The story of the dark knowledge
-- from model distillation to contrastive learning

Y.F. Liou

@NTU 2020/10/13

今天講題的兩大保證

- 保證講者沒有強大的數學背景
  - 請相信生物學家...
- 保證內容不會是全部正確的
  - 歡迎討論

Before we start

- 我們做個約定
  - 我做“民調”的時候，請各位盡量舉手
  - 希望每個到場的，可以盡量學到有趣的東西

- 我保證
  - 除非你想說話，不然我絕對不會拿麥克風讓你說話

About me

- Manager of the AI contexture center of Coretronic Corp.

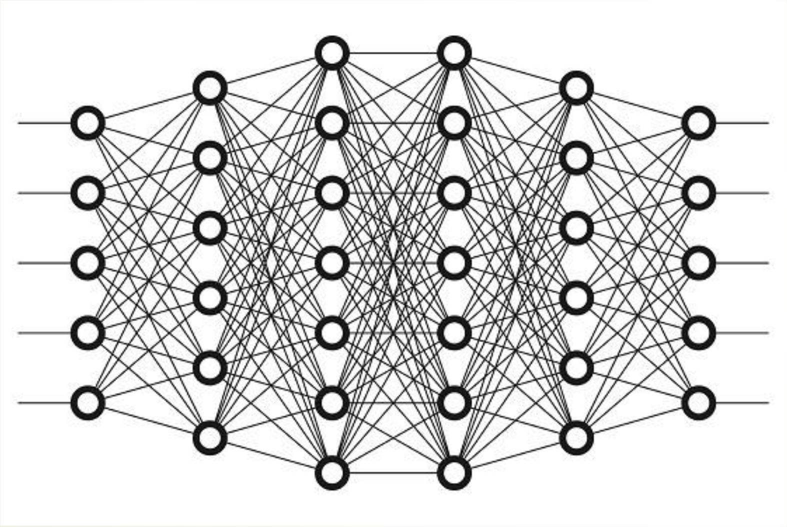
- BS., Life Science, Fu-Jen University
- MS., Biotechnology, National Taipei University of Technology
- PhD., Bioinformatics and Systems Biology, National Chaio Tung University

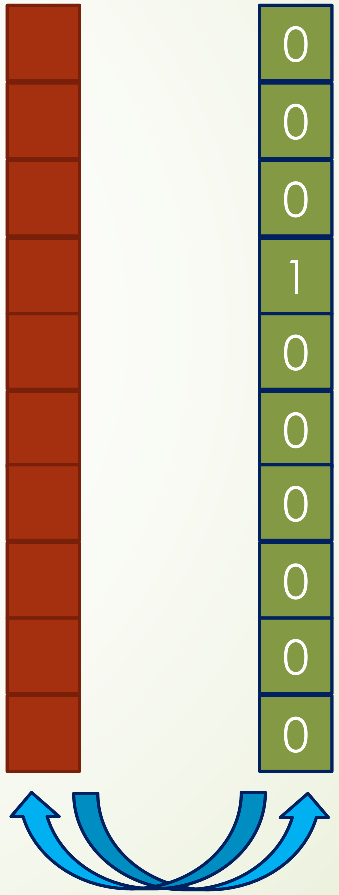
The issues we will talk

- The basic training process of neural network
- The model distillation
  - Why and how will do the model distillation
- The contrastive learning

How to training your neural network

How to train your model – the gradient decent method



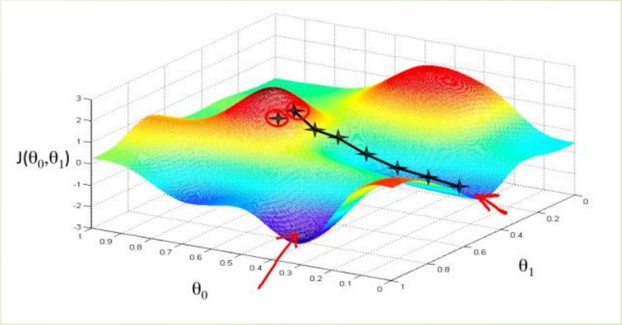


0
0
0
1
0
0
0
0
0
0

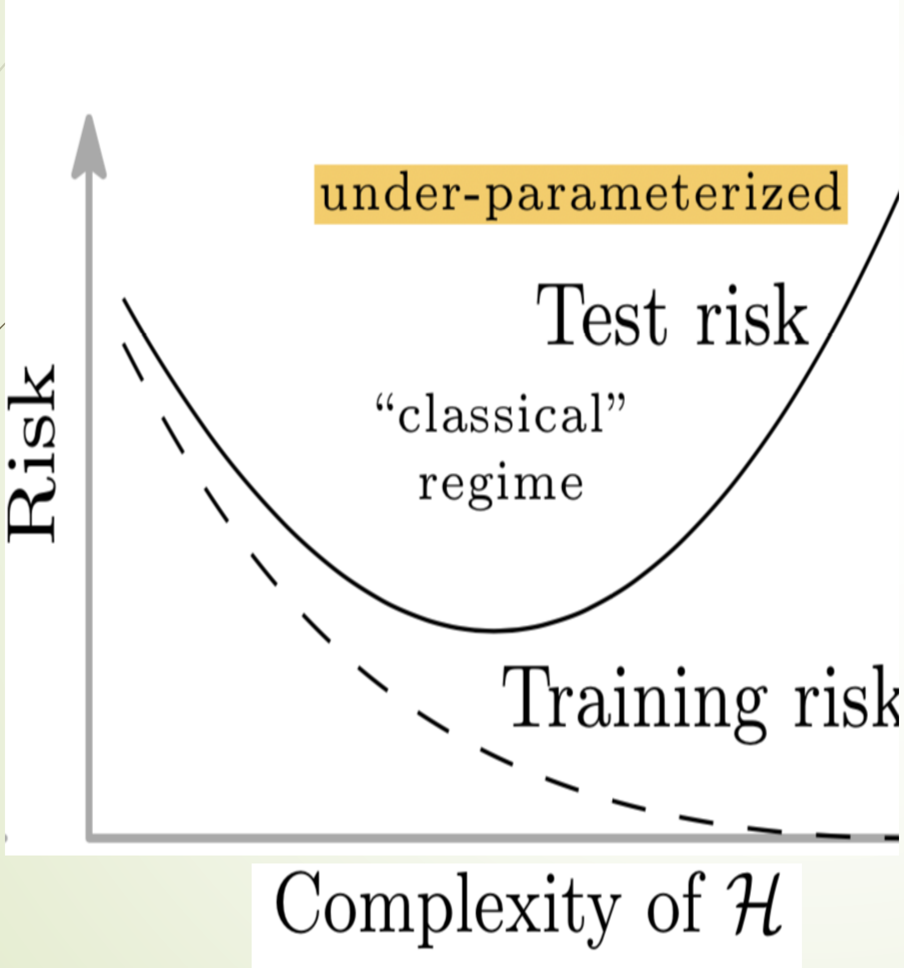
Cross entropy

Shannon's entropy equation:

$$H(X) = -\sum_{i=0}^{N-1} p_i \log_2 p_i$$



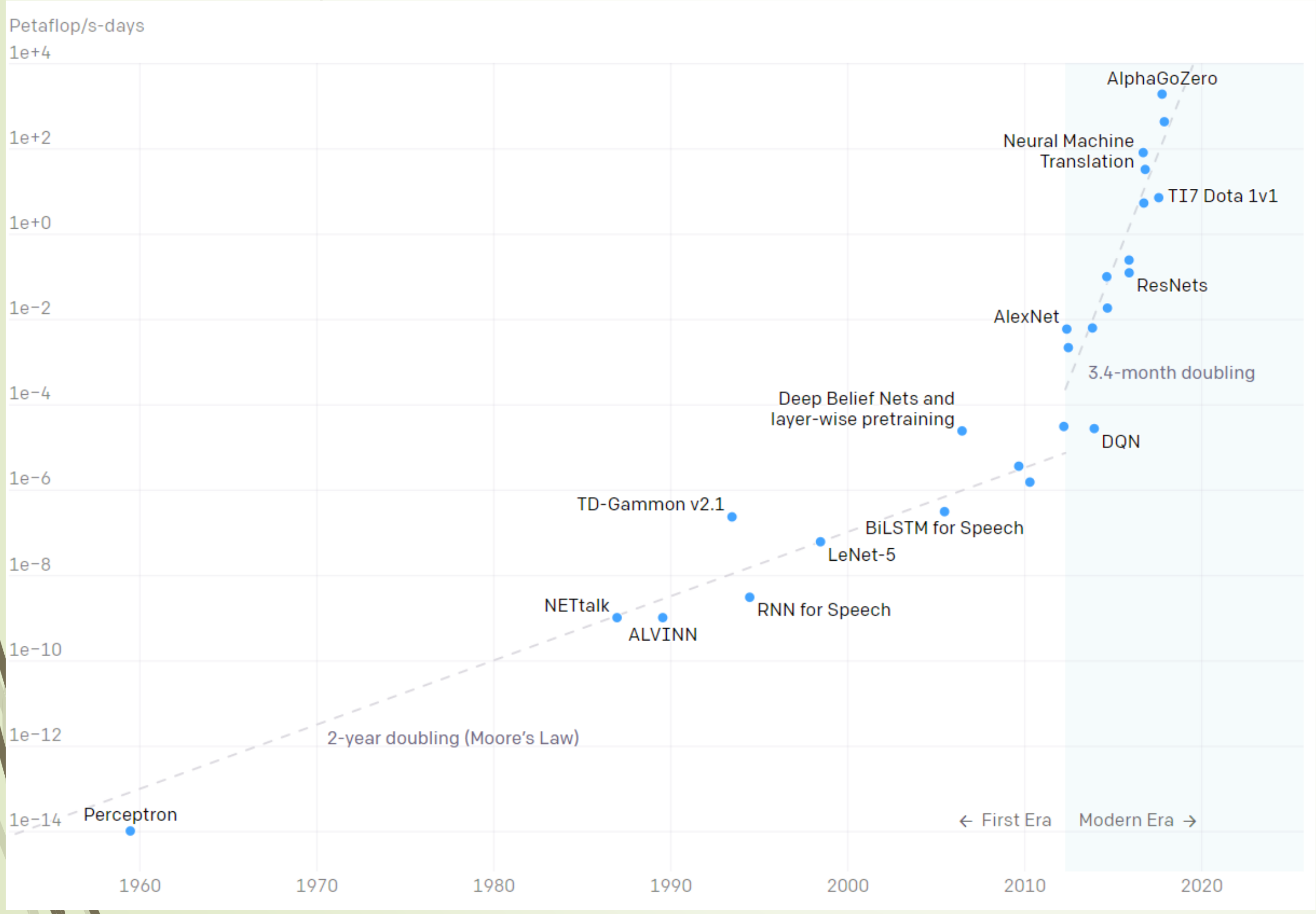
The minimal risk theory of the traditional machine learning



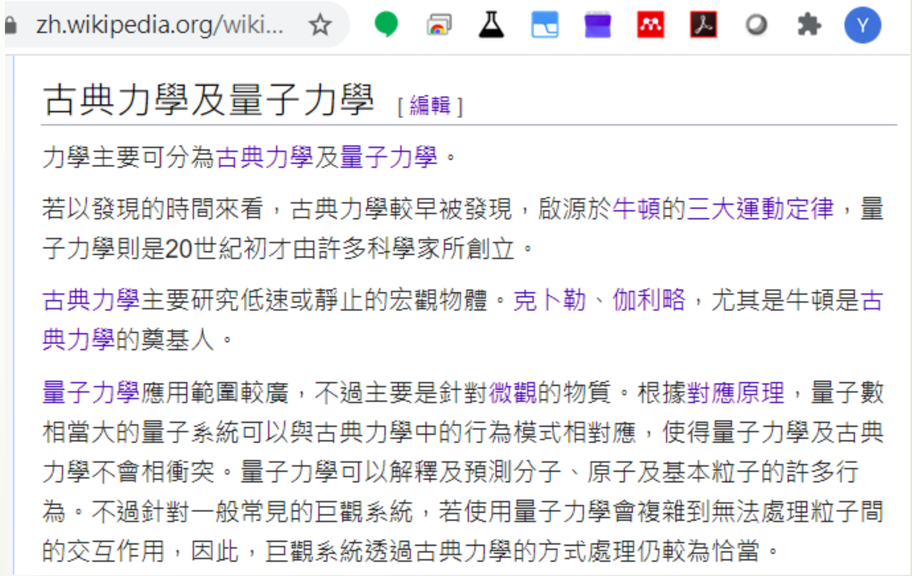
Whtat happen to recent machine learning area?

https://lilianweng.github.io/lil-log/2019/03/14/are-deep-neural-networks-dramatically-overfitted.html

Is the traditional theory is the everything ?



The knowledge revolution is not first time in the human history.



古典力學及量子力學 [編輯]

力學主要可分為古典力學及量子力學。

若以發現的時間來看，古典力學較早被發現，啟源於牛頓的三大運動定律，量子力學則是20世紀初才由許多科學家所創立。

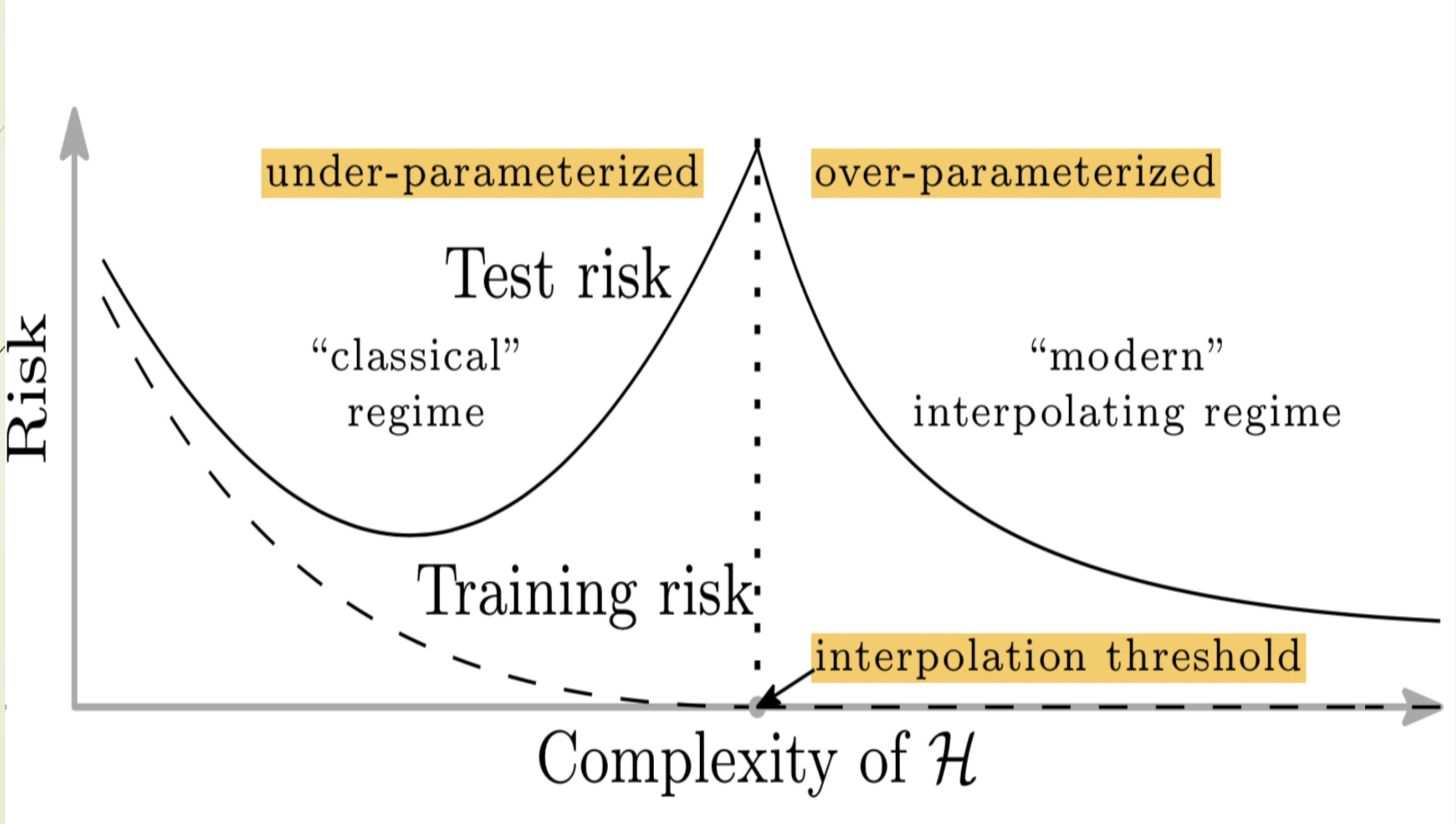
古典力學主要研究低速或靜止的宏觀物體。克卜勒、伽利略，尤其是牛頓是古典力學的奠基人。

