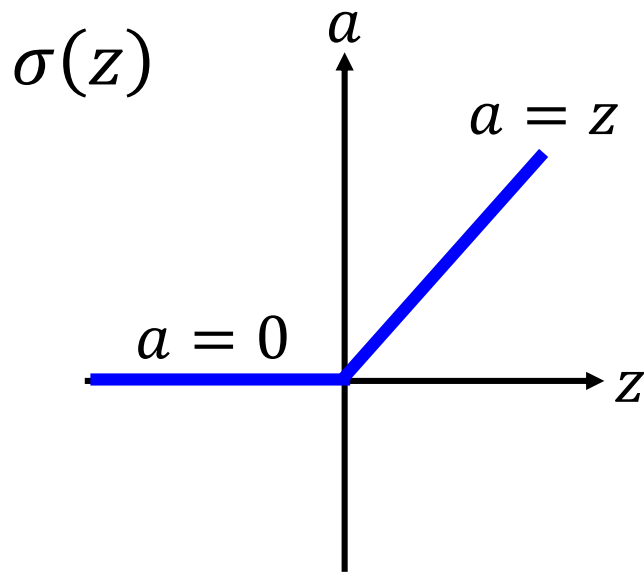


Activation Function: SELU

ReLU

- Rectified Linear Unit (ReLU)

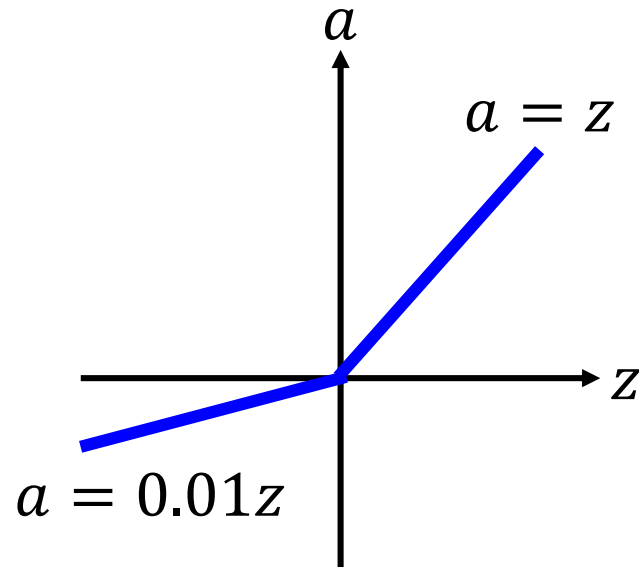
Reason:



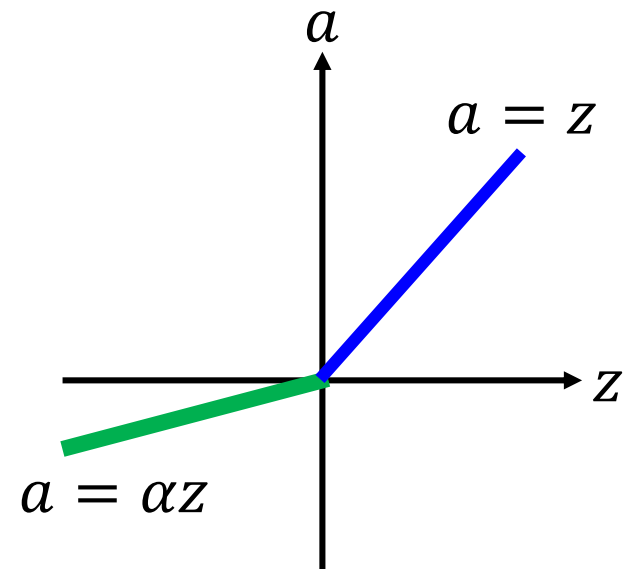
1. Fast to compute
2. Biological reason
3. Infinite sigmoid with different biases
4. Vanishing gradient problem

