Generative Adversarial Network and its Applications to Human Language Processing

Hung-yi Lee
Outline

Part I: General Introduction of Generative Adversarial Network (GAN)

Part II: Applications to Natural Language Processing

Part III: Applications to Speech Processing
All Kinds of GAN ... https://github.com/hindupuravinash/the-gan-zoo

GAN
ACGAN
BGAN
CGAN
DCGAN
EBGAN
fGAN
GoGAN
...

Cumulative number of named GAN papers by month

It is a wise choice to attend this tutorial.


2 We use the Greek $\alpha$ prefix for $\alpha$-GAN, as AEGAN and most other Latin prefixes seem to have been taken https://deephunt.in/the-gan-zoo-79597dc8c347.
Generative Adversarial Network (GAN)

• Anime face generation as example

vector → Generator → high dimensional vector

Discriminator → score

Larger score means real, smaller score means fake.
**Algorithm**

- Initialize generator and discriminator
- In each training iteration:

**Step 1**: Fix generator G, and update discriminator D

Discriminator learns to assign high scores to real objects and low scores to generated objects.
Algorithm

• Initialize generator and discriminator
• In each training iteration:

**Step 2**: Fix discriminator D, and update generator G

Generator learns to “fool” the discriminator
**Algorithm**

- Initialize generator and discriminator
- In each training iteration:
  
  **Learning D**
  - Sample some real objects:
  - Generate some fake objects:

  **Learning G**
  - Update
  - Fix

  **Update**
  - Fix

**Algorithm**

- Initialize generator and discriminator
- In each training iteration:
  
  **Learning D**
  - Sample some real objects:
  - Generate some fake objects:

  **Learning G**
  - Update
  - Fix

**Algorithm**

- Initialize generator and discriminator
- In each training iteration:
  
  **Learning D**
  - Sample some real objects:
  - Generate some fake objects:

  **Learning G**
  - Update
  - Fix

**Algorithm**

- Initialize generator and discriminator
- In each training iteration:
  
  **Learning D**
  - Sample some real objects:
  - Generate some fake objects:

  **Learning G**
  - Update
  - Fix

The faces generated by machine.

The images are generated by Yen-Hao Chen, Po-Chun Chien, Jun-Chen Xie, Tsung-Han Wu.
Conditional Generation

**Generation**

\[
\begin{bmatrix}
0.3 \\
-0.1 \\
\vdots \\
-0.7
\end{bmatrix}
\begin{bmatrix}
0.1 \\
-0.1 \\
\vdots \\
0.7
\end{bmatrix}
\begin{bmatrix}
-0.3 \\
0.1 \\
\vdots \\
0.9
\end{bmatrix}
\]

In a specific range

**Conditional Generation**

“Girl with red hair and red eyes”

“Girl with yellow ribbon”
Conditional GAN

Normal distribution \( z \)

\( c: \text{ red hair} \)

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

\( c \)

\( D \) (better)

\( x \)

\( \text{scalar} \)

\( x \) is realistic or not + \( c \) and \( x \) are matched or not

True text-image pairs:

- (red hair, ) 1
- (blue hair , ) 0
- (red hair, Image ) 0

[Scott Reed, et al, ICML, 2016]
Conditional GAN
[Scott Reed, et al, ICML, 2016]

The images are generated by Yen-Hao Chen, Po-Chun Chien, Jun-Chen Xie, Tsung-Han Wu.

$G(x) = G(c, z)$

c: text

paired data
blue eyes
red hair
short hair

red hair,
green eyes

blue hair,
red eyes

[Scott Reed, et al, ICML, 2016]
Conditional GAN

"a dog barking sound"

Training Data
Collection

video
Conditional GAN

- Audio-to-image

The images are generated by Chia-Hung Wan and Shun-Po Chuang.
https://wjohn1483.github.io/audio_to_scene/index.html
Conditional GAN - Image-to-label

Multi-label Image Classifier

Input condition

Generated output
Conditional GAN - Image-to-label

The classifiers can have different architectures.

The classifiers are trained as conditional GAN.

