC: - too expensive
  - $\Rightarrow$ Variational inference

Variational autoencoder

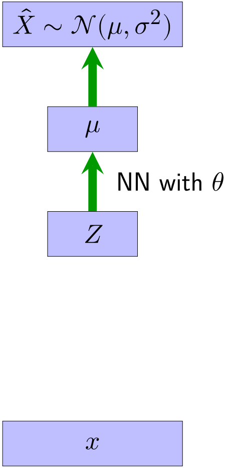
Kingma and Willing 2014



- $p(z|x) = \frac{p(z)p_\theta(x|z)}{p(x)} = \frac{p(z)p_\theta(x|z)}{\int p(z)p_\theta(x|z)dz}$
- For simplicity, pick $p(z) = \mathcal{N}(z; 0, 1)$ and $p_\theta(x|z) = \mathcal{N}(\mu, \sigma^2)$, the posterior $p(z|x)$ is still intractable since computing $p(x)$ needs to integrate over all possible z
- We might use MAP or Monte Carlo sampling (MCMC) to estimate $p(z|x)$ but
    - MAP: - too biased
    - MCMC: - too expensive
    - $\Rightarrow$ Variational inference

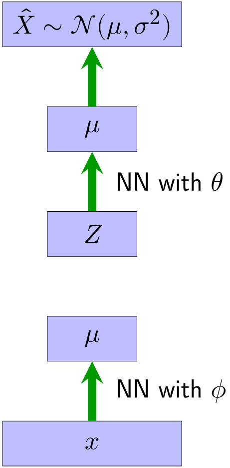
Variational autoencoder

Kingma and Willing 2014



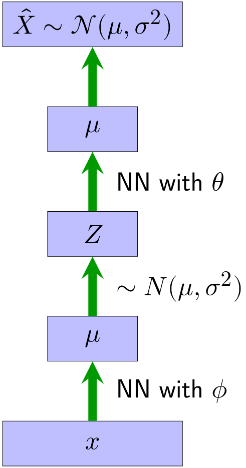
- Instead of trying to find the exact posterior $p(z|x)$, approximate it as a Gaussian distribution with parameters obtained through an NN
- Unfortunately, the loss $-\log p(x)$ is still intractable, but we can approximate $\log p(x)$ with a lower bound
- Instead of minimizing the loss, or maximizing $\log p(x)$ directly, we will maximize its lower bound instead

Variational autoencoder
Kingma and Willing 2014



- Instead of trying to find the exact posterior $p(z|x)$, approximate it as a Gaussian distribution with parameters obtained through an NN
- Unfortunately, the loss $-\log p(x)$ is still intractable, but we can approximate $\log p(x)$ with a lower bound
- Instead of minimizing the loss, or maximizing $\log p(x)$ directly, we will maximize its lower bound instead

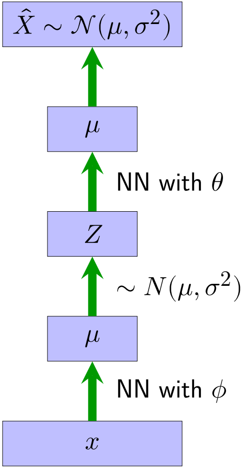
Variational autoencoder
Kingma and Willing 2014



- Instead of trying to find the exact posterior $p(z|x)$, approximate it as a Gaussian distribution with parameters obtained through an NN
- Unfortunately, the loss $-\log p(x)$ is still intractable, but we can approximate $\log p(x)$ with a lower bound
- Instead of minimizing the loss, or maximizing $\log p(x)$ directly, we will maximize its lower bound instead

Variational autoencoder

Kingma and Willing 2014



- Instead of trying to find the exact posterior $p(z|x)$, approximate it as a Gaussian distribution with parameters obtained through an NN
- Unfortunately, the loss $-\log p(x)$ is still intractable, but we can approximate $\log p(x)$ with a lower bound
- Instead of minimizing the loss, or maximizing $\log p(x)$ directly, we will maximize its lower bound instead

Variational lower bound (EBLO)

$$\log p(x) = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} \frac{q_\phi(z|x)}{q_\phi(z|x)}$$

$$= \log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}$$

Since the above is true for all z,

$$\log p(x) = E_{Z\sim q_\phi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}\right]$$

$$= \underbrace{E_{Z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)\|p(z))}_{\text{EBLO}(x,\theta,\phi)\ \text{“Evidence Lower BOund”}} + \underbrace{KL(q_\phi(z|x)\,\|p(z|x))}_{\geq 0}$$