量子力學應用範圍較廣，不過主要是針對微觀的物質。根據對應原理，量子數相當大的量子系統可以與古典力學中的行為模式相對應，使得量子力學及古典力學不會相衝突。量子力學可以解釋及預測分子、原子及基本粒子的許多行為。不過針對一般常見的巨觀系統，若使用量子力學會複雜到無法處理粒子間的交互作用，因此，巨觀系統透過古典力學的方式處理仍較為恰當。

Demo time ~~

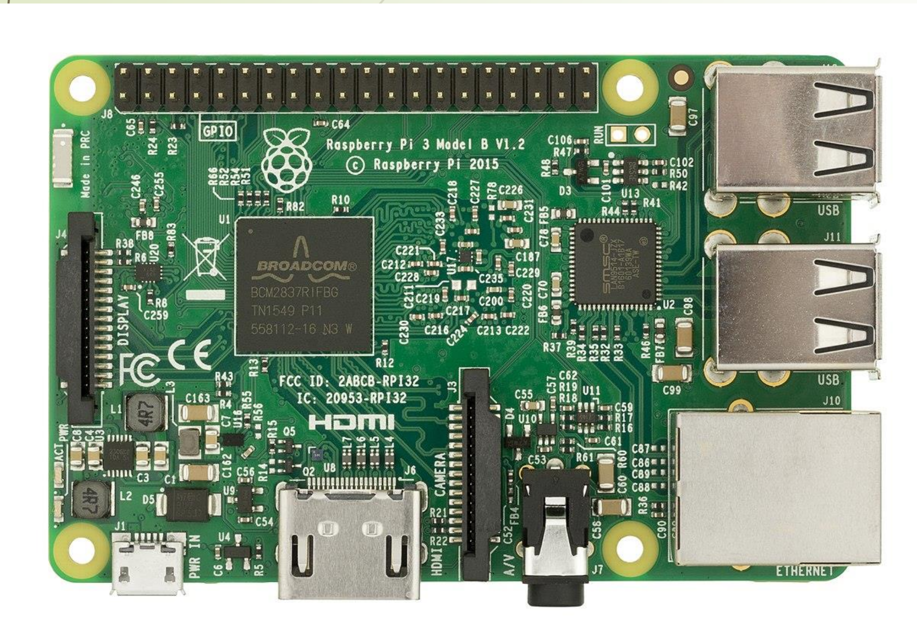
https://openai.com/blog/ai-and-compute/

What is happening to overparameterization?



https://lilianweng.github.io/lil-log/2019/03/14/are-deep-neural-networks-dramatically-overfitted.html

The issues of over-parameter model would cause



The huge model need huge computing power. But several applications could not provide such computing resource.

The ways to save the computing resources of neural network

- Physical ways
  - Neural network compiler
    - Operations on the chip
    - Pruning (maybe this can also be thought as a theoretical ways as mentioned below)
  - Quantization
    - The precision of the weights
- Theoretical ways
  - Distillation

- Since the distillation is a “theoretical way”, there are many interesting results can be found in such researches.

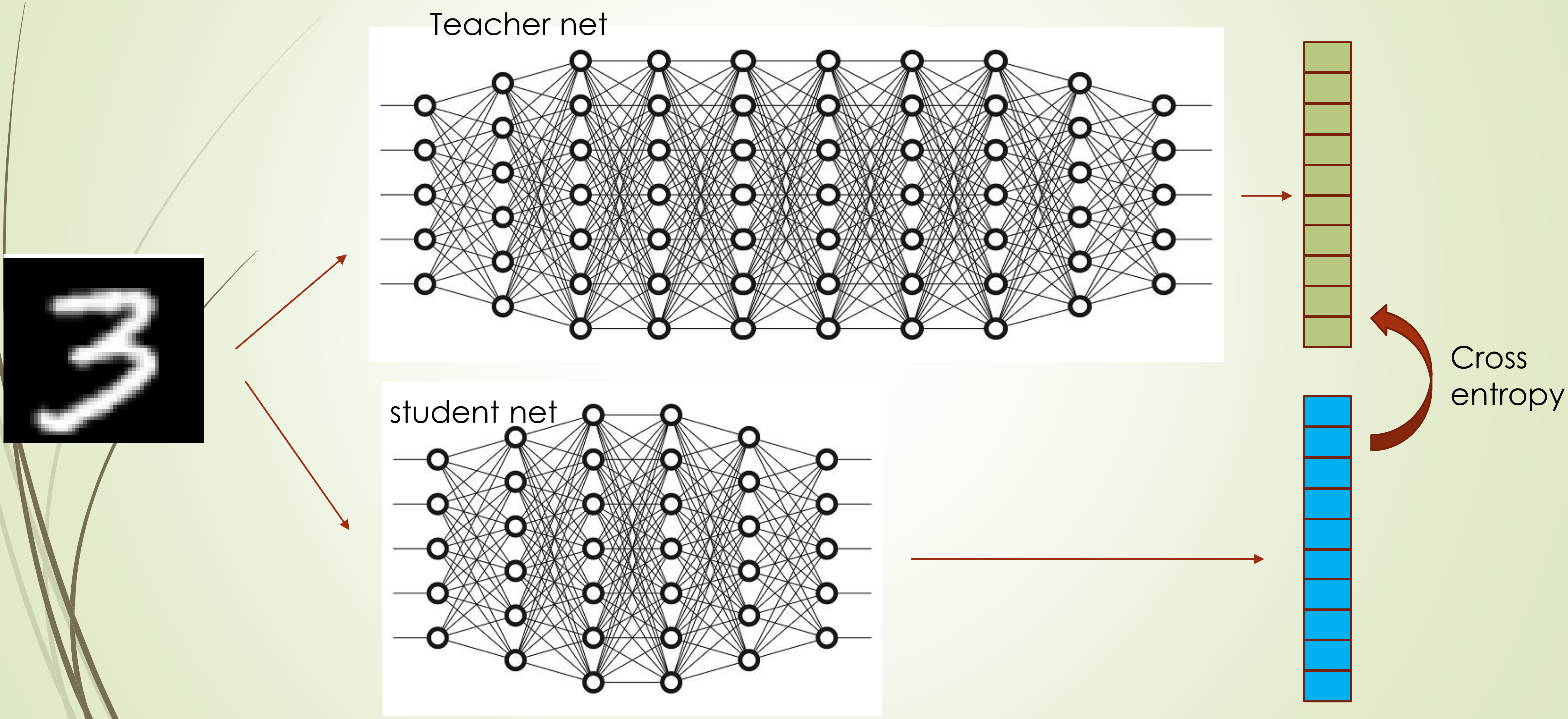
The model distillation -- Enhancing the dark knowledge

This would be the beginning of the contrastive learning

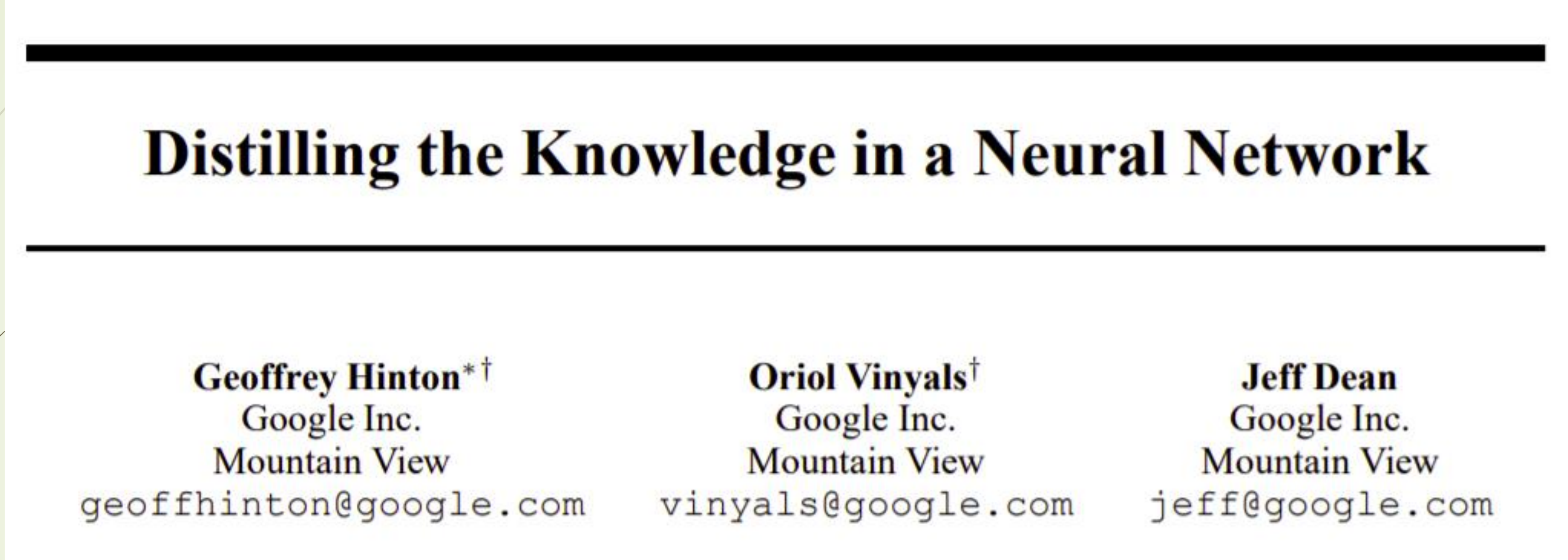
Usually,

The main logit will dominate most of output signals. If we want to enhancing the dark knowledge, softening the output would be a way …

The basic form of “model distillation”



When will the biologist be useful ?

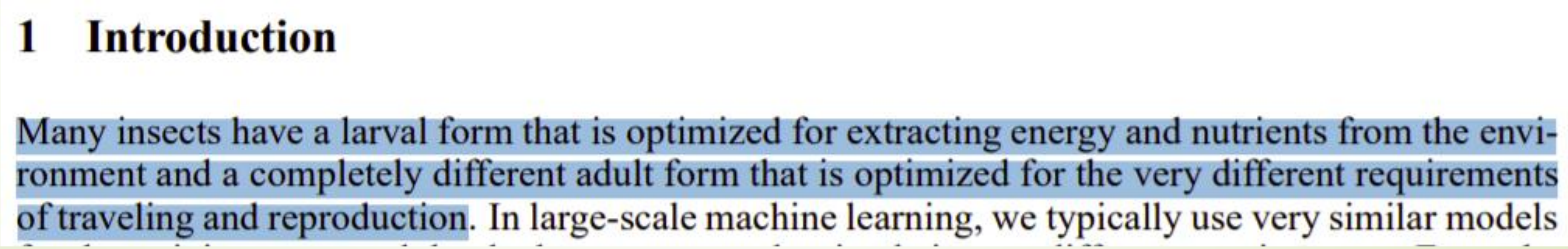


Distilling the Knowledge in a Neural Network

Geoffrey Hinton*†
Google Inc.
Mountain View
geoffhinton@google.com

Oriol Vinyals†
Google Inc.
Mountain View
vinyals@google.com

Jeff Dean
Google Inc.
Mountain View
jeff@google.com



1 Introduction

Many insects have a larval form that is optimized for extracting energy and nutrients from the environment and a completely different adult form that is optimized for the very different requirements of traveling and reproduction. In large-scale machine learning, we typically use very similar models

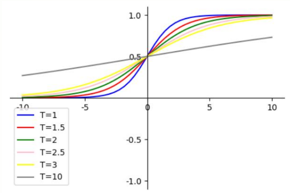
結論: 生物學家可能只有在introduction開頭有用，後面用處就沒有這麼明顯了。

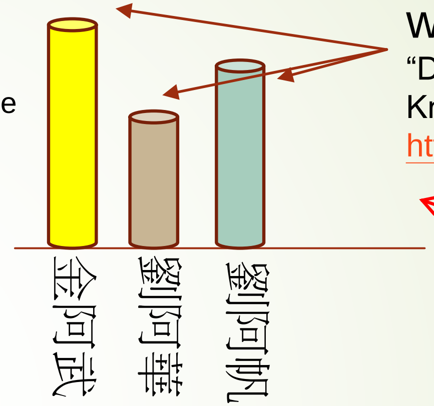
Distilling knowledge in a neural network

Teacher net

student net

Results can use:
1.Softmax with temperature
2.Logits (better)





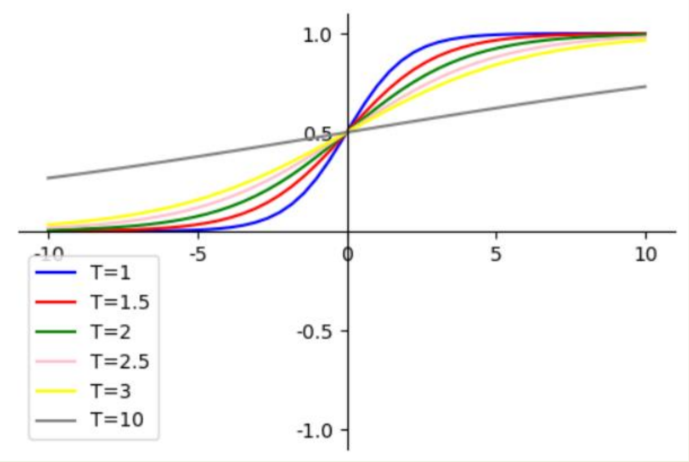
Perturbation the logits would be better.
“Deep Model Compression: Distilling Knowledge from Noisy Teachers”
https://arxiv.org/pdf/1610.09650.pdf

loss

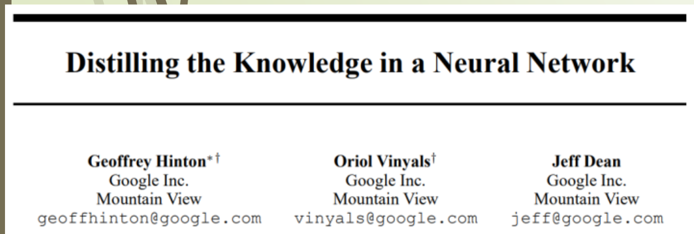
results

Ground truth

loss



Teacher nets are thought to lean much “empirical” information whic would contain in the categories no activated.



