Introduction of Generative Adversarial Network (GAN)

李宏毅
Hung-yi Lee

Generative Adversarial Network (GAN)

- How to pronounce “GAN”?

Google 小姐

Yann LeCun's comment

What are some recent and potentially upcoming breakthroughs in unsupervised learning?

Yann LeCun, Director of AI Research at Facebook and Professor at NYU

Written Jul 29 · Upvoted by Joaquin Quiñonero Candela, Director Applied Machine Learning at Facebook and Huang Xiao

有史以來

Adversarial training is the coolest thing since sliced bread.

I've listed a bunch of relevant papers in a previous answer.

Expect more impressive results with this technique in the coming years.

What's missing at the moment is a good understanding of it so we can make it work reliably. It's very finicky. Sort of like ConvNet were in the 1990s, when I had the reputation of being the only person who could make them work (which wasn't true).

https://www.quora.com/What-are-some-recent-and-potentially-upcoming-breakthroughs-in-unsupervised-learning

Yann LeCun’s comment

What are some recent and potentially upcoming breakthroughs in deep learning?

Yann LeCun, Director of AI Research at Facebook and Professor at NYU

Written Jul 29 · Upvoted by Joaquin Quiñonero Candela, Director Applied Machine Learning at Facebook and Nikhil Garg, I lead a team of Quora engineers working on ML/NLP problems

……

The most important one, in my opinion, is adversarial training (also called GAN for Generative Adversarial Networks). This is an idea that was originally proposed by Ian Goodfellow when he was a student with Yoshua Bengio at the University of Montreal (he since moved to Google Brain and recently to OpenAI).

This, and the variations that are now being proposed is the most interesting idea in the last 10 years in ML, in my opinion.

https://www.quora.com/What-are-some-recent-and-potentially-upcoming-breakthroughs-in-deep-learning

All Kinds of GAN …

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

GAN

ACGAN

BGAN

CGAN

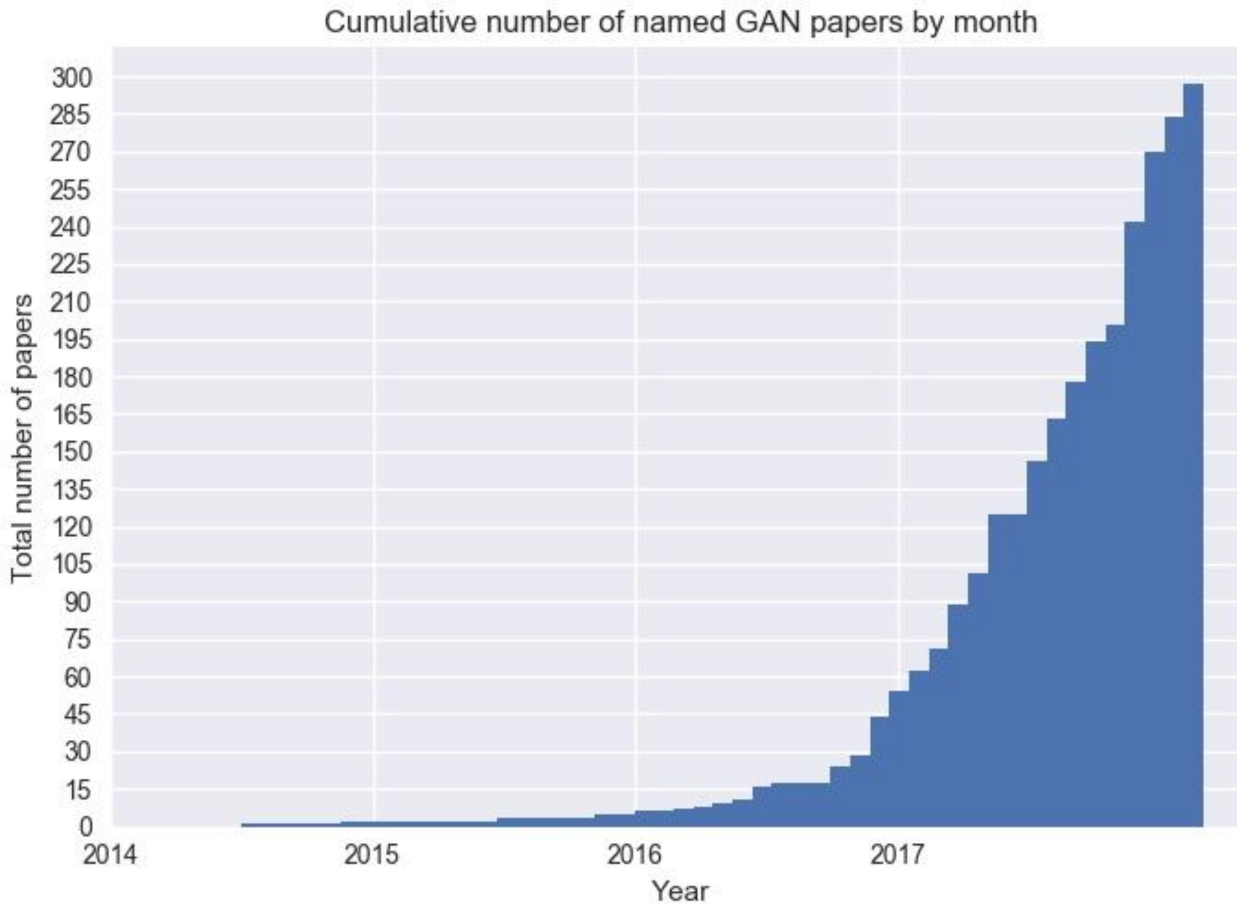
DCGAN

EBGAN

fGAN

GoGAN

⋮



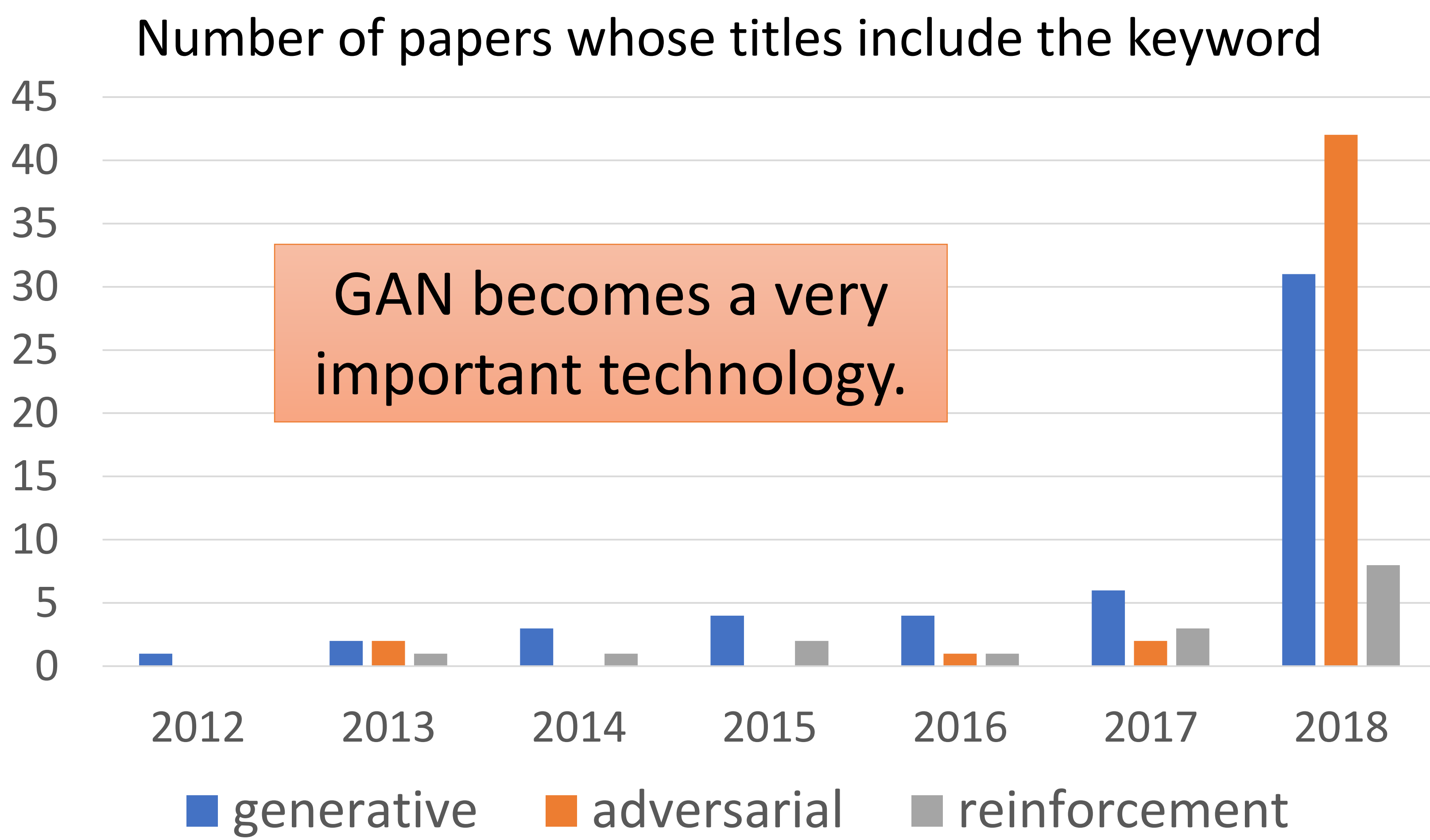
Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, Shakir Mohamed, "Variational Approaches for Auto-Encoding Generative Adversarial Networks", arXiv, 2017

[2]We use the Greek α prefix for α-GAN, as AEGAN and most other Latin prefixes seem to have been taken https://deephunt.in/the-gan-zoo-79597dc8c347.

Signal Processing Conference

ICASSP

Keyword search on session index page, so session names are included.



Outline

- Basic Idea of GAN
- GAN as structured learning
- Can Generator learn by itself?
- Can Discriminator generate?
- A little bit theory

Generation

We will control what to generate latter. → Conditional Generation

Image Generation

$$\begin{bmatrix} 0.3 \\ -0.1 \\ \vdots \\ -0.7 \end{bmatrix} \begin{bmatrix} 0.1 \\ -0.1 \\ \vdots \\ 0.7 \end{bmatrix} \begin{bmatrix} -0.3 \\ 0.1 \\ \vdots \\ 0.9 \end{bmatrix} \rightarrow \text{NN Generator} \rightarrow$$

In a specific range

Sentence Generation

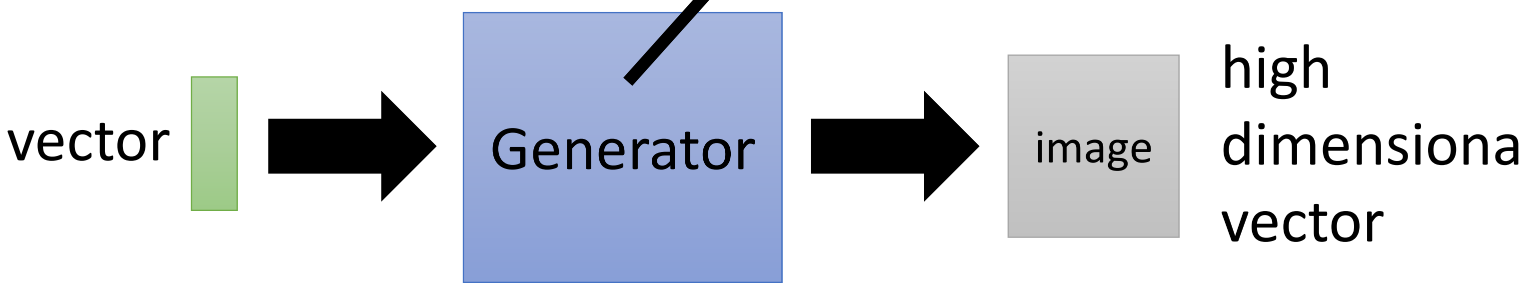
$$\begin{bmatrix} 0.3 \\ -0.1 \\ \vdots \\ -0.7 \end{bmatrix} \begin{bmatrix} 0.1 \\ -0.1 \\ \vdots \\ 0.2 \end{bmatrix} \begin{bmatrix} -0.3 \\ 0.1 \\ \vdots \\ 0.5 \end{bmatrix} \rightarrow \text{NN Generator} \rightarrow$$

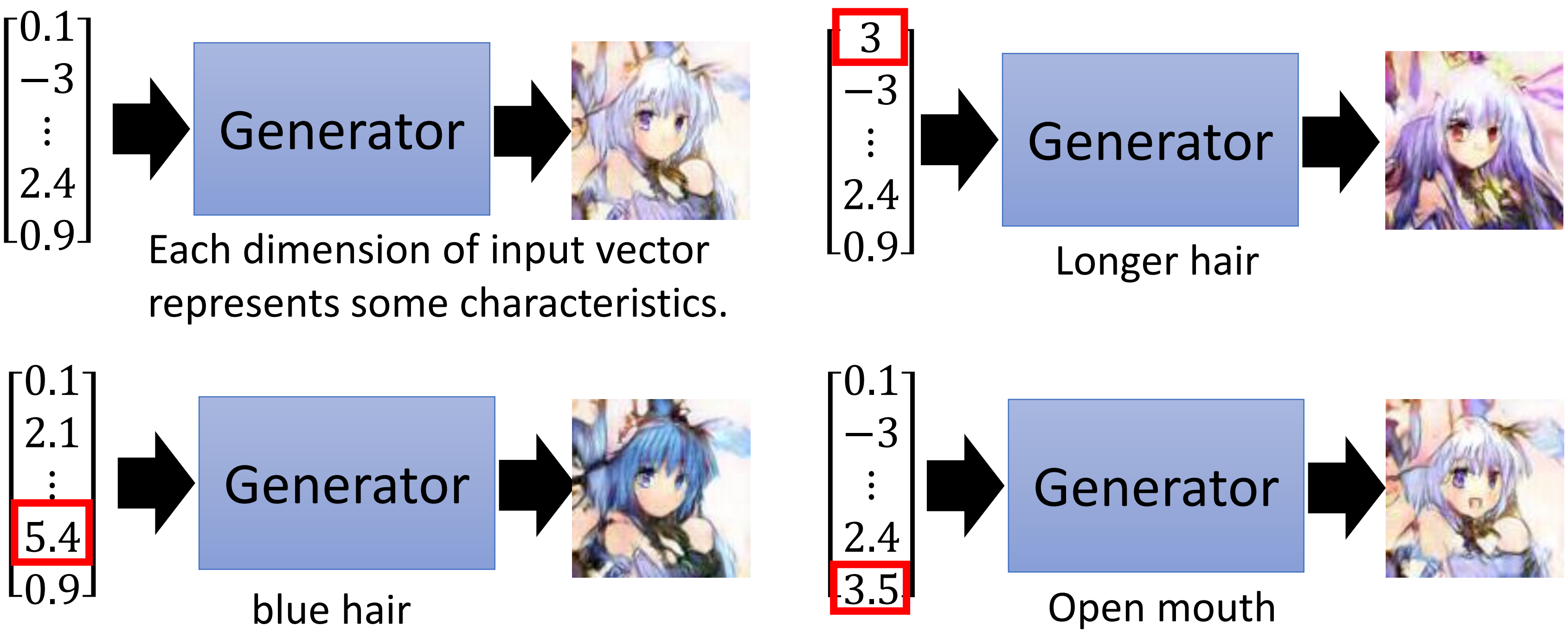
How are you?
Good morning.
Good afternoon.