Training: $\theta^*, \phi^* = \arg\max_{\theta,\phi} \sum_i \text{EBLO}(x^{(i)}, \theta, \phi)$

Variational lower bound (EBLO)

$$\log p(x) = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} \frac{q_\phi(z|x)}{q_\phi(z|x)}$$

$$= \log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}$$

Since the above is true for all z,

$$\log p(x) = E_{Z\sim q_\phi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}\right]$$

$$= \underbrace{E_{Z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)\|p(z))}_{\text{EBLO}(x,\theta,\phi)\ \text{“Evidence Lower BOund”}} + \underbrace{KL(q_\phi(z|x)\ \|p(z|x))}_{\geq 0}$$

Training: $\theta^*, \phi^* = \arg\max_{\theta,\phi} \sum_i \text{EBLO}(x^{(i)}, \theta, \phi)$

Variational lower bound (EBLO)

$$\log p(x) = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} \frac{q_\phi(z|x)}{q_\phi(z|x)}$$
$$= \log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}$$

Since the above is true for all z,

$$\log p(x) = E_{Z\sim q_\phi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}\right]$$
$$= \underbrace{E_{Z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)\|p(z))}_{\text{EBLO}(x,\theta,\phi)\text{ "Evidence Lower BOund"}} + \underbrace{KL(q_\phi(z|x)\,\|p(z|x))}_{\geq 0}$$

Training: $\theta^*, \phi^* = \arg\max_{\theta,\phi} \sum_i \text{EBLO}(x^{(i)}, \theta, \phi)$

Variational lower bound (EBLO)

$$\log p(x) = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} \frac{q_\phi(z|x)}{q_\phi(z|x)}$$
$$= \log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}$$

Since the above is true for all z,

$$\log p(x) = E_{Z\sim q_\phi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}\right]$$
$$= \underbrace{E_{Z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)\|p(z))}_{\text{EBLO}(x,\theta,\phi)\text{ ``Evidence Lower BOund''}} + \underbrace{KL(q_\phi(z|x)\,\|p(z|x))}_{\geq 0}$$

Training: $\theta^*, \phi^* = \arg\max_{\theta,\phi} \sum_i \text{EBLO}(x^{(i)}, \theta, \phi)$

Variational lower bound (EBLO)

$$\log p(x) = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} \frac{q_\phi(z|x)}{q_\phi(z|x)}$$
$$= \log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}$$

Since the above is true for all z,

$$\log p(x) = E_{Z\sim q_\phi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}\right]$$
$$= \underbrace{E_{Z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)\|p(z))}_{\text{EBLO}(x,\theta,\phi)\ \text{“Evidence Lower BOund”}} + \underbrace{KL(q_\phi(z|x)\ \|p(z|x))}_{\geq 0}$$

Training: $\theta^*, \phi^* = \arg\max_{\theta,\phi} \sum_i \text{EBLO}(x^{(i)}, \theta, \phi)$

Variational lower bound (EBLO)

$$\log p(x) = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} = \log \frac{p_\theta(x|z)p(z)}{p(z|x)} \frac{q_\phi(z|x)}{q_\phi(z|x)}$$
$$= \log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}$$

Since the above is true for all z,

$$\log p(x) = E_{Z\sim q_\phi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p(z)} + \log \frac{q_\phi(z|x)}{p(z|x)}\right]$$
$$= \underbrace{E_{Z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)\|p(z))}_{\text{EBLO}(x,\theta,\phi)\text{ ``Evidence Lower BOund''}} + \underbrace{KL(q_\phi(z|x)\ \|p(z|x))}_{\geq 0}$$

Training: $\theta^*, \phi^* = \arg\max_{\theta,\phi} \sum_i \text{EBLO}(x^{(i)}, \theta, \phi)$

Variational autoencoder
Kingma and Willing 2014

Maximizing EBLO means that:

- Want small $KL(q_\phi(z|x)\|p(z))$ (the difference between the approx distribution from $p(z)$)
  - This turns out to have closed-form solution since we are dealing with Gaussian distributions
- Want large $E_{Z\sim q_\phi(z|x)}[\log p_\theta(x|z)]$ (expected log prob of the evidence with approx distribution)
  - need to backprop through a random node z
  - can be solved by the "reparametrization trick"

