Generative models

Outline

1. Preview: Auto-Encoders, VAE
2. Generative models with GAN
3. GAN architectures for (conditional) image generation

Drawing? => learning from examples
Generative models

Outline

1. Preview: Auto-Encoders, VAE
2. Generative models with GAN
3. GAN architectures for (conditional) image generation
Recall Algo GAN

for number of training iterations do
  for k steps do
    • Sample minibatch of $m$ noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p_g(z)$.
    • Sample minibatch of $m$ examples $\{x^{(1)}, \ldots, x^{(m)}\}$ from data generating distribution $p_{data}(x)$.
    • Update the discriminator by ascending its stochastic gradient:
      $$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[ \log D\left(x^{(i)}\right) + \log \left(1 - D\left(G\left(z^{(i)}\right)\right)\right) \right].$$
  end for
  • Sample minibatch of $m$ noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p_g(z)$.
  • Update the generator by descending its stochastic gradient:
    $$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log \left(1 - D\left(G\left(z^{(i)}\right)\right)\right).$$
end for

Question: Which architectures for G and D?
Basic Archi for G and D and expe

Models
$G$ and $D$ fully connected nets
or convolutional for $D$, (de)convolutional for $G$ (as seen for segmentation nets)
ReLU and/or sigmoids, dropout

Datasets
MNIST, Toronto Face Database, CIFAR-10

GAN - Evaluation

- Approximate $p_g$ by fitting a Gaussian Parzen window on the generated images.
- Cross-validate $\sigma$ to maximize likelihood of validation set
- Compute the likelihood of the test set

Evaluation not trivial, can be done using generated images as inputs for deep nets $\Rightarrow$ inception scores
GAN - Qualitative results 1/2

Figure: Right col nearest from dataset. a) MNIST, b) TFD, c) CIFAR-10 (fully connected), d) CIFAR-10 (convolutional $D$, deconvolutional $G$)
GAN - Qualitative results 2/2

Figure: Linear interpolation between 2 points in z space

- Advantages:
  - Computational advantages (no complex likelihood inference)
  - Can represent sharper distributions

- Disadvantages:
  - $G$ and $D$ must be well synchronized for the algorithm to converge correctly
GAN architectures

• How to improve result quality?
  • Spatial resolution
    ⇒ Cascade of GAN
  • Object quality
    ⇒ Progressive growing of spatial resolution in G
Laplacian Pyramid GANs (LAPGANs)

- GANs do not work well for complex / high level / natural images.
- Idea: decompose the generation in successive tasks using Laplacian Pyramid (of GANs)

Let $d(l)$ and $u(l)$ be down-sampling and up-sampling operations. Gaussian pyramid:

$$\mathcal{G}(l) = [l_0, l_1, \ldots, l_K], l_k = d^{(k)}(l)$$

Laplacian pyramid:

$$h_k = \mathcal{L}_k(l) = l_k - u(l_{k+1})$$

Reconstruction:

$$l_k = u(l_{k+1}) + h_k$$
LAPGAN model - sampling

- Set of generative convnets: $G_0, ..., G_K$
- Generated details: $\tilde{h}_k = G_k(z_k, u(\tilde{l}_{k+1}))$
- Reconstructed image: $\tilde{l}_k = u(\tilde{l}_{k+1}) + \tilde{h}_k$ ($\tilde{l}_{K+1} = 0$)
LAPGAN model - training

- Low-pass version of $I_0$: $I_0 = u(d(I_0)$
- Either:
  - High-pass version of $I_0$: $h_0 = I_0 - I_0$
  - Generate $\tilde{h}_0 = G_0(z_0, I_0)$
- Forward $D_0(I_0 + h_0$ or $\tilde{h}_0$)
- Backward $D_0$ and $G_0$
- $G_K$ and $D_K$ are trained as a simple GAN
LAPGAN model - Experiments

- **Datasets:** CIFAR-10, STL
- **Initial scale:**
  - $G_K$ and $D_K$ have 2 hidden layers and ReLU
  - $z_K \sim U_{[-1,1]}^{100}$
  - Trained as GAN
- **Subsequent scales:**
  - $G_k$ and $D_k$ convnets with 3 and 2 layers
  - $z_k \sim U_{[-1,1]}^{100}$ ("color" layer)
  - Trained as CGAN
LAPGAN model - Results - CIFAR
Figure 4: STL samples: (a) Random 96x96 samples from our LAPGAN model. (b) Coarse-to-fine generation chain.
LAPGAN model - Results - LSUN
LAPGAN

• Good idea but Genetor structure too weak
• Recent improvement: Pix2PixHD described later
Progressive growing of spatial resolution in G

Deep Convolutional GANs (DCGANs)

GANs are hard to scale => Identify a family of architectures that gives stable training

- 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
Progressive growing of spatial resolution in G: DCGAN

Upsampling step by step

Combine with convolutional layers
DCGAN - Results - generated bedrooms
DCGAN
Walking in latent
DCGAN - Arithmetics in Z space

smiling woman

neutral woman

neutral man

= smiling man

man with glasses

man without glasses

woman without glasses

= woman with glasses
Conditional GAN
Motivation

Prior distribution $z$

c: train

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

Image

Generator

Generator

Generator

Generator
Conditional GAN

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

\[ c^1: \text{a dog is running} \quad \hat{x}^1: \]
\[ c^2: \text{a bird is flying} \quad \hat{x}^2: \]

Text: “train”

Target of NN output

A blurry image!
Conditional GAN

\[ x = G(c, z) \]

- **c**: train
- **\( z \)**: Prior distribution

Text: “train”

A blurry image!