Distilling the Knowledge in a Neural Network

Geoffrey Hinton[∗†]
Google Inc.
Mountain View
geoffhinton@google.com

Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

Jeff Dean
Google Inc.
Mountain View
jeff@google.com

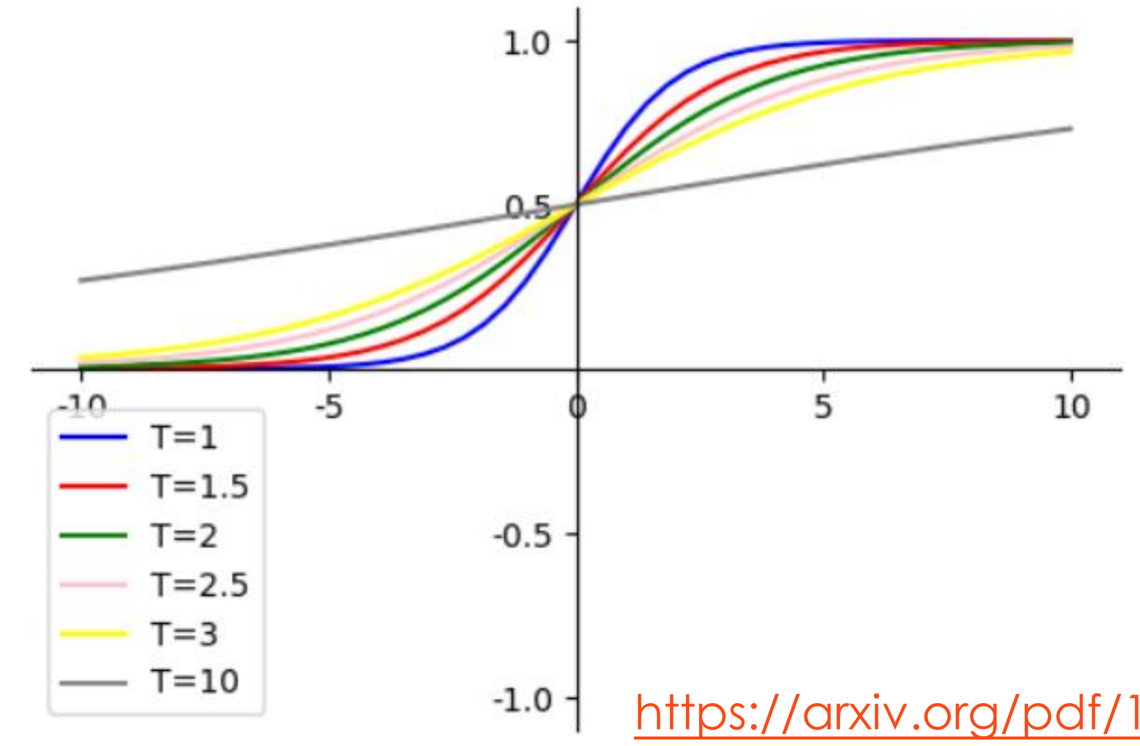
Temperature of activation function

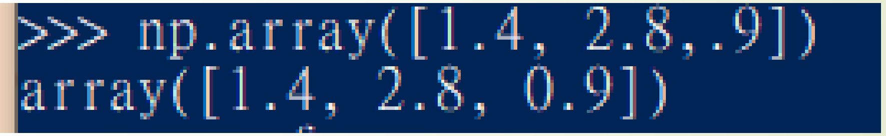
$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

z = logits
T = Temperature
j = index of class
q = new logits

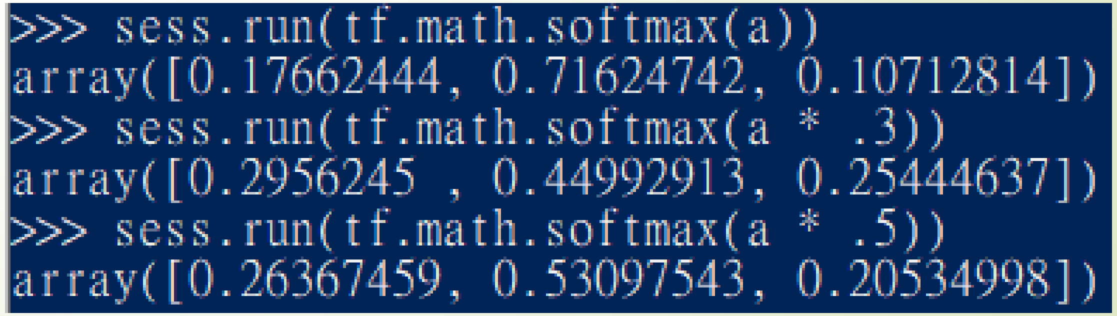
目標是把teacher 的label做soft。
可能的原因是teacher所做出來的答案不一定正確，而且內部有許多Dark knowledge。因此做soft後，希望student可以學到teacher所給予的所有資訊。

Here, "Temperature" enlarge the signal of dark knowledge.





```
>>> np.array([1.4, 2.8,.9])
array([1.4, 2.8, 0.9])
```

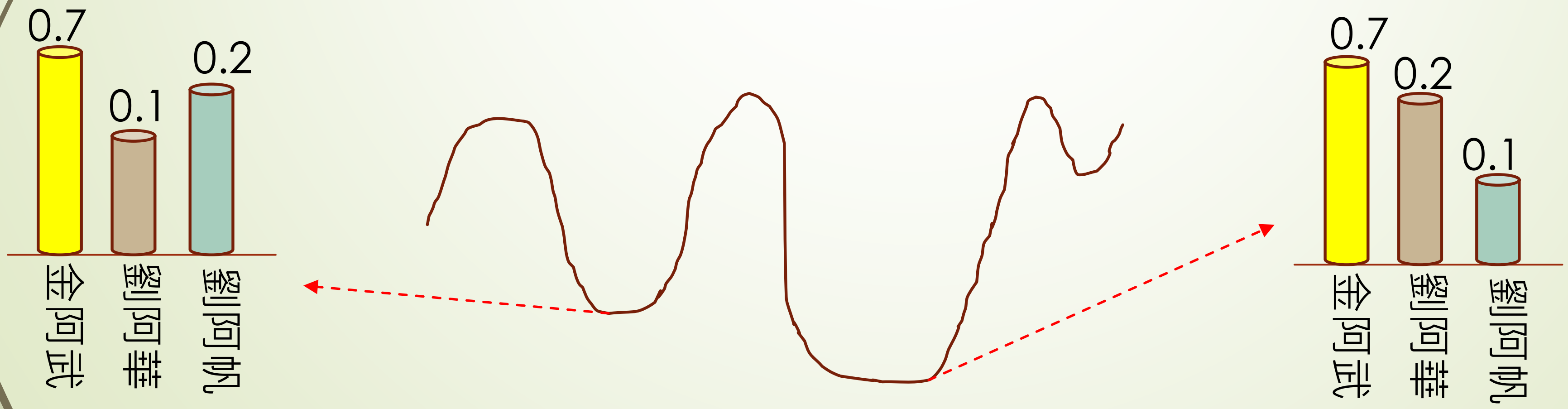


```
>>> sess.run(tf.math.softmax(a))
array([0.17662444, 0.71624742, 0.10712814])
>>> sess.run(tf.math.softmax(a * .3))
array([0.2956245 , 0.44992913, 0.25444637])
>>> sess.run(tf.math.softmax(a * .5))
array([0.26367459, 0.53097543, 0.20534998])
```

https://arxiv.org/pdf/1503.02531.pdf

The Dark knowledge

- 這些非label的部分，在knowledge distillation的課題中稱呼"Dark knowledge"
- 影響Dark knowledge的表現，有可能也能影響整個模型
  - What if we control the dark knowledge directly?



這兩個結果對於最後使用SGD更新，回傳的loss是相同的。但是在dark knowledge部分表現非常不同。

The “labeling way” would be a key

Label smoothing

$$p_k = \frac{exp(x^T w_k)}{\sum_{l=1}^{L} exp(x^T w_l))}$$

Predictions as a function of activations in penultimate layer

p_k: Likelihood the model assigns to the *k-th* class

w_k: Weights and biases of the last layer

x: Vector containing the activations of the penultimate layer

$$y_k^{LS} = y_k(1-\alpha) + \frac{\alpha}{K}$$

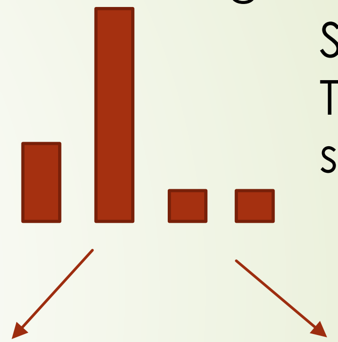
Applying label smoothing to hard targets

0	1	0	0	original

α= .8

.05	.85	.05	.05	smoothing

The purpose of the label smoothing is design to overcome the “overconfidence” which would also cause the overfitting.



Support :
The model give a softmax result.

0	1	0	0

The traditional method only considers the activated class

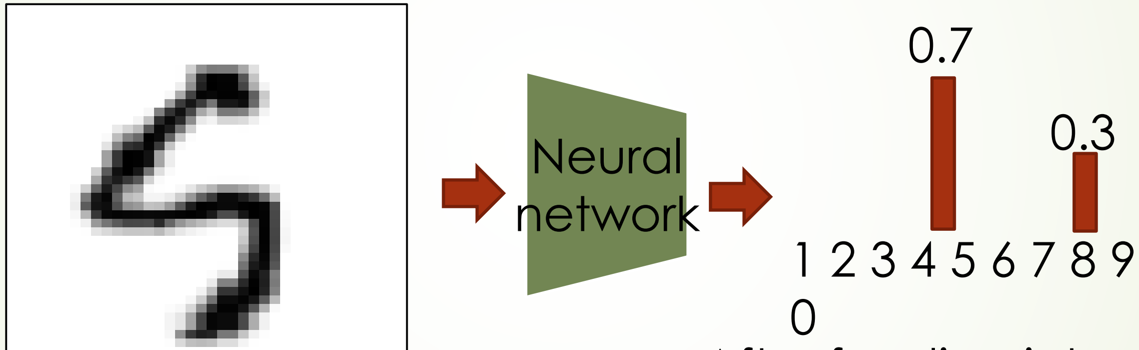
.05	.85	.05	.05

After label smoothing, all the logits are considered

https://medium.com/@nainaakash012/when-does-label-smoothing-help-89654ec75326

Considering the wrong or poor labeling

If there is a number like this



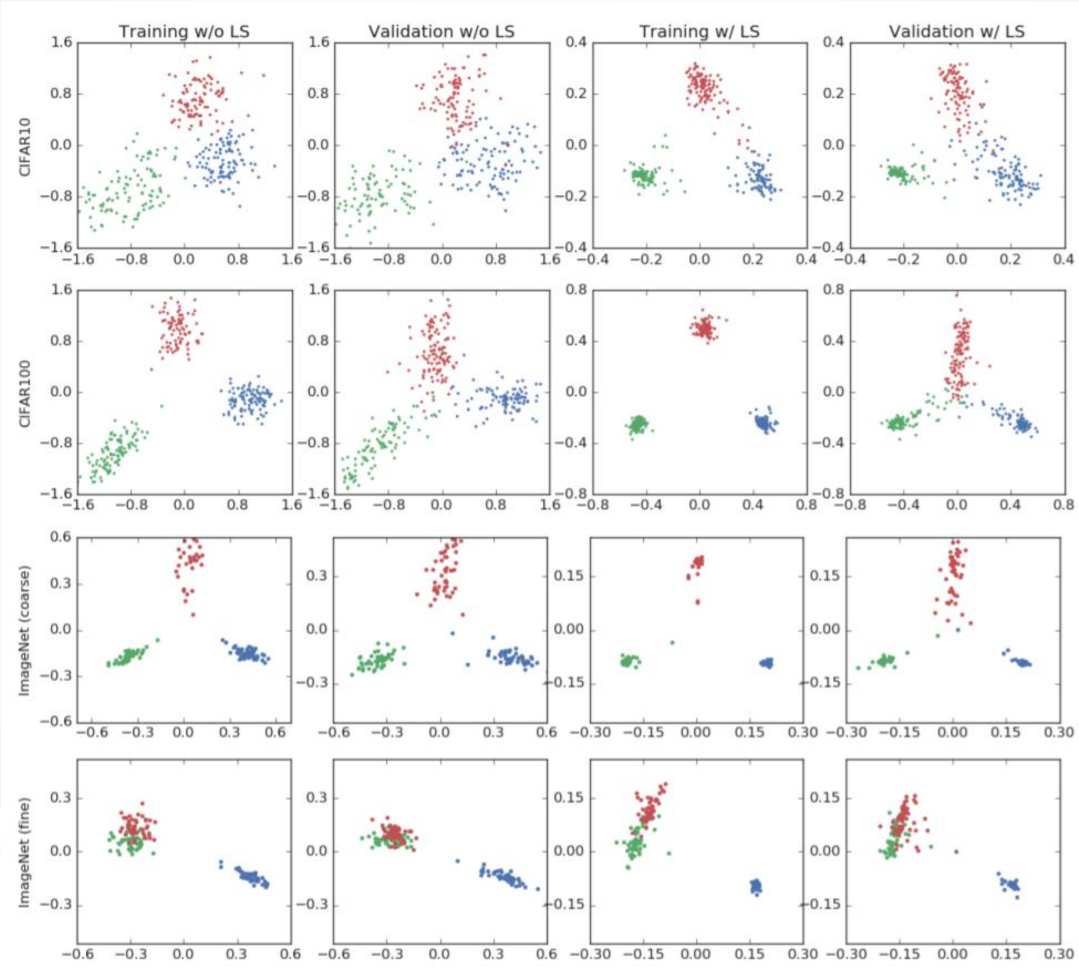
This number is 5, and it also look like 9.

After feeding into a neural network, The answer of 5 and 9 will both give a penalty. 但實際上選定這兩個答案也有道理。

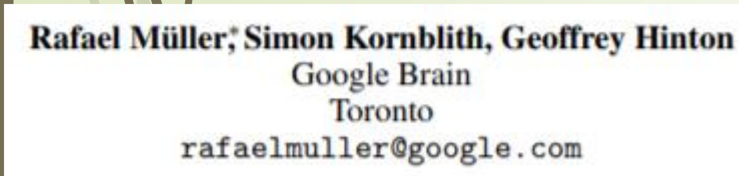
Label smoothing helps to find the dense probability

If we use the label smoothing, the penultimate layer representations give less spread results.

smoothing



The "spreading" of the teacher is also containing some information. Using label smoothing (which improve the training in many tasks) in distillation work would hurt the training process.

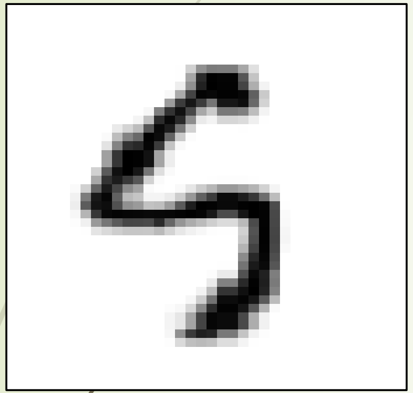


Rafael Müller*, Simon Kornblith, Geoffrey Hinton
Google Brain
Toronto
rafaelmuller@google.com

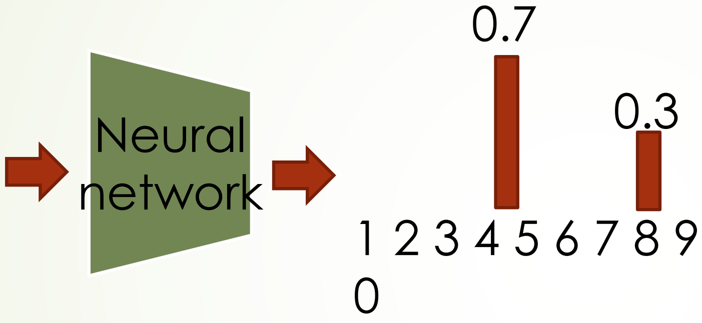
https://arxiv.org/pdf/1906.02629.pdf

Considering the minor output

If there is a number like this



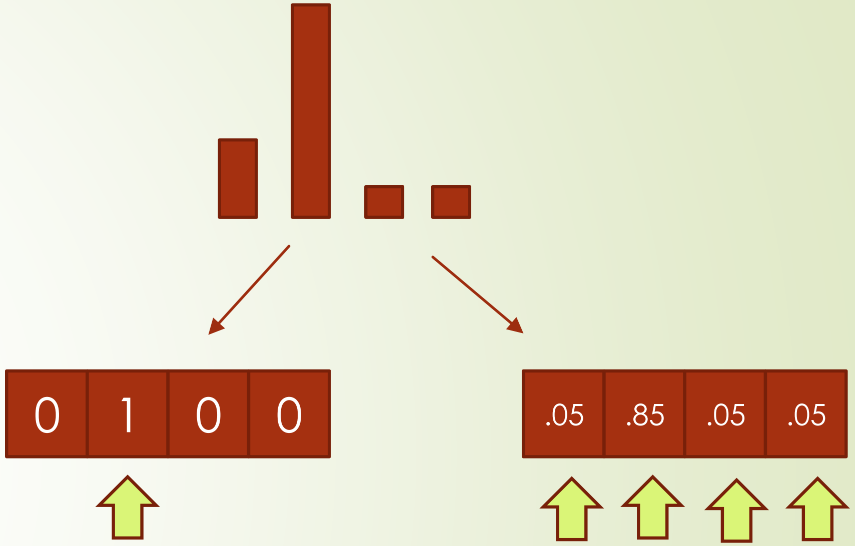
This number is 5, and it also look like 9.



