

Lecture 3 : Image as input

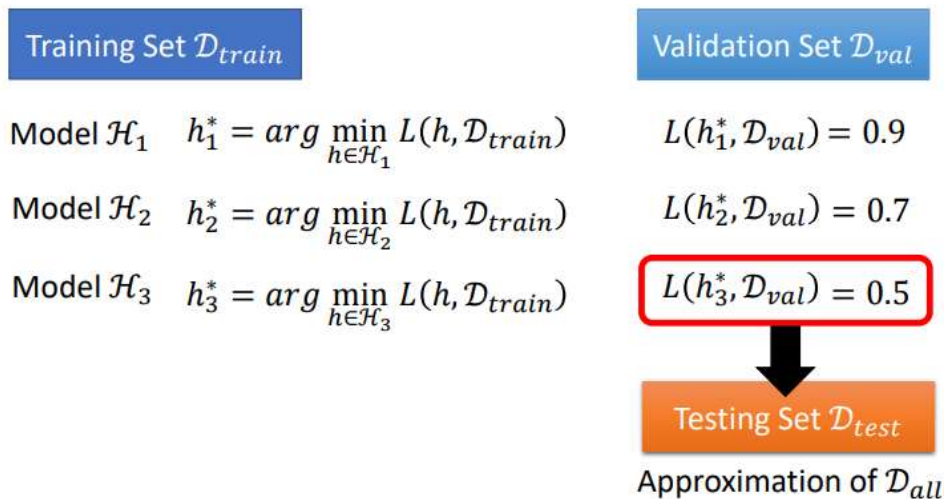
Create at 2022/06/01

- Lecture 3 : Image as input
 - 為甚麼用了 validation set 還是 overfitting 呢 ?
 - 魚與熊掌可以兼得的深度學習，深度學習到底好在哪裡 ?
 - Why we need deep ?
- 上課資源 :
 1. 為什麼用了驗證集 (validation set) 結果卻還是過擬合(overfitting)了呢 ?
(<https://www.youtube.com/watch?v=xQXh3fSvD1A>).
 2. 魚與熊掌可以兼得的深度學習 (<https://www.youtube.com/watch?v=yXd2D5J0QDU>).

為甚麼用了 validation set 還是 overfitting 呢 ?

- 延伸資料 : 【機器學習2021】機器學習任務攻略 (<https://www.youtube.com/watch?v=WeHM2xpYQpw>).

Validation Set



- 如何選擇要使用哪一個模型 ?
 - 不會用 training data 直接去決定 h_1^*, h_2^*, h_3^* 哪一個比較好
 - 會在 validation set 上去評估 h_1^*, h_2^*, h_3^* 各別的 *Loss*
 - 去看哪一個 function 在 validation set 上得到的 *Loss* 最低，就會選擇那一個 function 去用在 testing set 上

Training Set \mathcal{D}_{train}

$$\text{Model } \mathcal{H}_1 \quad h_1^* = \arg \min_{h \in \mathcal{H}_1} L(h, \mathcal{D}_{train})$$

$$\text{Model } \mathcal{H}_2 \quad h_2^* = \arg \min_{h \in \mathcal{H}_2} L(h, \mathcal{D}_{train})$$

$$\text{Model } \mathcal{H}_3 \quad h_3^* = \arg \min_{h \in \mathcal{H}_3} L(h, \mathcal{D}_{train})$$

Validation Set \mathcal{D}_{val}

$$L(h_1^*, \mathcal{D}_{val}) = 0.9$$

$$L(h_2^*, \mathcal{D}_{val}) = 0.7$$

$$L(h_3^*, \mathcal{D}_{val}) = 0.5$$

$$\mathcal{H}_{val} = \{h_1^*, h_2^*, h_3^*\} \quad h^* = \arg \min_{h \in \mathcal{H}_{val}} L(h, \mathcal{D}_{val})$$

Using validation set to select model =

considered as “training” by \mathcal{D}_{val}

Your model is $\mathcal{H}_{val} = \{h_1^*, h_2^*, h_3^*\}$

- 可以看成有一個 model 是 \mathcal{H}_{val} · 這個 model 裡面只有 3 個可能的 function h_1^*, h_2^*, h_3^*
- 只是可以選擇的 function 非常的少

