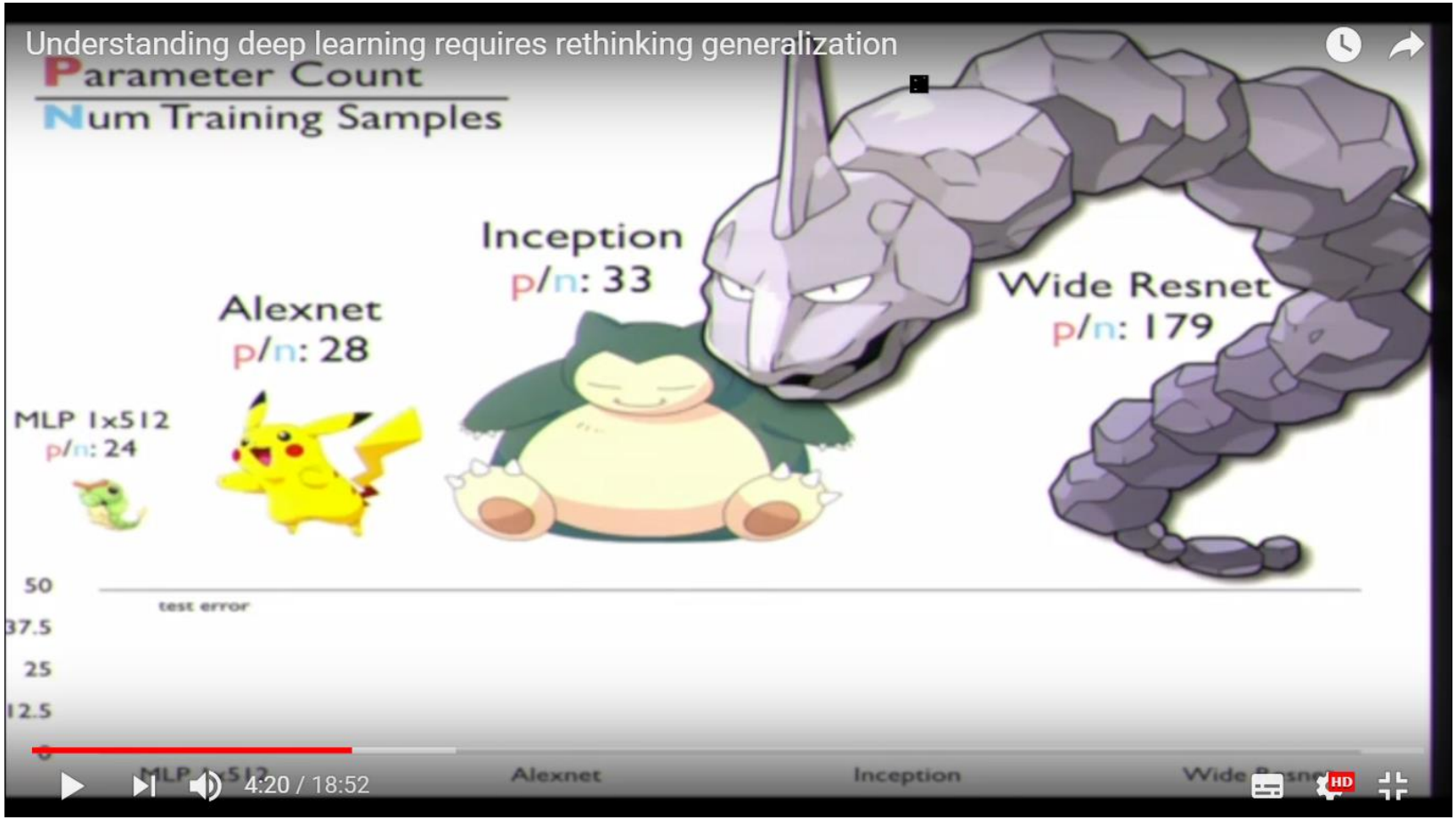
Generalization Ability

We use very large network today

參數量是遠大於training data的量的



Source of image: https://www.youtube.com/watch?v=kCj51pTQPKI

機器學習基石
https://youtu.be/XxPB9GlJEUk

Generalization Gap

R:training data 量

M: model capacity(function set)大小

No matter the data distribution

With probability $1-\delta$ 發生下面這個等式

正確率

$$E_{train} \leq E_{test} \leq E_{train} + \Omega(R, M, \delta)$$

Smaller δ, larger Ω

R is the number of training data → Larger R, smaller Ω

M is the "capacity" of your model → Larger M, larger Ω

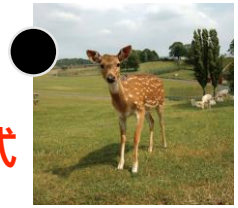
("size" of the function set) model越大(能力越強)越容易overfitting