Powered by: http://mattya.github.io/chainer-DCGAN/

Basic Idea of GAN

It is a neural network (NN), or a function.





Basic Idea of GAN

It is a neural network (NN), or a function.

image → Discri-minator → scalar image quality

Larger value means real, smaller value means fake.

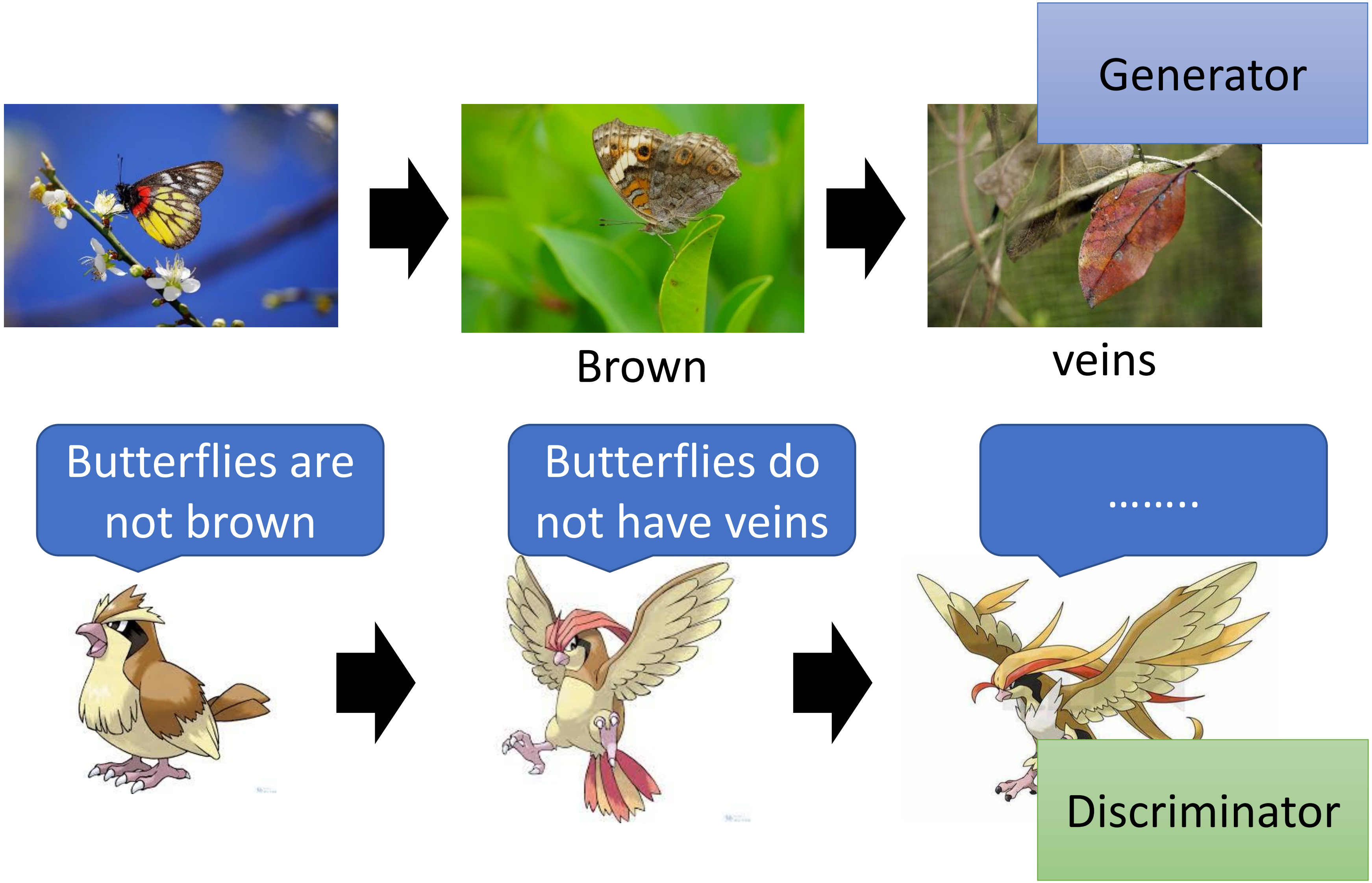
 → Discri-minator → 1.0

 → Discri-minator → 1.0

 → Discri-minator → 0.1

 → Discri-minator → 0.1

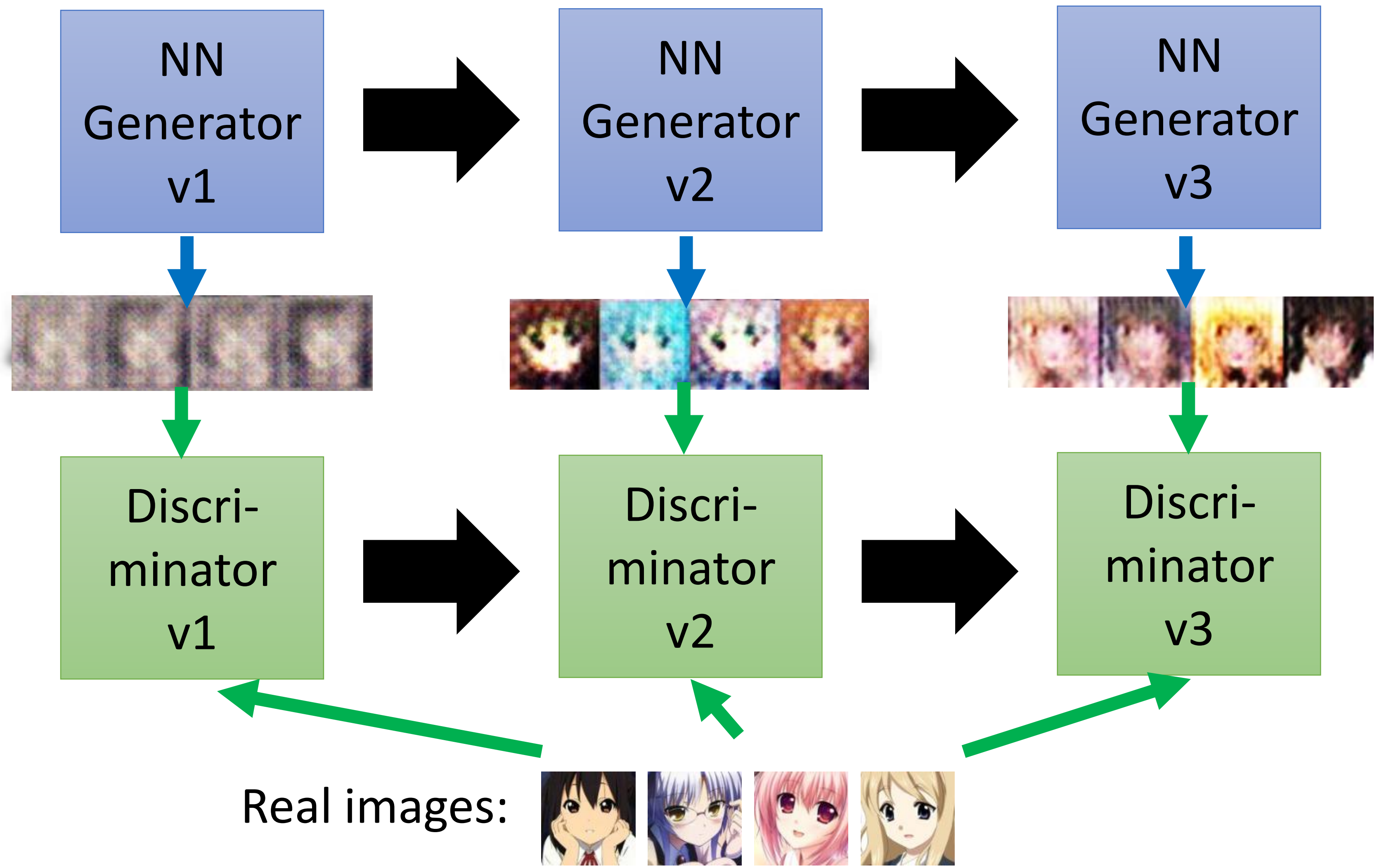
Basic Idea of GAN



Basic Idea of GAN

This is where the term "***adversarial***" comes from.

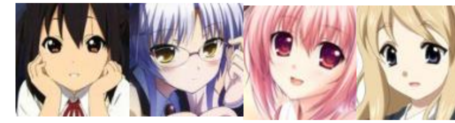
You can explain the process in different ways…….



Basic Idea of GAN (和平的比喻)

Generator (student)

Discriminator (teacher)



Generator v1

Discriminator v1

沒有兩個圈

Generator v2

Discriminator v2

沒有彩色

Generator v3

為什麼不自己學？

為什麼不自己做？

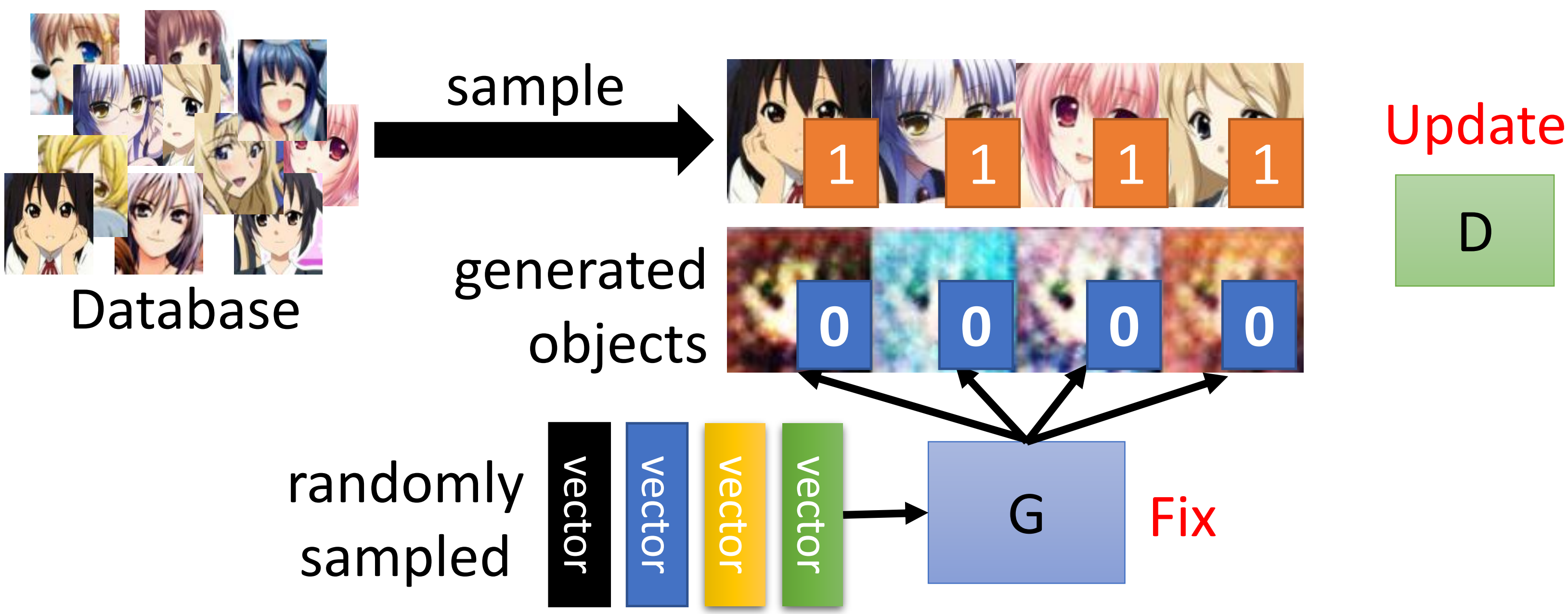
Generator v.s. Discriminator

- 寫作敵人，唸做朋友

Algorithm

- Initialize generator and discriminator G D
- In each training iteration:

Step 1: Fix generator G, and update discriminator D



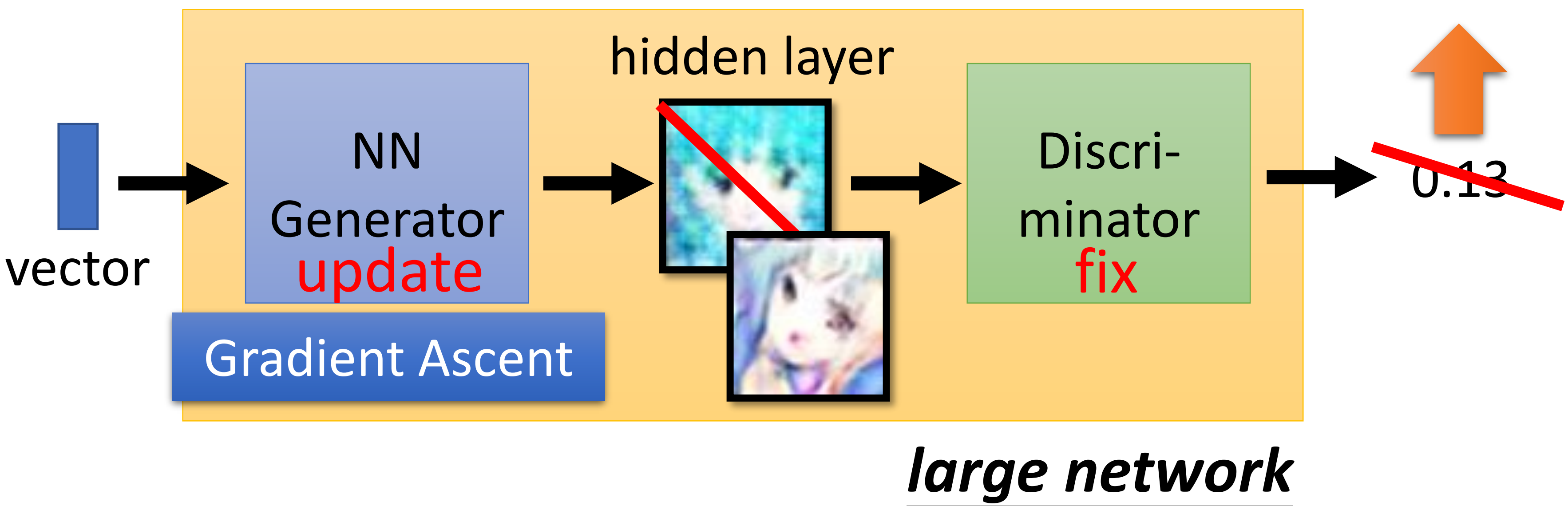
Discriminator learns to assign high scores to real objects and low scores to generated objects.