Variational autoencoder

Kingma and Willing 2014

Maximizing EBLO means that:

- Want small $KL(q_\phi(z|x)\|p(z))$ (the difference between the approx distribution from $p(z)$)
  - This turns out to have closed-form solution since we are dealing with Gaussian distributions
- Want large $E_{Z\sim q_\phi(z|x)}[\log p_\theta(x|z)]$ (expected log prob of the evidence with approx distribution)
  - need to backprop through a random node z
  - can be solved by the "reparametrization trick"

Variational autoencoder
Kingma and Willing 2014

Maximizing EBLO means that:

- Want small $KL(q_\phi(z|x)\|p(z))$ (the difference between the approx distribution from $p(z)$)
  - This turns out to have closed-form solution since we are dealing with Gaussian distributions
- Want large $E_{Z\sim q_\phi(z|x)}[\log p_\theta(x|z)]$ (expected log prob of the evidence with approx distribution)
  - need to backprop through a random node z
  - can be solved by the "reparametrization trick"

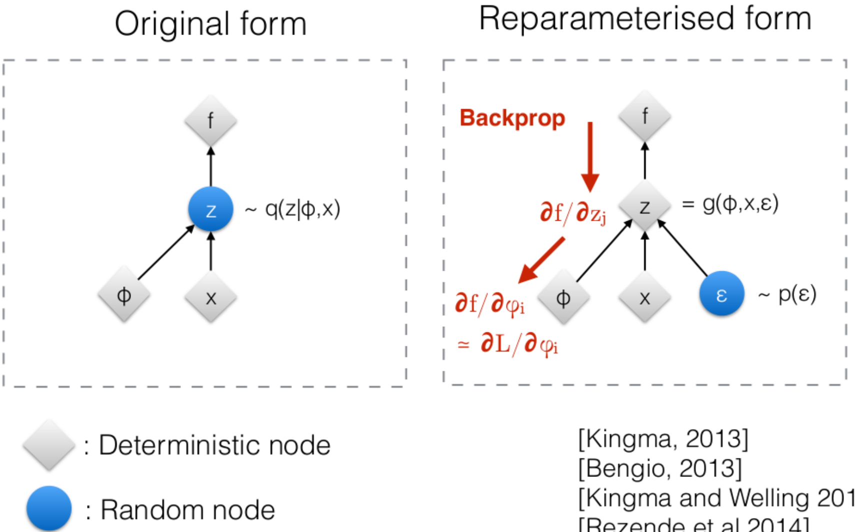
Variational autoencoder

Kingma and Willing 2014

Maximizing EBLO means that:

- Want small $KL(q_\phi(z|x)\|p(z))$ (the difference between the approx distribution from $p(z)$)
  - This turns out to have closed-form solution since we are dealing with Gaussian distributions
- Want large $E_{Z\sim q_\phi(z|x)}[\log p_\theta(x|z)]$ (expected log prob of the evidence with approx distribution)
  - need to backprop through a random node z
  - can be solved by the "reparametrization trick"

Reparametrization trick



[Kingma, 2013]
[Bengio, 2013]
[Kingma and Welling 2014]
[Rezende et al 2014]

Variational autoencoders

Variational Autoencoders

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

Variational autoencoders

Variational Autoencoders

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)},\theta,\phi)}$$

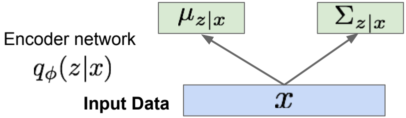
Let's look at computing the bound (forward pass) for a given minibatch of input data