How to measure the "capacity"? VC dimension (d_{VC})

利用vc dimension來evaluate model capacity

如果故意亂教而model都學的起來（overfitting），則代表ㄖ⨯他 vc dimension >=3

Given 3 data points



總共有八種不同的label方式

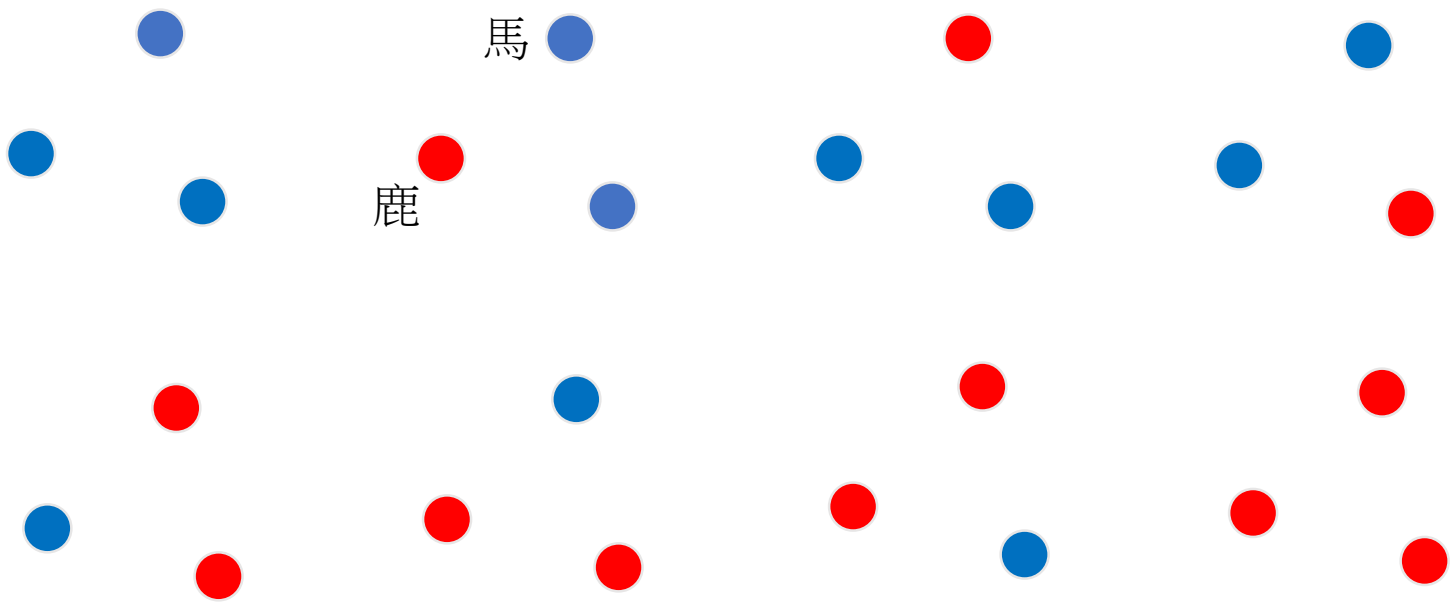
Random label (故意亂教)

Model M can always achieve 0% error rate

(亂教 Model M 都學得會)

VC dimension (d_{VC}) of Model M $\geq$ 3

e.g. linear model



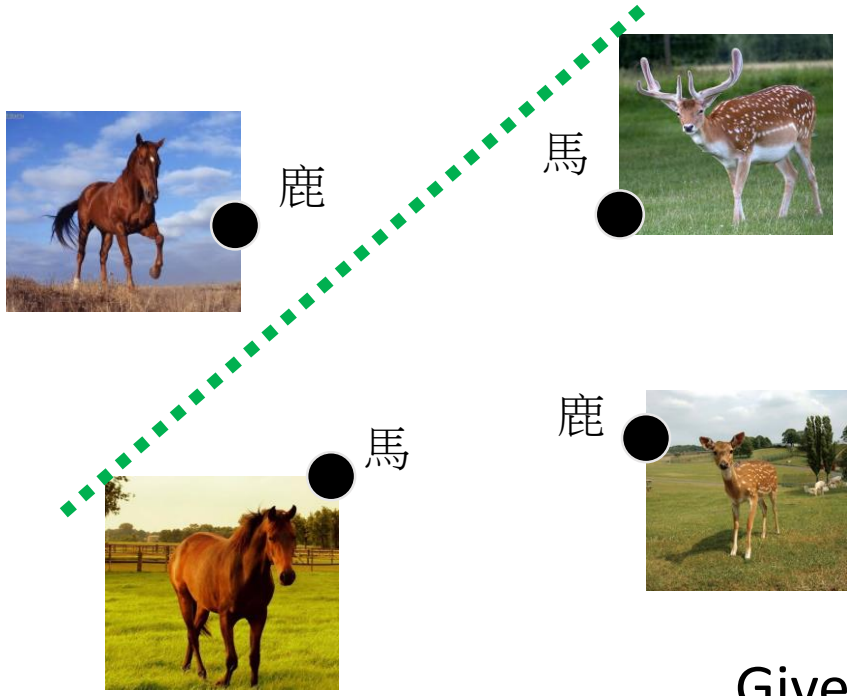
Random label (故意亂教)