After feeding into a neural network, The answer of 5 and 9 will both give a penalty.
但實際上選定這兩個答案也有道理。

從模型預測的其他label輸出似乎也代表特殊意義。

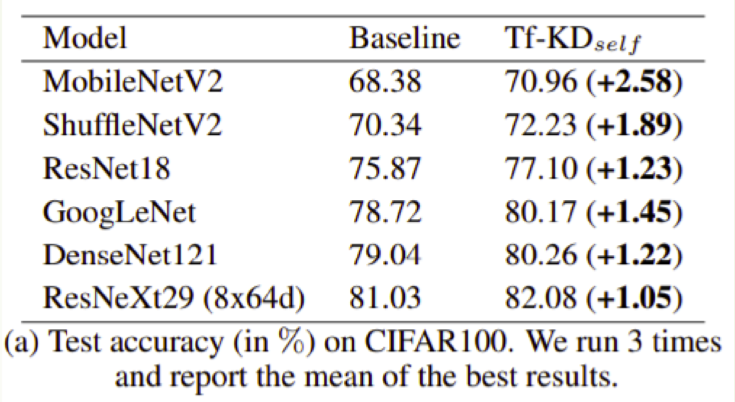
- 是否表示"在這些次要的輸出中，模型告訴我們看到了甚麼"。
- 也許某些資訊也可以透過控制這些次要的輸出，反饋給模型。



Considering the minor outputs of logits is quite similar to consider the “dark knowledge” of model distillation.

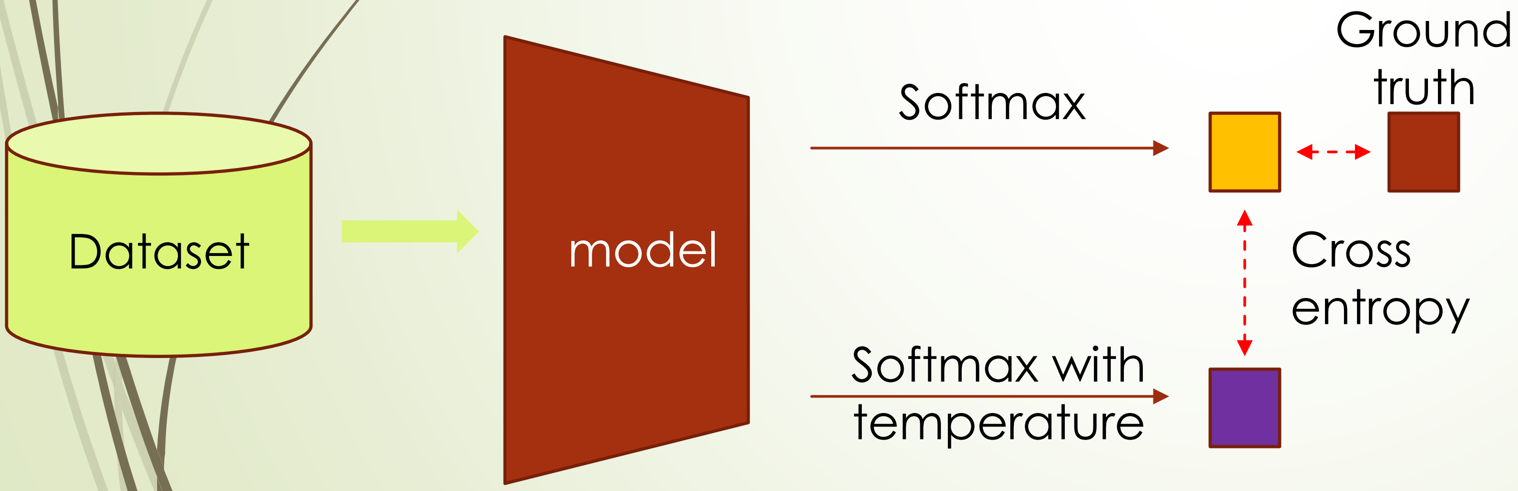
Teacher free knowledge distillation

- Label smoothing 跟knowledge distillation 都專注在minor logits
  - Label smoothing – 不依賴任何prior，直接設定超參數來看結果
  - Knowledge distillation – 依賴teacher 給予比重
  - 兩種想法結合以後，也許可以達到self-regularize。因為如果模型將錯誤的minor logits提高，也許就能更進一步讓模型來“理解”問題。

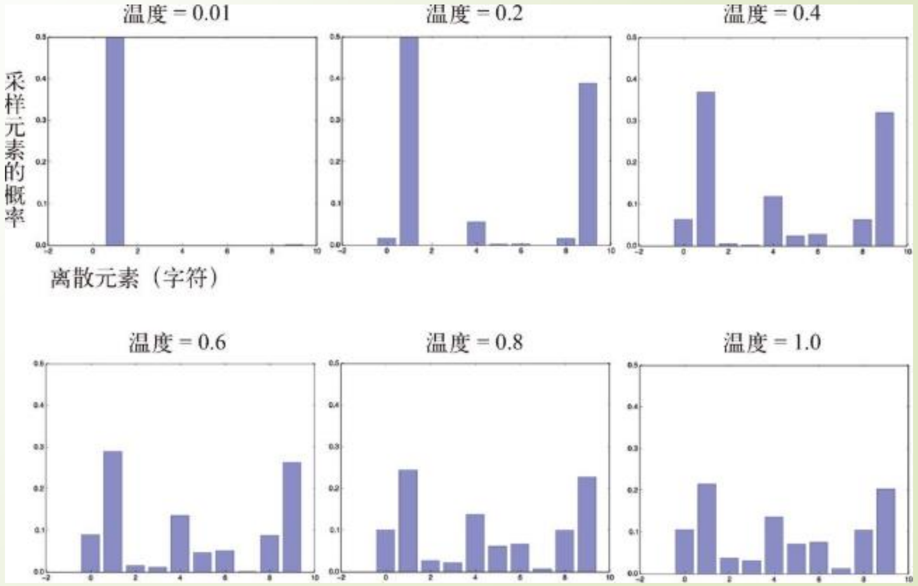


Model	Baseline	Tf-KD$_{self}$
MobileNetV2	68.38	70.96 (**+2.58**)
ShuffleNetV2	70.34	72.23 (**+1.89**)
ResNet18	75.87	77.10 (**+1.23**)
GoogLeNet	78.72	80.17 (**+1.45**)
DenseNet121	79.04	80.26 (**+1.22**)
ResNeXt29 (8x64d)	81.03	82.08 (**+1.05**)

(a) Test accuracy (in %) on CIFAR100. We run 3 times and report the mean of the best results.



使用不同溫度，可以凸顯出來的minor logits強度也不同。就看希望模型可以注意到多小的程度。



http://static.kancloud.cn/mikl_maple/python/1726331

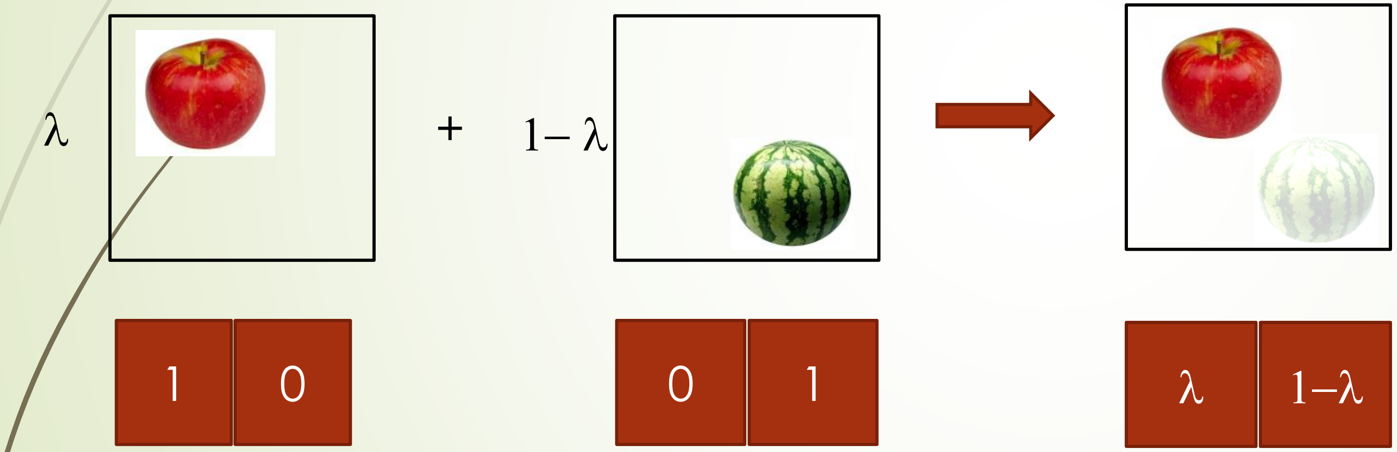
https://arxiv.org/pdf/1909.11723.pdf [CVPR2020]

What if controlling the dark knowledge directly

mixup

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j, \qquad \text{where } x_i, x_j \text{ are raw input vectors}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j, \qquad \text{where } y_i, y_j \text{ are one-hot label encodings}$$

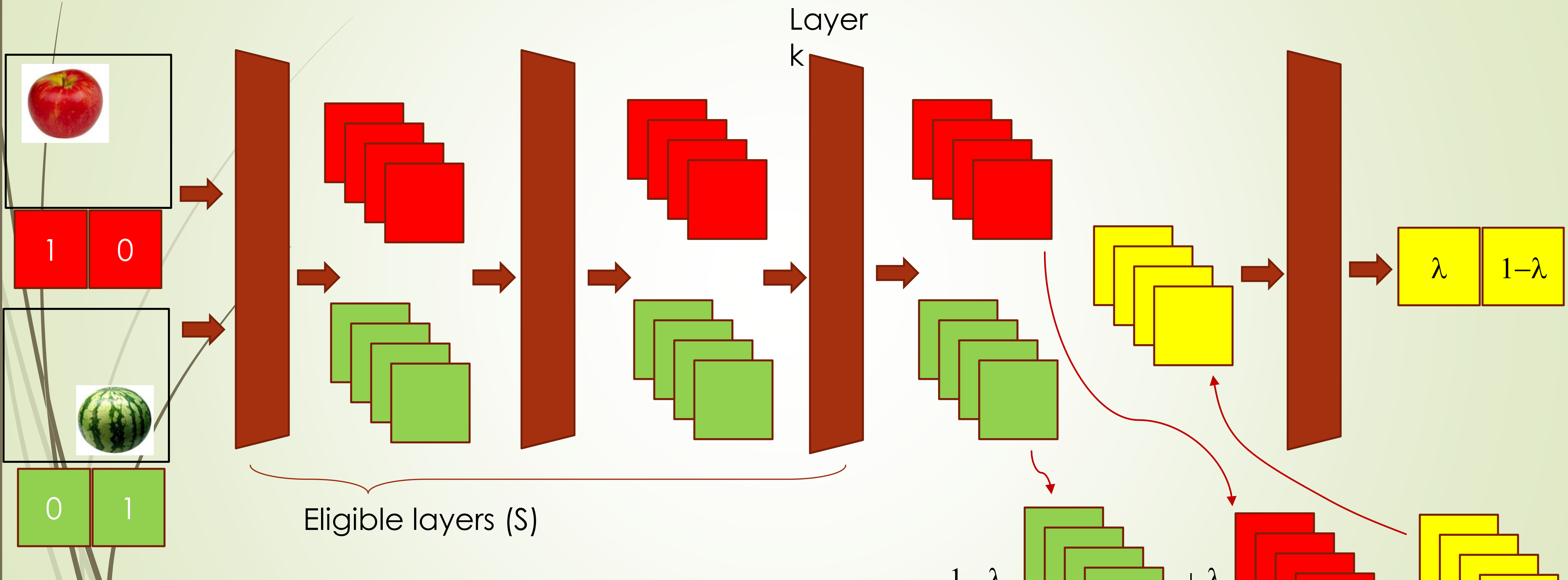


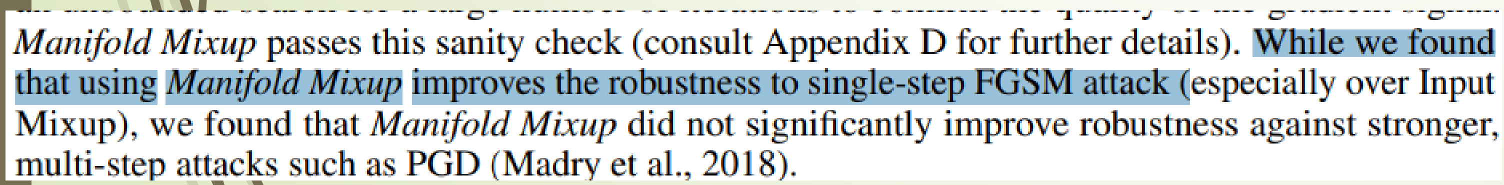
Main purpose:

1. Data augmentation
2. Maybe … Dark knowledge

ICLR2018
https://arxiv.org/pdf/1710.09412.pdf

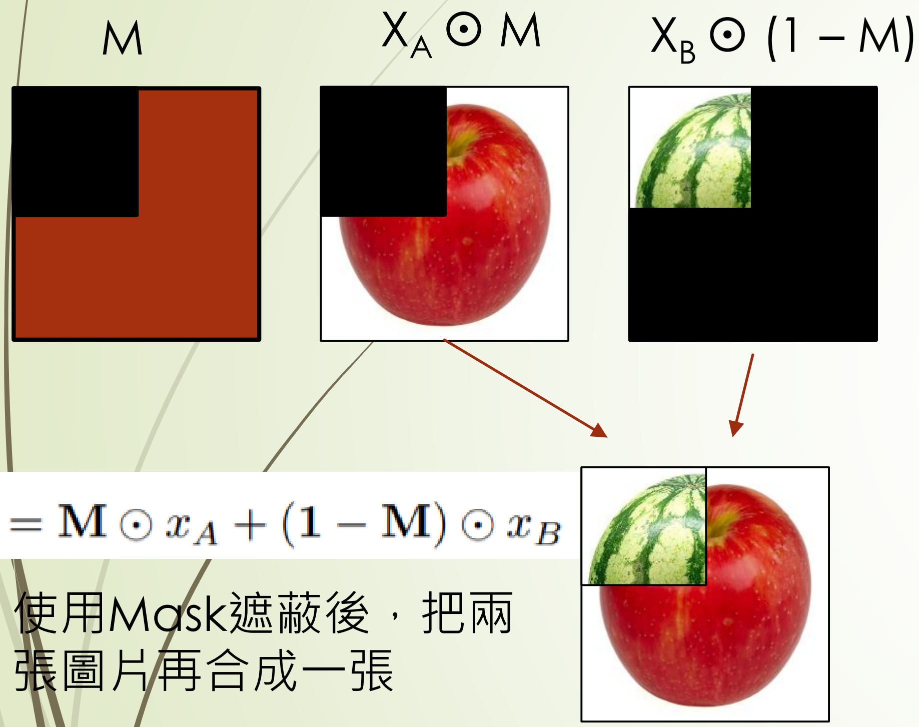
Manifold mixup




Manifold Mixup passes this sanity check (consult Appendix D for further details). While we found that using *Manifold Mixup* improves the robustness to single-step FGSM attack (especially over Input Mixup), we found that *Manifold Mixup* did not significantly improve robustness against stronger, multi-step attacks such as PGD (Madry et al., 2018).