[Tsai, et al., submitted to ICASSP 2019]

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>MS-COCO</th>
<th>NUS-WIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-16</td>
<td>56.0</td>
<td>33.9</td>
<td></td>
</tr>
<tr>
<td>+ GAN</td>
<td>60.4</td>
<td>41.2</td>
<td></td>
</tr>
<tr>
<td>Inception</td>
<td>62.4</td>
<td>53.5</td>
<td></td>
</tr>
<tr>
<td>+GAN</td>
<td>63.8</td>
<td>55.8</td>
<td></td>
</tr>
<tr>
<td>Resnet-101</td>
<td>62.8</td>
<td>53.1</td>
<td></td>
</tr>
<tr>
<td>+GAN</td>
<td>64.0</td>
<td>55.4</td>
<td></td>
</tr>
<tr>
<td>Resnet-152</td>
<td>63.3</td>
<td>52.1</td>
<td></td>
</tr>
<tr>
<td>+GAN</td>
<td>63.9</td>
<td>54.1</td>
<td></td>
</tr>
<tr>
<td>Att-RNN</td>
<td>62.1</td>
<td>54.7</td>
<td></td>
</tr>
<tr>
<td>RLSD</td>
<td>62.0</td>
<td>46.9</td>
<td></td>
</tr>
</tbody>
</table>
The classifiers can have different architectures.

The classifiers are trained as conditional GAN.

Conditional GAN outperforms other models designed for multi-label.

### Conditional GAN - Image-to-label

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>MS-COCO</th>
<th>NUS-WIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-16</td>
<td>56.0</td>
<td>33.9</td>
<td></td>
</tr>
<tr>
<td>+ GAN</td>
<td>60.4</td>
<td>41.2</td>
<td></td>
</tr>
<tr>
<td>Inception</td>
<td>62.4</td>
<td>53.5</td>
<td></td>
</tr>
<tr>
<td>+GAN</td>
<td>63.8</td>
<td>55.8</td>
<td></td>
</tr>
<tr>
<td>Resnet-101</td>
<td>62.8</td>
<td>53.1</td>
<td></td>
</tr>
<tr>
<td>+GAN</td>
<td>64.0</td>
<td>55.4</td>
<td></td>
</tr>
<tr>
<td>Resnet-152</td>
<td>63.3</td>
<td>52.1</td>
<td></td>
</tr>
<tr>
<td>+GAN</td>
<td>63.9</td>
<td>54.1</td>
<td></td>
</tr>
<tr>
<td>Att-RNN</td>
<td>62.1</td>
<td>54.7</td>
<td></td>
</tr>
<tr>
<td>RLSD</td>
<td>62.0</td>
<td>46.9</td>
<td></td>
</tr>
</tbody>
</table>
Conditional GAN – Speech Recognition

Table 1. Speech recognition performance. "+LM" refers to shallow fusion decoding jointly with RNN-LM [13], "+AT" refers to the adversarial training proposed here, "+Both" indicates training with AT and joint decoding with RNN-LM, and BT is the prior work of back-translation [21].

<table>
<thead>
<tr>
<th>Data</th>
<th>Method</th>
<th>CER/WER (%)</th>
<th>WER Δ†</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) w/o unpair text</td>
<td>(a) Baseline</td>
<td>10.5 / 21.6</td>
<td>10.5 / 21.7</td>
</tr>
<tr>
<td></td>
<td>(b) +LM</td>
<td>10.9 / 20.0</td>
<td>11.1 / 20.3</td>
</tr>
<tr>
<td></td>
<td>(c) +AT</td>
<td><strong>9.5</strong> / <strong>19.9</strong></td>
<td><strong>9.6</strong> / <strong>20.1</strong></td>
</tr>
<tr>
<td></td>
<td>(d) +Both</td>
<td>9.4 / 17.9</td>
<td>9.7 / 18.3</td>
</tr>
<tr>
<td>(B) w/ 360hrs text</td>
<td>(e) +LM</td>
<td>10.5 / 19.6</td>
<td>10.6 / 19.6</td>
</tr>
<tr>
<td></td>
<td>(f) +AT</td>
<td><strong>9.1</strong> / <strong>19.1</strong></td>
<td><strong>9.5</strong> / <strong>19.2</strong></td>
</tr>
<tr>
<td></td>
<td>(g) +Both</td>
<td>9.0 / 17.1</td>
<td>9.1 / 17.3</td>
</tr>
<tr>
<td></td>
<td>(h) BT‡</td>
<td>10.3 / 23.5</td>
<td>10.3 / 23.6</td>
</tr>
<tr>
<td></td>
<td>(i) BT+LM‡</td>
<td>9.8 / 21.6</td>
<td>10.0 / 22.0</td>
</tr>
<tr>
<td>(C) w/ 860hrs text</td>
<td>(j) +LM</td>
<td>9.9 / 18.6</td>
<td>10.2 / 18.8</td>
</tr>
<tr>
<td></td>
<td>(k) +AT</td>
<td><strong>8.6</strong> / <strong>18.5</strong></td>
<td><strong>8.8</strong> / <strong>18.7</strong></td>
</tr>
<tr>
<td></td>
<td>(l) +Both</td>
<td>7.9 / 15.3</td>
<td>8.2 / 15.8</td>
</tr>
</tbody>
</table>