There are some cases linear model can not learn.

(知道是來亂的，所以不學)

VC dimension (d_{VC}) of Linear Model < 4

linear model解不了，但deep network是可以解的

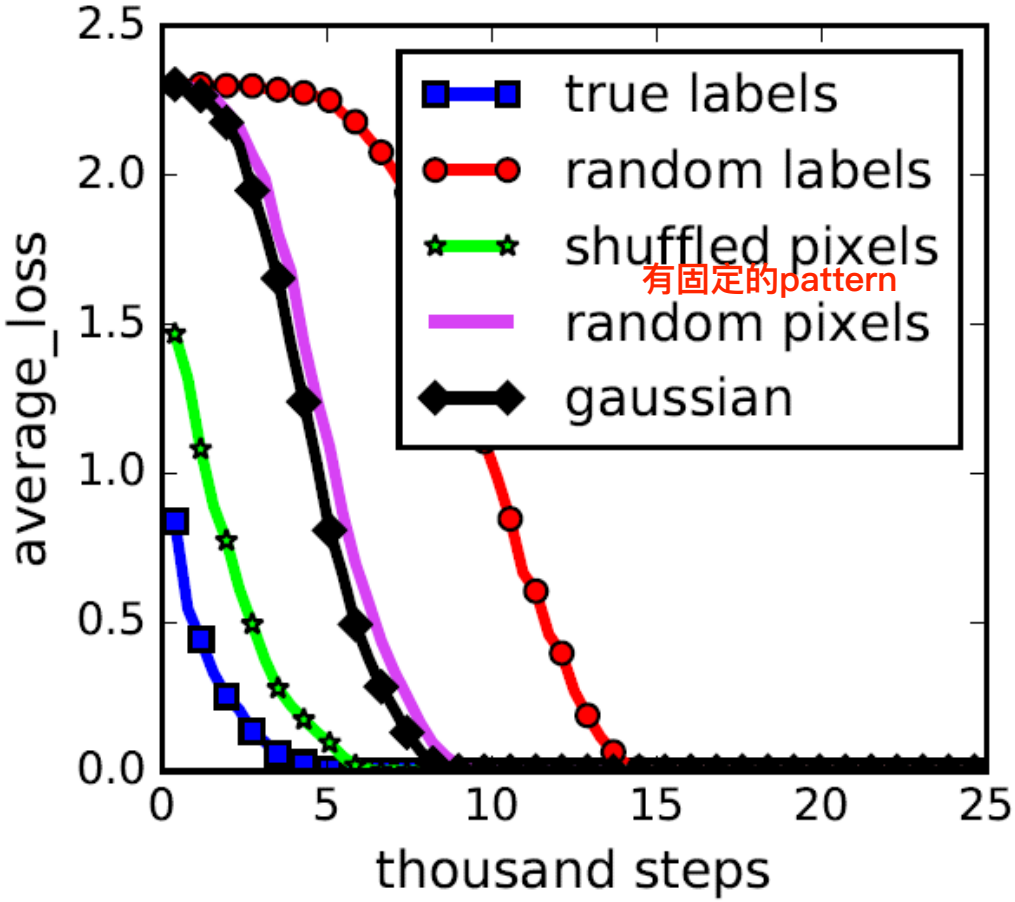


Given 4 data points

What is the capacity of deep models?

如果所有data都random label，network都學的起來(loss = 0)，假設training data = 50000筆，則vc dimension>=50000

Inception model on the CIFAR10



Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals, "Understanding deep learning requires rethinking generalization", ICLR 2017

如果今天training accu已經到100%了M這時如果要提高test accu，則應該降低momdel capacity，ex降低unit數目
但反而提高capacity（提高unit數目）會讓test accu 在上升

Overparameterized Network?