https://arxiv.org/pdf/1806.05236.pdf

CutMix

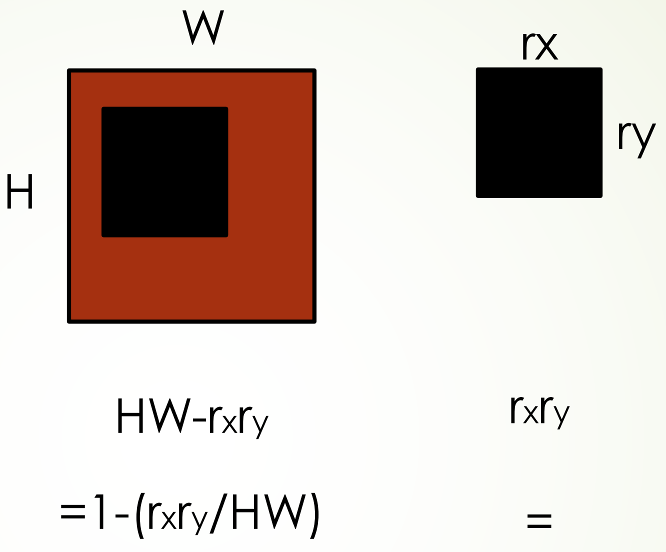


$$\tilde{x} = \mathbf{M} \odot x_A + (\mathbf{1} - \mathbf{M}) \odot x_B$$

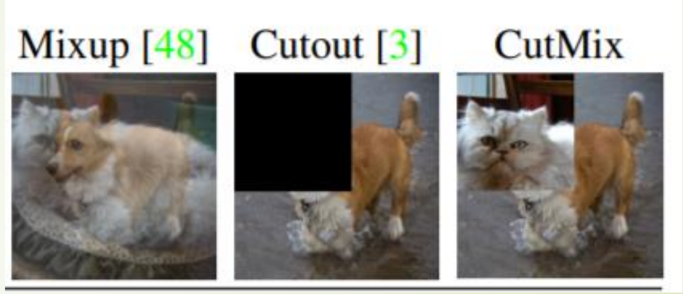
使用Mask遮蔽後，把兩張圖片再合成一張

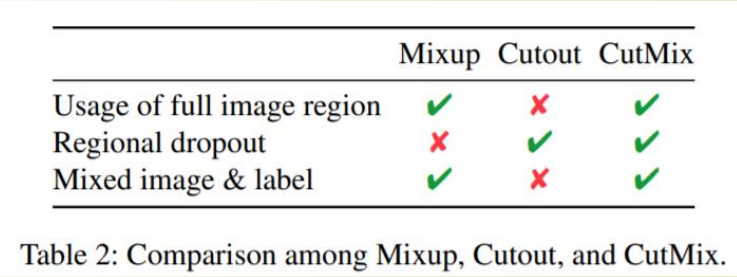
https://arxiv.org/pdf/1905.04899.pdf

label部分直接使用圖像所佔的比例，舉例:



如果Mask掉大小為r_xr_y，則圖A像原始比例剩下$HW-r_xr_y$，另外一張就是r_xr_y 。
Label比例也調整為$(HW-r_xr_y)$跟(r_xr_y)





	Mixup	Cutout	CutMix
Usage of full image region	✓	✗	✓
Regional dropout	✗	✓	✓
Mixed image & label	✓	✗	✓

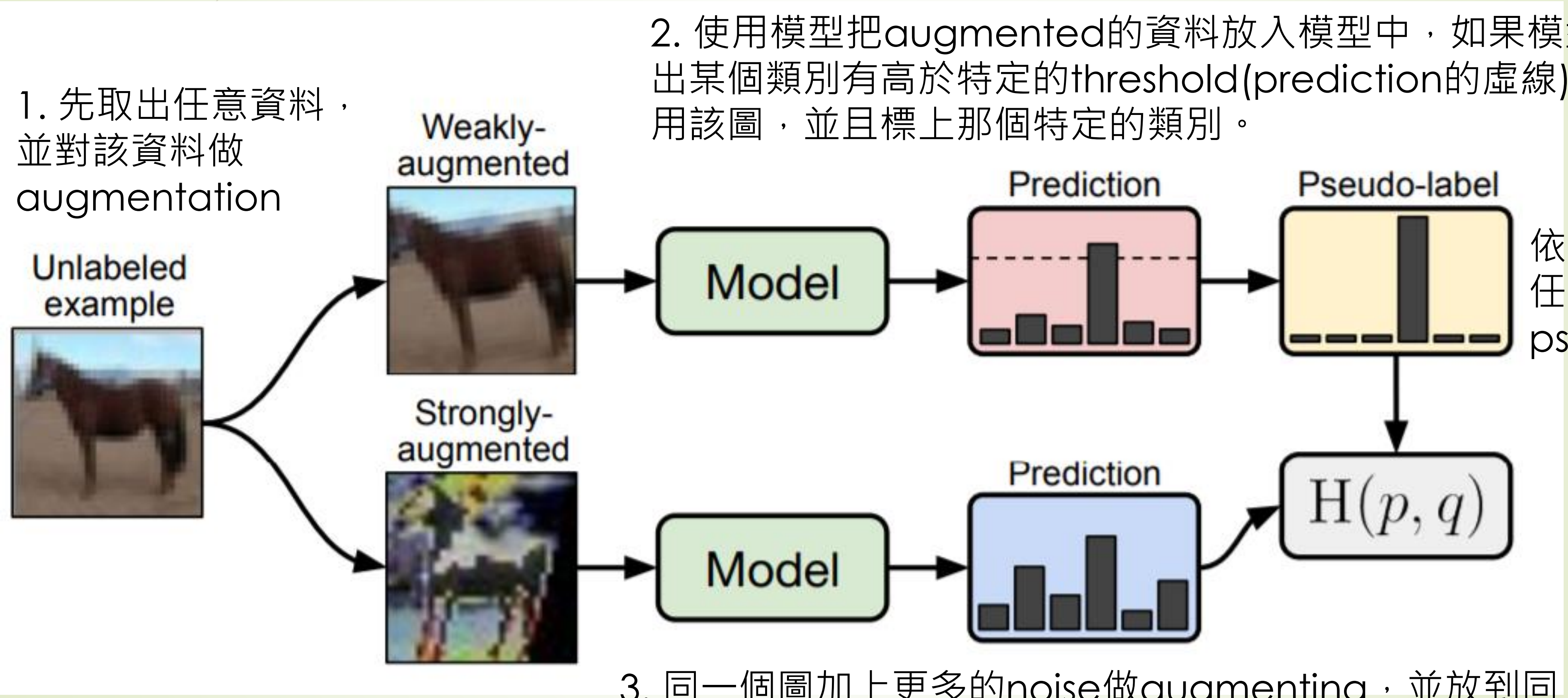
Table 2: Comparison among Mixup, Cutout, and CutMix.

內涵:
Dark knowledge部分，如果一開始無法從pretrain model中取得做蒸餾，也許我們可以自己製作這些dark knowledge。

Extract the knowledge of unlabeled data

If you have lots of unlabeled data and you want to leverage the information inside them

FixMatch



1. 先取出任意資料，並對該資料做augmentation

2. 使用模型把augmented的資料放入模型中，如果模型有算出某個類別有高於特定的threshold(prediction的虛線)，就使用該圖，並且標上那個特定的類別。

依據接下來的任務來指定pseudo-label

3. 同一個圖加上更多的noise做augmenting，並放到同一個模型當中，最後把算出來的結果進行cross entropy的運算

https://arxiv.org/pdf/2001.07685.pdf

MixMatch

After various objections, the most important result will be get after average.

Consistency regularization

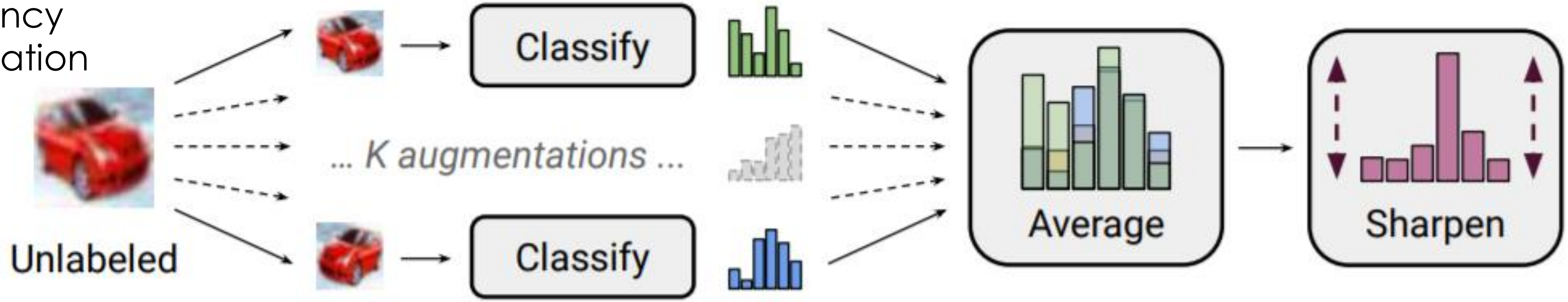
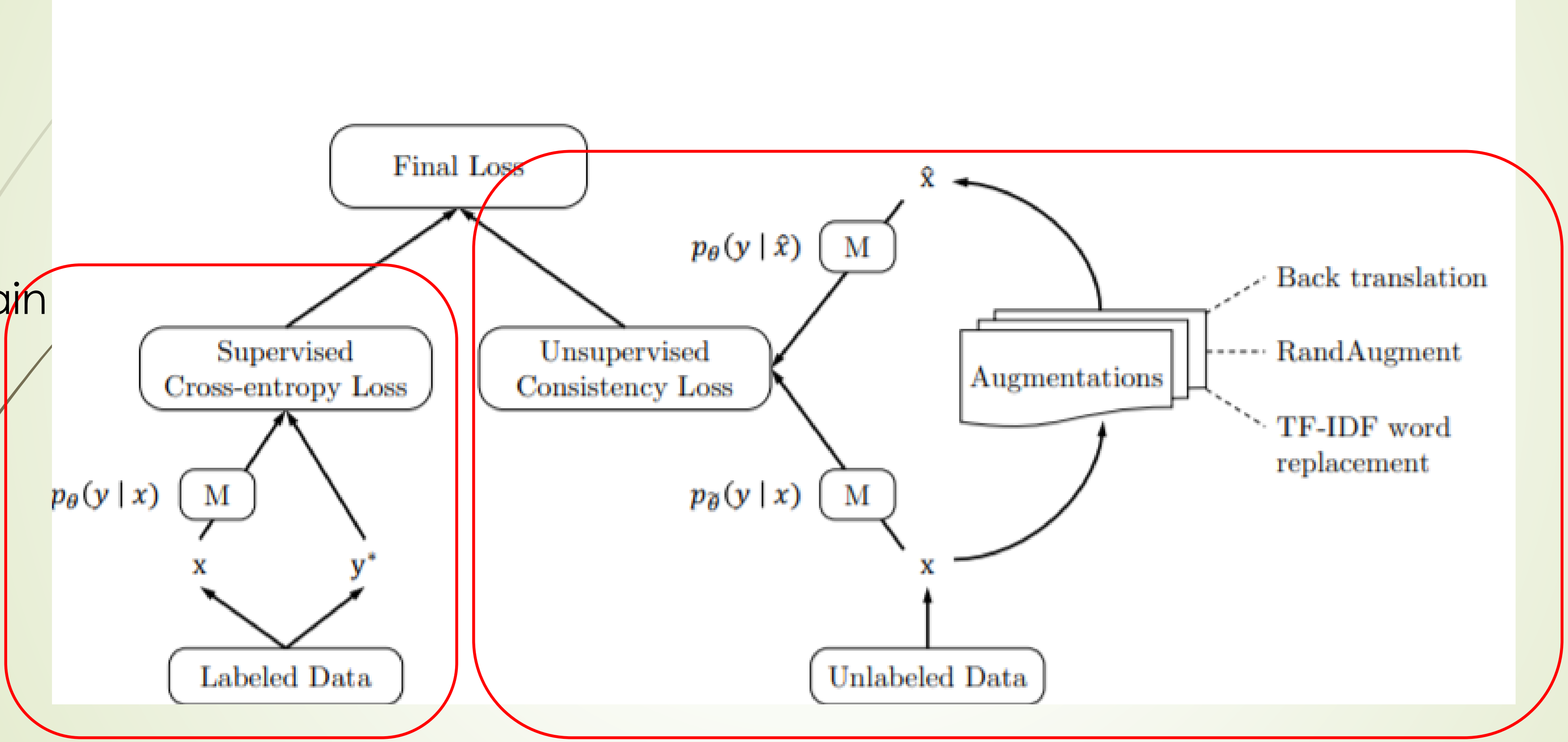


Figure 1: Diagram of the label guessing process used in MixMatch. Stochastic data augmentation is applied to an unlabeled image K times, and each augmented image is fed through the classifier. Then, the average of these K predictions is "sharpened" by adjusting the distribution's temperature. See algorithm 1 for a full description.

https://arxiv.org/pdf/1905.02249.pdf

Unsupervised data augmentation (UDA)

