Topics in AI (CPSC 532S): Multimodal Learning with Vision, Language and Sound

Lecture 15: Generative Models
Supervised vs. Unsupervised Learning

**Supervised Learning**

**Data:** \((x, y)\)
- \(x\) is data, \(y\) is label

**Goal:** Learn a function to map \(x \rightarrow y\)

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

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, [cs231n Stanford](https://cs231n.stanford.edu)
Supervised vs. Unsupervised Learning

**Supervised Learning**

**Data:** \((x, y)\)
- \(x\) is data, \(y\) is label

**Goal:** Learn a function to map \(x \rightarrow y\)

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

Object Detection

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, [cs231n Stanford](https://cs231n.stanford.edu)
Supervised vs. Unsupervised Learning

**Supervised Learning**

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

Goal: Learn a function to map \(x \rightarrow y\)

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

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

**Goal:** Learn a function to map $x \rightarrow y$

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

---

A cat sitting on a suitcase on the floor

---

* This image is CC0 public domain

---

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, [cs231n Stanford](https://cs231n.stanford.edu)
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

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs213n Stanford
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.

dimensionality reduction

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
**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.*

---

*Figure copyright Ian Goodfellow, 2016. Reproduced with permission.*

---

1-dim density estimation

2-dim density estimation

---

* 2-d density images *left* and *right* are CC0 public domain

---

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, [cs231n Stanford](https://www.cs231n.com)
## Supervised vs. Unsupervised Learning

<table>
<thead>
<tr>
<th>Supervised Learning</th>
<th>Unsupervised Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data:</strong> (x, y)</td>
<td><strong>Data:</strong> x</td>
</tr>
<tr>
<td>x is data, y is label</td>
<td>Just data, no labels!</td>
</tr>
<tr>
<td><strong>Goal:</strong> Learn a function to map x→y</td>
<td><strong>Goal:</strong> Learn some underlying hidden structure of the data</td>
</tr>
<tr>
<td><strong>Examples:</strong> Classification, regression, object detection, semantic segmentation, image captioning, etc.</td>
<td><strong>Examples:</strong> Clustering, dimensionality reduction, feature learning, density estimation, etc.</td>
</tr>
</tbody>
</table>

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Generative Models

Given training data, generate new samples from the same distribution

Training data \( \sim p_{\text{data}}(x) \)

Generated samples \( \sim p_{\text{model}}(x) \)

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

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Generative Models

Given training data, generate new samples from the same distribution

Training data $\sim p_{\text{data}}(x)$

Generated samples $\sim p_{\text{model}}(x)$

Addresses density estimation, a core problem in unsupervised learning

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

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
**Taxonomy of Generative Models**

- **Generative models**
  - **Explicit density**
    - Tractable density
      - Fully Visible Belief Nets
        - NADE
        - MADE
        - PixelRNN/CNN
        - Change of variables models (nonlinear ICA)
  - Approximate density
    - Variational
    - Markov Chain
      - Variational Autoencoder
  - Implicit density
    - Markov Chain
      - GSN
      - GAN

* Figure copyright and adapted from Ian Goodfellow, Tutorial on Generative Adversarial Networks, 2017.
* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Taxonomy of Generative Models

Generative models

Explicit density

Tractable density
- Fully Visible Belief Nets
  - NADE
  - MADE
- PixelRNN/CNN
  Change of variables models (nonlinear ICA)

Approximate density
- Variational Autoencoder

Implicit density

Markov Chain
- Variational
- Markov Chain
  - GSN
- Boltzmann Machine
- GAN

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

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Why **Generative** Models?

— Realistic **samples** for artwork, super-resolution, colorization, *etc.*
Why **Generative** Models?

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

- Generative models of time-series data can be used for **simulation**, **predictions** and planning (reinforcement learning applications)

- Training generative models can also enable inference of latent representation that can be useful as **general features**

- **Dreaming** / hypothesis visualization

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, cs231n Stanford
PixelRNN and PixelCNN
Explicit Density model

Use chain rule to decompose likelihood of an image $x$ into product of (many) 1-d distributions

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

then maximize likelihood of training data

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
PixelRNN

R

G

B

Histograms of pixel values for R, G, and B channels.
**Explicit** Density model

Use chain rule to decompose likelihood of an image $x$ into product of (many) 1-d distributions

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

- Likelihood of image $x$
- Probability of i’th pixel value given all previous pixels

then maximize likelihood of training data

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford

[ van der Oord et al., 2016 ]
Explicit Density model

Use chain rule to decompose likelihood of an image $x$ into product of (many) 1-d distributions

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

then maximize likelihood of training data

Complex distribution over pixel values, so let's model using neural network

Also requires defining ordering of “previous pixels”
PixelRNN

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford

[ van der Oord et al., 2016 ]
Generate image pixels starting from the corner

Dependency on previous pixels model using an RNN (LSTM)

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Generate image pixels starting from the corner

Dependency on previous pixels model using an RNN (LSTM)

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford

[ van der Oord et al., 2016 ]
Generate image pixels starting from the corner

Dependency on previous pixels model using an RNN (LSTM)

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
PixelRNN

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford

[ van der Oord et al., 2016 ]
Generate image pixels starting from the corner

Dependency on previous pixels model using an RNN (LSTM)
Generate image pixels starting from the corner

Dependency on previous pixels model using an RNN (LSTM)

**Problem:** sequential generation is slow

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, *cs231n Stanford*
Still generate image pixels starting from the corner

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

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Still generate image pixels starting from the corner

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

**Training:** maximize likelihood of training images

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

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Still generate image pixels starting from the corner

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

**Training**: maximize likelihood of training images

\[ p(x) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1}) \]

Generation is still slow (sequential), but learning is faster

*slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford*
Generated Samples

32x32 CIFAR-10

32x32 ImageNet

[ van der Oord et al., 2016 ]

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
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…

* slide from Fei-Fei Li, Justin Johnson, Serena Yeung, cs231n Stanford
Multi-scale PixelRNN

Take sub-sampled pixels as additional input pixels

Can capture better global information (more visually coherent)

[ van der Oord et al., 2016 ]

* slide from Hsiao-Ching Chang, Ameya Patil, Anand Bhattad
Multi-scale PixelRNN

[ van der Oord et al., 2016 ]

* slide from Hsiao-Ching Chang, Ameya Patil, Anand Bhattad
Conditional Image Generation

Similar to PixelRNN/CNN but conditioned on a high-level image description vector $h$

$$p(x) = p(x_1, x_2, \ldots, x_{n^2})$$

$$p(x|h) = p(x_1, x_2, \ldots, x_{n^2}|h)$$
Conditional Image Generation

[ van der Oord et al., 2016 ]

African elephant

Sandbar

* slide from Hsiao-Ching Chang, Ameya Patil, Anand Bhattad