Variational autoencoders

Variational Autoencoders

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)},\theta,\phi)}$$

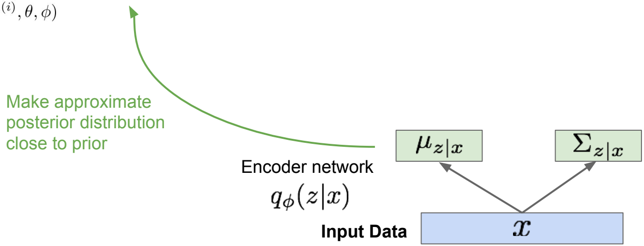


Variational autoencoders

Variational Autoencoders

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$



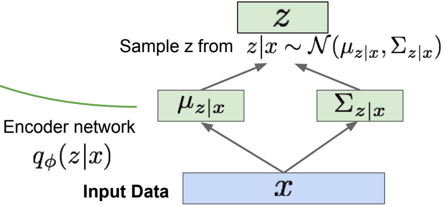
Variational autoencoders

Variational Autoencoders

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

Make approximate posterior distribution close to prior



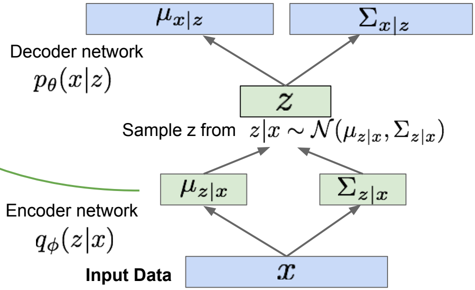
Variational autoencoders

Variational Autoencoders

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

Make approximate posterior distribution close to prior



Variational autoencoders

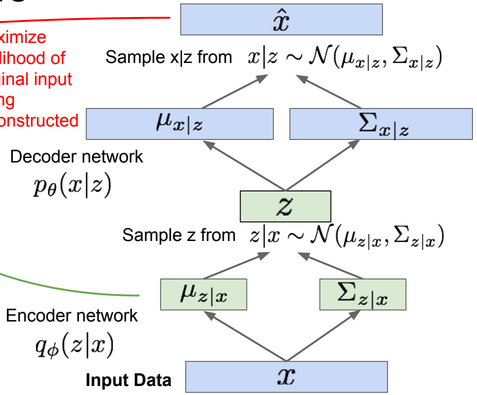
Variational Autoencoders

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)},\theta,\phi)}$$

Maximize likelihood of original input being reconstructed

Make approximate posterior distribution close to prior



Fei-Fei Li & Justin Johnson & Serena Yeung Lecture 13 - 89 May 18, 2017

Variational autoencoders

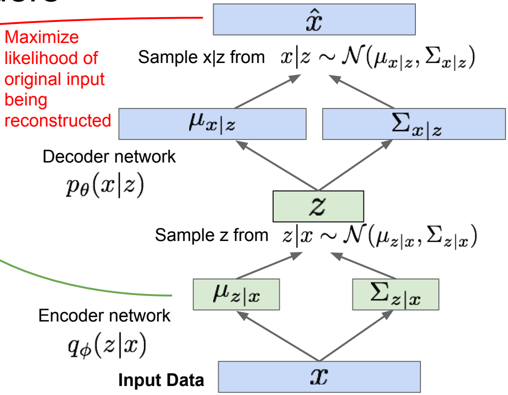
Variational Autoencoders

Putting it all together: maximizing the likelihood lower bound

$$\underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\mathcal{L}(x^{(i)},\theta,\phi)}$$

Make approximate posterior distribution close to prior

For every minibatch of input data: compute this forward pass, and then backprop!

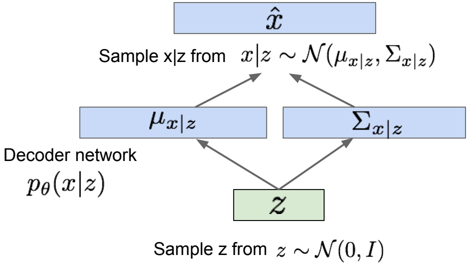


Fei-Fei Li & Justin Johnson & Serena Yeung Lecture 13 - 90 May 18, 2017

Variational autoencoders

Variational Autoencoders: Generating Data!

Use decoder network. Now sample z from prior!



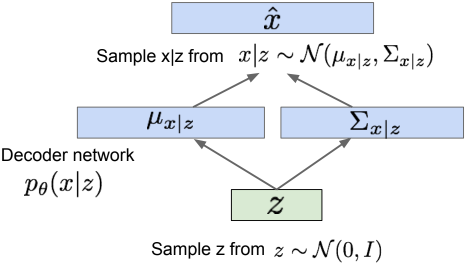
Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

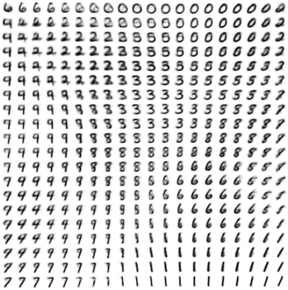
Fei-Fei Li & Justin Johnson & Serena Yeung Lecture 13 - 91 May 18, 2017

Variational autoencoders

Variational Autoencoders: Generating Data!