Algorithm

- Initialize generator and discriminator G D
- In each training iteration:

Step 2: Fix discriminator D, and update generator G

Generator learns to "fool" the discriminator



Algorithm

Initialize θ_d for D and θ_g for G

- In each training iteration:

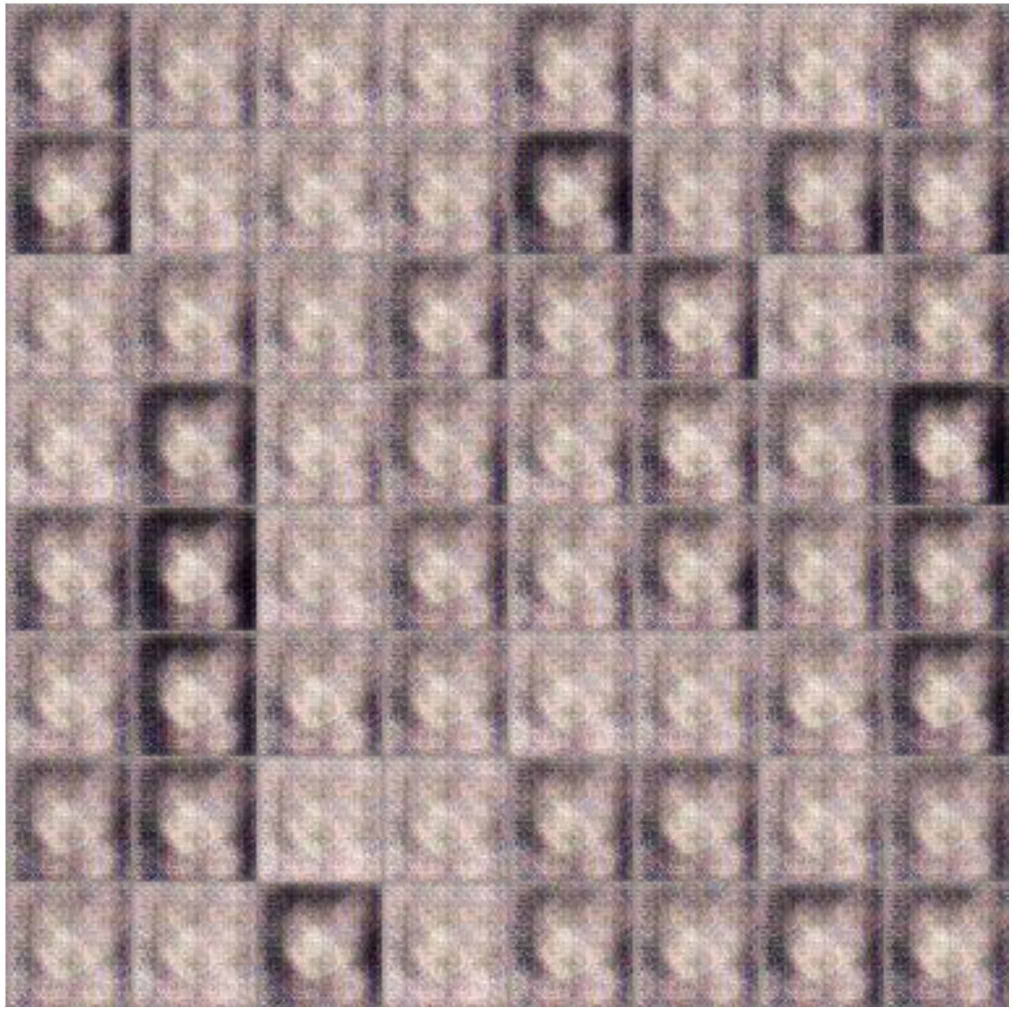
Learning D

- Sample m examples $\{x^1, x^2, \dots, x^m\}$ from database
- Sample m noise samples $\{z^1, z^2, \dots, z^m\}$ from a distribution
- Obtaining generated data $\{\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m\}$, $\tilde{x}^i = G\left(z^i\right)$
- Update discriminator parameters θ_d to maximize
  - $\tilde{V} = \frac{1}{m}\sum_{i=1}^{m} logD\left(x^i\right) + \frac{1}{m}\sum_{i=1}^{m} log\left(1 - D\left(\tilde{x}^i\right)\right)$
  - $\theta_d \leftarrow \theta_d + \eta \nabla \tilde{V}(\theta_d)$

Learning G

- Sample m noise samples$\{z^1, z^2, \dots, z^m\}$ from a distribution
- Update generator parameters θ_g to maximize
  - $\tilde{V} = \frac{1}{m}\sum_{i=1}^{m} log\left(D\left(G\left(z^i\right)\right)\right)$ 分數越大越好
  - $\theta_g \leftarrow \theta_g + \eta \nabla \tilde{V}\left(\theta_g\right)$

Anime Face Generation



100 updates

Source of training data: https://zhuanlan.zhihu.com/p/24767059

Anime Face Generation

1000 updates

Anime Face Generation

2000 updates

Anime Face Generation

5000 updates

Anime Face Generation

10,000 updates

Anime Face Generation

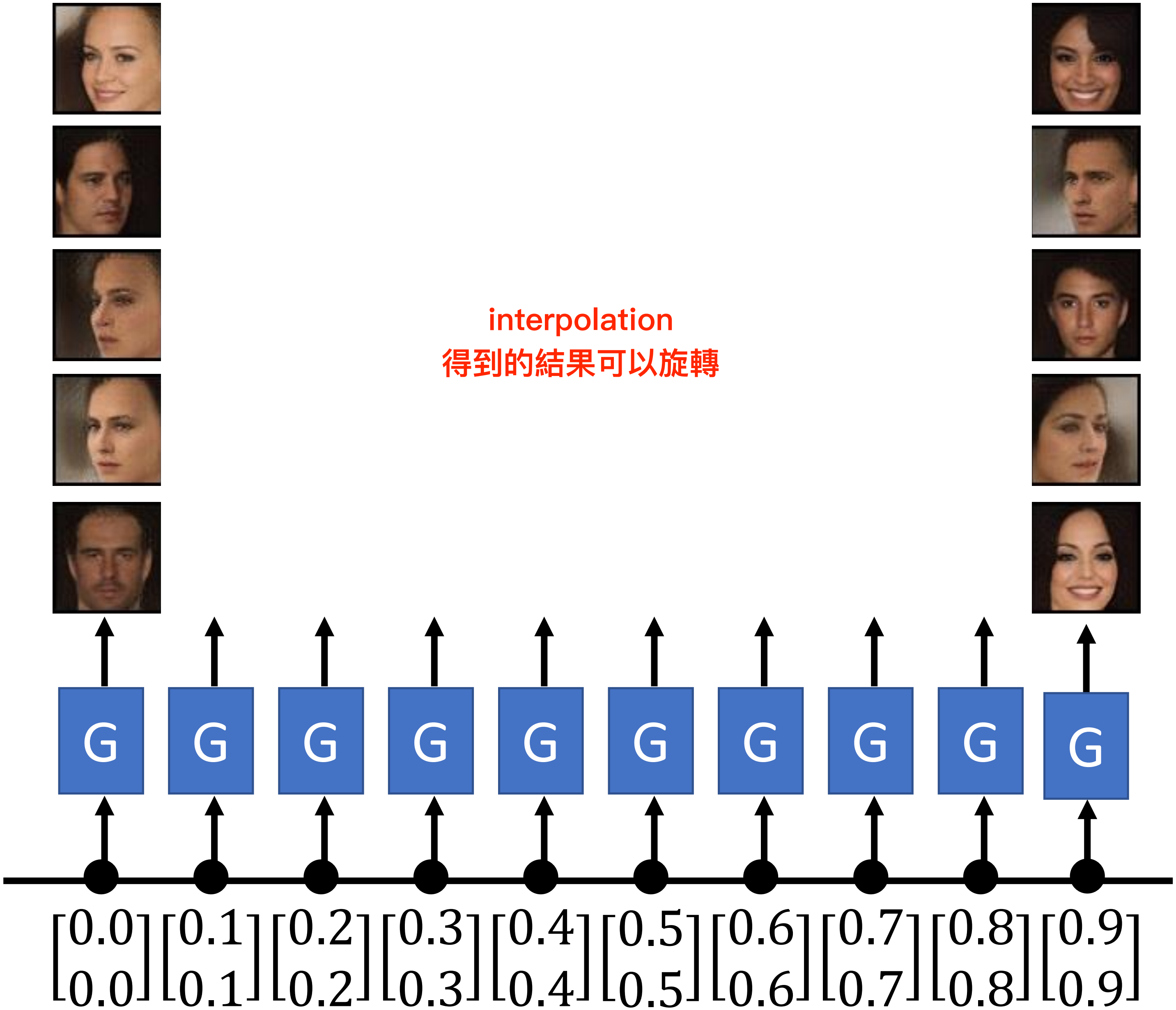
20,000 updates

Anime Face Generation

50,000 updates

The faces generated by machine.

圖片生成：
吳宗翰、謝濬丞、
陳延昊、錢柏均



感謝陳柏文同學提供實驗結果

Outline

- Basic Idea of GAN
- GAN as structured learning
- Can Generator learn by itself?
- Can Discriminator generate?
- A little bit theory

Structured Learning

Machine learning is to find a function f

$$f : X \rightarrow Y$$

Regression: output a scalar

Classification: output a "class" (one-hot vector)

1	0	0

Class 1

0	1	0

Class 2

0	0	1

Class 3

Structured Learning/Prediction: output a

output更複雜的情況

sequence, a matrix, a graph, a tree ……

Output is composed of components with dependency

Output Sequence $f : X \rightarrow Y$

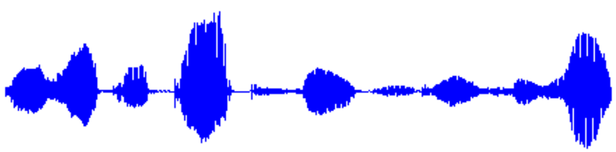
Machine Translation

X : “機器學習及其深層與結構化”
(sentence of language 1)

Y : “Machine learning and having it deep and structured”
(sentence of language 2)

Speech Recognition

X :



(speech)

Y : 感謝大家來上課”
(transcription)

Chat-bot 可以加GAN

X : “How are you?”
(what a user says)

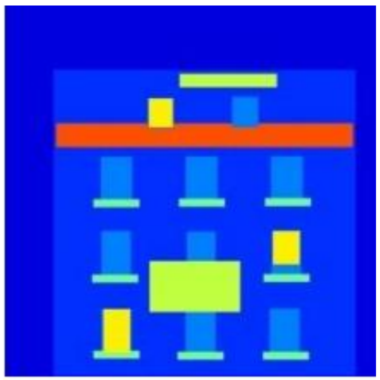
Y : “I’m fine.”
(response of machine)

Output Matrix

$$f : X \rightarrow Y$$

Image to Image

X :



Y :

Colorization:

Ref: https://arxiv.org/pdf/1611.07004v1.pdf

Text to Image

X : "this white and yellow flower have thin white petals and a round yellow stamen"

Y :

ref: https://arxiv.org/pdf/1605.05396.pdf

Why Structured Learning Challenging?

只有少數的paired training data是否能做起來

- **One-shot/Zero-shot Learning**:

假設有些類別根本沒有麼範例，依然能夠做起來嗎

  - In classification, each class has some examples.
  - In structured learning,
    - If you consider each possible output as a "class" ......

把每一個output視為一種class，則structure learning可以視為一種one shot learning

    - Since the output space is huge, most "classes" do not have any training data.
    - Machine has to create new stuff during testing.

如何學到generalization ability

    - Need more intelligence

在testing的時候想要辨別的分類問題，或許在training data中根本沒有看過或是很少個data
這時候要做得好就有難度

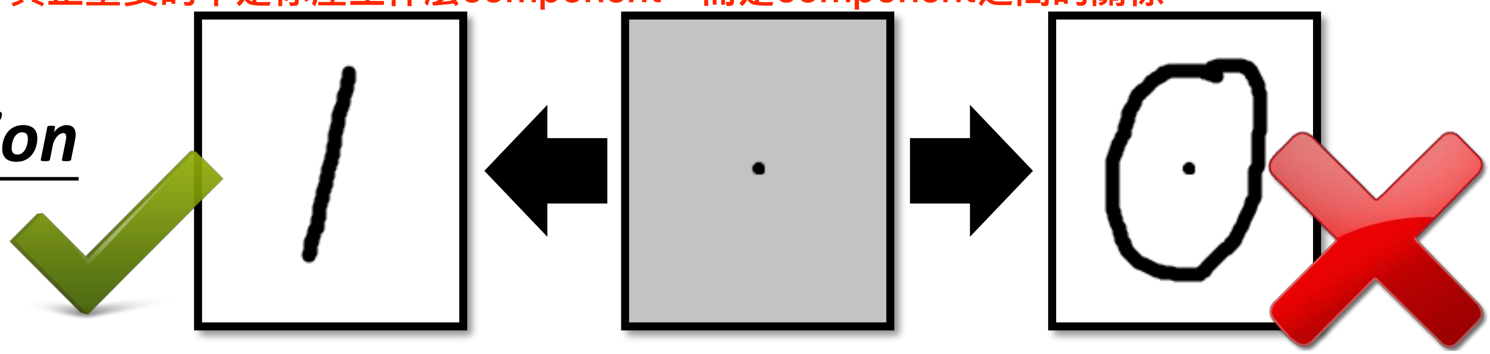
Why Structured Learning Challenging?

要讓機器有大局觀

- Machine has to learn to do ***planning***
  - Machine generates objects component-by-component, but it should have a big picture in its mind.
  - Because the output components have dependency, they should be considered globally.

真正重要的不是你產生什麼component，而是component之間的關係

Image Generation



Sentence Generation

這個婆娘不是人

九天玄女下凡塵

Structured Learning Approach

component分開去產生的物件

Generator

Learn to generate the object at the component level

Bottom Up

一個一個產生，不知道大局觀

+

從整體的方向來看到底好不好

Discriminator

Evaluating the whole object, and find the best one

Top Down

大局觀，判斷整體而言是不是正確的

Outline

Basic Idea of GAN

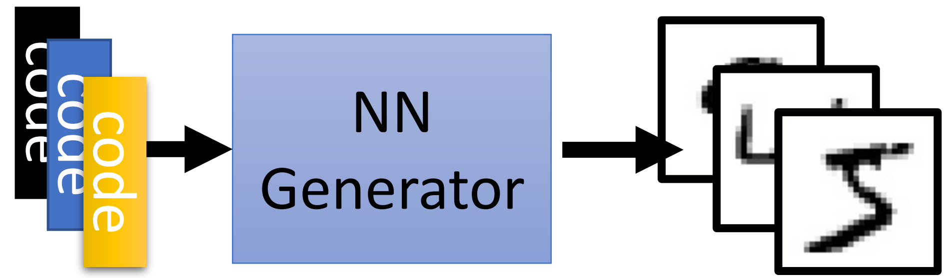
GAN as structured learning

Can Generator learn by itself?

Can Discriminator generate?

A little bit theory

Generator

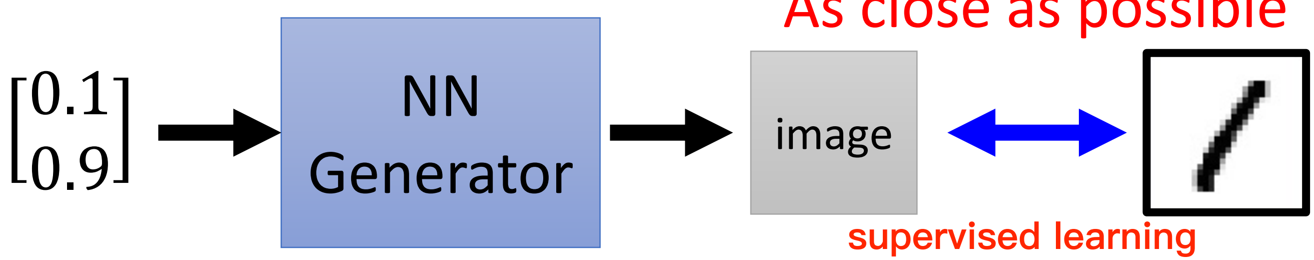


code: $\begin{bmatrix} 0.1 \\ -0.5 \end{bmatrix}$ $\begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix}$ $\begin{bmatrix} 0.2 \\ -0.1 \end{bmatrix}$ $\begin{bmatrix} 0.3 \\ 0.2 \end{bmatrix}$

(where does they come from?)