Using validation set to select model =

considered as “training” by \mathcal{D}_{val}

Your model is $\mathcal{H}_{val} = \{h_1^*, h_2^*, h_3^*\}$

$$L(h^{train}, \mathcal{D}_{all}) - L(h^{all}, \mathcal{D}_{all}) \leq \delta$$

$$P(\mathcal{D}_{train} \text{ is bad}) \leq |\mathcal{H}| \cdot 2\exp(-2N\varepsilon^2)$$

$$L(h^{val}, \mathcal{D}_{all}) - L(h^{all}, \mathcal{D}_{all}) \leq \delta$$

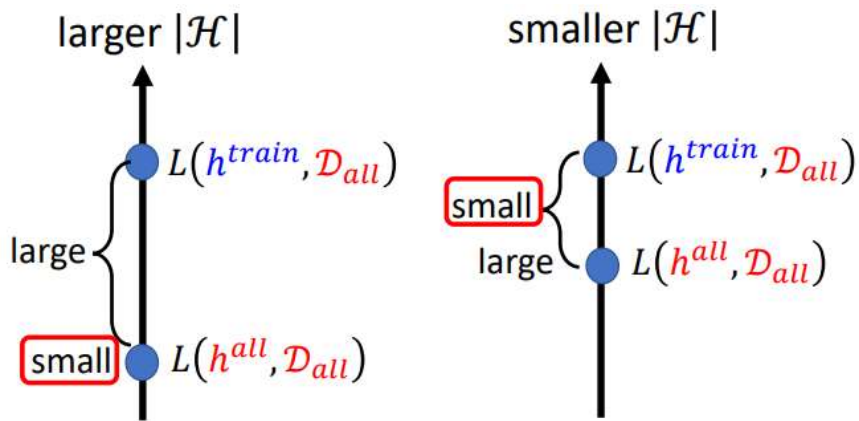
$$P(\mathcal{D}_{val} \text{ is bad}) \leq |\mathcal{H}_{val}| \cdot 2\exp(-2N_{val}\varepsilon^2)$$

↑
It is small.

Hopefully ☺

- 為甚麼有了 validation set 還是有可能會 overfitting
 - 如果 $|\mathcal{H}|$ 仍然很大 · 有可能還是有很高的 $P(\mathcal{D}_{val} \text{ is bad})$

魚與熊掌可以兼得的深度學習，深度學習到底好在哪裡？



魚與熊掌可以兼得嗎？

$$h^{all} = \arg \min_{h \in \mathcal{H}} L(h, \mathcal{D}_{all})$$

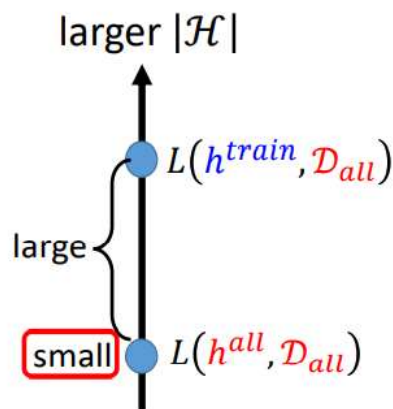
Still small loss

Small (fewer candidates)

- 有沒有一個 $Loss$ 很低的理想，同時現實跟理想又很接近
 - 找到一個 H 很少，同時這個 H 都是精英讓 $Loss$ 很低

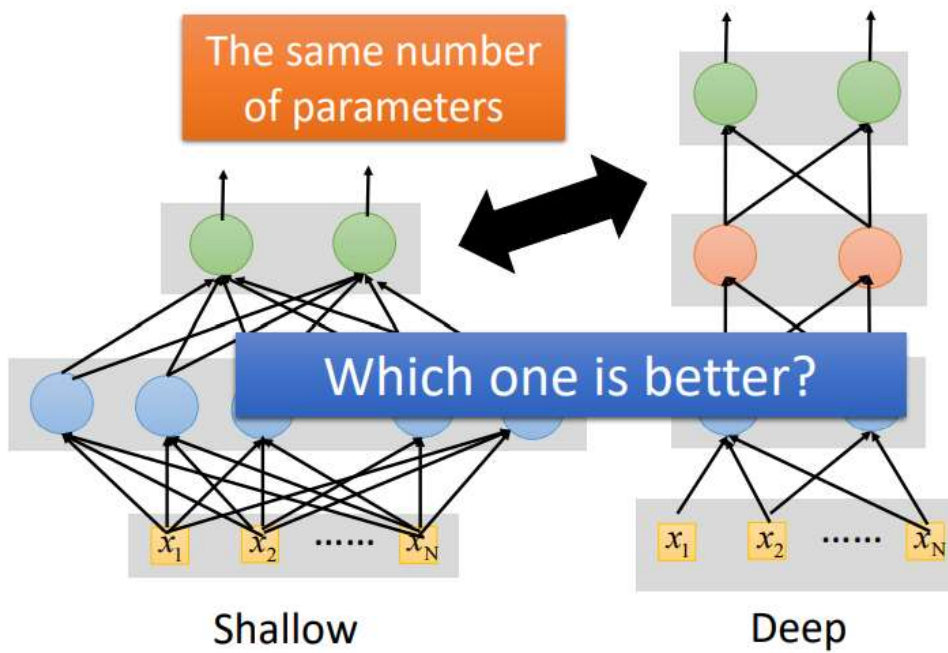
Deeper is Better?