It is a distribution. Approximate the distribution below.
Conditional GAN

Prior distribution $z$ \rightarrow G \rightarrow \text{Image} \quad x = G(c,z)$

$c$: train

$x$ is realistic or not

Positive example:

Negative example:

$D$ (type 1) \rightarrow \text{scalar}

$c$ \rightarrow D (type 2) \rightarrow \text{scalar}

Positive example: (train, Image)

Negative example: (train, Image) (cat, Image)
Conditional GAN (CGAN model)

\[
\min_G \max_D \left( \mathbb{E}_{x,y \sim p_{\text{data}}(x,y)} \left[ \log D(x, y) \right] + \mathbb{E}_{y \sim p_y, z \sim p_z(z)} \left[ \log (1 - D(G(z, y), y)) \right] \right)
\]
Applications

- Text2Im
- Im2Im
- Unpaired transformation
- ...

Text2Image

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

\[ \varphi \]

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

\[ \varphi(t) \]

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

Generator Network

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.

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

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

this white and yellow flower have thin white petals and a round yellow stamen
<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></td>
</tr>
<tr>
<td>the center is yellow surrounded by wavy dark purple petals</td>
<td></td>
</tr>
<tr>
<td>this flower has lots of small round pink petals</td>
<td></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></td>
</tr>
<tr>
<td>a group of people on skis stand in the snow</td>
<td></td>
</tr>
<tr>
<td>a man in a wet suit riding a surfboard on a wave</td>
<td></td>
</tr>
</tbody>
</table>
Image-to-Image Translation \textit{pix2pix}

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

\cite{Isola2017CVPR}

\textbf{Isola et al. CVPR 2017}
Image-to-image pix2pix

\[ x = G(z | c) \]

https://arxiv.org/pdf/1611.07004
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?

\[
\begin{array}{c}
\text{D} \\
\text{} \quad \text{\quad} \quad \text{} \\
\text{} \quad \text{\quad} \quad \text{\quad} \\
\end{array}
\]

\[
\begin{array}{c}
\text{\quad} \\
\text{\quad} \\
\text{\quad} \\
\end{array}
\]

\[
\begin{array}{c}
\text{\quad} \\
\text{\quad} \\
\text{\quad} \\
\end{array}
\]

\[
\begin{array}{c}
\text{\quad} \\
\text{\quad} \\
\text{\quad} \\
\end{array}
\]

\[
\begin{array}{c}
\text{\quad} \\
\text{\quad} \\
\text{\quad} \\
\end{array}
\]

Negative examples

Real or fake pair?

\[
\begin{array}{c}
\text{D} \\
\text{} \quad \text{\quad} \quad \text{} \\
\text{} \quad \text{\quad} \quad \text{\quad} \\
\end{array}
\]

\[
\begin{array}{c}
\text{\quad} \\
\text{\quad} \\
\text{\quad} \\
\end{array}
\]

\[
\begin{array}{c}
\text{\quad} \\
\text{\quad} \\
\text{\quad} \\
\end{array}
\]

\[
\begin{array}{c}
\text{\quad} \\
\text{\quad} \\
\text{\quad} \\
\end{array}
\]

G tries to synthesize fake images that fool D

D tries to identify the fakes

[Isola et al. CVPR 2017]
Label2Image

Input       Ground truth       L1       cGAN       L1 + cGAN

[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

Multi-scale Discriminators

Robust Objective

Residual blocks

real

synthesized

D

real

synthesized

Di

real

match

Di

match
Pix2pixHD [CVPR 2018]

Our Method

- Boundary improvement
**Feature Embedding: Inference**

- **Labels** and **Features** from **Reals**
- **Image generation network**
- **Synthesized** images
- **Feature encoder network**
- **Features**
- **Instance-wise average pooling**
Unpaired Transformation
Unpaired Transformation - Cycle GAN, Disco GAN

Transform an object from one domain to another without paired data

paired data

photo

van Gogh

Domain X

Domain Y
Cycle GAN

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

Domain X

\(G_{X \rightarrow Y}\)

Become similar to domain Y

Not what we want

Input image belongs to domain Y or not

Domain Y

ignore input

scalar
Cycle GAN

Lack of information for reconstruction

Input image belongs to domain Y or not

as close as possible
Cycle GAN

as close as possible

$G_{X \rightarrow Y}$

$D_X$

$D_Y$

$G_{Y \rightarrow X}$

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
Video GAN


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

[Wu et al. NIPS 2016]
GAN for Inpainting
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:
  - $\mathcal{L}_{rec}$ **L2 reconstruction loss:** learn the structure of the missing region (average multiple modes in prediction)
  - $\mathcal{L}_{adv}$ **Adversarial loss:** make it look real (pick a mode from the distribution)

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

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

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

- 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

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

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

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

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