Normal process to train the neural network

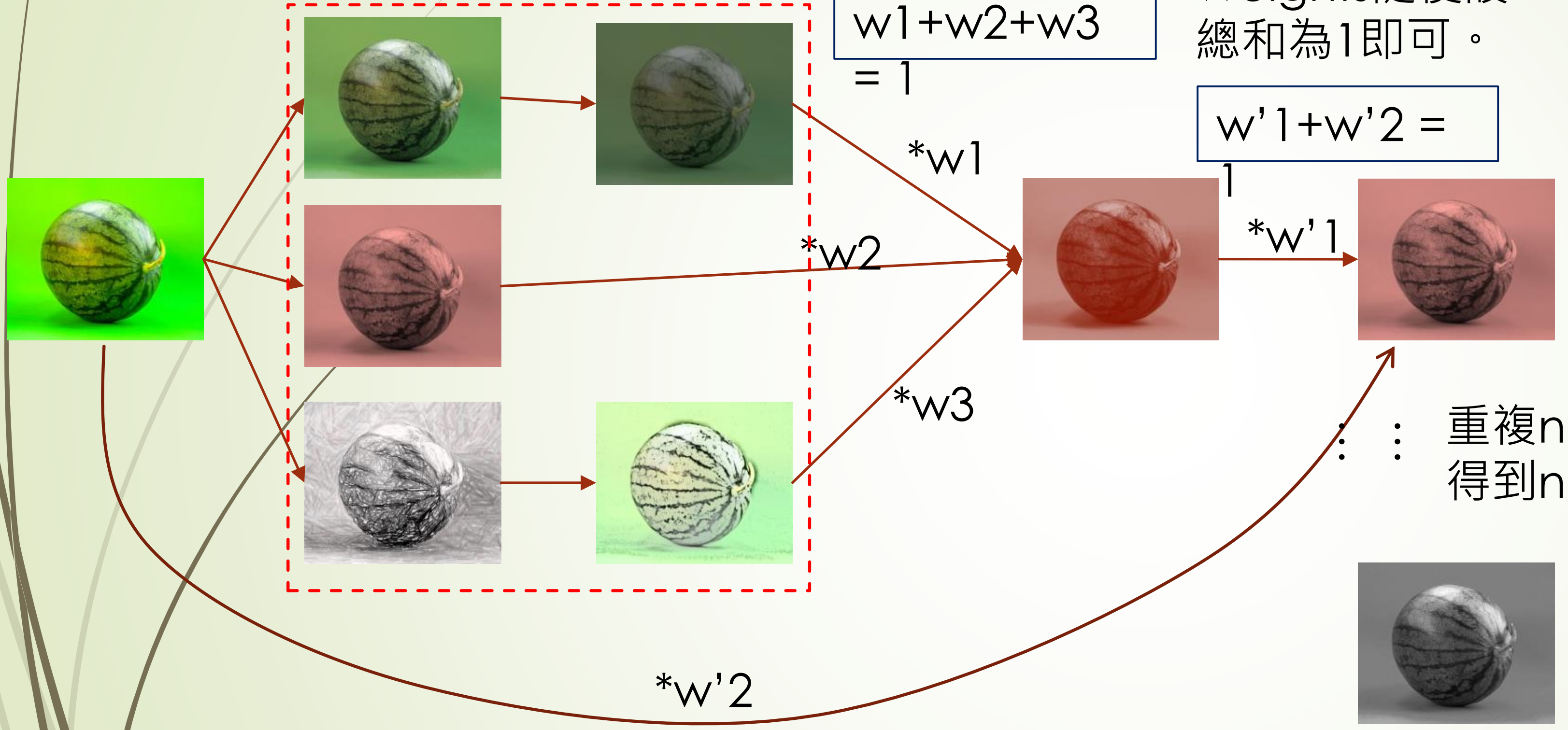


1. 把原始資料進行augmentation(加噪音等等的處理)
2. 以原始圖透過模型預測出來的數值作為答案，augmentation的圖最後答案應該要跟原始圖的答案一模一樣。

https://arxiv.org/pdf/1904.12848.pdf

AugMix

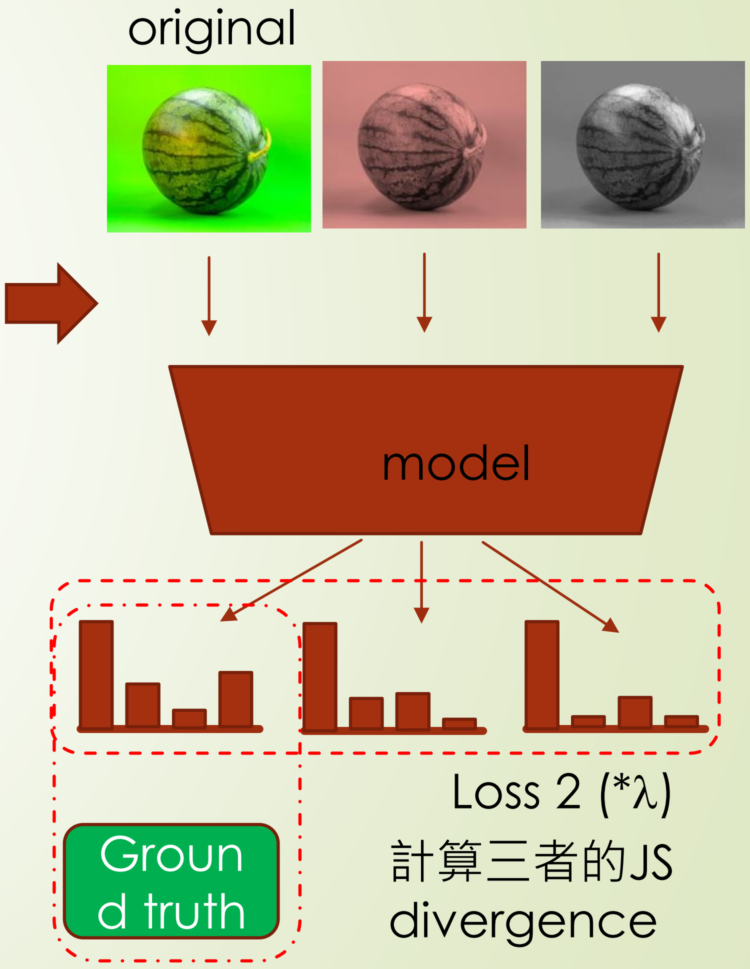
發揮你的想像力!!愛怎麼au就怎麼au



Weights隨便設，
總和為1即可。

$$M = (p_{\text{orig}} + p_{\text{augmix1}} + p_{\text{augmix2}})/3$$

$$\text{JS}(p_{\text{orig}}; p_{\text{augmix1}}; p_{\text{augmix2}}) = \frac{1}{3}\Big(\text{KL}[p_{\text{orig}} \| M] + \text{KL}[p_{\text{augmix1}} \| M] + \text{KL}[p_{\text{augmix2}} \| M]\Big).$$



Loss 2 (*λ)
計算三者的JS
divergence

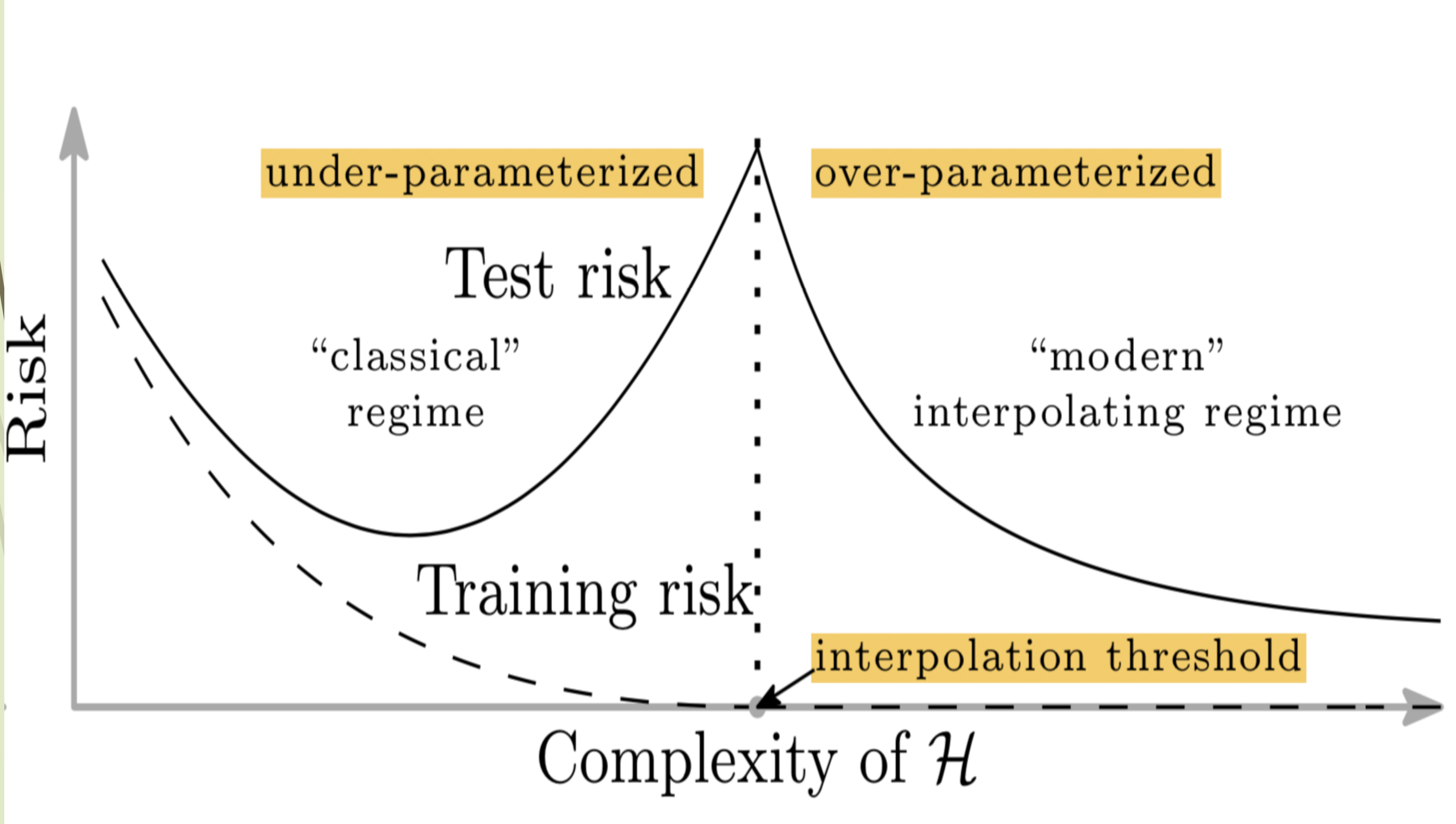
Loss 1 $\mathcal{L}(p_{\text{orig}}, y) + \lambda\, \text{JS}(p_{\text{orig}}; p_{\text{augmix1}}; p_{\text{augmix2}}).$

https://arxiv.org/pdf/1912.02781.pdf

Since the unlabeled data could give “information”

Using “data” for regularization would be a way to control the overparameterization models …

What is happening to overparameterization?

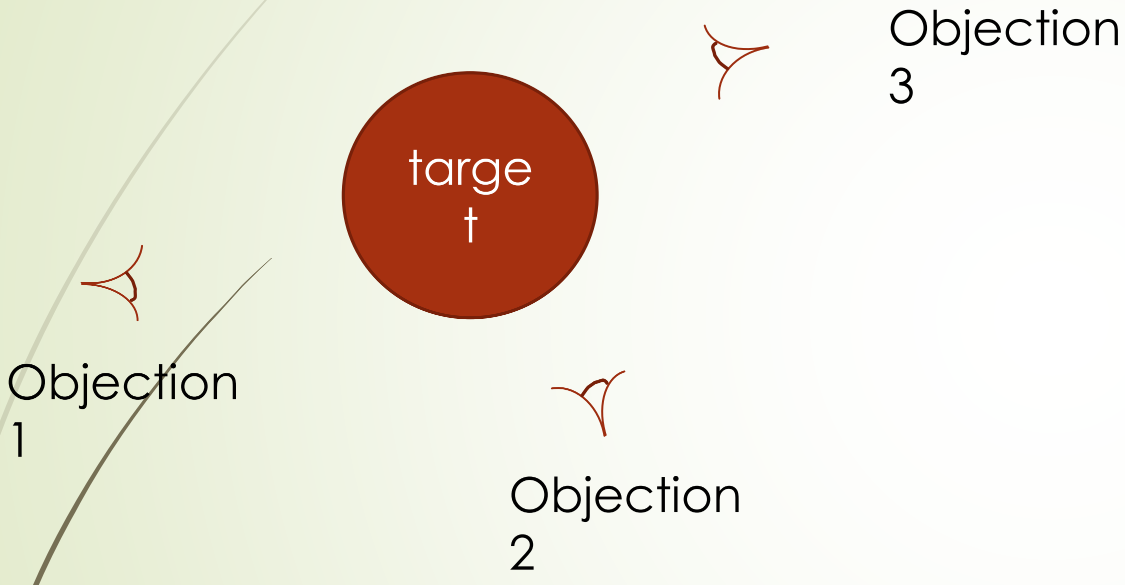


使用over-parameterization的model，基礎的machine learning concept多會不管用。
For example:
L1, L2 regularizations。在基礎的機器學習中是透過降低參數量達到generalized purpose。但是在over-parameter mode上，降低很多參數仍然會讓model處於over-parameterized的狀態。因此可能不會太管用。

解決方式:
目前最好的regularization method仍然是使用training data來達到目的。

https://lilianweng.github.io/lil-log/2019/03/14/are-deep-neural-networks-dramatically-overfitted.html

Contrastive learning



target

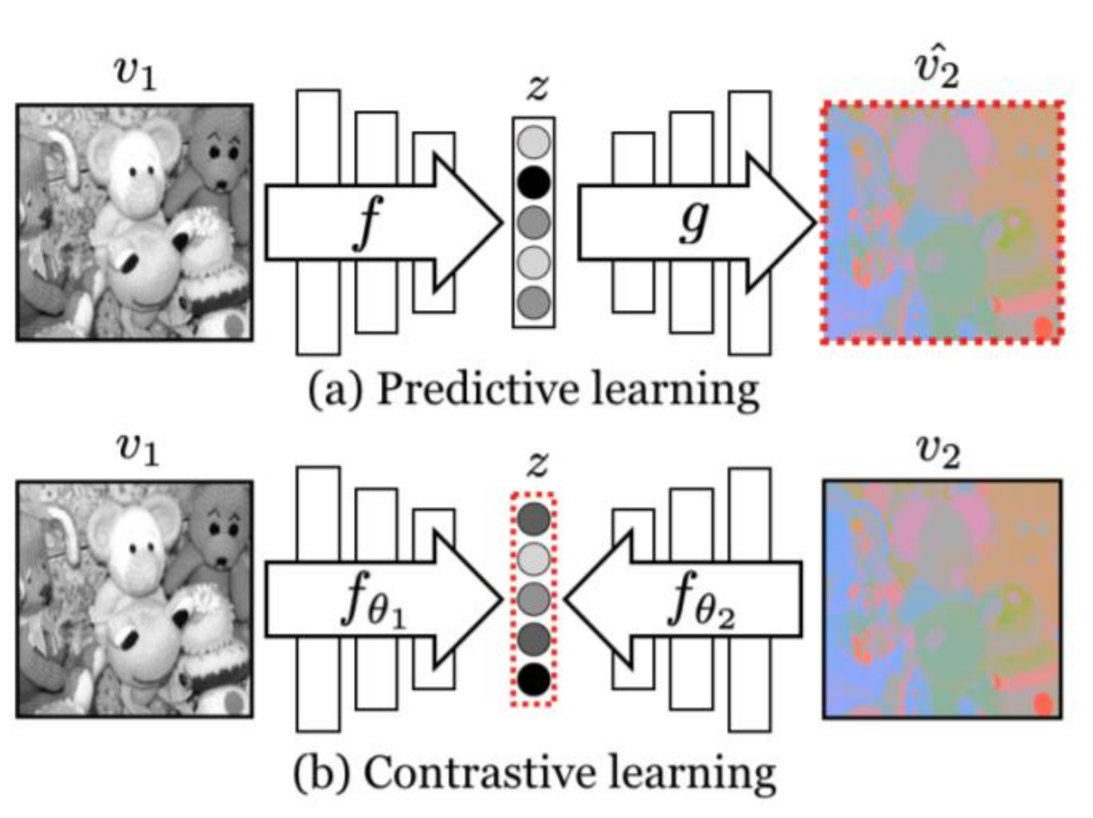
Objection 3

Objection 1

Objection 2

If we use different objection for the same target, we would get different presenting which are belong to the same objects

We need to tell the neural network that “these are the same thing”.