† Relative improvement with respect to the baseline.
‡ Prior work [21], baseline WER 25.2% on test set reported.
Unsupervised Conditional GAN

Transform an object from one domain to another **without paired data** (e.g. style transfer)
Unsupervised Conditional Generation

• **Approach 1: Direct Transformation**
  
  ![Diagram of direct transformation](image)

  For texture or color change

• **Approach 2: Projection to Common Space**
  
  ![Diagram of projection to common space](image)

  Larger change, only keep the semantics
Direct Transformation

Domain X \rightarrow G_{X \rightarrow Y} \rightarrow \text{Become similar to domain Y} \rightarrow D_Y \rightarrow \text{scalar}

Input image belongs to domain Y or not
Direct Transformation

$G_{X \rightarrow Y}$

Domain X → $G_{X \rightarrow Y}$ → Domain Y

Become similar to domain Y

Not what we want!

ignore input

$D_Y$ → scalar

Input image belongs to domain Y or not

Domain X

Domain Y
Direct Transformation

\[ G_X \rightarrow Y \rightarrow G_Y \rightarrow X \]

as close as possible

Cycle consistency

Lack of information for reconstruction

\[ D_Y \rightarrow \text{scalar} \]

Input image belongs to domain Y or not

Domain Y

Cycle GAN

$G_X \rightarrow Y$, $G_Y \rightarrow X$, $D_X$, $D_Y$ as close as possible

scalar: belongs to domain X or not

$G_Y \rightarrow X$, $G_X \rightarrow Y$, $D_Y$, $D_X$ as close as possible

scalar: belongs to domain Y or not
Unsupervised Conditional Generation

• Approach 1: Direct Transformation

• Approach 2: Projection to Common Space

For texture or color change

Larger change, only keep the semantics
Projection to Common Space

Target

EN_X -> DE_X

EN_Y -> DE_Y

Face Attribute

Domain X

Domain Y
Projection to Common Space

Minimizing reconstruction error

Training
Projection to Common Space

Training

Minimizing reconstruction error

Because we train two auto-encoders separately ...

The images with the same attribute may not project to the same position in the latent space.
Minimizing reconstruction error

The domain discriminator forces the output of $EN_X$ and $EN_Y$ to have the same distribution.

$EN_X$ and $EN_Y$ fool the domain discriminator

The domain discriminator forces the output of $EN_X$ and $EN_Y$ to have the same distribution. [Guillaume Lample, et al., NIPS, 2017]
**Projection to Common Space**

*Training*

Sharing the parameters of encoders and decoders

Couple GAN [Ming-Yu Liu, et al., NIPS, 2016]

UNIT [Ming-Yu Liu, et al., NIPS, 2017]
Projection to Common Space

Training

Minimizing reconstruction error

Discriminator of X domain

Discriminator of Y domain

Cycle Consistency:

Used in ComboGAN [Asha Anoosheh, et al., arXiv, 017]
**Projection to Common Space**

**Training**

To the same latent space

Discriminator of X domain

Discriminator of Y domain

Semantic Consistency:

Outline

- Part I: General Introduction of Generative Adversarial Network (GAN)
- Part II: Applications to Natural Language Processing
- Part III: Applications to Speech Processing
Unsupervised Conditional Generation

**Image Style Transfer**

- Photos
- Vincent van Gogh's paintings
- Not Paired

**Text Style Transfer**

- It is good.
  - It’s a good day.
  - I love you.
- Not Paired
- It is bad.
  - It’s a bad day.
  - I don’t love you.

**positive**

**negative**
Cycle GAN

$$G_{X \rightarrow Y} \quad \text{as close as possible} \quad G_{Y \rightarrow X}$$

scalar: belongs to domain X or not

$$D_X$$

scalar: belongs to domain Y or not

$$D_Y$$

$$G_{Y \rightarrow X} \quad \text{as close as possible} \quad G_{X \rightarrow Y}$$
Cycle GAN

It is bad.
It is good.
It is bad.
I love you.
I hate you.
I love you.