Image:

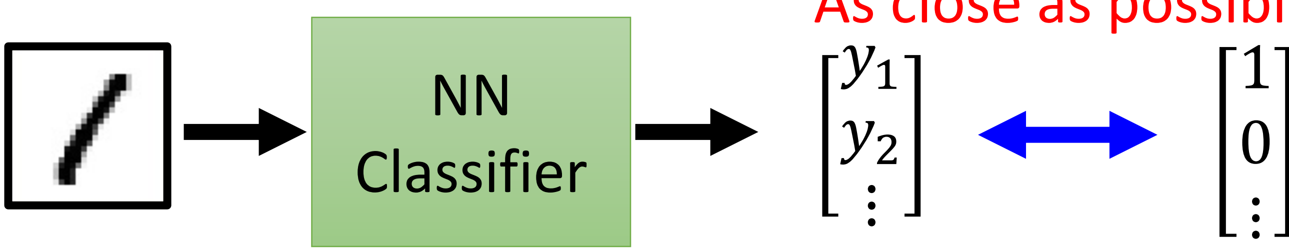
As close as possible



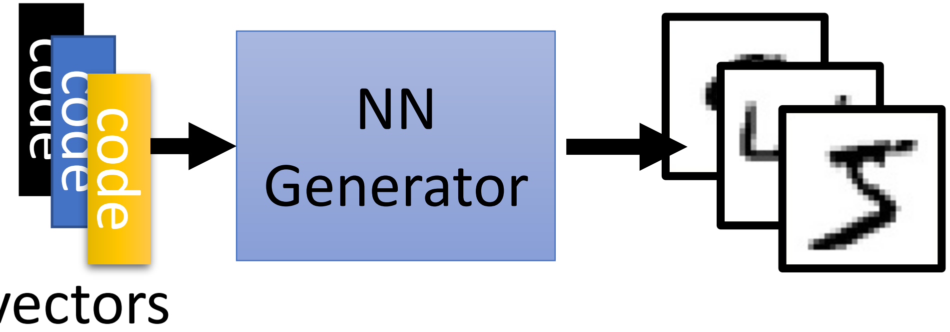
supervised learning

As close as possible

c.f.



Generator

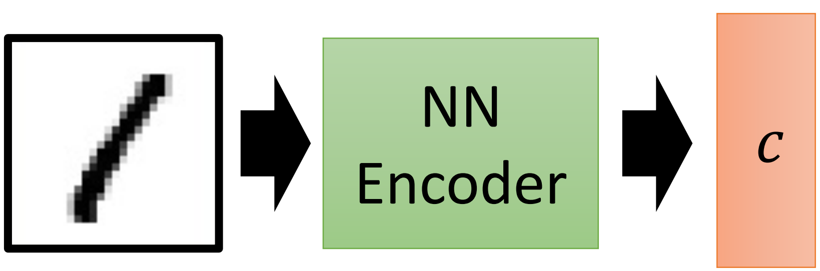


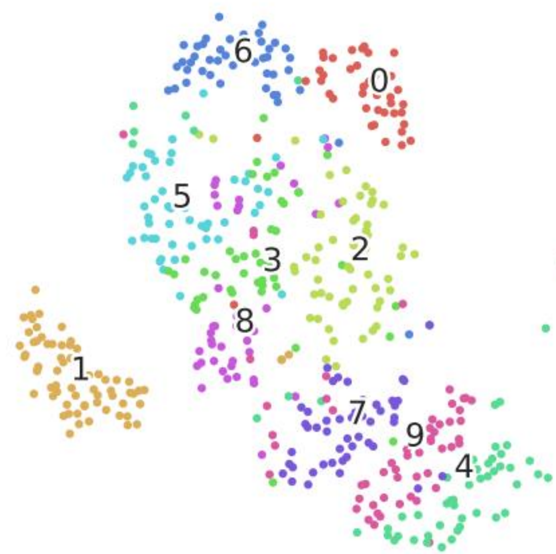
code: $\begin{bmatrix} 0.1 \\ -0.5 \end{bmatrix}$ $\begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix}$ $\begin{bmatrix} 0.2 \\ -0.1 \end{bmatrix}$ $\begin{bmatrix} 0.3 \\ 0.2 \end{bmatrix}$

(where does they come from?)

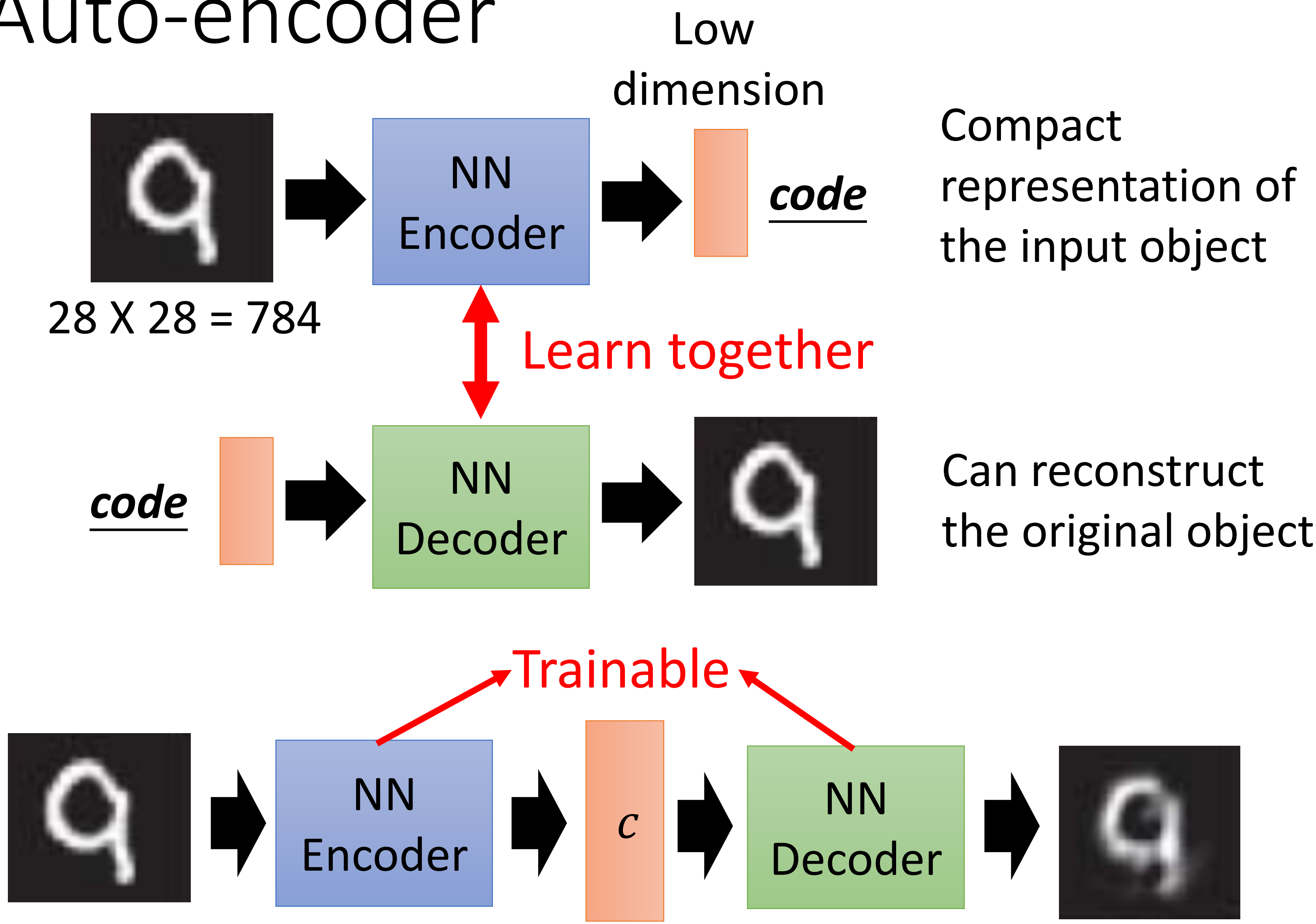
Image:

Encoder in auto-encoder provides the code ☺

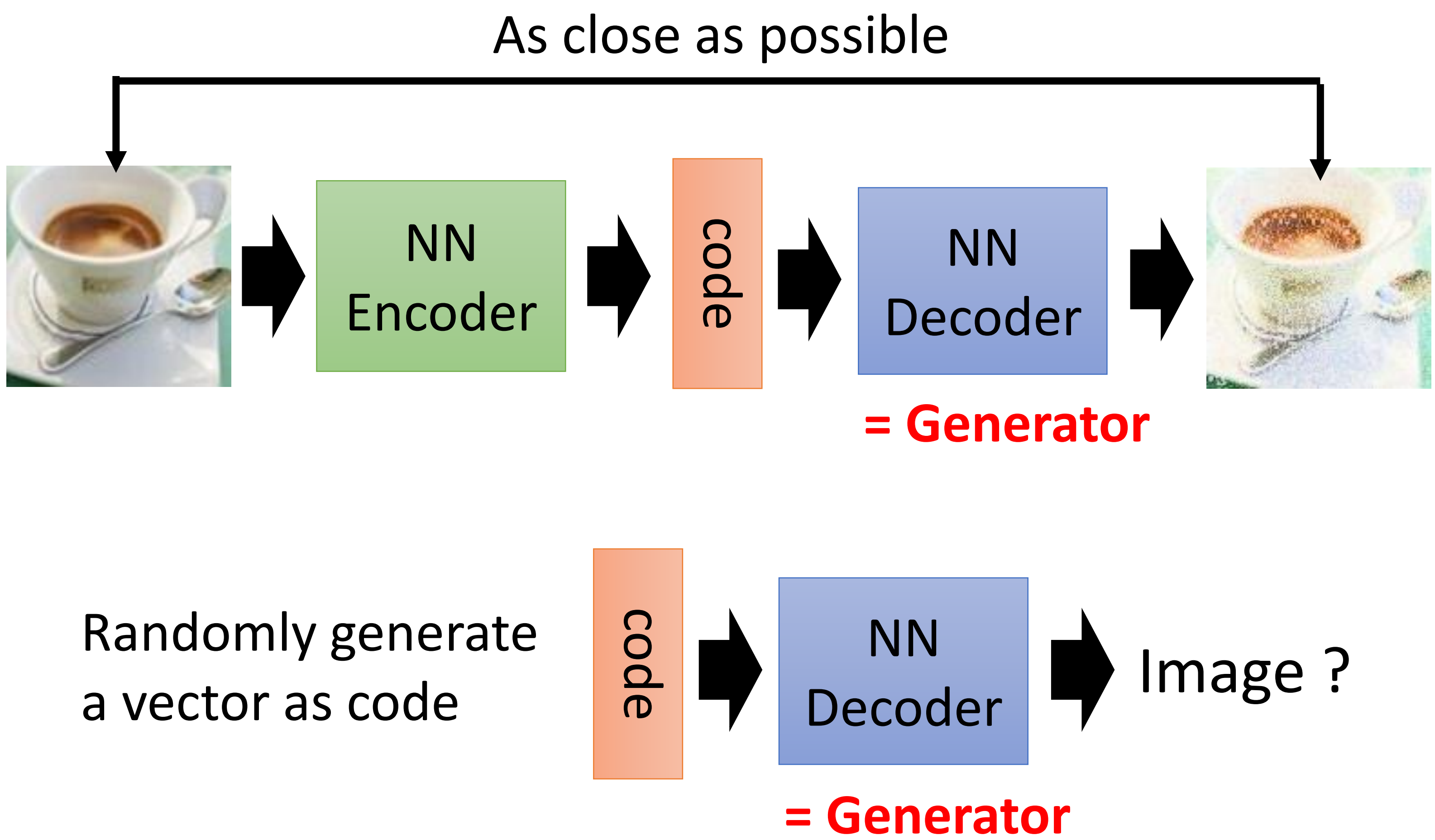




Auto-encoder

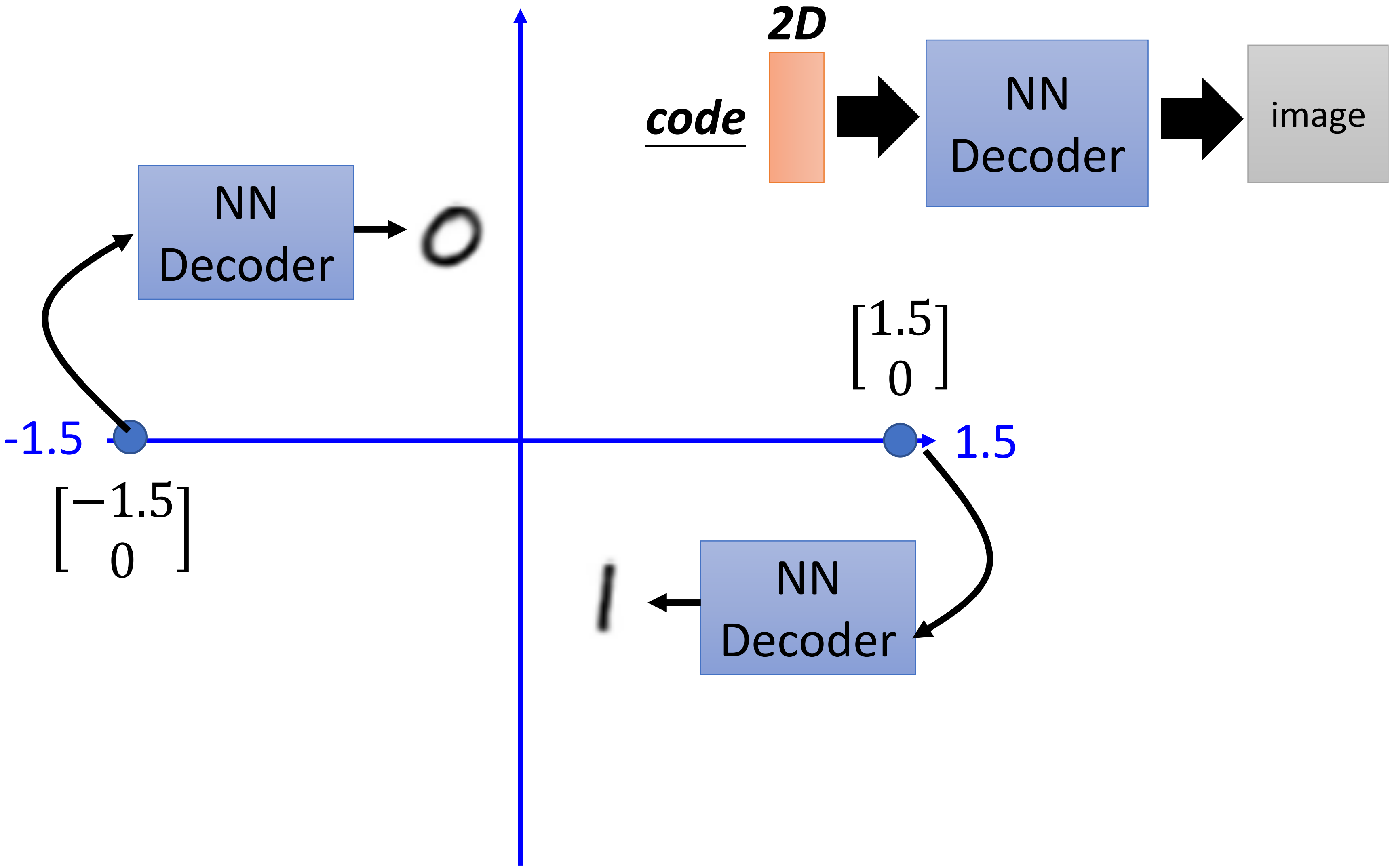


Auto-encoder



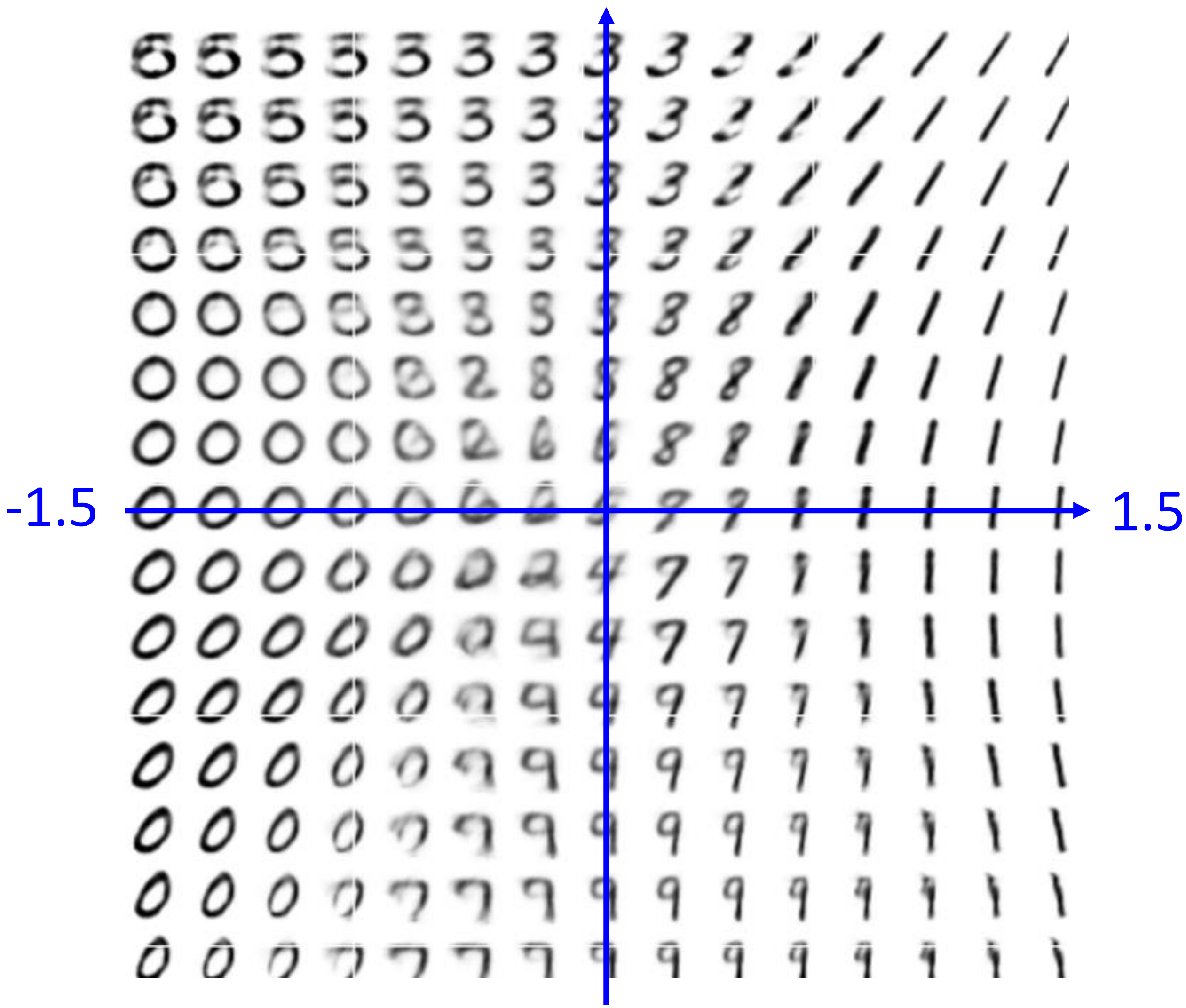
(real examples)

Auto-encoder

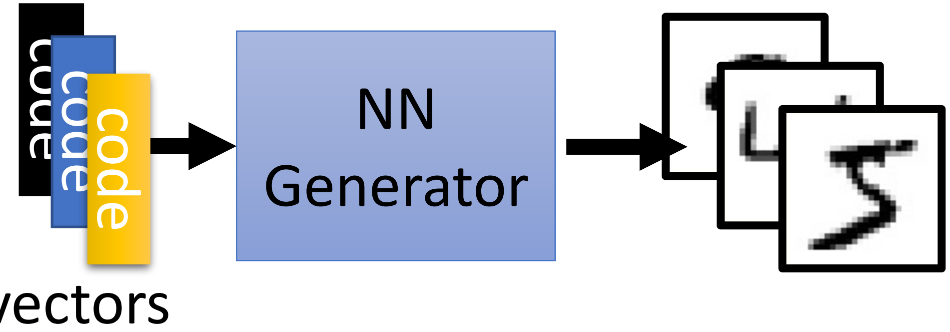


(real examples)

Auto-encoder



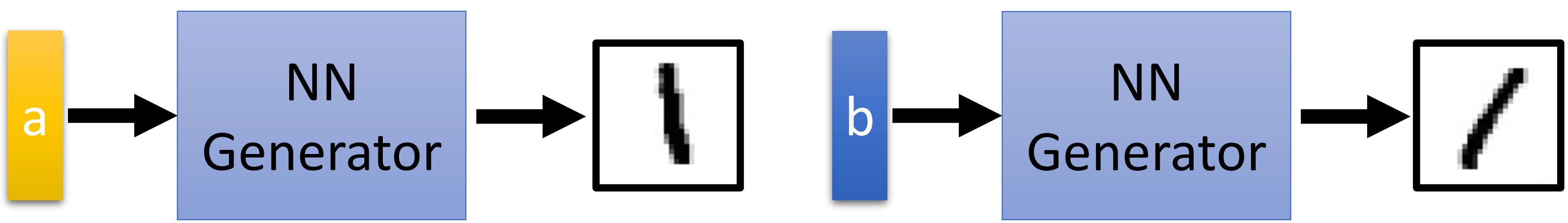
Auto-encoder



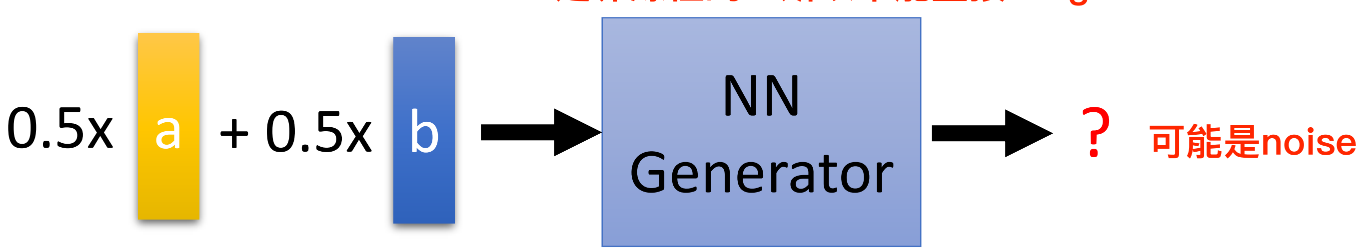
code: $\begin{bmatrix} 0.1 \\ -0.5 \end{bmatrix}$ $\begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix}$ $\begin{bmatrix} 0.2 \\ -0.1 \end{bmatrix}$ $\begin{bmatrix} 0.3 \\ 0.2 \end{bmatrix}$

(where does them come from?)

Image:



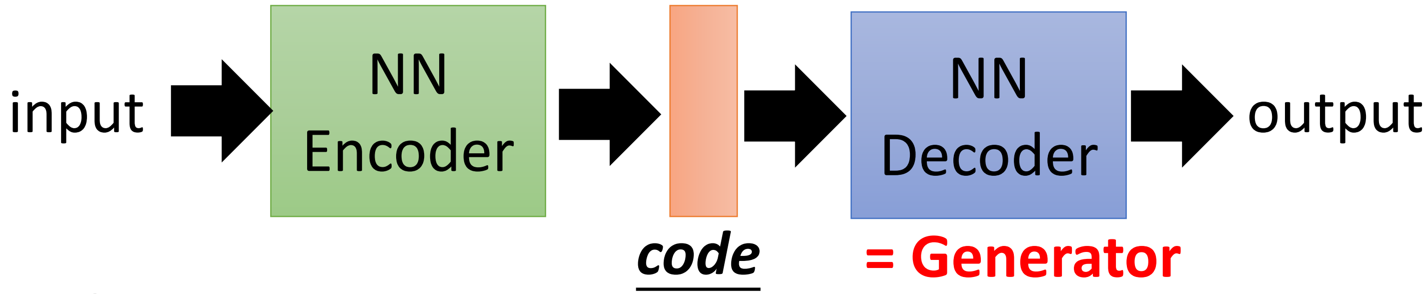
network是非線性的，所以不能直接weight sum



Auto-encoder

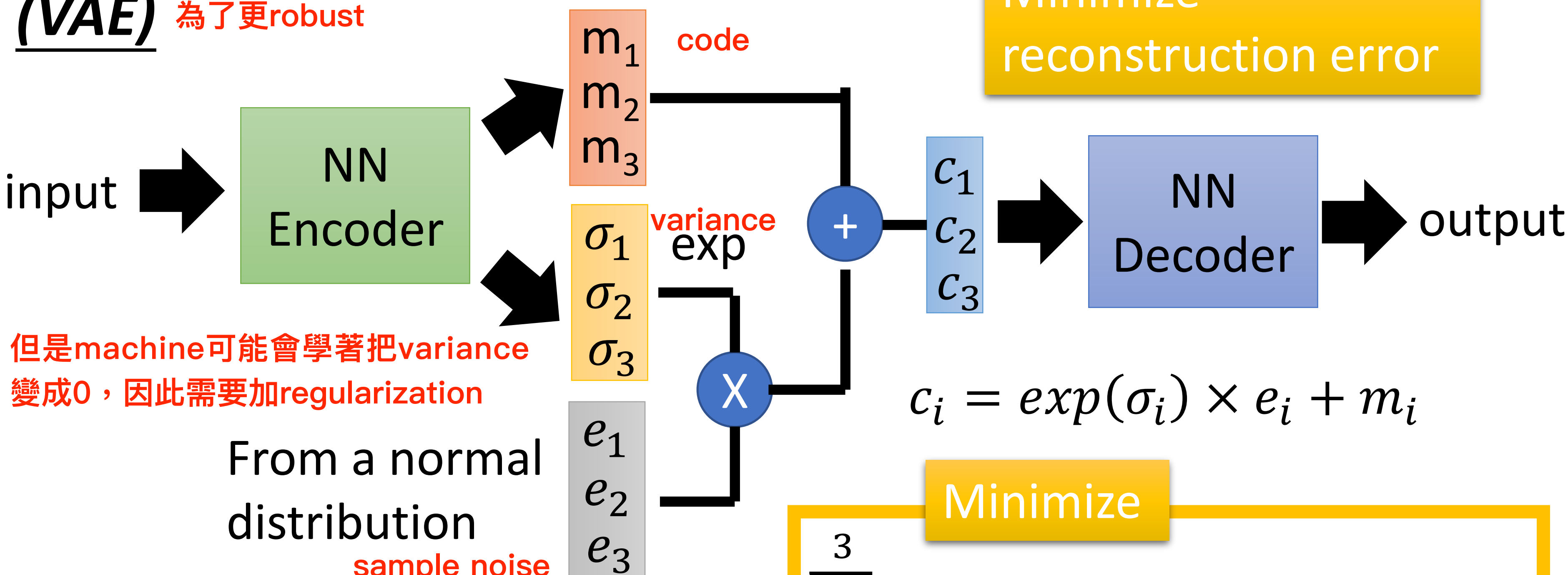
相信這些high dimension的影像的manifold其實就是這個code

如果code太高維雖然好train，但是machine可能會學著一直copy即可



Variational Auto-encoder (VAE)

為了更robust



但是machine可能會學著把variance變成0，因此需要加regularization

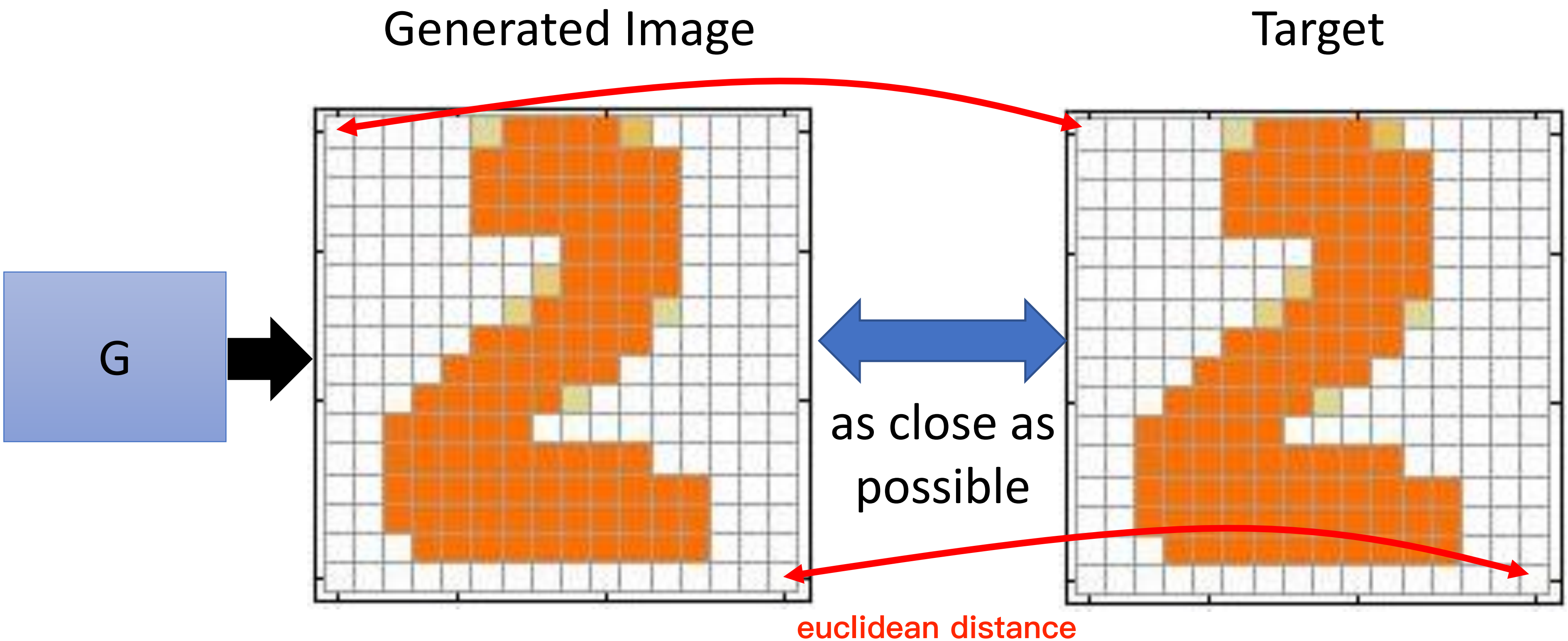
$$c_i = exp(\sigma_i) \times e_i + m_i$$

Minimize

$$\sum_{i=1}^{3}(exp(\sigma_i) - (1 + \sigma_i) + (m_i)^2)$$

machine不只要能看到vector a 產生1，還會學到看到vector a+上一堆noise，也要能夠產生1

What do we miss?