(a) Predictive learning

(b) Contrastive learning

https://arxiv.org/pdf/1906.05849.p

Mean teacher

Already proposing the concepts of self-supervise (consistency regularization)

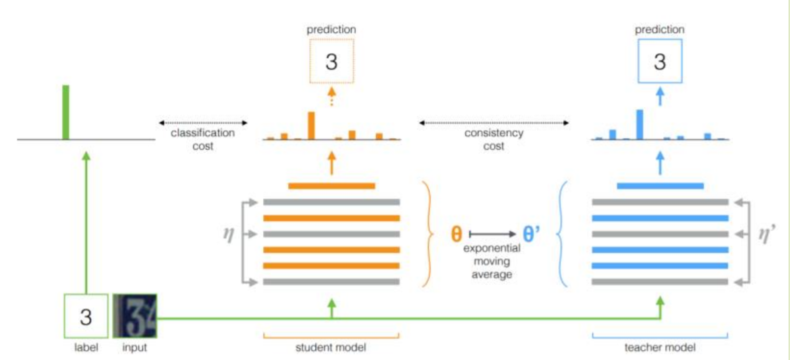
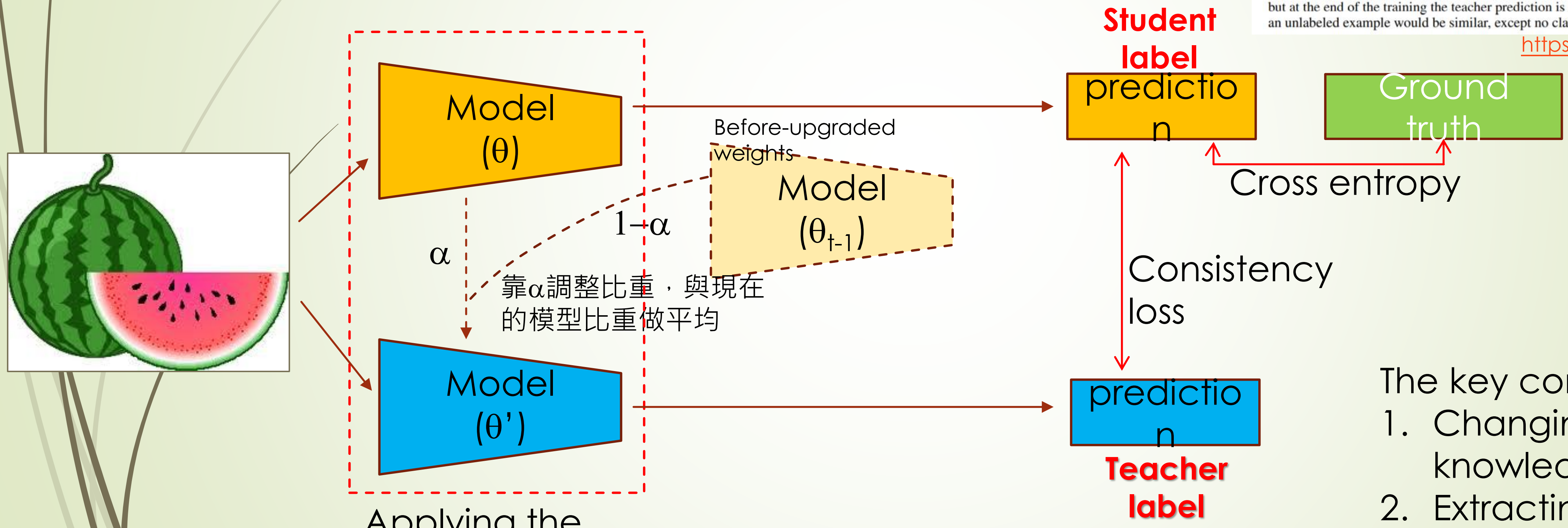


Figure 2: The Mean Teacher method. The figure depicts a training batch with a single labeled example. Both the student and the teacher model evaluate the input applying noise (η, η') within their computation. The softmax output of the student model is compared with the one-hot label using classification cost and with the teacher output using consistency cost. After the weights of the student model have been updated with gradient descent, the teacher model weights are updated as an exponential moving average of the student weights. Both model outputs can be used for prediction, but at the end of the training the teacher prediction is more likely to be correct. A training step with an unlabeled example would be similar, except no classification cost would be applied.

https://arxiv.org/pdf/1703.01780.p



靠α調整比重，與現在的模型比重做平均

Applying the noise to the weights (η and η')

(Dropout would be a good

α would not to be large during initially training stage

The key concept:

1. Changing the "dark knowledge"
2. Extracting the ensemble characteristics.

自己跟自己比還不夠

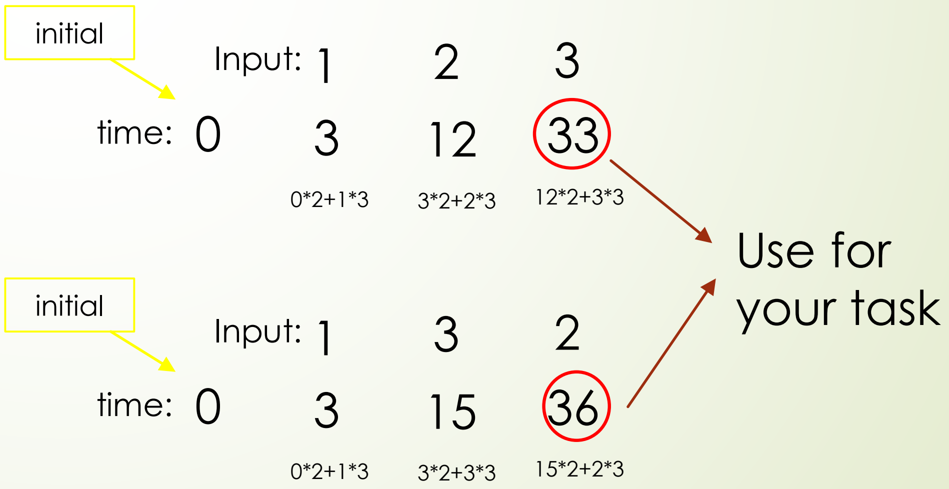
透過跟別人比較，更了解自己的定位

Metric learning (meta-learning)

Recurrent neural network

- RNN(recurrent neural network)
  - Use the same tensor for sequence data
  - For example
    - time * 2
    - input * 3

Learning target



initial

Input:		1	2	3
time:	0	3	12	33
		0*2+1*3	3*2+2*3	12*2+3*3

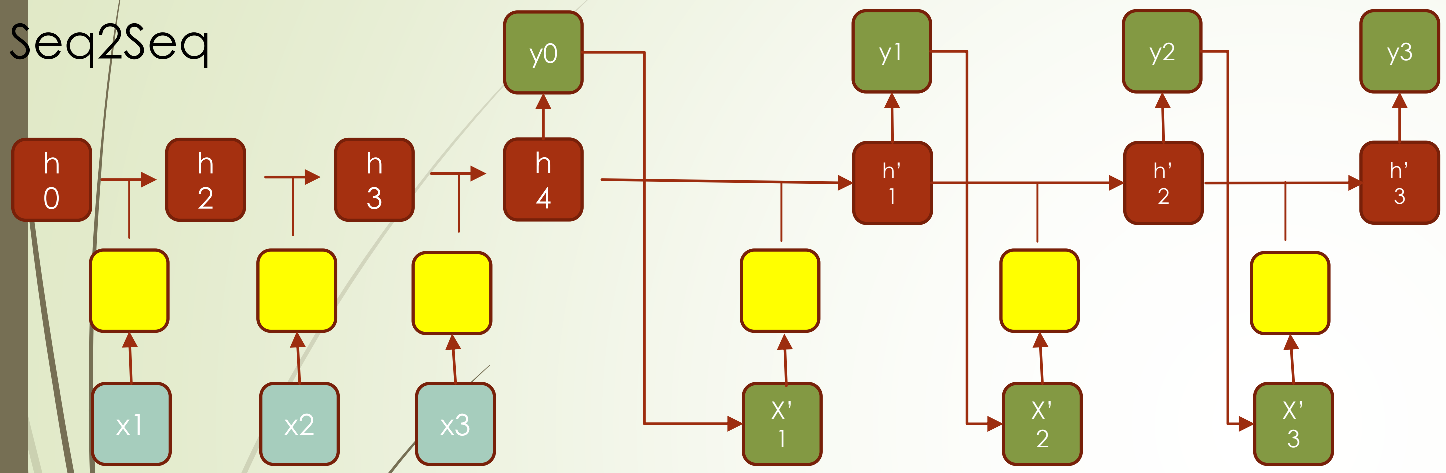
initial

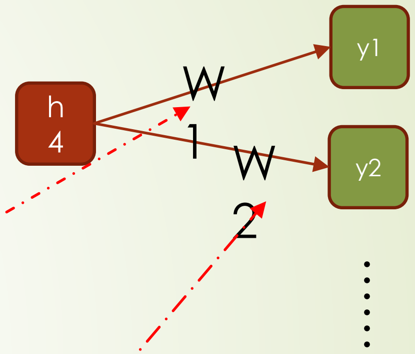
Input:		1	3	2
time:	0	3	15	36
		0*2+1*3	3*2+3*3	15*2+2*3

Use for your task

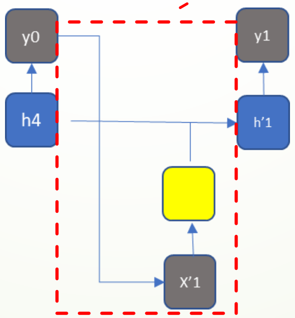
Rethinking of RNN

問題:
是否能學到一堆矩陣就當成不同位置的開關?

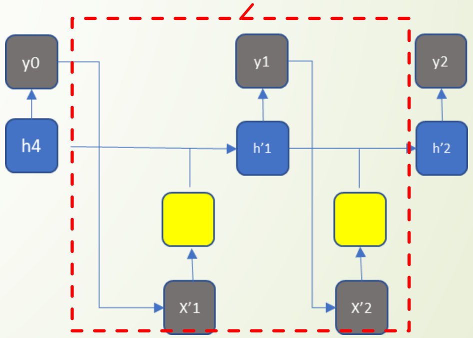




1. h'系列全部都是由h4所得到，換句話說，h4早就隱含了h'系列的所有資訊
2. h'系列都是h4透過某些方式"打開"開關。例如h'1接收了x'1；h'2等於接收了x'1及x'2，以此類推。因此可以把這些"接收訊息"(例如x'1, x'1+x'2, ...)當成"矩陣"，而這矩陣專門用來打開h4對於特定位置的開關。當h4接受到這些特定開關以後，就能把特定數值輸出即可。



Y1根本是自己產生，只是透過x'1打開自己

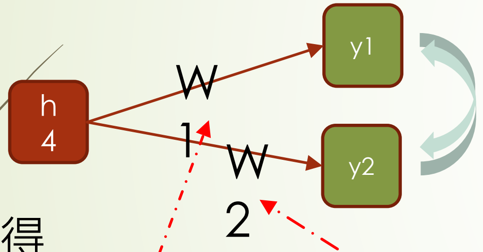


Y2是透過x'1和x'2的訊息累加後做為開關。但x'1和x'2也是由h4自己產生

Contrastive prediction code (CPC)

What is the “Contrastive”?
如果有“同一個”物件，我們用“不同角度”來看它，就能找出不同處。

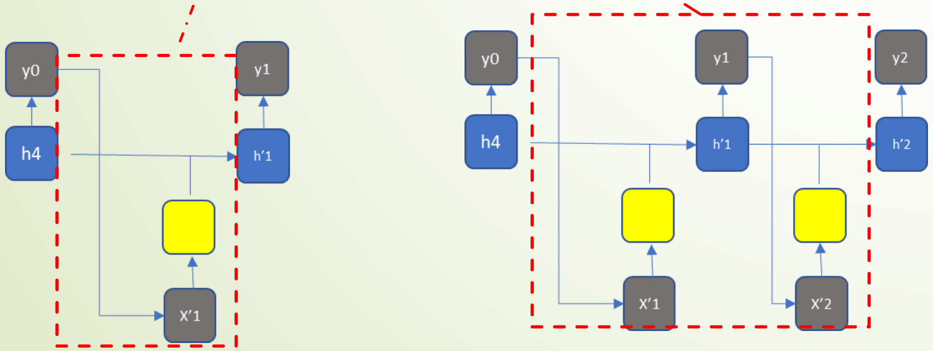
目前常用的方法就是加上noise或甚至直接乘以矩陣。



同一個h4，用不同的W處理，同等于用不同角度在看h4。

因此把h4投射到不同位置，等於在計算mutual information。

假設從RNN部分取得的資訊可以拆解(參閱上一頁的投影片)



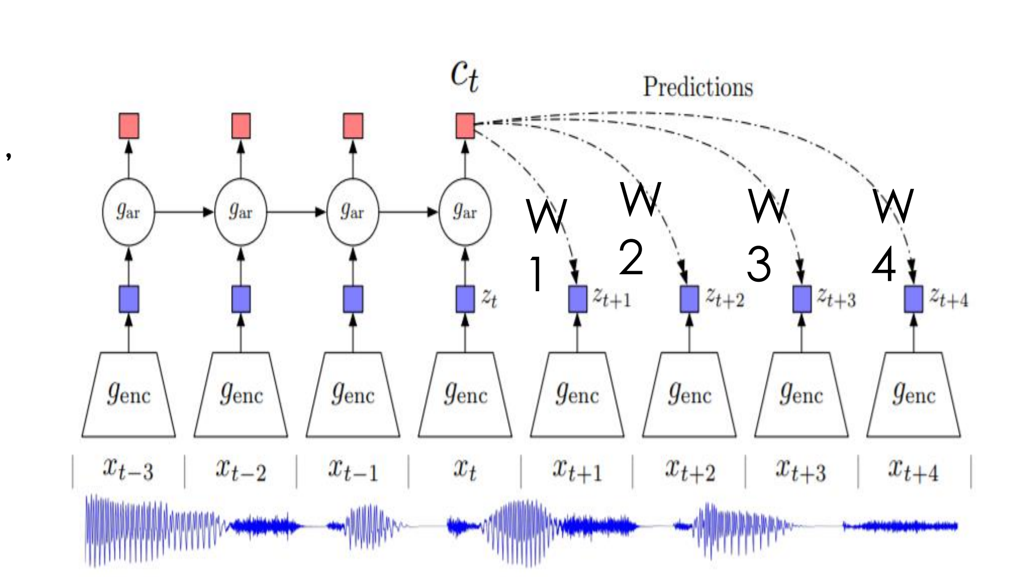
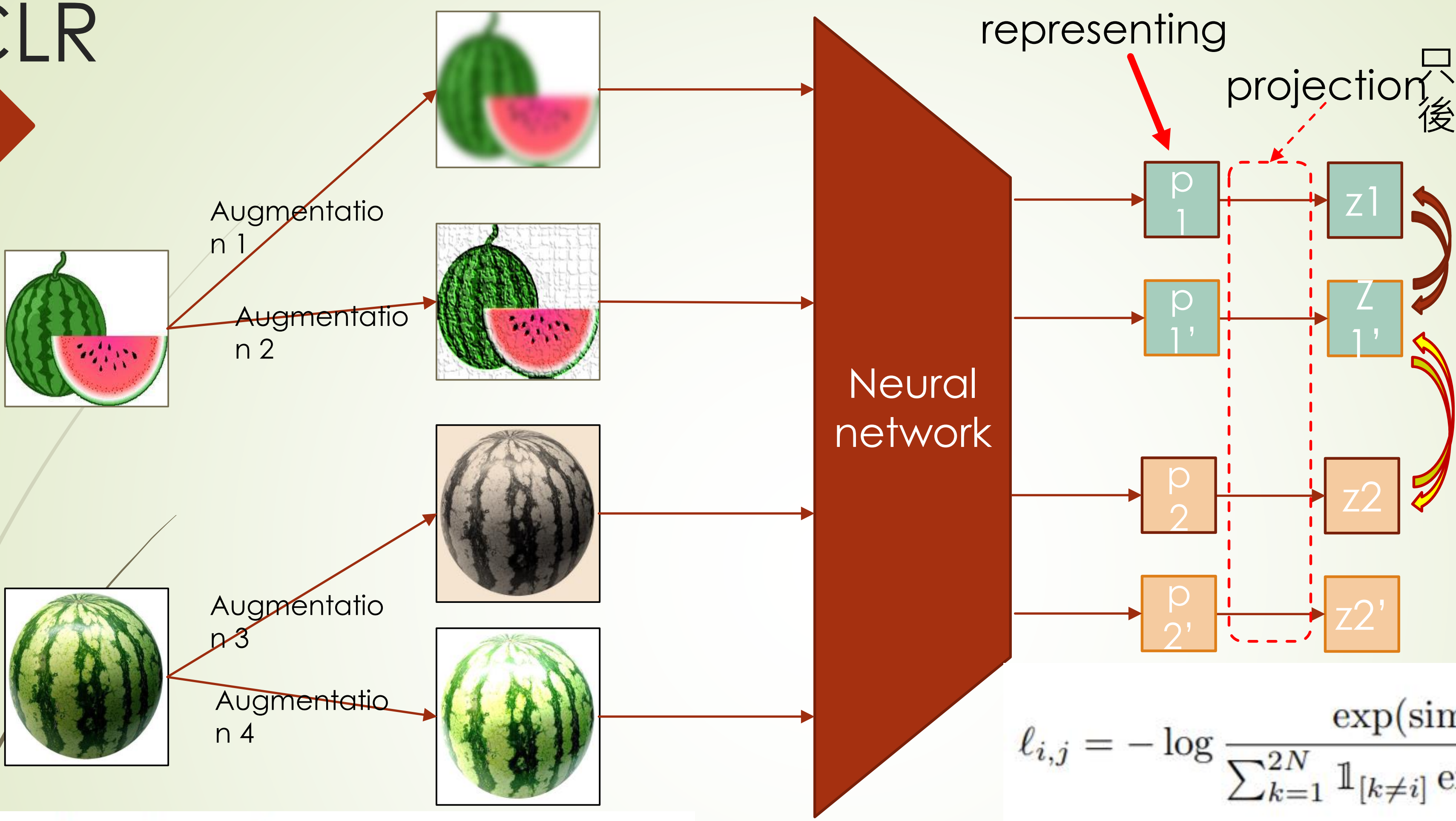