Layer X Size	Word Error Rate (%)
1 X 2k	24.2
2 X 2k	20.4
3 X 2k	18.4
4 X 2k	17.8
5 X 2k	17.2
7 X 2k	17.1



- Network 越深，參數量變多，可以讓理想越來越美好
- 如果資料量夠多，理想跟現實的差距越來越少
- 深度學習需要一個大模型，大模型伴隨著需要大量的資料，如果沒有大量資料就會 overfitting
- 沒有大量資料就不適合用深度學習

Fat + Short v.s. Thin + Tall



- 有一樣的參數量時，哪一個會比較好？

Fat + Short v.s. Thin + Tall

Layer X Size	Word Error Rate (%)	Layer X Size	Word Error Rate (%)
1 X 2k	24.2		
2 X 2k	20.4		
3 X 2k	18.4		
4 X 2k	17.8		
5 X 2k	17.2	1 X 3772	22.5
7 X 2k	17.1	1 X 4634	22.6
		1 X 16k	22.1

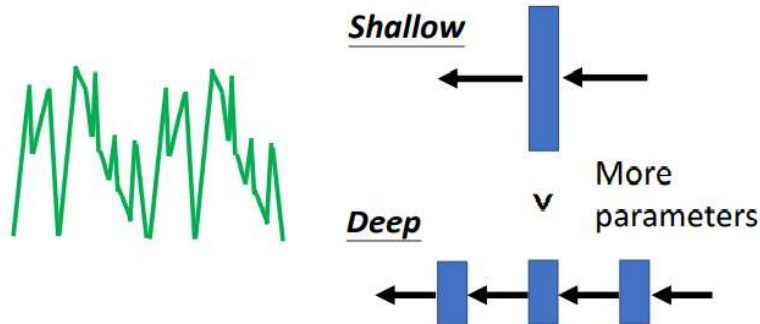
- 實驗結果，Layer 加深會比肥胖模型的效果好
- 與其把 network 變胖不如把 network 變高

Why we need deep ?

Why we need deep?

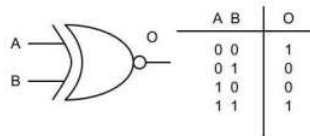
Yes, one hidden layer can represent any function.

However, using deep structure is more effective.

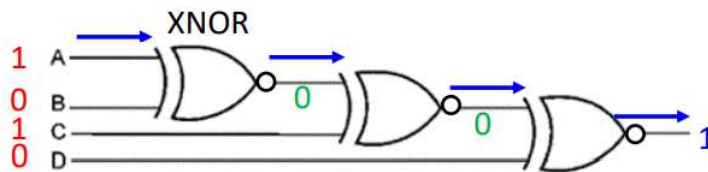
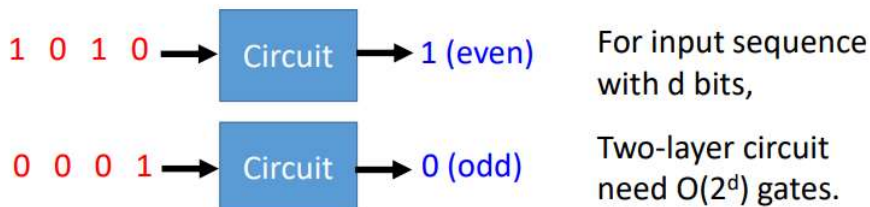