$G_{X \rightarrow Y}$
$G_{Y \rightarrow X}$
$D_X$
$D_Y$

negative sentence?
positive sentence?

as close as possible
Discrete Issue

Seq2seq model

hidden layer
with discrete output

positive sentence?

large network

$G_{X \to Y}$

update

It is bad.

negative

It is good.

positive

$D_Y$

fix

Backpropagation

positive sentence?
Three Categories of Solutions

Gumbel-softmax


Continuous Input for Discriminator


“Reinforcement Learning”

Cycle GAN

\[ G_{X \rightarrow Y} \rightarrow \text{is positive.} \rightarrow D_X \rightarrow \text{is positive.} \rightarrow G_{Y \rightarrow X} \rightarrow \text{is positive.} \]

\[ G_{Y \rightarrow X} \rightarrow \text{is positive.} \rightarrow D_Y \rightarrow \text{is positive.} \rightarrow G_{X \rightarrow Y} \rightarrow \text{is positive.} \]

\[ \text{I love you.} \rightarrow \text{is positive.} \rightarrow G_{Y \rightarrow X} \rightarrow \text{is positive.} \rightarrow \text{I love you.} \]

\[ \text{I hate you.} \rightarrow \text{is negative.} \rightarrow G_{X \rightarrow Y} \rightarrow \text{is negative.} \rightarrow \text{I hate you.} \]

\[ \text{Discrete?} \]

Word embedding

[Lee, et al., ICASSP, 2018]
Cycle GAN

- **Negative** sentence to **positive** sentence:
  - it's a crappy day → it's a great day
  - i wish you could be here → you could be here
  - it's not a good idea → it's good idea
  - i miss you → i love you
  - i don't love you → i love you
  - i can't do that → i can do that
  - i feel so sad → i happy
  - it's a bad day → it's a good day
  - it's a dummy day → it's a great day
  - sorry for doing such a horrible thing → thanks for doing a great thing
  - my doggy is sick → my doggy is my doggy
  - my little doggy is sick → my little doggy is my little doggy
Projection to Common Space

Decoder hidden layer as discriminator input

[Shen, et al., NIPS, 2017]

$EN_X$ and $EN_Y$ fool the domain discriminator

[Zhao, et al., ICML 2018]

[Fu, et al., AAAI, 2018]
Unsupervised Conditional Generation

**Image Style Transfer**

- Photos
- Not Paired
- Vincent van Gogh’s paintings

**Text Style Transfer**

- Document
- Not Paired
- Summary

This is **unsupervised abstractive summarization**.
Abstractive Summarization

• Now machine can do **abstractive summary** by seq2seq (write summaries in its own words)

Supervised: We need lots of labelled training data.
Unsupervised Abstractive Summarization

• Now machine can do **abstractive summary** by seq2seq (write summaries in its own words)
Unsupervised Abstractive Summarization

Human written summaries → Real or not

Discriminator

document → word sequence → Summary?

Seq2seq

Real or not

Summary?
Unsupervised Abstractive Summarization

Human written summaries → Real or not

Discriminator

document → Seq2seq → word sequence → Seq2seq → document

minimize the reconstruction error
Unsupervised Abstractive Summarization

Only need a lot of documents to train the model.

This is a *seq2seq2seq auto-encoder*.

Using a sequence of words as latent representation.
Unsupervised Abstractive Summarization

Human written summaries

Let Discriminator considers my output as real

Real or not

REINFORCE algorithm to deal with the discrete issue

document

word sequence

Readable

Summary?

Seq2seq

G

Discriminator

D

document

Seq2seq

R

Readable

Summary?
Experimental results

English Gigaword (Document title as summary)

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>33.2</td>
<td>14.2</td>
<td>30.5</td>
</tr>
<tr>
<td>Trivial</td>
<td>21.9</td>
<td>7.7</td>
<td>20.5</td>
</tr>
<tr>
<td>Unsupervised (matched data)</td>
<td>28.1</td>
<td>10.0</td>
<td>25.4</td>
</tr>
<tr>
<td>Unsupervised (no matched data)</td>
<td>27.2</td>
<td>9.1</td>
<td>24.1</td>
</tr>
</tbody>
</table>

- Matched data: using the title of English Gigaword to train Discriminator
- No matched data: using the title of CNN/Diary Mail to train Discriminator
Semi-supervised Learning

Using matched data

Approaches to deal with the discrete issue.

3.8M pairs are used.
Outline

Part I: General Introduction of Generative Adversarial Network (GAN)

Part II: Applications to Natural Language Processing

Part III: Applications to Speech Processing
Unsupervised Conditional Generation

**Image Style Transfer**

photos

Not Paired

Vincent van Gogh’s paintings

**Speech Style Transfer**

Speaker A

Not Paired

Speaker B

This is **unsupervised voice conversion**.
Voice Conversion
In the past

How are you?