ReLU - variant

Leaky ReLU



Parametric ReLU



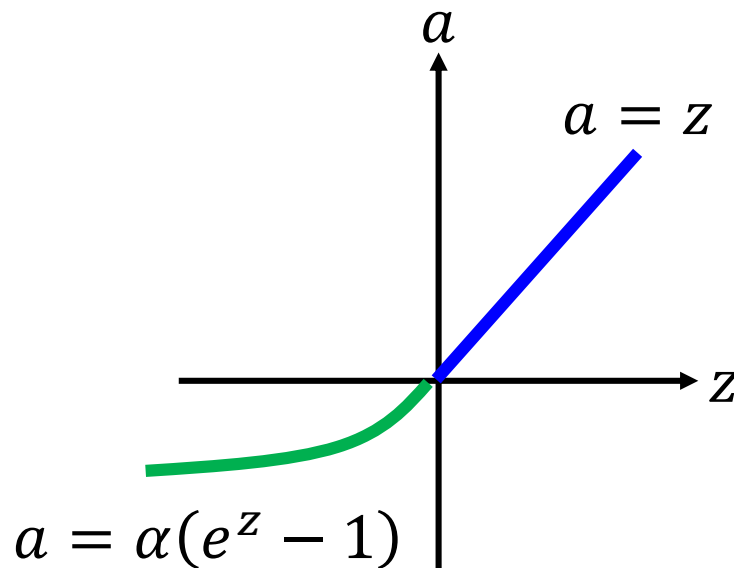
α also learned by
gradient descent

(1) Definition of scaled exponential linear units (SELUs)

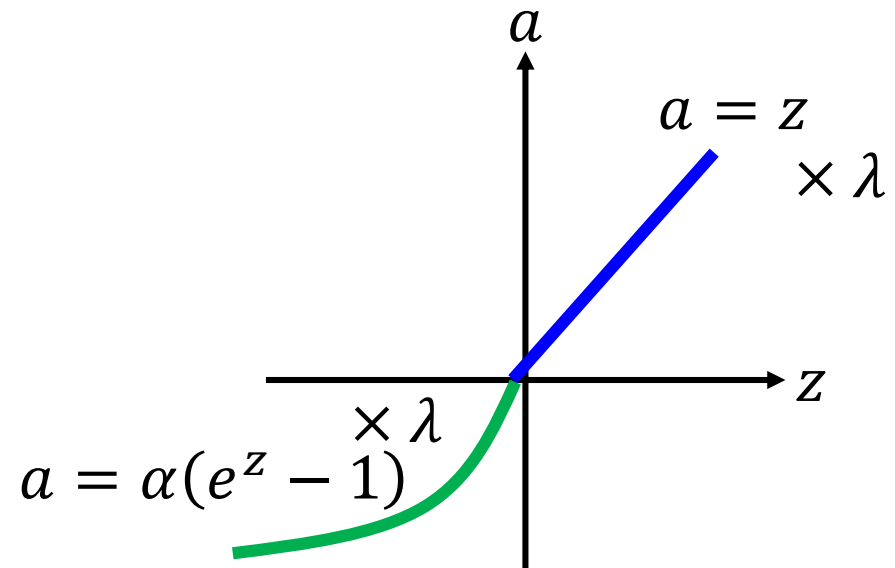
```
In [3]: def selu(x):  
    with ops.name_scope('elu') as scope:  
        alpha = 1.6732632423543772848170429916717  
        scale = 1.0507009873554804934193349852946  
        return scale*tf.where(x>=0.0, x, alpha*tf.nn.elu(x))
```

<https://github.com/bioinf-jku/SNNs>

Exponential Linear Unit (ELU)