雖然 one hidden layer network 可以表示任何 function，但是用一個 deep 的架構往往是會比較有效率的

Analogy – Logic Circuits



• E.g., parity check

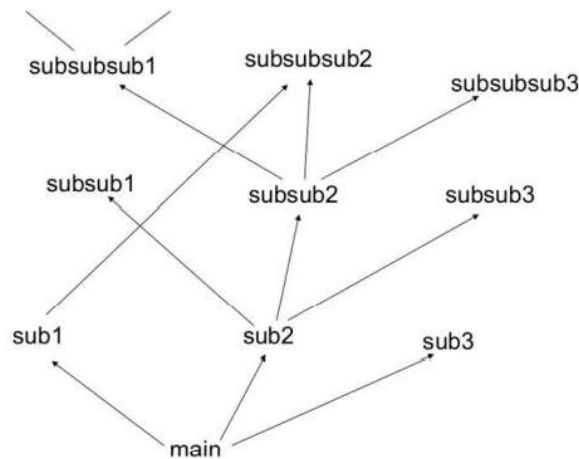


With multiple layers, we need only $O(d)$ gates.

結構深的運算方式比只有一層的運算方式來的有效率

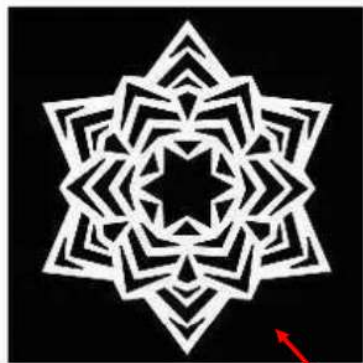
Analogy – Programming

Don't put everything in your main function.



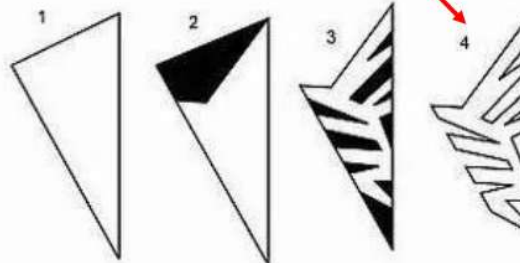
- 寫程式的時候也是，會分成很多個 module，不會全部寫在同一個 function 裡面
- 避免程式太過攏長，也會用到 deep 的結構