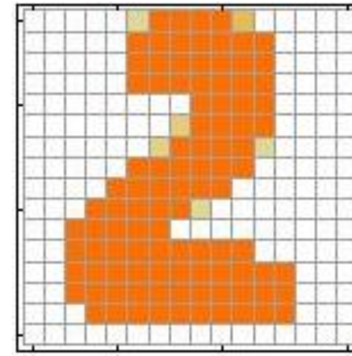
It will be fine if the generator can truly copy the target image.

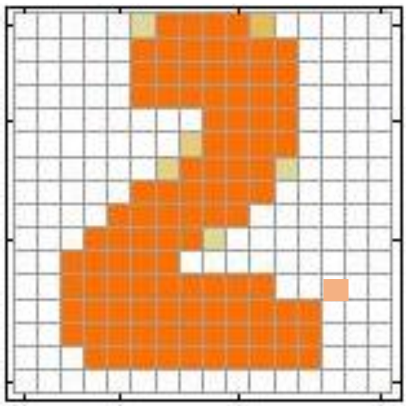
What if the generator makes some mistakes …….

Some mistakes are serious, while some are fine.

What do we miss?

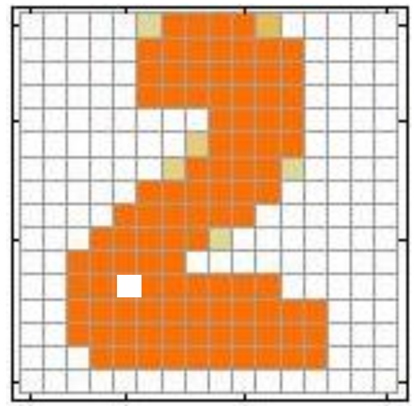
Target





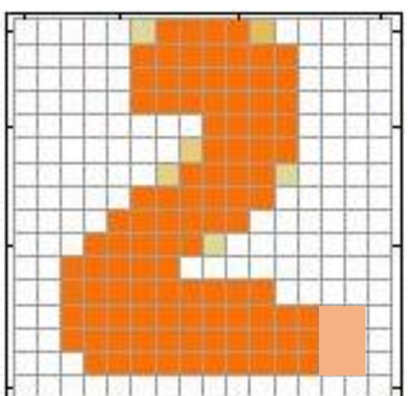
1 pixel error

我覺得不行



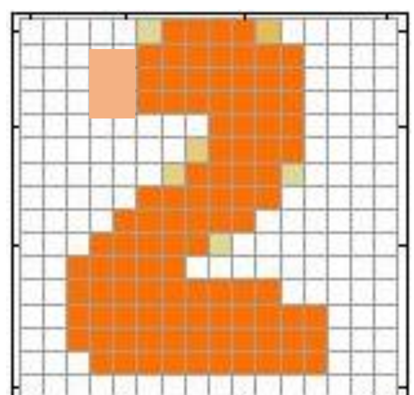
1 pixel error

我覺得不行



6 pixel errors

我覺得其實可以



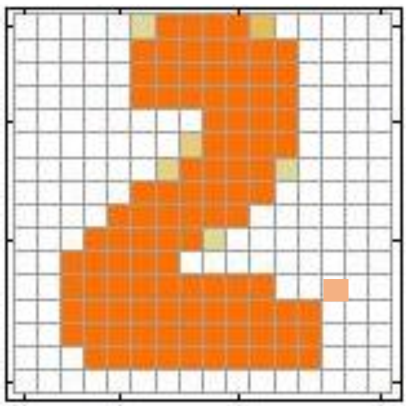
6 pixel errors

我覺得其實可以

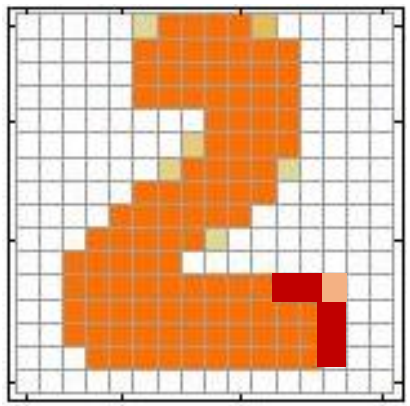
network架構很難考慮到component之間的關係

What do we miss?

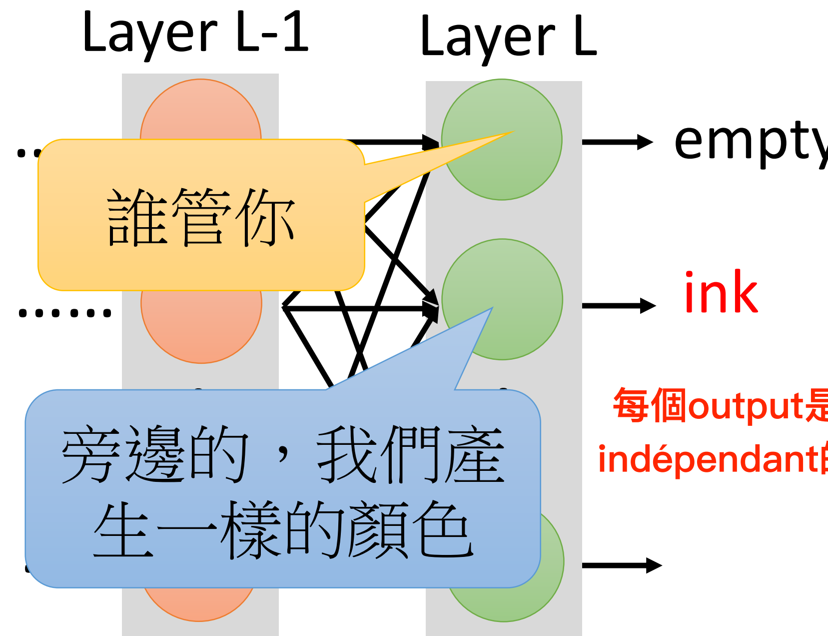
Each neural in output layer corresponds to a pixel.



我覺得不行



我覺得其實可以



每個output是 indépendant的

The relation between the components are critical.

Although highly correlated, they cannot influence each other.

Need deep structure to catch the relation between components.

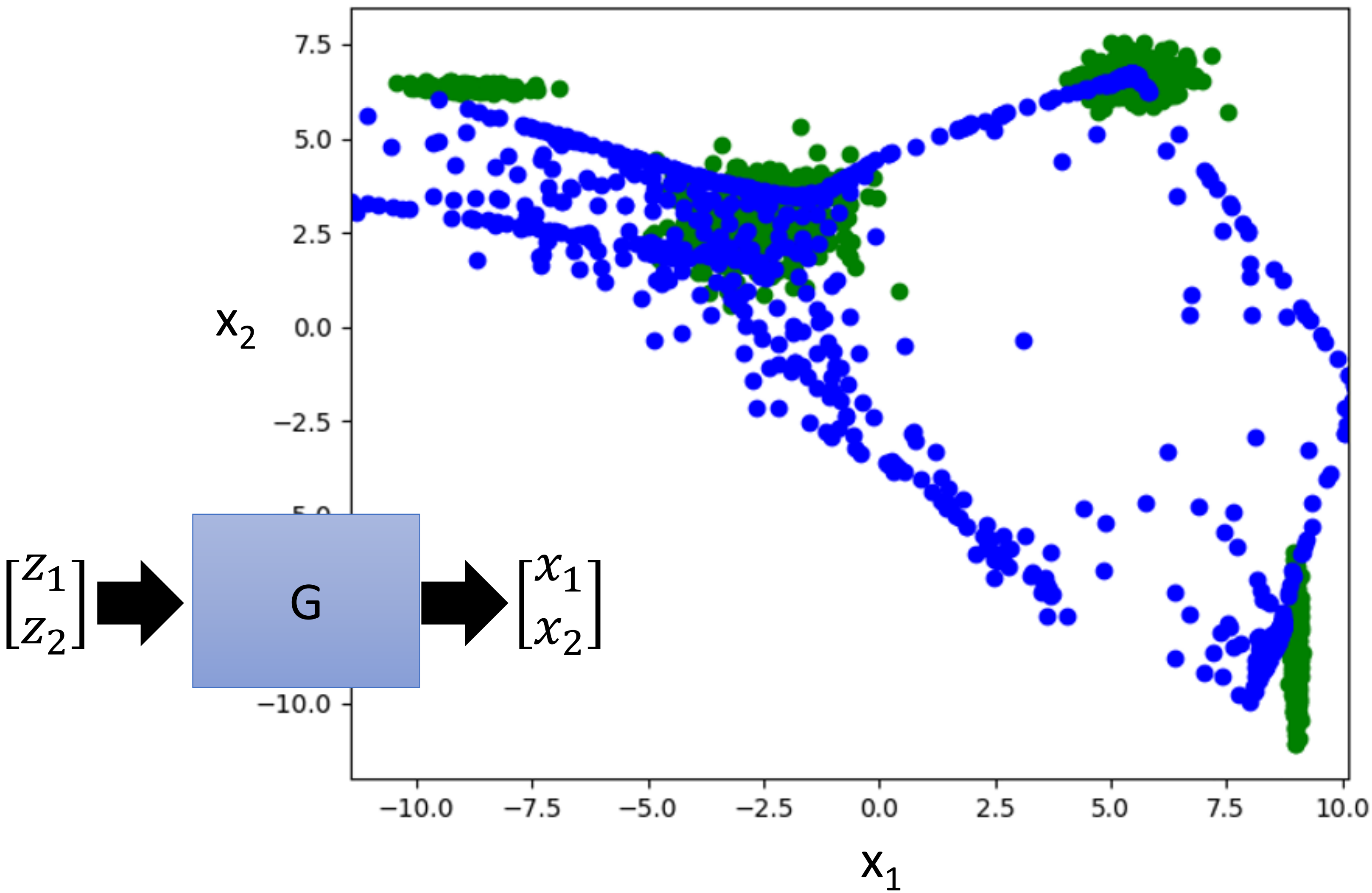
如果今天train一個GAN vs auto-encoder，往往auto-encoder的架構要更深才能達到接近GAN的結果

藍色是auto-encoder的generator產生出來的distribution
綠色是ground truth的distribution

感謝 黃淞楓 同學提供結果

(Variational) Auto-encoder

對generator要考慮pixel之間的correlation是很難的



Outline

- Basic Idea of GAN
- GAN as structured learning
- Can Generator learn by itself?
- Can Discriminator generate?
- A little bit theory

Discriminator

Evaluation function, Potential Function, Energy Function …

- Discriminator is a function D (network, can deep)

$$D : X \rightarrow R$$

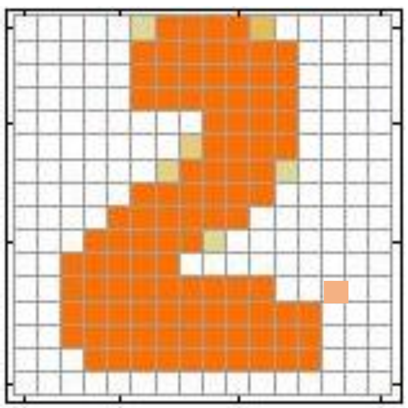
  - Input x: an object x (e.g. an image)
  - Output D(x): scalar which represents how “good” an object x is

Can we use the discriminator to generate objects?

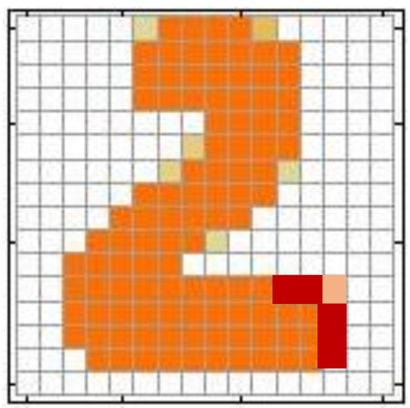
Yes.

Discriminator

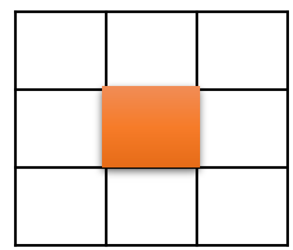
- It is easier to catch the relation between the components by top-down evaluation.



我覺得不行



我覺得其實 OK



This CNN filter is good enough.

Discriminator

- Suppose we already have a good discriminator D(x) …

窮舉所有image丟進去D看誰大，就sample他～

Inference

- Generate object $\tilde{x}$ that

$$\tilde{x} = \arg\max_{x \in X} D(x)$$

Enumerate all possible x !!!

It is feasible ???

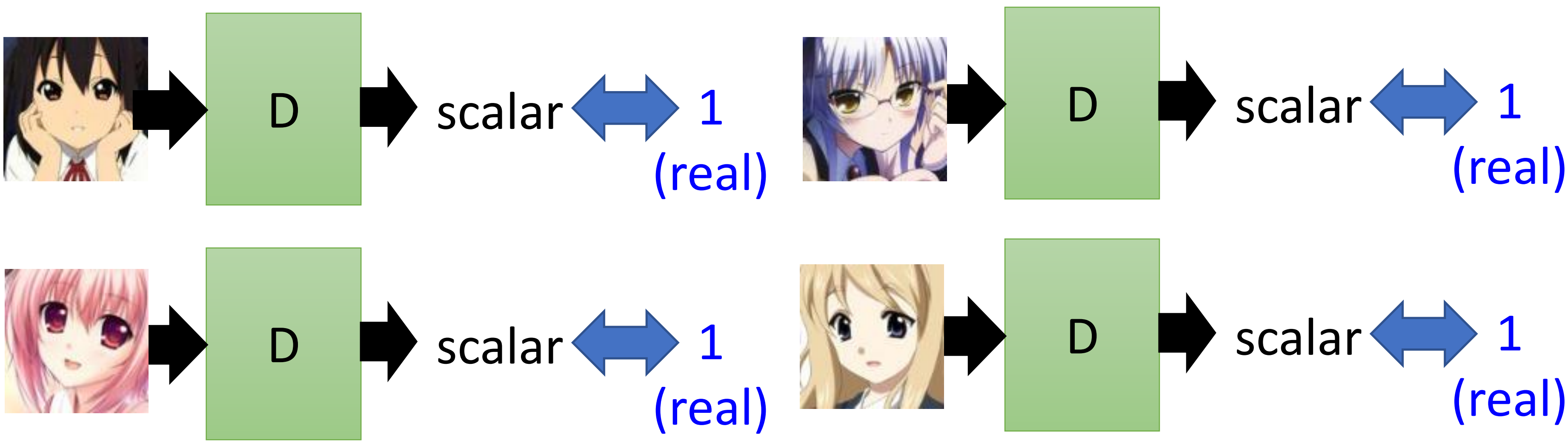
How to learn the discriminator?