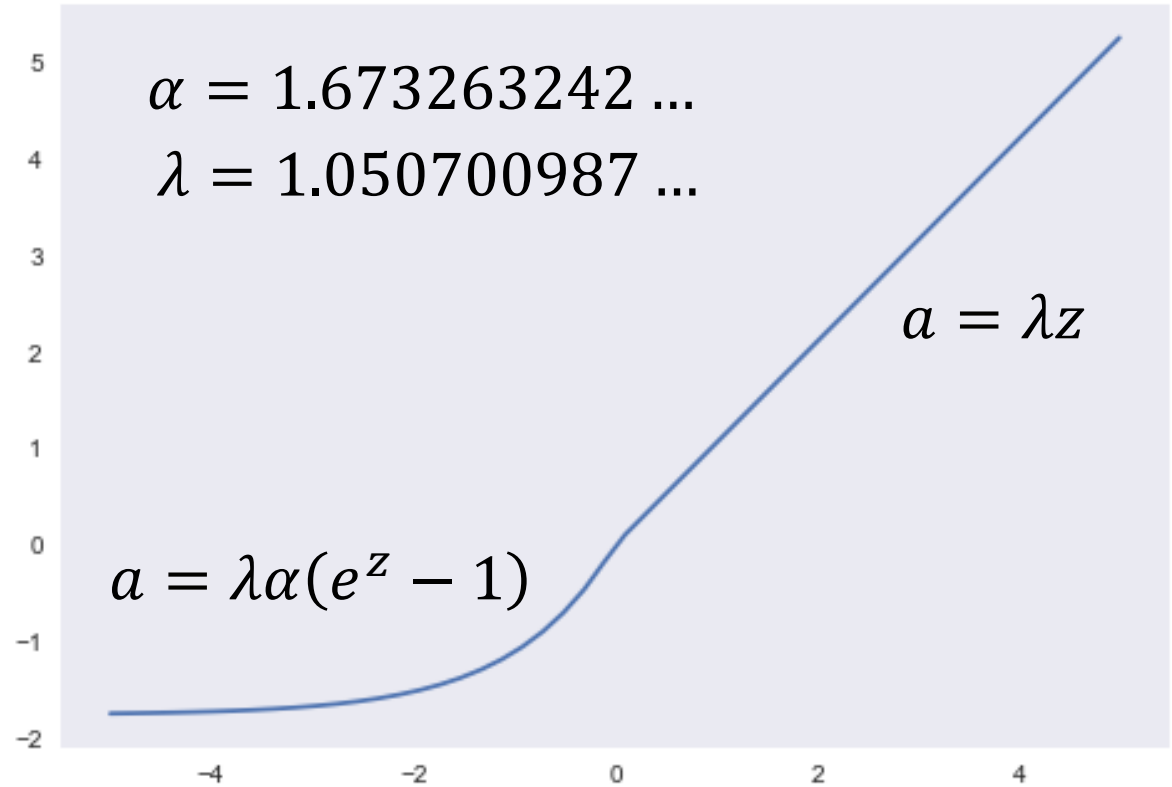


Scaled ELU (SELU)



$$\alpha = 1.6732632423543772848170429916717$$
$$\lambda = 1.0507009873554804934193349852946$$

SELU



Positive and negative values

➡ The whole ReLU family has this property except the original ReLU.