No matter the data distribution

With probability $1-\delta$

$$E_{test} \leq E_{train} + \Omega(R, M, \delta)$$

假設今天有兩個model都可以達到training error = 0
則我們當然選擇model capacity較小的來做為我們的model

Smaller δ, larger Ω

R is the number of training data → Larger R, smaller Ω

M is the "capacity" of your model → Larger M, larger Ω

("size" of the function set)

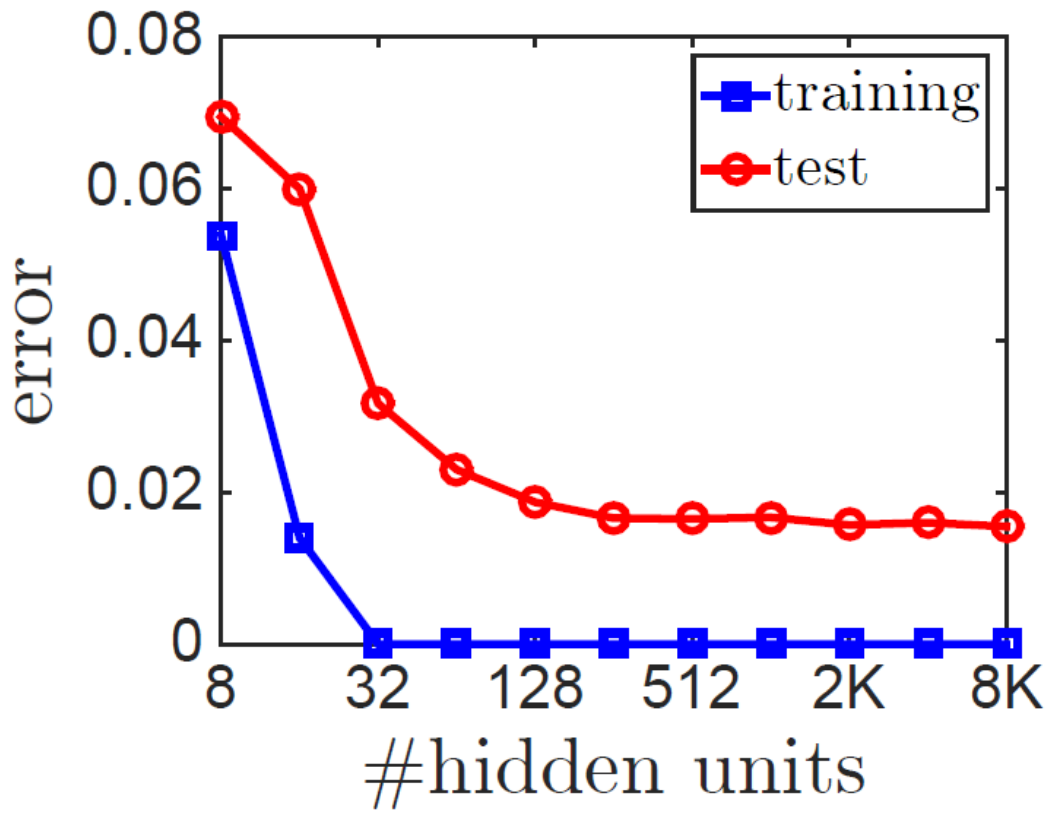
If two models have the same E_{train} → Select the one with smaller capacity

Demo

Overparameterized Network?