Good morning

With GAN

天気真好

再見囉

 Speakers A and B are talking about completely different things.

How are you?

Good morning
Cycle GAN

\[ G_X \rightarrow Y \quad \text{as close as possible} \quad G_Y \rightarrow X \]

\[ D_Y \quad \text{scalar: belongs to domain Y or not} \quad D_X \quad \text{scalar: belongs to domain X or not} \]

\[ G_Y \rightarrow X \quad \text{as close as possible} \quad G_X \rightarrow Y \]
Cycle GAN for Voice Conversion

$G_{X \rightarrow Y}$ as close as possible $G_{Y \rightarrow X}$

$D_X$: scalar: belongs to domain X or not

$G_{Y \rightarrow X}$ as close as possible $G_{X \rightarrow Y}$

$D_Y$: scalar: belongs to domain Y or not

X: Speaker A, Y: Speaker B

Projection to Common Space

$EN_X$  

$DE_X$  

$EN_Y$  

$DE_Y$
Projection to Common Space

- All the speakers share the same encoder.
- The model can deal with the speakers never seen during training.
Projection to Common Space

Use a vector (one-hot) to represent speaker identity.

All the speakers also share the same decoder.

The encoder fools the discriminator.

We hope that encoder can extract the phonetic information while removing the speaker information.
Projection to Common Space

**Training**

- **A** is reading the sentence of **B**
- **A** says: How are you?
- **Decoder**
- **Discriminator**
- Which speakers?
- **Decoder**
- **Discriminator**
- **A** is reconstructing the voice.

**Testing**

- **B** says: Hello
- **A** says: Hello
- **A** is reconstructing the voice.
How are you?

Different colors: different words
Different colors: different speakers

Discriminator

Which speakers?

Does it contain phonetic information?
"Audio" Word to Vector
Issues

Training

A is reading the sentence of B

Testing

A is reading the sentence of B

Which speakers?

reconstructed

Hello

Low Quality
2nd Stage Training

- Encoder
- Decoder
- Discriminator
- Speaker Classifier

Extra Criterion for Training

Different Speakers

No learning target???

real or generated?

which speaker?
Experimental Results

• Subjective evaluations (20 speakers in VCTK)

[Chou et al., INTERSPEECH, 2018]
Demo

Source: Hello

Target: Hello

Source to Target: Hello

Thanks Ju-chieh Chou for providing the results.
https://jjery2243542.github.io/voice_conversion_demo/
Thanks Ju-chieh Chou for providing the results.
https://jjery2243542.github.io/voice_conversion_demo/
Unsupervised Conditional Generation

This is **unsupervised speech recognition**.
Supervised Speech Recognition

(I believe you have seen similar figures before.)

• Supervised learning needs lots of annotated speech.
• However, most of the languages are low resourced.
Speech Recognition in the Future

Learning human language with very little supervision

http://www.parenting.com/article/teach-baby-to-talk
Unsupervised Speech Recognition

• Machine learns to recognize speech from unparallel speech and text.

This idea was too crazy to be realized in the past. However, it becomes possible with GAN recently.

[Liu, et al., INTERSPEECH, 2018]
[Chen, et al., arXiv, 2018]
Acoustic Token Discovery

Acoustic tokens can be discovered from audio collection without text annotation.

Acoustic tokens: chunks of acoustically similar audio segments with token IDs

[Zhang & Glass, ASRU 09]
[Huijbregts, ICASSP 11]
[Chan & Lee, Interspeech 11]
Acoustic Token Discovery

Acoustic tokens can be discovered from audio collection without text annotation.

Acoustic tokens: chunks of acoustically similar audio segments with token IDs

[Zhang & Glass, ASRU 09]
[Huijbregts, ICASSP 11]
[Chan & Lee, Interspeech 11]
Acoustic Token Discovery

**Phonetic-level acoustic tokens** are obtained by segmental sequence-to-sequence autoencoder.

[Wang, et al., ICASSP, 2018]
Unsupervised Speech Recognition

Phone-level Acoustic Pattern Discovery

Phoneme sequences from Text

[1] Liu, et al., INTERSPEECH, 2018

Librispeech (Word recognition)

TIMIT (Phoneme recognition), audio and text are unmatched

TIMIT (Phoneme recognition), audio and text are matched
Concluding Remarks

Part I: General Introduction of Generative Adversarial Network (GAN)

Part II: Applications to Natural Language Processing

Part III: Applications to Speech Processing
To Learn More ..... 

(My YouTube Channel, 30K subscribers, 2.4M total views)