Use decoder network. Now sample z from prior!



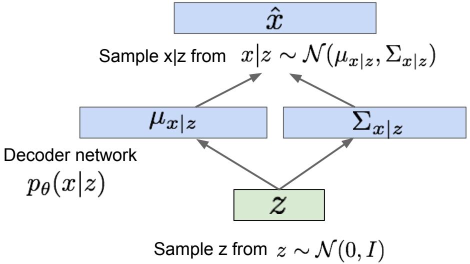


Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

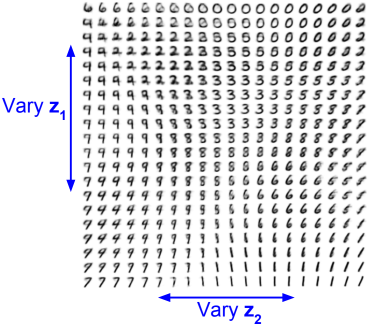
Variational autoencoders

Variational Autoencoders: Generating Data!

Use decoder network. Now sample z from prior!



Data manifold for 2-d **z**



Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

Variational autoencoders

Variational Autoencoders: Generating Data!

Diagonal prior on **z** => independent latent variables

Different dimensions of **z** encode interpretable factors of variation

Degree of smile

Vary $\mathbf{z}_1$

Vary $\mathbf{z}_2$

Head pose

Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

Fei-Fei Li & Justin Johnson & Serena Yeung Lecture 13 - 94 May 18, 2017

Variational autoencoders

Variational Autoencoders: Generating Data!

Diagonal prior on **z** => independent latent variables

Different dimensions of **z** encode interpretable factors of variation

Also good feature representation that can be computed using $q_\phi(z|x)$!

Degree of smile

Vary $\mathbf{z}_1$

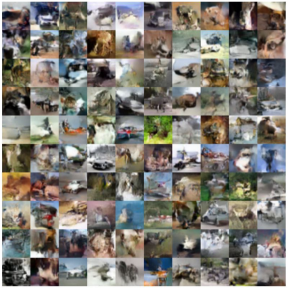
Vary $\mathbf{z}_2$

Head pose

Kingma and Welling, "Auto-Encoding Variational Bayes", ICLR 2014

Variational autoencoders

Variational Autoencoders: Generating Data!



32x32 CIFAR-10

Labeled Faces in the Wild

Summary of variational autoencoders

- Probabilistic spin to traditional autoencoders to allow data generation. Use variational lower bound to workaround intractable density estimation

  Pros
  - Systematic approach to generative models (train end-to-end)
  - Allows inference of $q_\phi(z|x)$ that can be used for feature representation

  Cons
  - Maximizes lower bound rather than exact cost function. Less direct than say PixelRNN/PixelCNN
  - Samples generated are lower quality compared to the state-of-the-art (GANs)
- Follow-up research:
  - More flexible approximations, e.g., richer model in approximating the posterior (typically just use diagonal Gaussian in the basic model)
  - Incorporating structure in latent variables
  - Disentangled variational autoencoder

Conclusions

- Conventional autoencoders are important tools for dimension reduction and data representation in general
- Generative models are some very exciting hot topics in deep learning
  - Especially useful for datasets with few or no labels
  - Many other possible applications yet to be discovered
- We discuss several generative models, in particular
  - Variational autoencoders: autoencoders + variational inference
  - Generative adversarial networks (GANs): more recent and gaining lots of interests

Conclusions

- Conventional autoencoders are important tools for dimension reduction and data representation in general
- Generative models are some very exciting hot topics in deep learning
    - Especially useful for datasets with few or no labels
    - Many other possible applications yet to be discovered
- We discuss several generative models, in particular
    - Variational autoencoders: autoencoders + variational inference
    - Generative adversarial networks (GANs): more recent and gaining lots of interests

Conclusions

- Conventional autoencoders are important tools for dimension reduction and data representation in general
- Generative models are some very exciting hot topics in deep learning
  - Especially useful for datasets with few or no labels
  - Many other possible applications yet to be discovered
- We discuss several generative models, in particular
  - Variational autoencoders: autoencoders + variational inference
  - Generative adversarial networks (GANs): more recent and gaining lots of interests