loss已經=0的情況下

當hidden layer unit增加的時候, 竟然還可以降低testing error

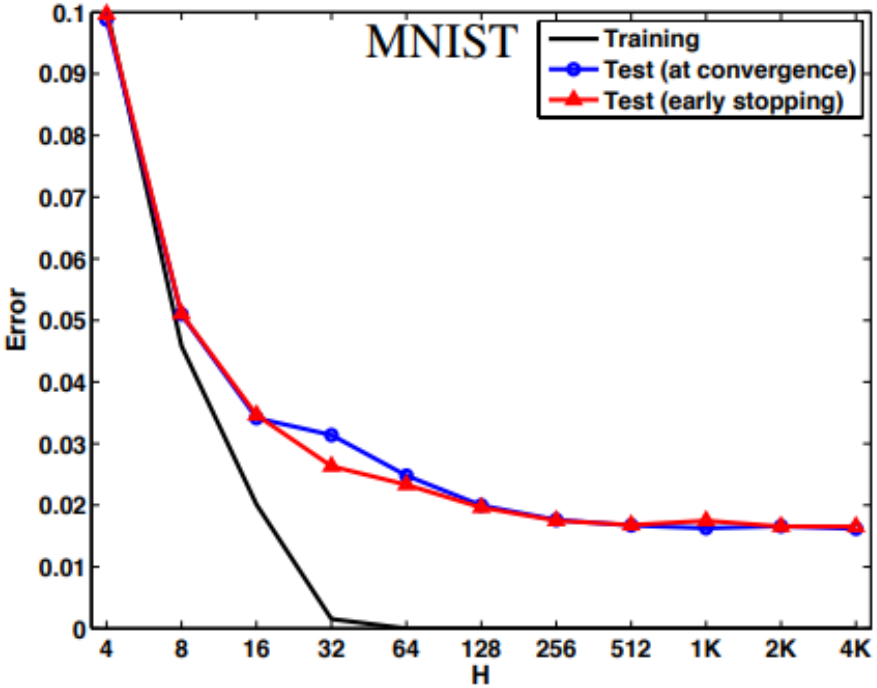


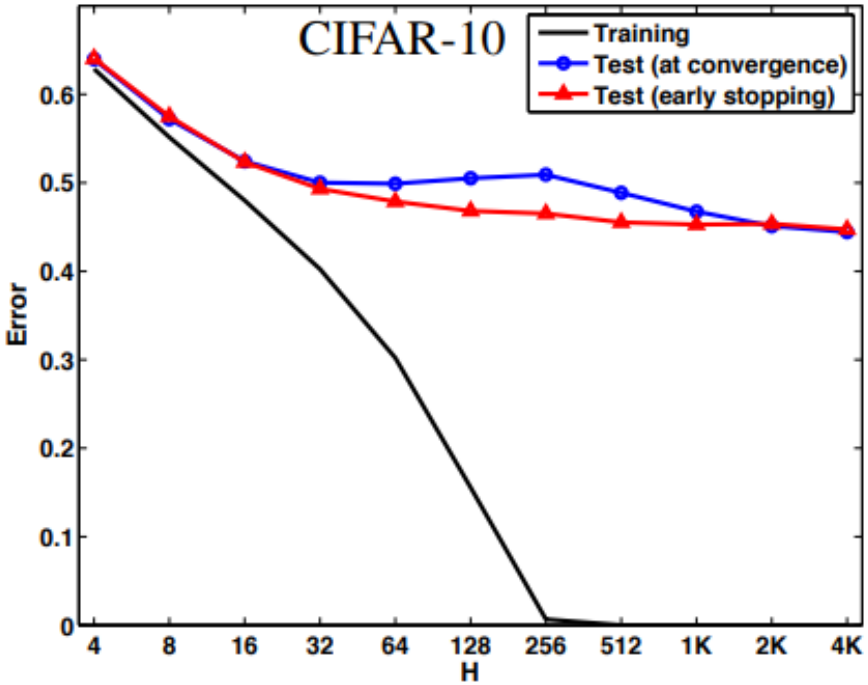
MNIST

https://arxiv.org/abs/1706.08947

Overparameterized Network?