More Analogy



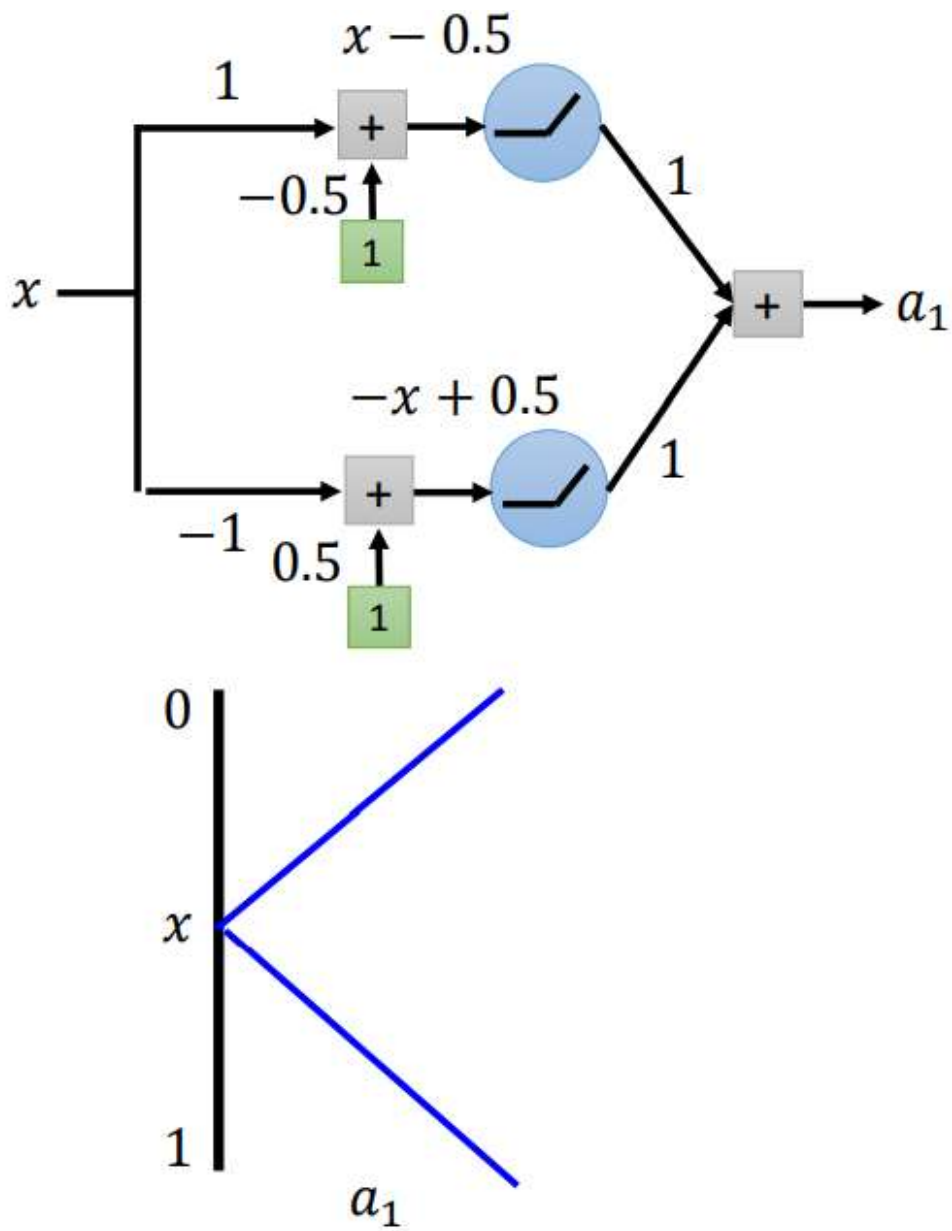
剪很多刀

比較有效率

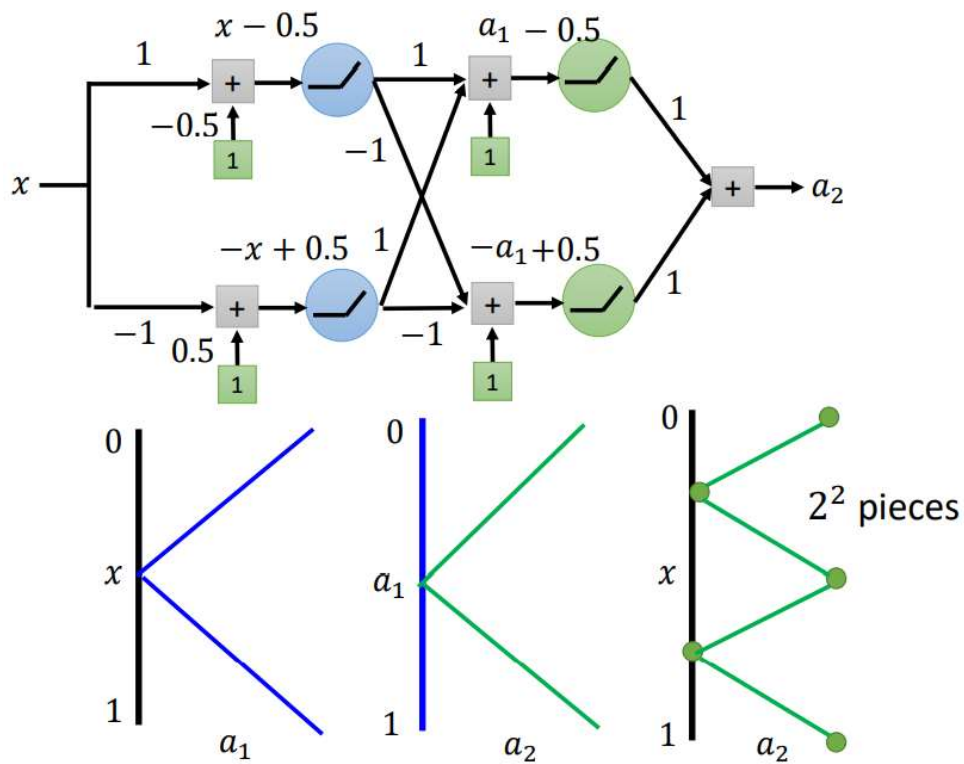


头条号 / 幼师宝典

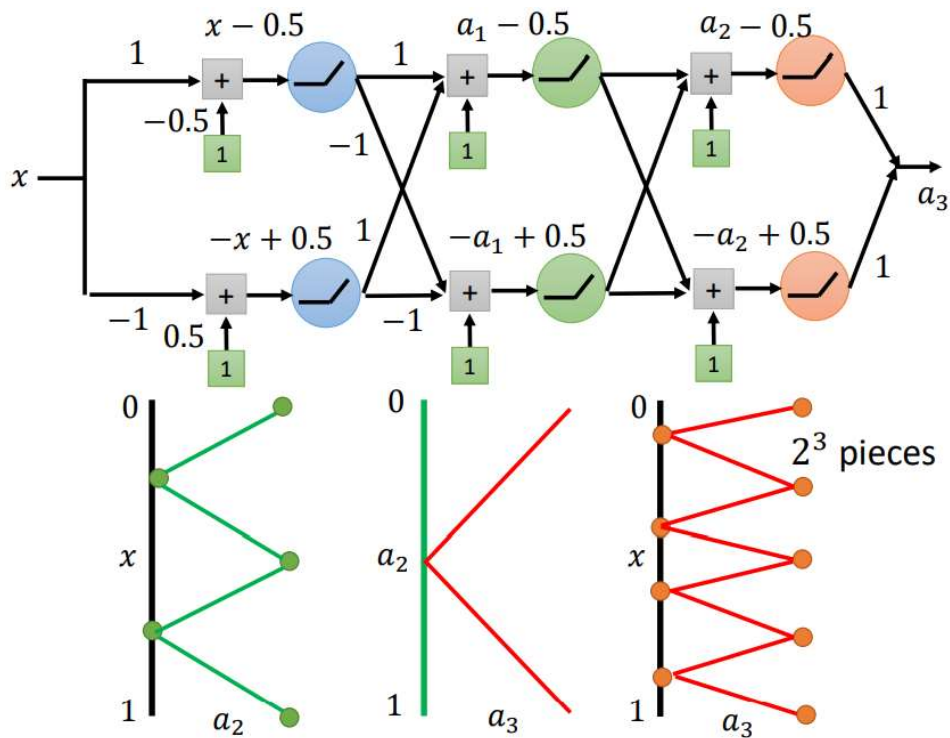
有結構的比較有效率



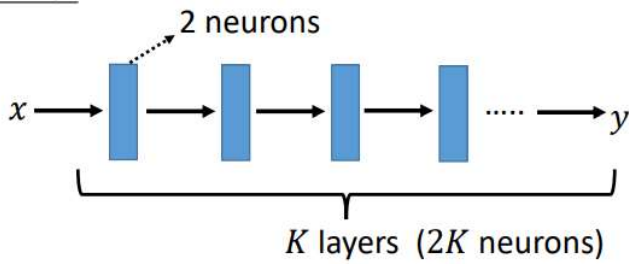
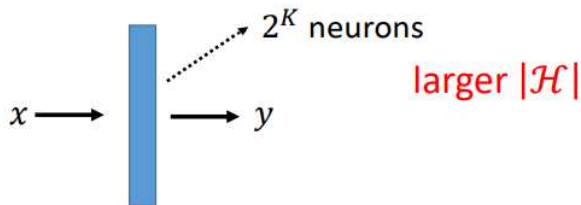
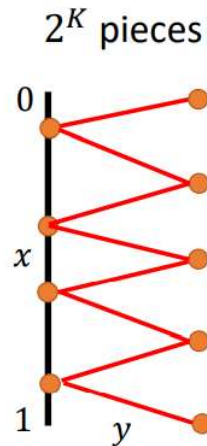
只有一層的 network



兩層 network



三層 network

Deepsmaller $|\mathcal{H}|$ **Shallow**larger $|\mathcal{H}|$ 

- 證明使用 deep 的結構效率較好
- 要產生同一個 function，deep network 使用的參數量較小，有比較簡單的模型
- shallow network 參數量較大，有比較複雜的模型，而複雜的模型比較容易 overfitting，所以會需要更大量的資料

Thinks more

- Deep networks outperforms shallow ones when the required functions are complex and regular.
Image, speech, etc. have this characteristics.
- Deep is exponentially better than shallow even when $y = x^2$.

當所需的功能複雜且規則時，deep network 的效率優於 shallow network

- 延伸資料：
 - Deep Learning Theory 1-2: Potential of Deep (<https://www.youtube.com/watch?v=FN8jdCrgY0>).
 - Deep Learning Theory 1-3: Is Deep better than Shallow? (<https://www.youtube.com/watch?v=qpuLxXrHQB4>).

課程網頁 (<https://speech.ee.ntu.edu.tw/~hylee/ml/2022-spring.php>).

tags: 2022 李宏毅_機器學習