D是擅長批評的，因此生成東西很痛苦

Discriminator - Training

data只有positive example的情况

- I have some real images



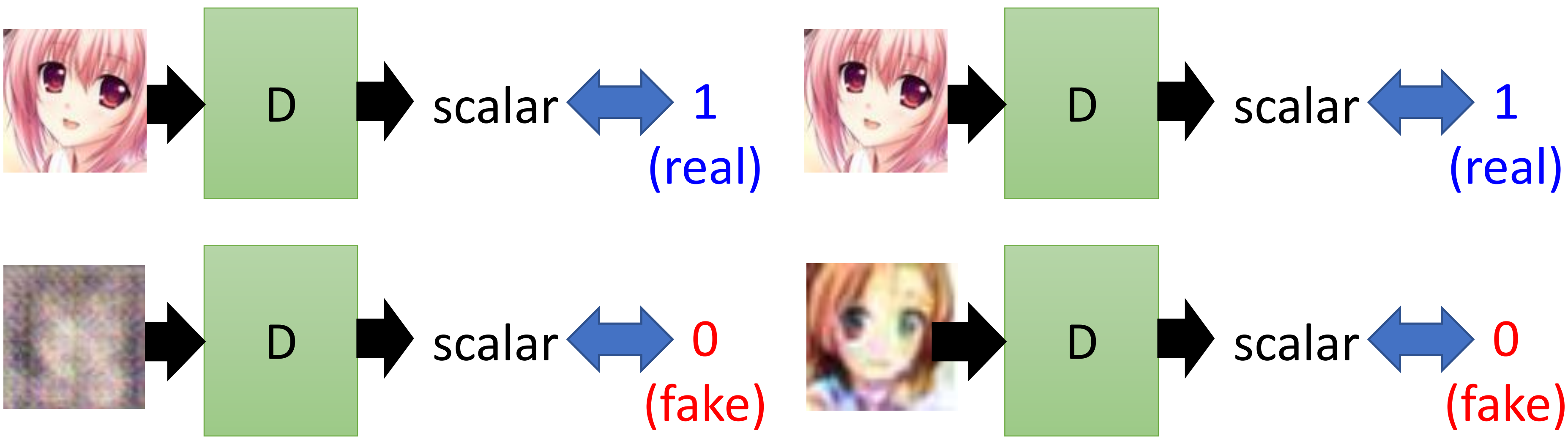
Discriminator only learns to output “1” (real).

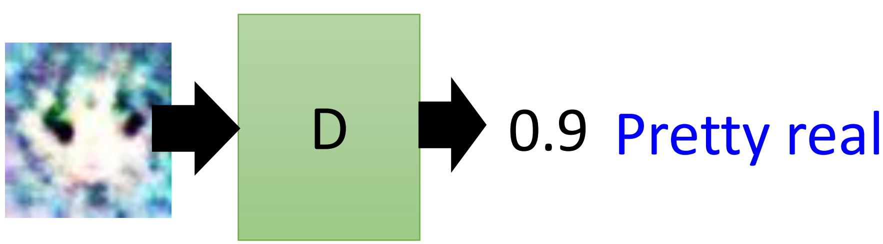
Discriminator training needs some negative examples.

Discriminator - Training

如何產生好的negative example，這樣D才能真的學會辨別
好的negative example就是逼近真實的negative example

- Negative examples are critical.

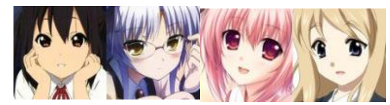




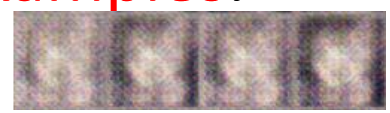
How to generate realistic negative examples?

Discriminator - Training

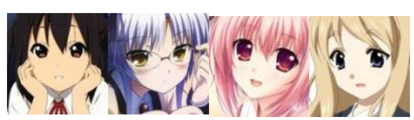
- **General Algorithm**



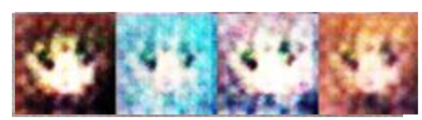
  - Given a set of positive examples, randomly generate a set of negative examples.

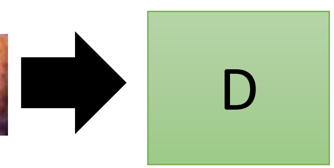


  - In each iteration
    - Learn a discriminator D that can discriminate positive and negative examples.



v.s.

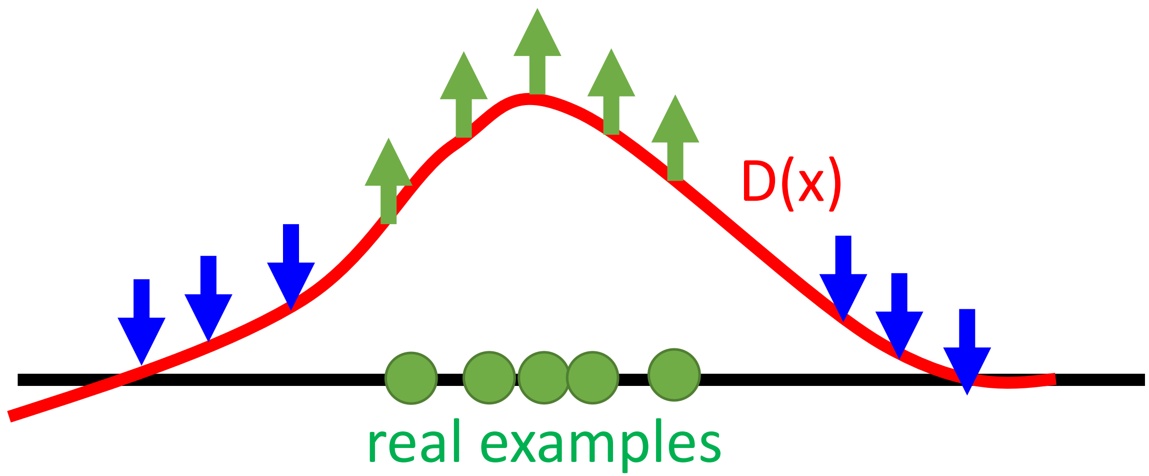




    - Generate negative examples by discriminator D

$$\tilde{x} = \arg\max_{x \in X} D(x)$$

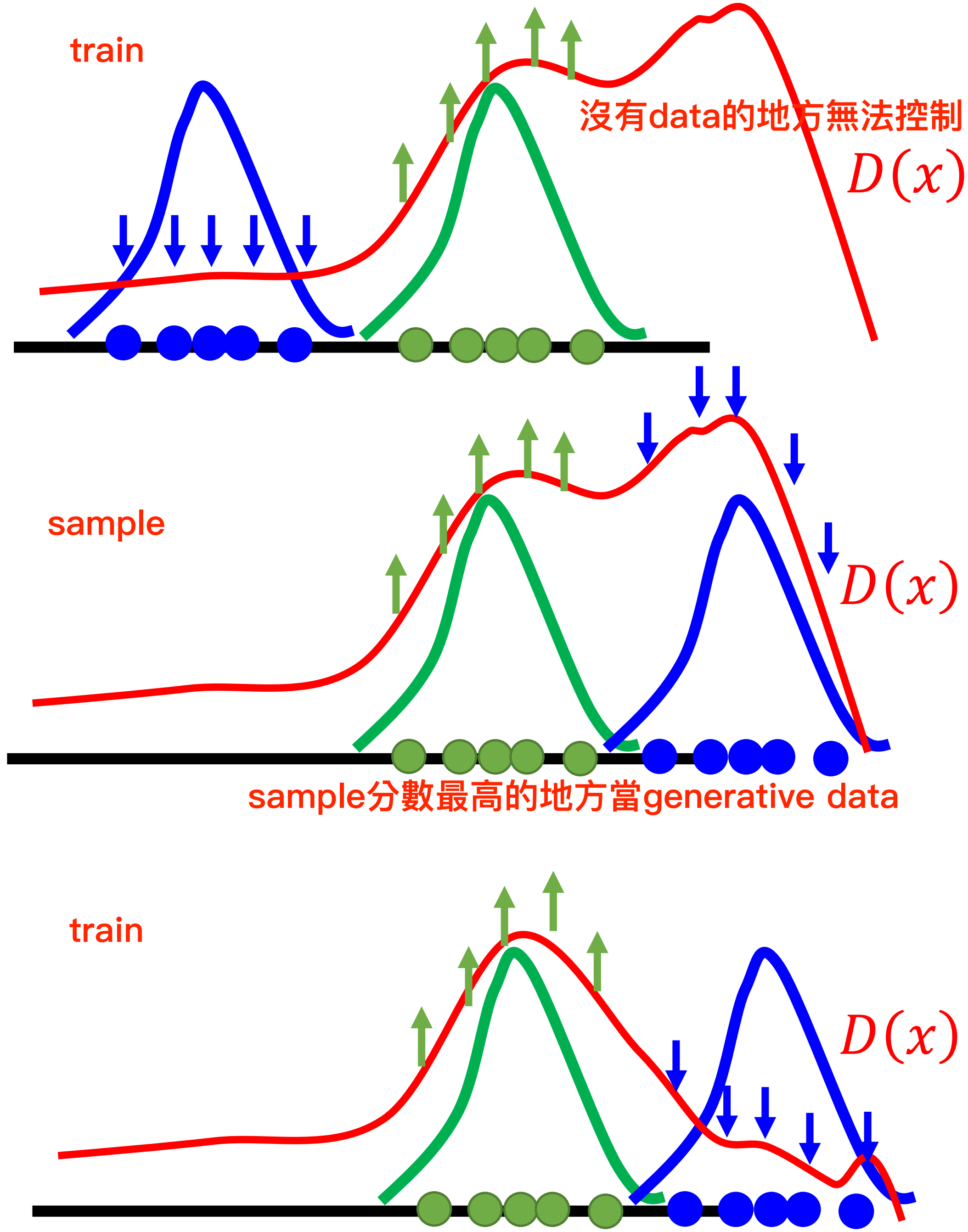
Discriminator - Training



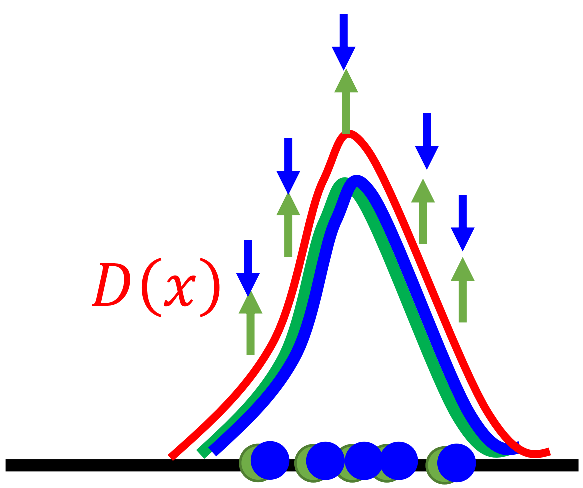
In practice, you cannot decrease all the x
other than real examples.

Discriminator - Training

- real
- generated



In the end ......



Structured Learning

- Structured Perceptron
- Structured SVM
- Gibbs sampling
- Hidden information
- Application: sequence labelling, summarization

上述train D的方法

Graphical Model

structure learning的一種

常見到的方法

Bayesian Network (Directed Graph)

Markov Random Field (Undirected Graph)

Energy-based Model: http://www.cs.nyu.edu/~yann/research/ebm/

Conditional Random Field

Markov Logic Network

Boltzmann Machine

Segmental CRF

(Only list some of the approaches)

Restricted Boltzmann Machine

Generator v.s. Discriminator

- ***Generator***
- Pros: 容易生成
  - Easy to generate even with deep model
- Cons:
  不容易考慮component之間的correlation
  - Imitate the appearance
  - Hard to learn the correlation between components

- ***Discriminator***
- Pros: 可以考慮大局
  - Considering the big picture
- Cons: 生個東西真的千難萬難
  - Generation is not always feasible
    - Especially when your model is deep
  - How to do negative sampling?

要假設D是線性的才能去解argmax問題
但是假設線性的又限制他的能力

Generator + Discriminator

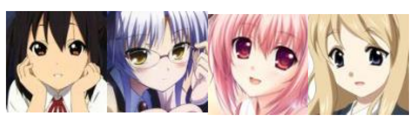
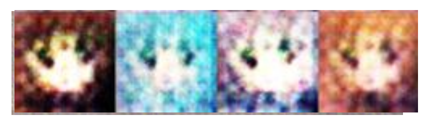
generator就是在學怎麼解argmax

Generator可以取代解argmax這個problem

- **General Algorithm**

  - Given a set of positive examples, randomly generate a set of negative examples.

  - In each iteration
    - Learn a discriminator D that can discriminate positive and negative examples.

 v.s.  → D

    - Generate negative examples by discriminator D

$$\boxed{G \rightarrow \tilde{x}} = \boxed{\tilde{x} = \arg\max_{x \in X} D(x)}$$

Benefit of GAN

- From Discriminator's point of view
  - Using generator to generate negative samples

$$\boxed{G \longrightarrow \tilde{x}} \quad = \quad \boxed{\tilde{x} = \arg\max_{x \in X} D(x)}$$

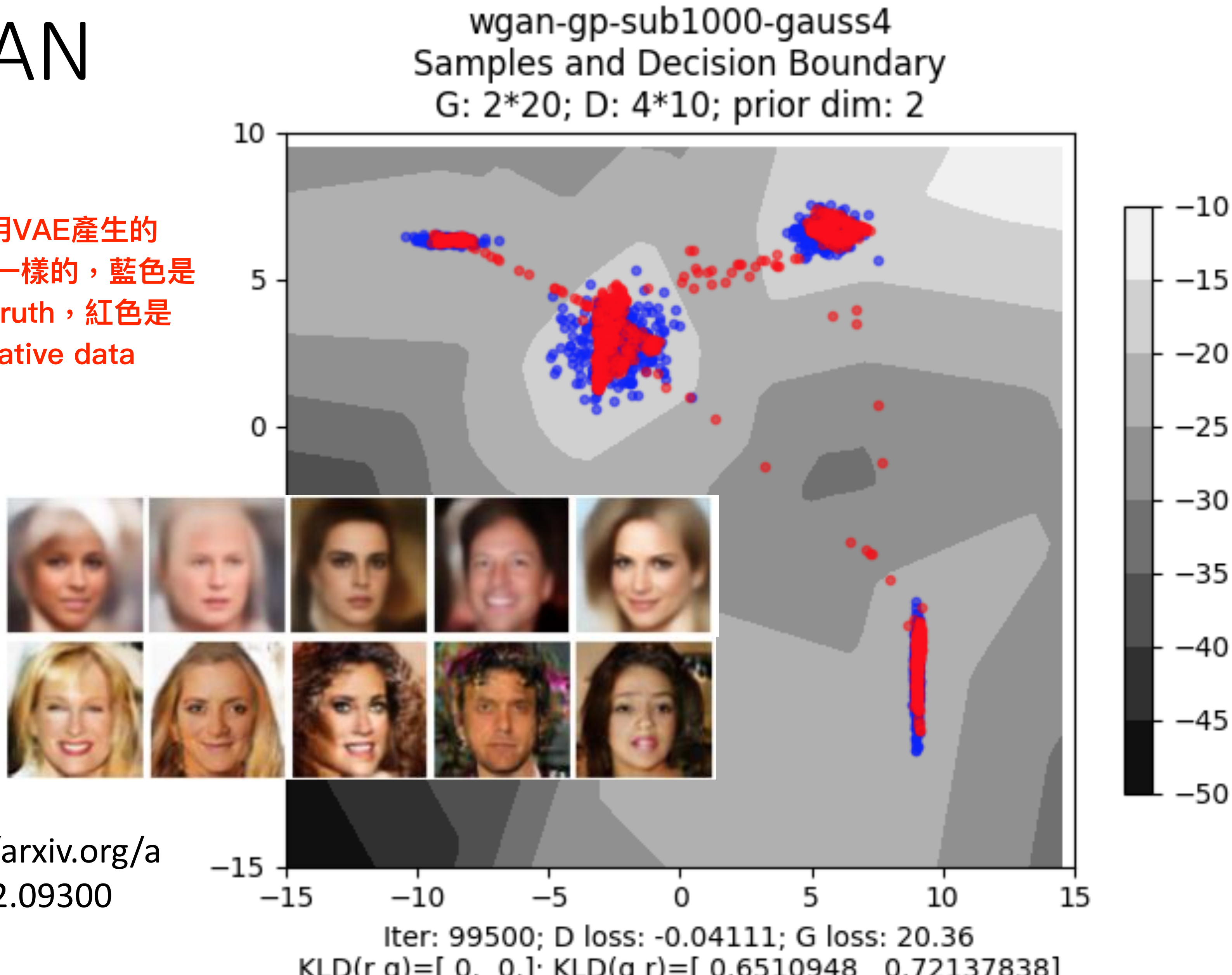
efficient

- From Generator's point of view
  - Still generate the object component-by-component
  - But it is learned from the discriminator with global view.