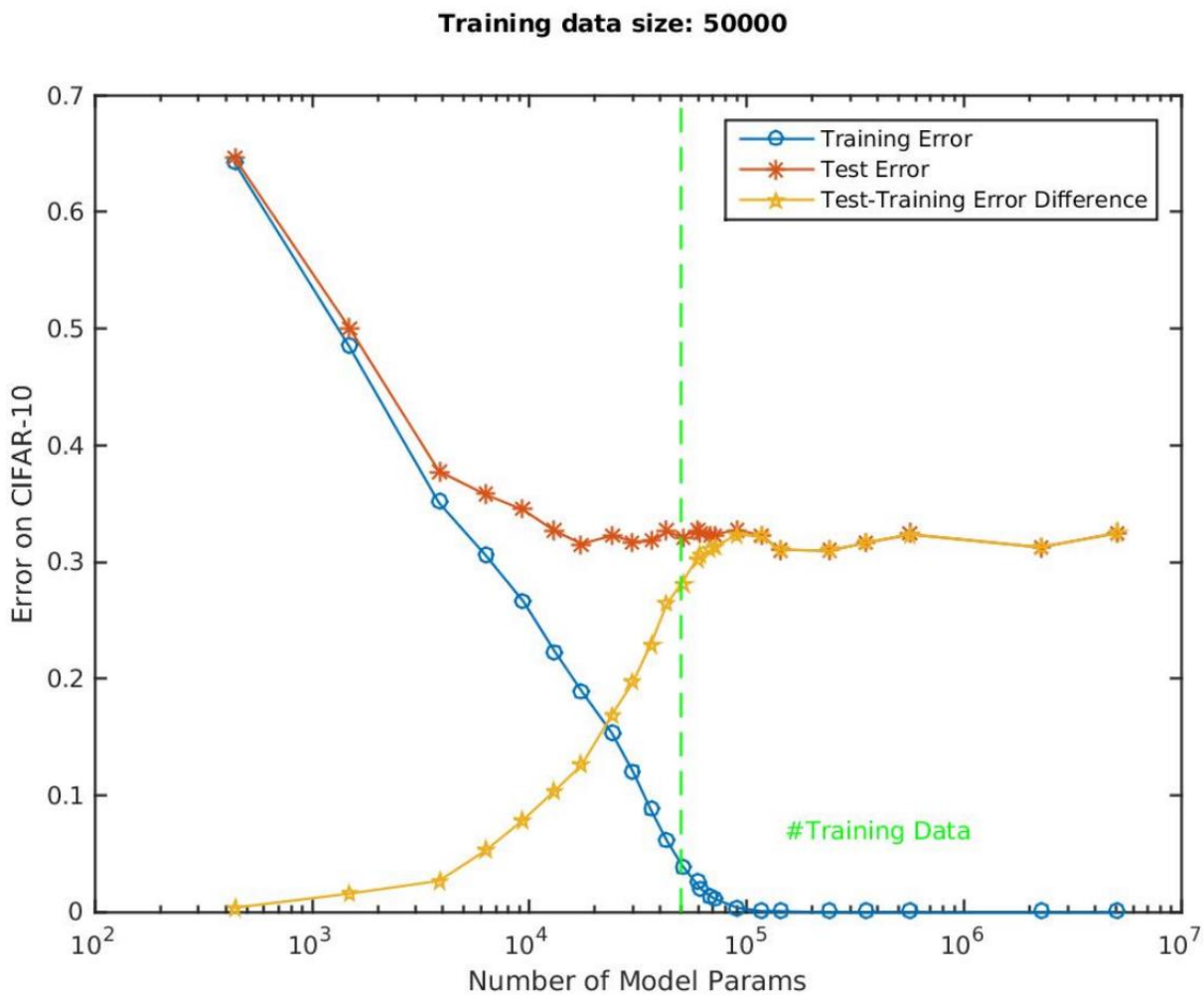
當hidden layer unit增加的時候, 即使達到training error = 0, 竟然還可以降低testing error