Saturation region ➡ ELU also has this property

Slope larger than 1 ➡ Only SELU also has this property

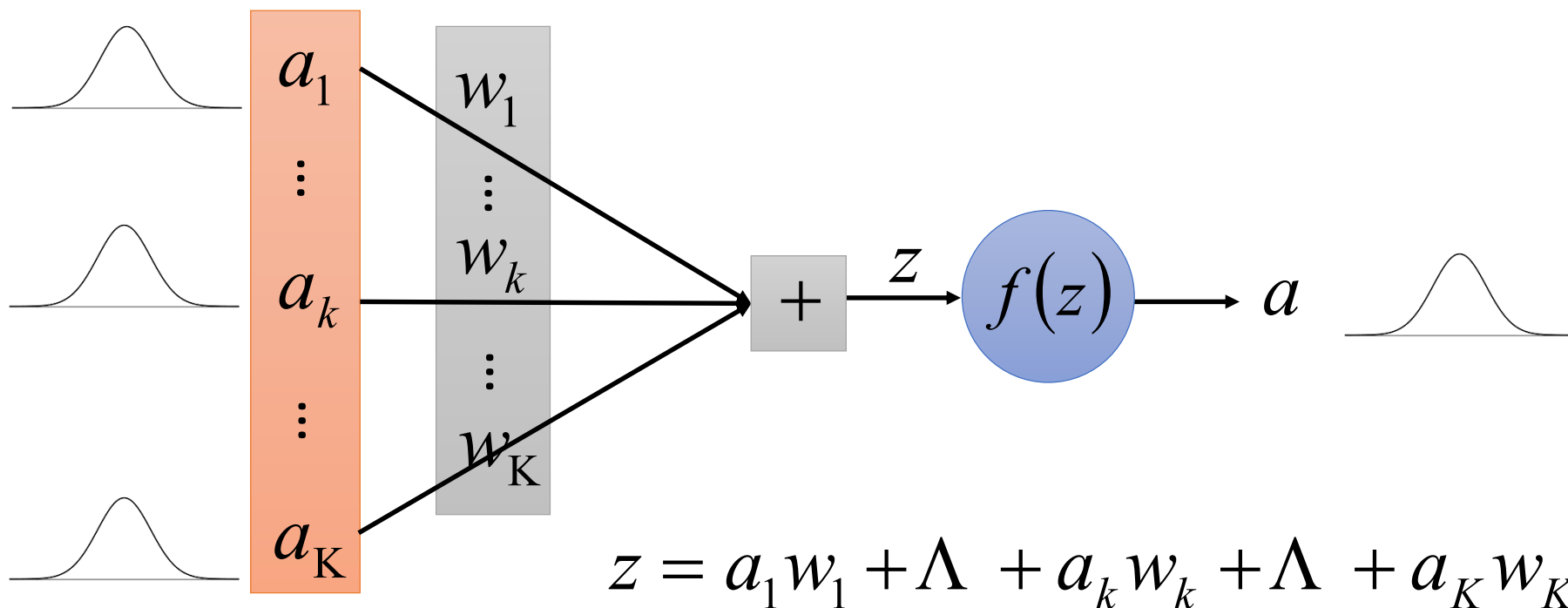
SELU

自帶normalization

The inputs are i.i.d random variables with mean μ and variance σ^2 . $\mu=0$ $\sigma^2=1$

$$\begin{aligned}\mu_z &= E[z] \\ &= \sum_{k=1}^K \frac{E[a_k]}{\mu} w_k = \mu \sum_{k=1}^K w_k = \mu \cdot K \mu_w \\ &\quad \quad \quad =0 \quad \quad =0\end{aligned}$$

希望最後output的a也是mean=0, variance=1



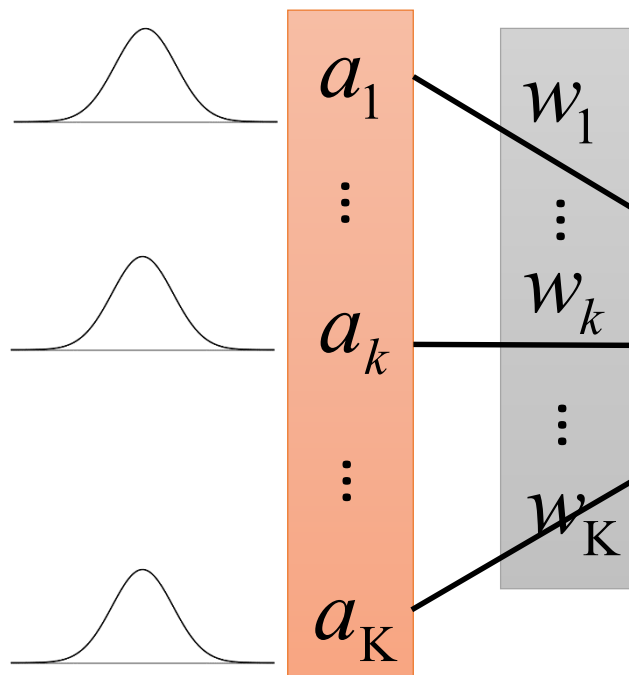
Do not have to be Gaussian

keras在default時sigmoid平方就是1/k
 如果不符合這個設定的話，selu就沒有用了！

SELU

希望他output的mean是0，output是1

The inputs are i.i.d random variables with mean μ and variance σ^2 . $\mu=0$, $\sigma^2=1$



假設 $\mu_w = 0$, $\sigma_w^2 = 1/k$

$$\mu_z = 0 \quad \mu_w = 0$$

$$\sigma_z^2 = E[(z - \mu_z)^2] = E[z^2]$$

$$= E[(a_1 w_1 + a_2 w_2 + \dots)^2]$$

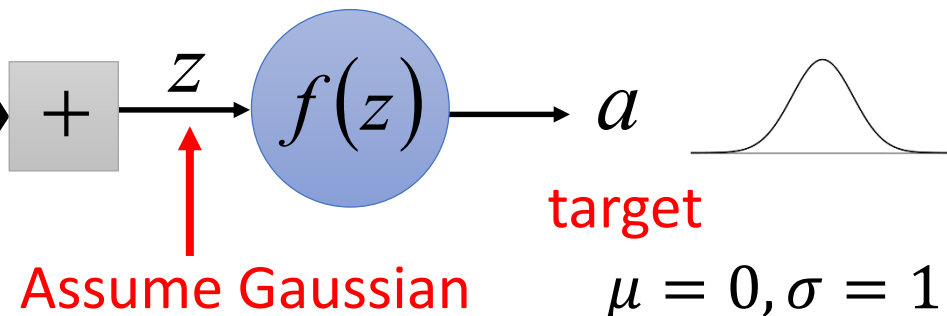
$$= \sum_{k=1}^K (w_k)^2 \sigma^2 = \sigma^2 \cdot K \sigma_w^2 = 1$$

k是hidden layer的寬度

$\sigma^2=1$ $\sigma_w^2=1$

$$E[(a_k w_k)^2] = (w_k)^2 E[(a_k)^2] = (w_k)^2 \sigma^2$$

$$E[a_i a_j w_i w_j] = w_i w_j E[a_i] E[a_j] = 0$$



$$z = a_1 w_1 + \Lambda + a_k w_k + \Lambda + a_K w_K$$

Demo

93 頁的證明

Source of joke:

<https://zhuanlan.zhihu.com/p/27336839>

SELU is actually more general.

$$\begin{aligned} & \left. \frac{2(2x-y)(2x+y)2.911}{(\sqrt{2}\sqrt{x}) \left(\sqrt{\pi \left(\frac{2x+y}{\sqrt{2}\sqrt{x}} \right)^2 + 2.911^2 + \frac{(2.911-1)\sqrt{\pi}(2x+y)}{\sqrt{2}\sqrt{x}}} \right)} \right) \sqrt{\pi} - 0.0003 - \\ & (3x-y) + \left(\frac{(\sqrt{2}\sqrt{2.911})(x-y)(x+y)}{(\sqrt{\pi(x+y)^2 + 2 \cdot 2.911^2 x + (2.911-1)(x+y)\sqrt{\pi}})(\sqrt{2}\sqrt{x})} - \right. \\ & \left. \frac{2(2x-y)(2x+y)(\sqrt{2}\sqrt{2.911})}{(\sqrt{2}\sqrt{x}) \left(\sqrt{\pi(2x+y)^2 + 2 \cdot 2.911^2 x + (2.911-1)(2x+y)\sqrt{\pi}} \right)} \right) \sqrt{\pi} - 0.0003 - \\ & (3x-y) + 2.911 \left(\frac{(x-y)(x+y)}{(2.911-1)(x+y) + \sqrt{(x+y)^2 + \frac{2 \cdot 2.911^2 x}{\pi}}} - \right. \\ & \left. \frac{2(2x-y)(2x+y)}{(2.911-1)(2x+y) + \sqrt{(2x+y)^2 + \frac{2 \cdot 2.911^2 x}{\pi}}} \right) - 0.0003 \geq \\ & (3x-y) + 2.911 \left(\frac{(x-y)(x+y)}{(2.911-1)(x+y) + \sqrt{\left(\frac{2.911^2}{\pi} \right)^2 + (x+y)^2 + \frac{2 \cdot 2.911^2 x}{\pi} + \frac{2 \cdot 2.911^2 x}{\pi}}} - \right. \\ & \left. \frac{2(2x-y)(2x+y)}{(2.911-1)(2x+y) + \sqrt{(2x+y)^2 + \frac{2 \cdot 2.911^2 x}{\pi}}} \right) - 0.0003 - \\ & (3x-y) + 2.911 \left(\frac{(x-y)(x+y)}{(2.911-1)(x+y) + \sqrt{(x+y + \frac{2.911^2}{\pi})^2}} - \right. \\ & \left. \frac{2(2x-y)(2x+y)}{(2.911-1)(2x+y) + \sqrt{(2x+y)^2 + \frac{2 \cdot 2.911^2 x}{\pi}}} \right) - 0.0003 - \end{aligned}$$



Andrej Karpathy ✓

@karpathy

Following

maybe it's all generated by a char-rnn. I suspect we will never know.

