COURS RDFIA deep Image

Matthieu Cord
Sorbonne University
Course Outline – Week timeline

1. Computer Vision basics: Visual (local) feature detection and description, Bag of Word Image representation
2. Supervised learning: Introduction to Neural Networks (NNs)
3. Machine Learning basics: Risk, Classification, Datasets, benchmarks and evaluation, Linear classification (SVM)
4. Convolutional Nets for visual classification
5. Large deep convnets
6. Beyond ImageNet
7. Transfer learning and domain adaptation
8. Attention and Vision Transformers
9. Generative models with GANs
10. **Generative models with conditional GANs**
11. Control
12/14 Bayesian deep learning
Generative models

Outline

1. Preview: Auto-Encoders, VAE
2. Generative models with GAN
3. GAN architectures
4. Editing
5. Conditional GANs
Generative models

Outline

1. Preview: Auto-Encoders, VAE
2. Generative models with GAN
3. GAN architectures
4. Editing
5. Conditional GANs
   1. Principle
Motivation

Prior distribution $z$

c: train

Generator

Text

this white and yellow flower have thin white petals and a round yellow stamen

Generator

Image

Image

Generator

Image
Conditional GAN

- **Text to image** by traditional supervised learning

**c^1**: a dog is running  \( \hat{x}^1: \)

**c^2**: a bird is flying  \( \hat{x}^2: \)

Text: “train”

**Target of NN output**

A blurry image!
Conditional GAN

Prior distribution $z$ $\xrightarrow{\text{c: train}}$ $G$ $\xrightarrow{\text{Image}}$ $x = G(c,z)$

Text: “train”

It is a distribution
Approximate the distribution of real data
Conditional GAN

Prior distribution $z$ → $G$ → Image $x = G(c,z)$

$x$ is realistic or not

$x$ is realistic or not + $c$ and $x$ are matched or not

Positive example: $(\text{train, Image})$

Negative example: $(\text{train, Image}) (\text{cat, Image})$

$x$ → $D$ (type 1) → scalar

$c$ → $D$ (type 2) → scalar

$c$: train

Positive example: train

Negative example: Image

Positive example: $(\text{train, Image})$

Negative example: $(\text{train, Image}) (\text{cat, Image})$
Conditional GAN (CGAN model)

$$\min_{G} \max_{D} \left( \mathbb{E}_{x,y \sim p_{data}(x,y)} \left[ \log D(x, y) \right] + \mathbb{E}_{y \sim p_y, z \sim p_z(z)} \left[ \log(1 - D(G(z, y), y)) \right] \right)$$
Generative models

Outline

1. Preview: Auto-Encoders, VAE
2. Generative models with GAN
3. GAN architectures
4. Editing
5. Conditional GANs
   1. Principle
   2. Text2Image
This flower has small, round violet petals with a dark purple center

\[ \varphi \]  
\[ \varphi(t) \]  
\[ z \sim \mathcal{N}(0, 1) \]  

\[ \hat{x} := G(z, \varphi(t)) \]

Generator Network

This flower has small, round violet petals with a dark purple center

\[ \varphi \]  

\[ D(\hat{x}, \varphi(t)) \]

Discriminator Network

- **Positive samples:**
  - real image + right texts

- **Negative samples:**
  - fake image + right texts
  - Real image + wrong texts

[Reed et al. ICML 2016]
this small bird has a pink breast and crown, and black primaries and secondaries.

the flower has petals that are bright pinkish purple with white stigma

this magnificent fellow is almost all black with a red crest, and white cheek patch.

this white and yellow flower have thin white petals and a round yellow stamen

[Reed et al. ICML 2016]
<table>
<thead>
<tr>
<th>Caption</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>this flower has white petals and a yellow stamen</td>
<td>![Images of white flowers with yellow centers]</td>
</tr>
<tr>
<td>the center is yellow surrounded by wavy dark purple petals</td>
<td>![Images of purple flowers with wavy petals]</td>
</tr>
<tr>
<td>this flower has lots of small round pink petals</td>
<td>![Images of pink flowers with round petals]</td>
</tr>
</tbody>
</table>
StackGAN: similar idea with LapGAN

Figure 2. The architecture of the proposed StackGAN. The Stage-I generator draws a low-resolution image by sketching rough shape and basic colors of the object from the given text and painting the background from a random noise vector. Conditioned on Stage-I results, the Stage-II generator corrects defects and adds compelling details into Stage-I results, yielding a more realistic high-resolution image.