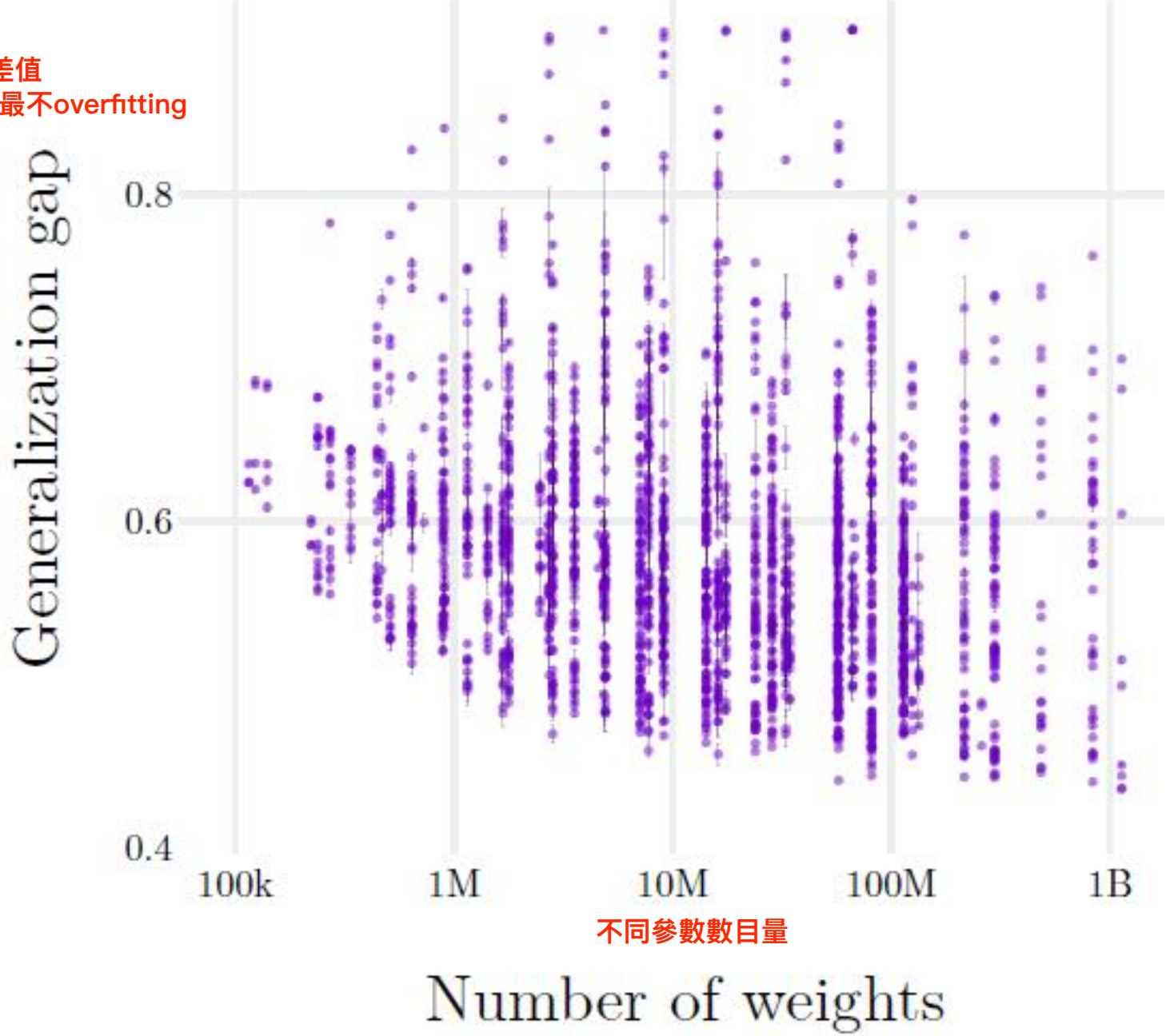
https://arxiv.org/pdf/1412.6614.pdf

Overparameterized Network?



https://arxiv.org/abs/1801.00173

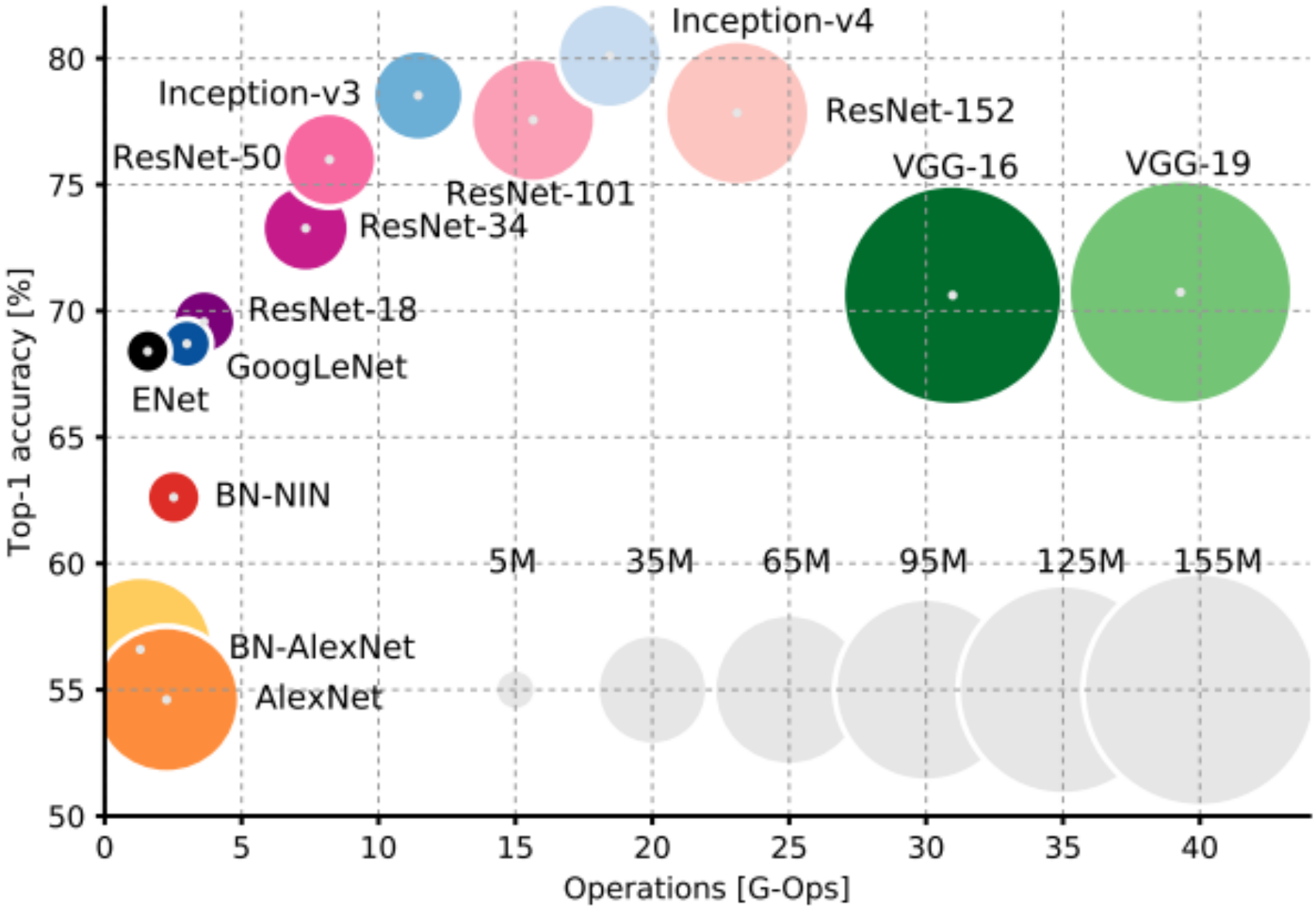
train set/test set差值
最大的network反而最不overfitting