RETWEETS

4

LIKES

41



2:54 AM - 10 Jun 2017



5



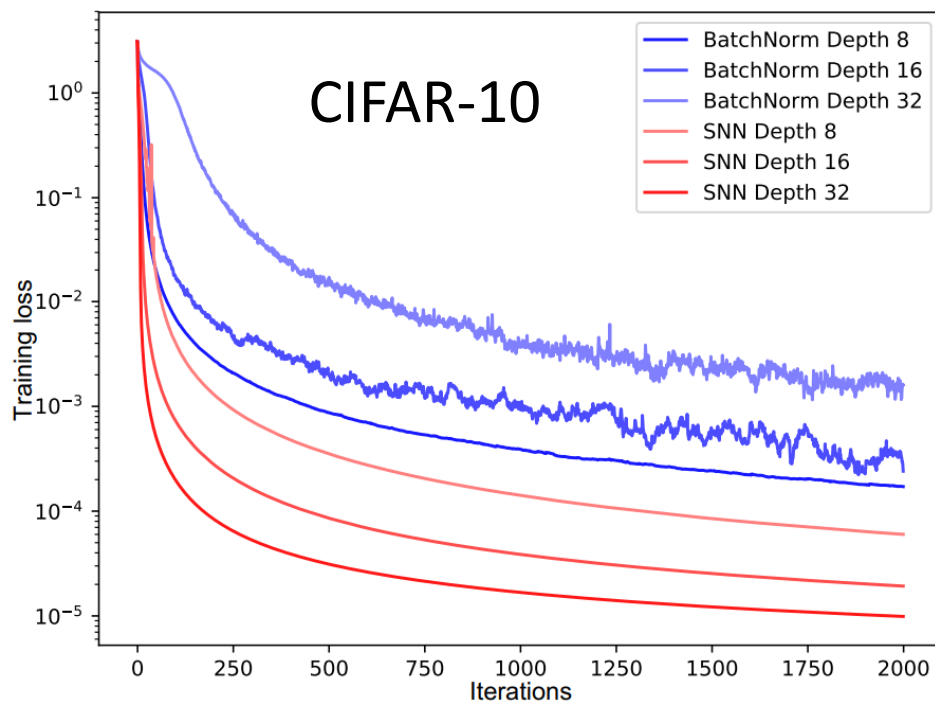
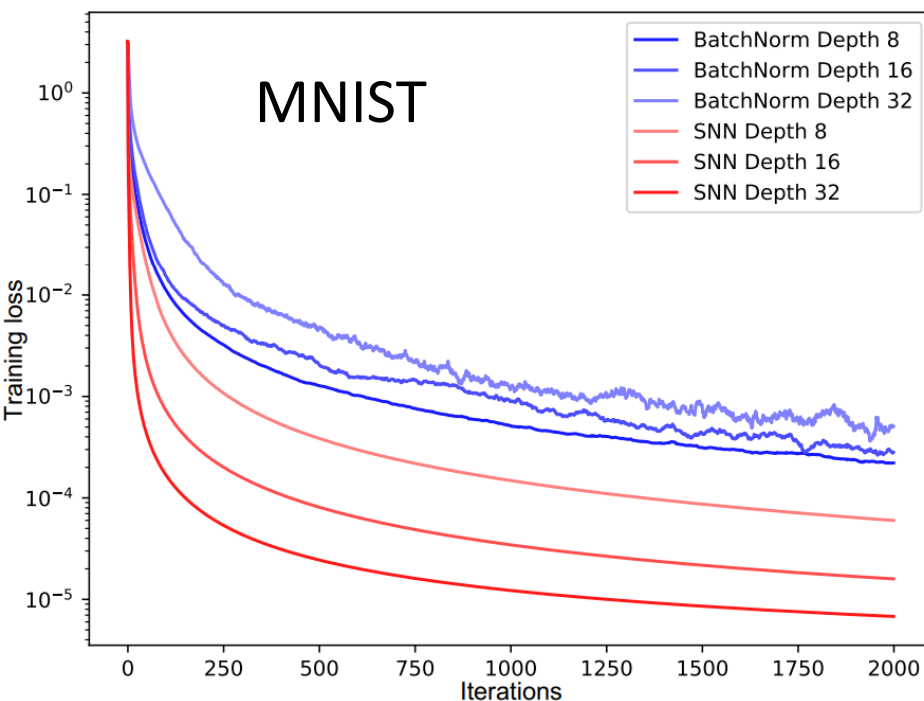
4



41



batch norm上下起伏會比較大，因為每次算出來的mean/variance會不一樣



在一般簡單的case如果用DNN不見得比傳統ML來得強，但是SELU最猛了

FNN method comparison

Method	avg. rank diff.	<i>p</i> -value
SNN	-0.756	
MSRAinit	-0.240*	2.7e-02
LayerNorm	-0.198*	1.5e-02
Highway	0.021*	1.9e-03
ResNet	0.273*	5.4e-04
WeightNorm	0.397*	7.8e-07
BatchNorm	0.504*	3.5e-06

ML method comparison

Method	avg. rank diff.	<i>p</i> -value
SNN	-6.7	
SVM	-6.4	5.8e-01
RandomForest	-5.9	2.1e-01
MSRAinit	-5.4*	4.5e-03
LayerNorm	-5.3	7.1e-02
Highway	-4.6*	1.7e-03
...

Demo