[Zhang et al. 2016]
StackGAN

This bird has a yellow belly and tarsus, grey back, wings, and brown throat, nape with a black face

This bird is white with some black on its head and wings, and has a long orange beak

This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments

(a) Stage-I images

(b) Stage-II images

[Zhang et al. 2016]
<table>
<thead>
<tr>
<th>Caption</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>a pitcher is about to throw the ball to the batter</td>
<td><img src="images/batting.jpg" alt="Images of baseball scenes" /></td>
</tr>
<tr>
<td>a group of people on skis stand in the snow</td>
<td><img src="images/skiing.jpg" alt="Images of skiing scenes" /></td>
</tr>
<tr>
<td>a man in a wet suit riding a surfboard on a wave</td>
<td><img src="images/surfing.jpg" alt="Images of surfing scenes" /></td>
</tr>
</tbody>
</table>
Generative models

Outline

1. Preview: Auto-Encoders, VAE
2. Generative models with GAN
3. GAN architectures
4. Editing
5. Conditional GANs
   1. Principle
   2. Text2Image
   3. Image2Image
Image-to-Image Translation *pix2pix*

- Conditioned on an image of different modality
- No need to specify the loss function

[Isola et al. CVPR 2017]
Image-to-image pix2pix

\[ x = G(z | c) \]
Image-to-image pix2pix

- Traditional supervised approach

Testing:

It is blurry because it is the average of several images.
Image-to-image

- Conditional GAN

Testing:

input  close  GAN  GAN + close  GT
Positive examples

Real or fake pair?

D

D tries to identify the fakes

G tries to synthesize fake images that fool D

Negative examples

Real or fake pair?

D

G

[Isola et al. CVPR 2017]
Label2Image

[Isola et al. CVPR 2017]
Edges2Image

[Isola et al. CVPR 2017]
Pix2pixHD [CVPR 2018]

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs
Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, Bryan Catanzaro
Pix2pixHD [CVPR 2018]

**Coarse-to-fine Generator**

- Residual blocks

**Multi-scale Discriminators**

- $D_1$
- $D_2$
- $D_3$

**Robust Objective**

- $D_i$
- real
- synthesized
- match
Results

• Qualitative comparisons
Improving Segmentation2Image strategy

[SPADE: Semantic Image Synthesis with Spatially-Adaptive Normalization, CVPR19]
Improving Segmentation2Image strategy

Previous approach:

Directly feed the semantic layout as input to the deep network, which is processed through stacks of convolution, normalization, and nonlinearity layers.

However, this is suboptimal as the normalization layers tend to “wash away” semantic information in input semantic segmentation masks.
Improving Segmentation2Image strategy

Conditional Normalization Layers

Proven effective for recent generative adversarial networks such as StyleGAN
Improving Segmentation2Image strategy

SPADE block: spatially-adaptive denormalization

\[
\gamma_{c,y,x}(m) \frac{h_{n,c,y,x} - \mu_{c}^{i}}{\sigma_{c}^{i}} + \beta_{c,y,x}(m)
\]

\[
\mu_{c}^{i} = \frac{1}{NH^{i}W^{i}} \sum_{n,y,x} h_{n,c,y,x}^{i}
\]

\[
\sigma_{c}^{i} = \sqrt{\frac{1}{NH^{i}W^{i}} \sum_{n,y,x} (h_{n,c,y,x}^{i})^{2} - (\mu_{c}^{i})^{2}}
\]
SPADE Generator
SPADE Generator

Better preserve semantic information against common normalization layers
SPADE results
Generative models

Outline

1. Preview: Auto-Encoders, VAE
2. Generative models with GAN
3. GAN architectures
4. Editing
5. Conditional GANs
   1. Principle
   2. Text2Image
   3. Image2Image
   4. Inpainting and general missing data encoder
Inpainting task

• Complete the missing part
Inpainting as unsupervised learning with GAN loss

Reconstruct missing pixels by decoding using context
Loss defined on the predicted patch and the real one (known at training time)
First proposition -- Architecture

- Architecture: Encoder/Fully connected/Decoder

- **DC-GAN** for inpainting task
- **Input:** $227 \times 227 \times 3$ image
- **Output:** encoder context features $(6 \times 6 \times 256)$
Channel-wise fully-connected layer

- **Input / output:** $6 \times 6 \times 256$ channels
- **First layer:** Channel-wise fully-connected  
  (each $6 \times 6$ input connected to the corresponding $6 \times 6$ output)
- **Second layer:** Stride 1 convolution to mix channels