CIFIR-10, 100%
training accuracy

https://arxiv.org/pdf/1802.08760.pdf

隨著參數越來越多 正確率越來越高

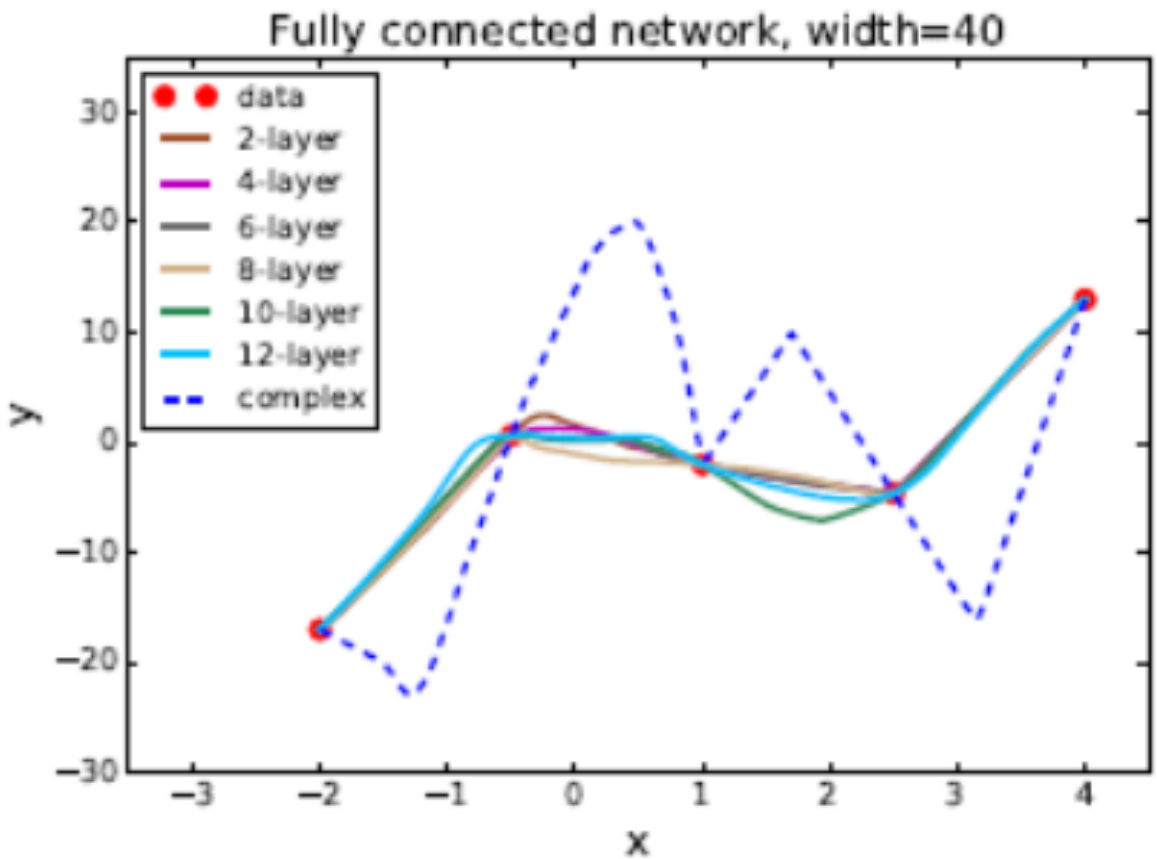


https://arxiv.org/abs/1605.07678

network自帶regularization

Network regularizes itself?

如果說今天用polymorphism function來逼近，則可能會震盪的非常大，但是用neural net反而都很平滑



即使model capacity變大，他仍然regular在平滑的曲線上

https://arxiv.org/pdf/1706.10239.pdf

Concluding Remarks

- The capacity of deep model is large.
- However, it does not overfit!
- The reason is not clear yet.

如果用gradient based來train model可能會自帶regularization，因為一開始的initial參數都很小（接近原點），而regularization就是希望參數能夠接近原點