感謝 段逸林 同學提供結果

GAN

跟剛剛用VAE產生的dataset是一樣的，藍色是ground truth，紅色是generative data



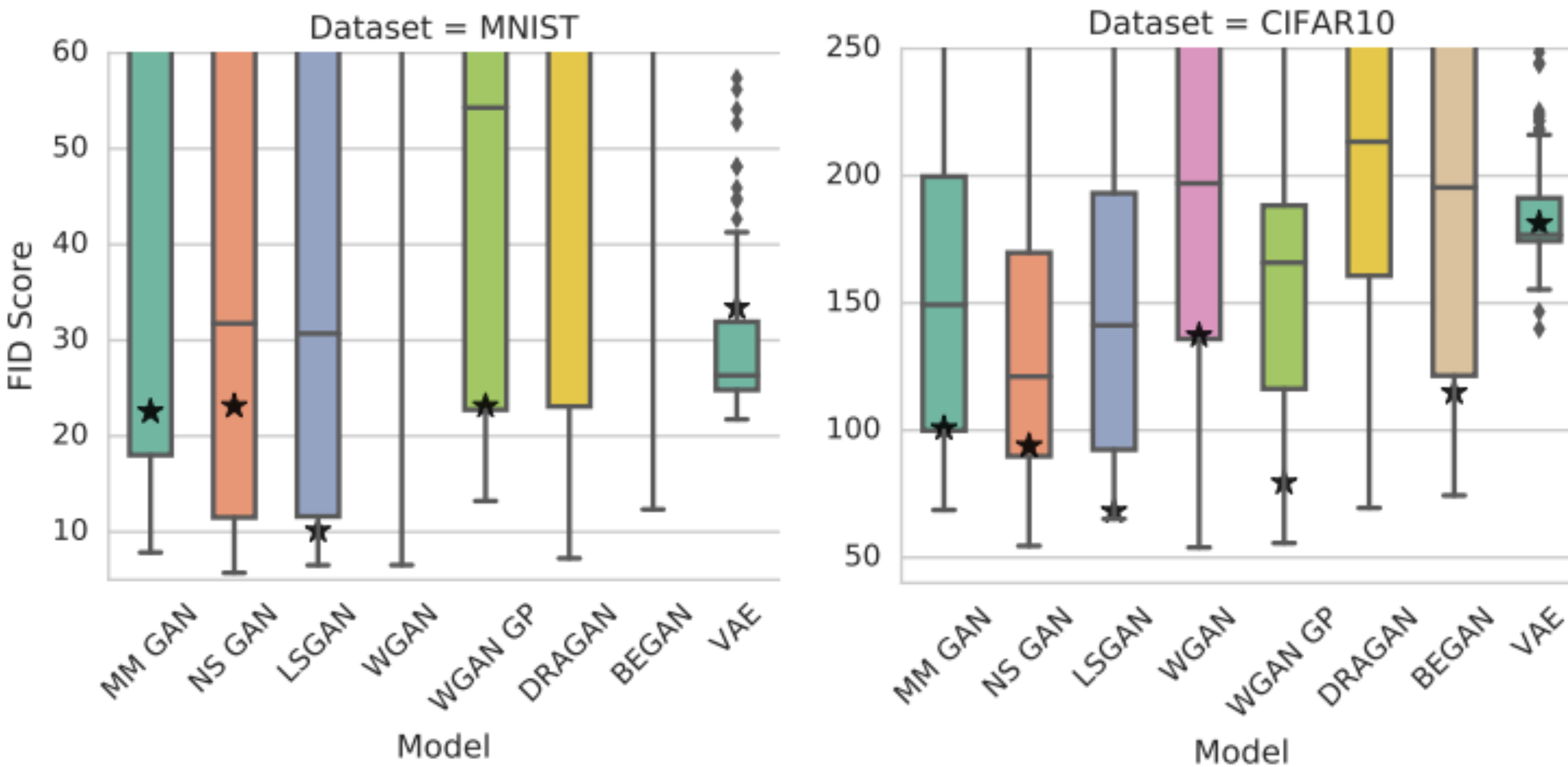
VAE

GAN

https://arxiv.org/abs/1512.09300

[Mario Lucic, et al. arXiv, 2017]

調整不同參數後得到的結果，可以發現GAN的變化是很大的但是upper bound比較好
值得一提的是ＶＡＥ相較起來穩定許多



比較各種不同的ＧＡＮ的結果　　結論是結果都差不多

FID[Martin Heusel, et al., NIPS, 2017]: Smaller is better

google 的paper　　FID越小代表越像真實的圖片

Next Time

- Preview
  - https://youtu.be/0CKeqXl5IY0
  - https://youtu.be/KSN4QYgAtao

Outline

Basic Idea of GAN

GAN as structured learning

Can Generator learn by itself?

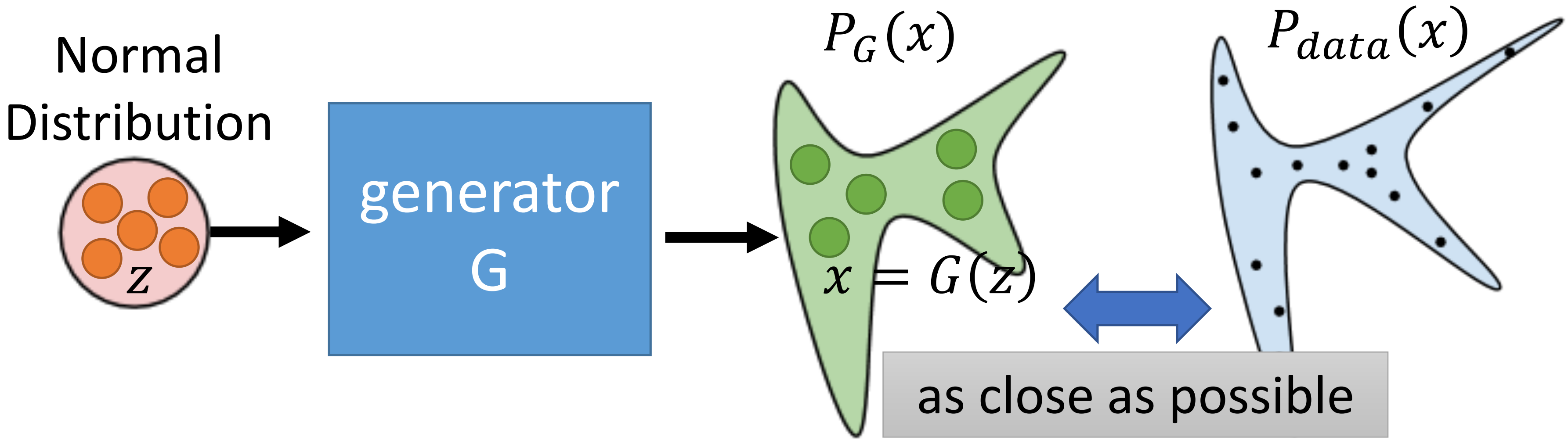
Can Discriminator generate?

A little bit theory

Generator

x: an image (a high-dimensional vector)

- A generator G is a network. The network defines a probability distribution P_G



$$G^* = arg \min_G \underline{Div(P_G, P_{data})}$$

Divergence between distributions P_G and P_{data}

How to compute the divergence?

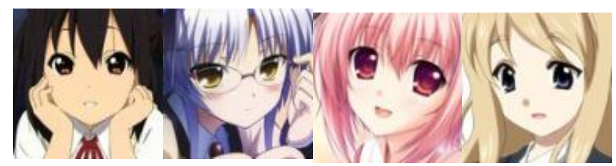
Discriminator

$$G^* = arg \min_G Div(P_G, P_{data})$$

Although we do not know the distributions of P_G and P_{data}, we can sample from them.

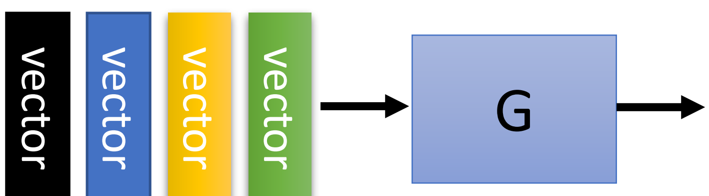
Database

sample



Sampling from $\boldsymbol{P_{data}}$

sample from normal



Sampling from $\boldsymbol{P_G}$

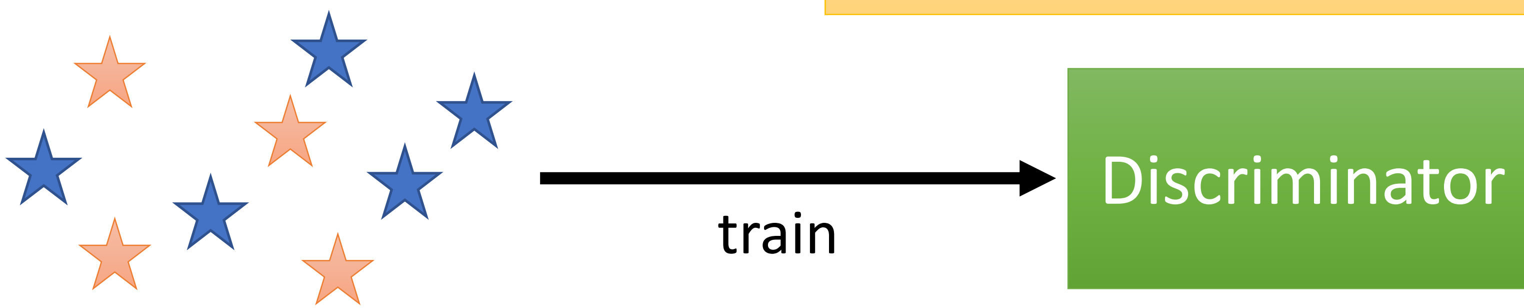
Discriminator

$$G^* = arg \min_G Div(P_G, P_{data})$$

: data sampled from P_{data}

: data sampled from P_G

Using the example objective function is exactly the same as training a binary classifier.



Example Objective Function for D

$$V(G,D) = E_{x \sim P_{data}}[logD(x)] + E_{x \sim P_G}[log(1 - D(x))]$$

(G is fixed)

Training: $D^* = arg \boxed{\max_D V(D,G)}$

The maximum objective value is related to JS divergence.

Discriminator

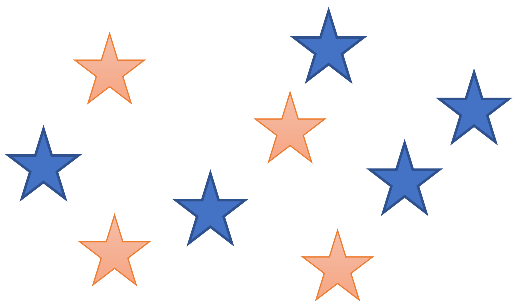
$$G^* = arg \min_G Div(P_G, P_{data})$$

★ : data sampled from P_{data}

★ : data sampled from P_G

Training:

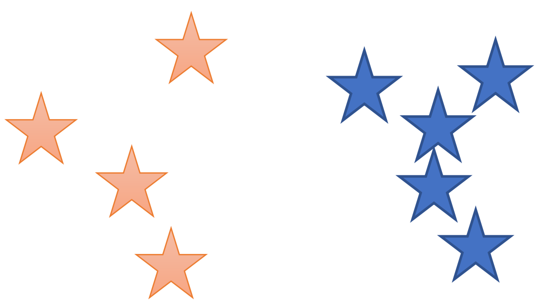
$$D^* = arg \boxed{\max_D V(D, G)}$$



train → Discriminator

small divergence

hard to discriminate
(cannot make objective large)



train → Discriminator

large divergence

easy to discriminate

[Goodfellow, et al., NIPS, 2014]

$$G^* = arg \min_G \max_D V(G,D)$$

$$D^* = arg \boxed{\max_D V(D,G)}$$

The maximum objective value is related to JS divergence.

- Initialize generator and discriminator
- In each training iteration:

Step 1: Fix generator G, and update discriminator D

Step 2: Fix discriminator D, and update generator G

Can we use other divergence?

Name	$D_f(P\|Q)$	Generator $f(u)$
Total variation	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert \,\mathrm{d}x$	$\frac{1}{2}\lvert u-1\rvert$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x$	$u\log u$
Reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\,\mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)}\,\mathrm{d}x$	$(u-1)^2$
Neyman χ^2	$\int \frac{(p(x)-q(x))^2}{q(x)}\,\mathrm{d}x$	$\frac{(1-u)^2}{u}$
Squared Hellinger	$\int \left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2\,\mathrm{d}x$	$(\sqrt{u}-1)^2$
Jeffrey	$\int (p(x)-q(x))\log\left(\frac{p(x)}{q(x)}\right)\,\mathrm{d}x$	$(u-1)\log u$
Jensen-Shannon	$\frac{1}{2}\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x$	$-(u+1)\log\frac{1+u}{2}+u\log u$
Jensen-Shannon-weighted	$\int p(x)\pi\log\frac{p(x)}{\pi p(x)+(1-\pi)q(x)}+(1-\pi)q(x)\log\frac{q(x)}{\pi p(x)+(1-\pi)q(x)}\,\mathrm{d}x$	$\pi u\log u-(1-\pi+\pi u)\log(1-\pi+\pi u)$
GAN	$\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x-\log(4)$	$u\log u-(u+1)\log(u+1)$

Using the divergence you like ☺

Name	Conjugate $f^*(t)$
Total variation	t
Kullback-Leibler (KL)	$\exp(t-1)$
Reverse KL	$-1-\log(-t)$
Pearson χ^2	$\frac{1}{4}t^2+t$
Neyman χ^2	$2-2\sqrt{1-t}$
Squared Hellinger	$\frac{t}{1-t}$
Jeffrey	$W(e^{1-t})+\frac{1}{W(e^{1-t})}+t-2$
Jensen-Shannon	$-\log(2-\exp(t))$
Jensen-Shannon-weighted	$(1-\pi)\log\frac{1-\pi}{1-\pi e^{t/\pi}}$
GAN	$-\log(1-\exp(t))$