Decoder

- **Architecture:** Same as DC-GAN: 5 up-convolutional layers  
  ("deconv" + ReLU)
- **Input:** decoder context features $6 \times 6 \times 256$
- **Output:** $227 \times 227 \times 3$ image
Training: Masking the images

• How to define the mask?
  • Center region of the image
  • Random regions (chosen solution)
  • Random segmentation mask from VOC (said to be equivalent to random regions)

• Formal definition: Defined by a mask \( \hat{M} \in \{0, 1\}^{227 \times 227} \) with 1 if the pixel should be masked
Training: Loss - Overview

- Trained completely from scratch to fill-up the masked areas
- **Problem**: multiple plausible solutions
- **Solution**: combining 2 losses:
  - $\mathcal{L}_{\text{rec}}$ **L2 reconstruction loss**: learn the structure of the missing region (average multiple modes in prediction)
  - $\mathcal{L}_{\text{adv}}$ **Adversarial loss**: make it look real (pick a mode from the distribution)

\[
\min_F \mathcal{L} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}
\]

\[
\mathcal{L}_{\text{rec}}(x) = \left\| \hat{M} \odot \left( x - F \left( (1 - \hat{M}) \odot x \right) \right) \right\|_2
\]

\[
\mathcal{L}_{\text{adv}} = \max_D \mathbb{E}_{x \in \mathcal{X}} \left[ \log(D(x)) + \log \left( 1 - D(F((1 - \hat{M}) \odot x)) \right) \right]
\]

- **Rq**: The encoder-decoder is the generator, $D$ is a CNN
Results

Dataset: StreetView Paris and ImageNet
Semantic inpainting - Qualitative results
Generalizing inpainting: missing data encoder
Adding perceptual loss, BB regression loss

\[ r^c = \left( \frac{x^c}{W}, \frac{y^c}{H}, \frac{x^c + w^c}{W}, \frac{y^c + h^c}{H} \right) \]

\[ \mathcal{L}_{\text{disc}}^{HnS}(\theta_d) = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \| r_i^c - \hat{r}_i^c(\theta_g, \theta_d) \| \]

\[ \mathcal{L}_{\text{gen}}^{HnS}(\theta_g) = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \| q_i^c - \hat{r}_i^c(\theta_g, \theta_d) \| \]

\[ \mathcal{L}_{\text{tot}}(\theta_g, \theta_d) = \mathcal{L}_{\text{rec}}(\theta_g) + \lambda_{\text{compl}} \mathcal{L}_{\text{compl}}^{vgg}(\theta_g) + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}(\theta_g, \theta_d) + \lambda_{HnS} \mathcal{L}_{\text{coord}}^{HnS}(\theta_g, \theta_d) \]
Results
Results
Generative models

Outline

1. Preview: Auto-Encoders, VAE
2. Generative models with GAN
3. GAN architectures
4. Editing
5. Conditional GANs
   1. Principle
   2. Text2Image
   3. Image2Image
   4. Inpainting and general missing data encoder
   5. Learning unpaired Transformation
Unpaired Transformation - Cycle GAN, Disco GAN

Transform an object from one domain to another \textit{without paired data}

\textit{paired data}

\begin{itemize}
  \item photo
  \item van Gogh
  \item Domain X
  \item Domain Y
\end{itemize}
Cycle GAN

https://arxiv.org/abs/1703.10593
https://junyanz.github.io/CycleGAN/

Domain X

Ignore input

$G_{X \rightarrow Y}$

Become similar to domain Y

Domain Y

Input image belongs to domain Y or not

$D_Y$ → scalar

Not what we want
Cycle GAN

\[ G_{X \rightarrow Y} \]  \[ G_{Y \rightarrow X} \]  

as close as possible

Domain X  \quad Domain Y

Lack of information for reconstruction

Input image belongs to domain Y or not

scalar
Cycle GAN

Domain X

as close as possible

Domain Y

Scalar: belongs to domain X or not

Scalar: belongs to domain Y or not

as close as possible
Results -- Cycle GAN

Photo → van Gogh

Domain X

Domain Y

Monet ↔ Photos

Zebras ↔ Horses

Summer ↔ Winter

Monet → photo

horse → zebra

photo → Monet

summer → winter

winter → summer
Appendix

GANs for Video, 3D, etc.
Video GAN


[Vondrick et al. NIPS 2016]
Shape modeling using 3D Generative Adversarial Network

Wu et al. NIPS 2016