Figure 1: Overview of Contrastive Predictive Coding, the proposed representation learning approach. Although this figure shows audio as input, we use the same setup for images, text and reinforcement learning.

https://arxiv.org/pdf/1807.03748.p

SimCLR



只是為了訓練，最後可以丟掉

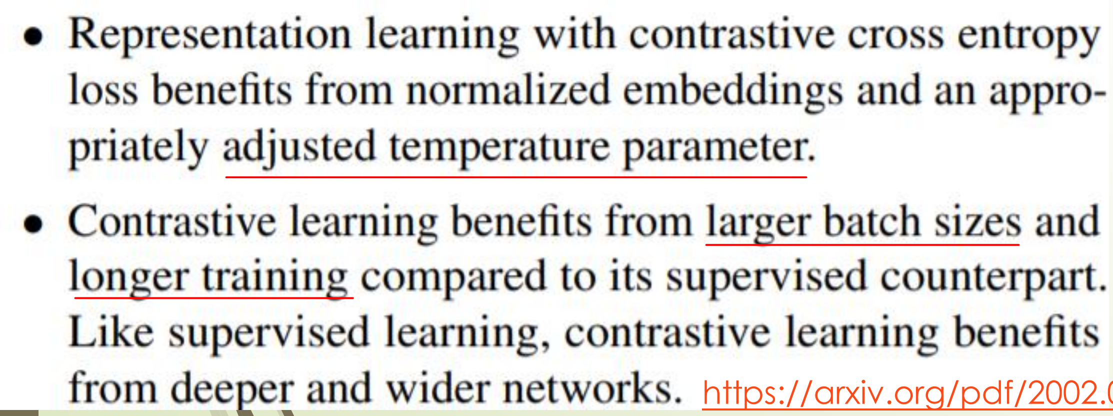
As close as possible (分子)

As far as possible (分母)

Use cos similarity in SimCLR

temperature

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}, \quad (1)$$



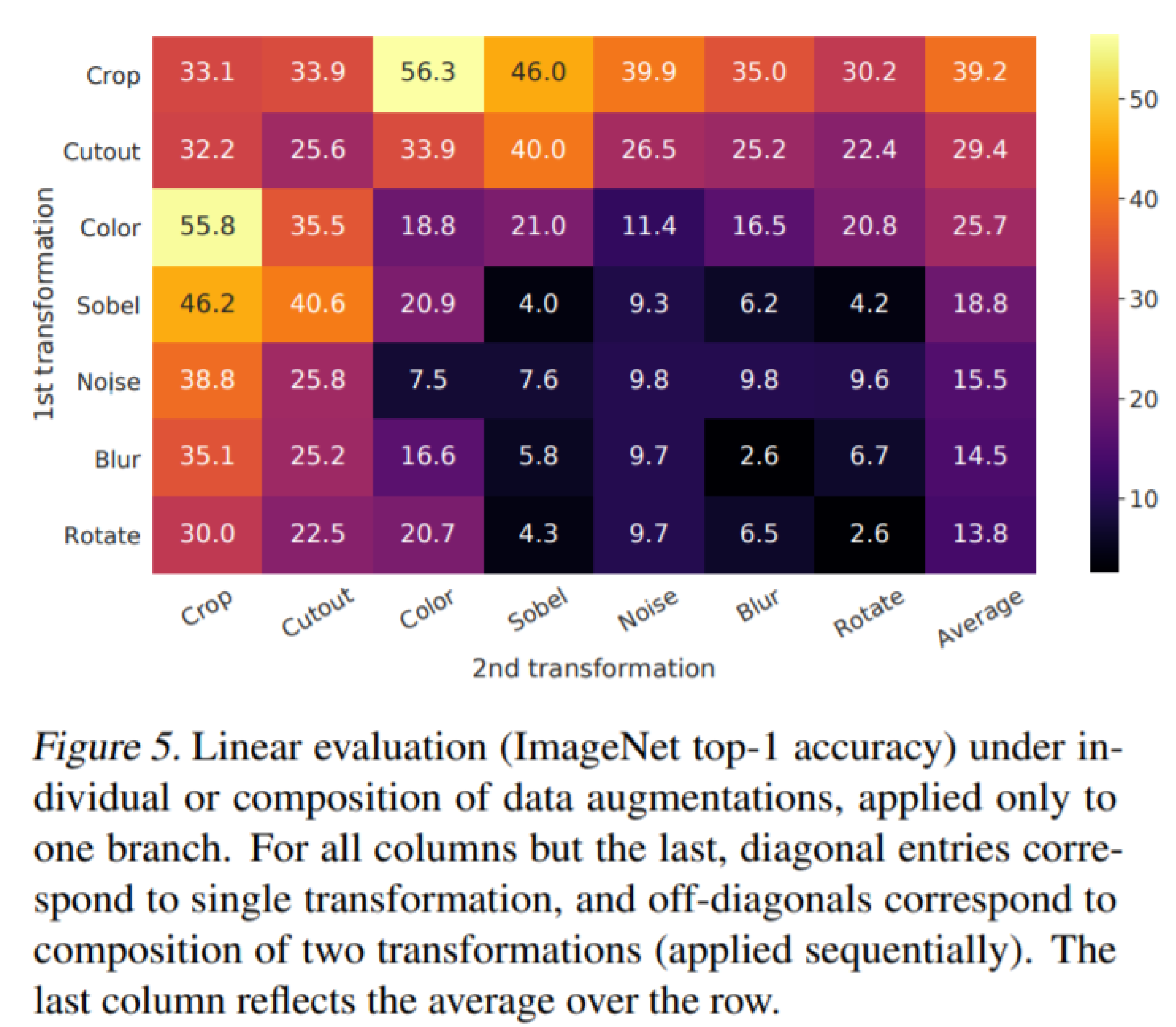
- Representation learning with contrastive cross entropy loss benefits from normalized embeddings and an appropriately adjusted temperature parameter.
- Contrastive learning benefits from larger batch sizes and longer training compared to its supervised counterpart. Like supervised learning, contrastive learning benefits from deeper and wider networks.

https://arxiv.org/pdf/2002.05709.pdf

Tips

1. Because the sample size will be huge, LARS is recommended in such tasks.
2. Using layer normalization or global batch normalization.

More in SimCLR – crop and color distortions get better performance



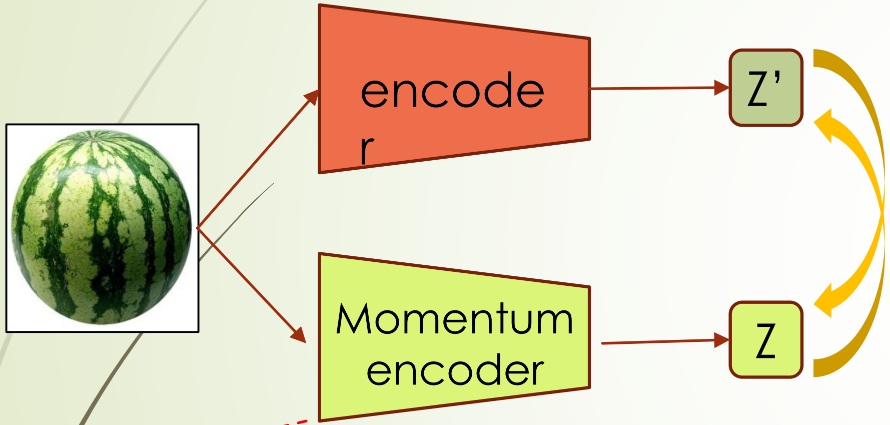
	Crop	Cutout	Color	Sobel	Noise	Blur	Rotate	Average
Crop	33.1	33.9	56.3	46.0	39.9	35.0	30.2	39.2
Cutout	32.2	25.6	33.9	40.0	26.5	25.2	22.4	29.4
Color	55.8	35.5	18.8	21.0	11.4	16.5	20.8	25.7
Sobel	46.2	40.6	20.9	4.0	9.3	6.2	4.2	18.8
Noise	38.8	25.8	7.5	7.6	9.8	9.8	9.6	15.5
Blur	35.1	25.2	16.6	5.8	9.7	2.6	6.7	14.5
Rotate	30.0	22.5	20.7	4.3	9.7	6.5	2.6	13.8

Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.

https://arxiv.org/pdf/2002.05709.pdf

MOCO

Knowledge can be transfer directly



Info-NCE

$$\mathcal{L}_q = -\log \frac{\exp(q{\cdot}k_+/\tau)}{\sum_{i=0}^{K} \exp(q{\cdot}k_i/\tau)}$$

$$\theta_Z \leftarrow \alpha\theta_Z + (1-\alpha)\theta_{Z'}$$

直接使用Z’比重更新Z。
其內涵為:
1. 兩個不同比重對同一件事情都能完整表達，表示這樣的表達是正確的。
2. 使用的info-NCE的並且搭配了temperature，表示在dark knowledge的部分模型也必須重視。

https://arxiv.org/pdf/1911.05722.pdf

Tips:

- Batch normalization(BN)必須要進行shuffling。

因為使用info-NCE的狀況下，BN會洩漏正樣本跟負樣本之間的訊息造成模型只看BN輸出。

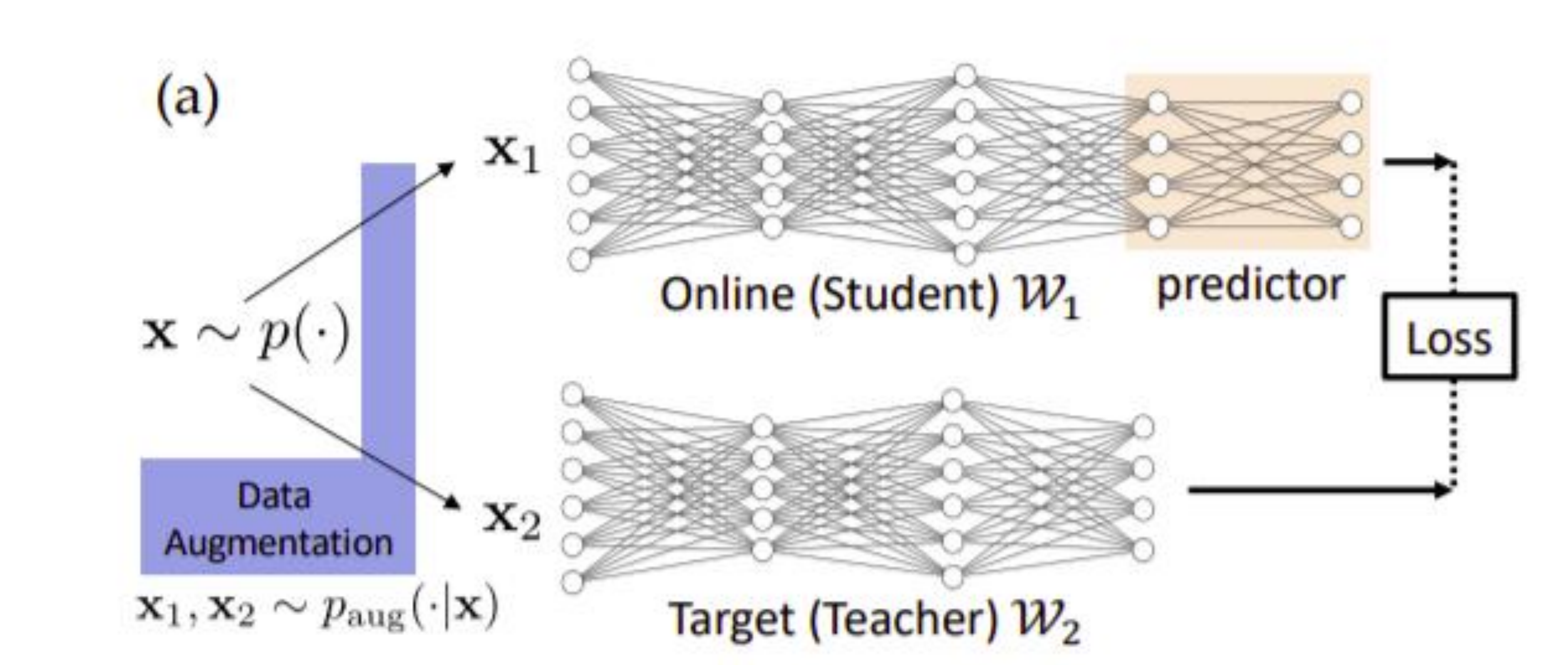
- α越大越好，即momentum encoder更新幅度越小越好

MOCO v2?

1. 補上projection (from SimCLR)
2. 補上augmentation (from SimCLR)

https://arxiv.org/pdf/2003.04297.pdf

Self-supervised learning could be thought as dual deep neural network


(a)
$\mathbf{x}_1$
$\mathbf{x} \sim p(\cdot)$
Data Augmentation
$\mathbf{x}_2$
$\mathbf{x}_1, \mathbf{x}_2 \sim p_{\text{aug}}(\cdot|\mathbf{x})$
Online (Student) $\mathcal{W}_1$
predictor
Loss
Target (Teacher) $\mathcal{W}_2$

https://arxiv.org/pdf/2010.00578.pdf

Conclusion and suggestions

- Handling the dark knowledge would be a way to handling the overparameterization model
- In practical, we would need to learn some skill that the we might not notice in fashion way.
  - LARS ?

Questions ?

If you like us, you can join our clubs: (of course, they are all FREE)

1. 北Bio : https://www.facebook.com/groups/446434039038963
2. 台灣人工智慧社團: https://www.facebook.com/groups/525579498272187/
3. Deep learning 101 : https://www.facebook.com/TWMANDeepLearning/