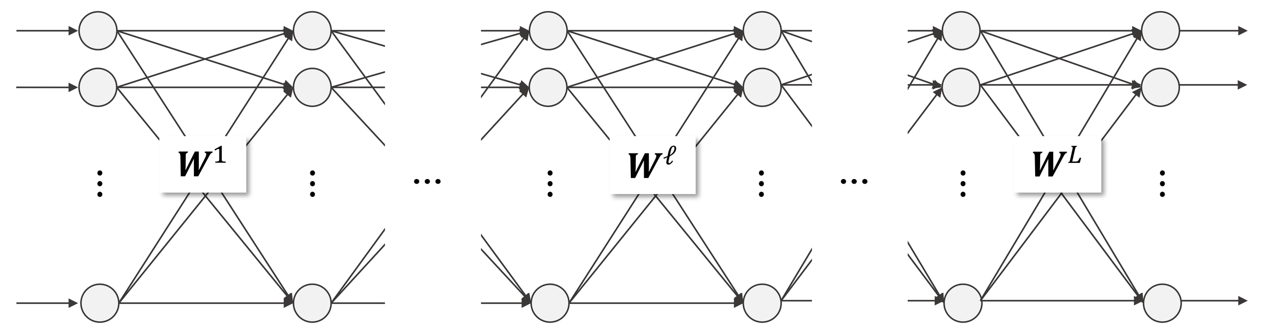
# Layer-wise Interpretation of Deep Neural Networks Using Identity Initialization

Shohei Kubota[1], Hideaki Hayashi[1], Tomohiro Hayase[2], and Seiichi Uchida[1]
[1]Kyushu University, Japan, [2] Fujitsu lab., Japan

ICASSP 2021 TORONTO Canada
June 6-11, 2021
Metro Toronto Convention Centre

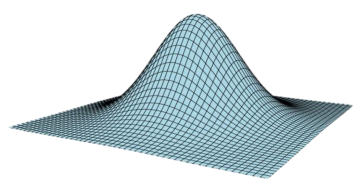
## Motivation

Can identity initialized deep neural networks be trained?




**Standard initialization**

- Random matrix

$W^\ell \sim$ 

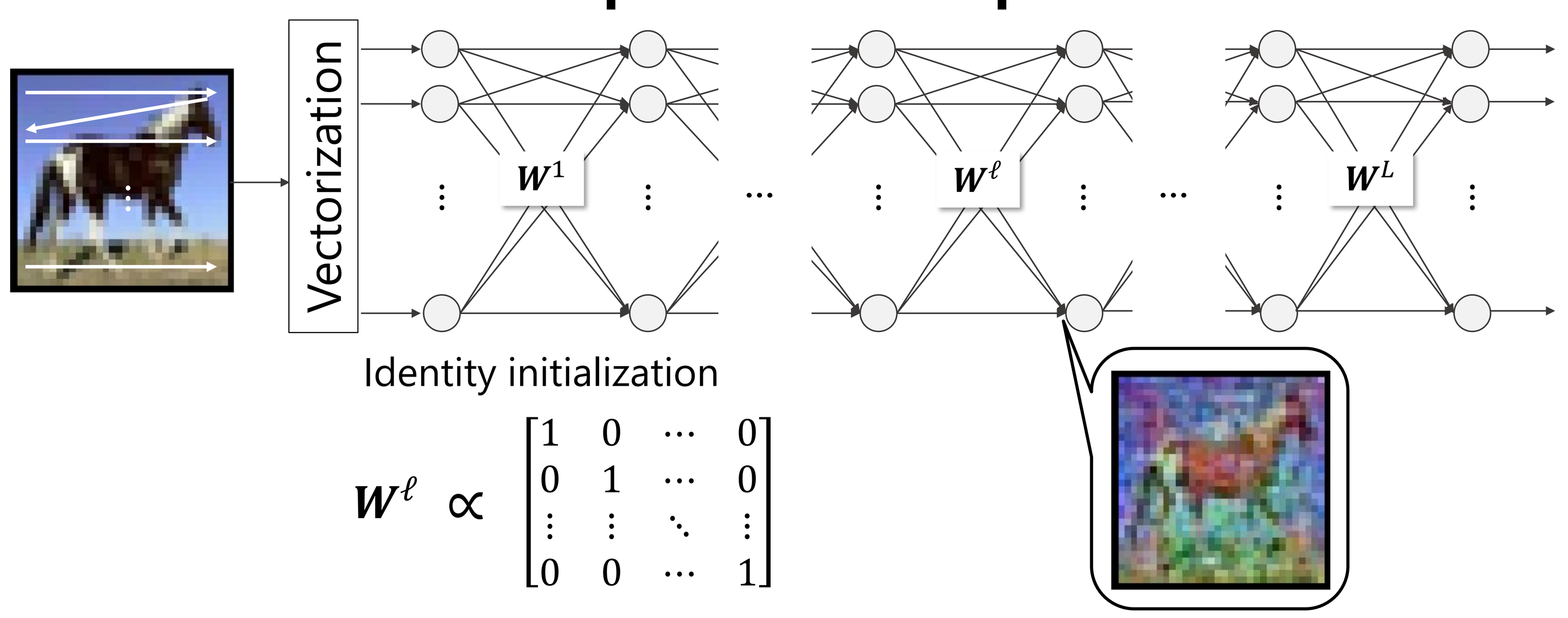
**Initialization in this study**

- Identity matrix

$$W^\ell \propto \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

## Merit | Interpretability

- **Learned weights are close to the identity matrix**
- **Intermediate features preserve the input structure**




Identity initialization

$$W^\ell \propto \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

## Difficulty | Gradient vanishing/exploding

- **Identity-initialized deep neural networks suffer from the gradient vanishing/exploding problems**

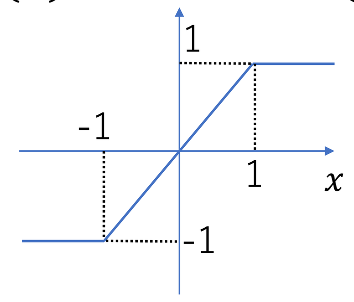
## Contribution

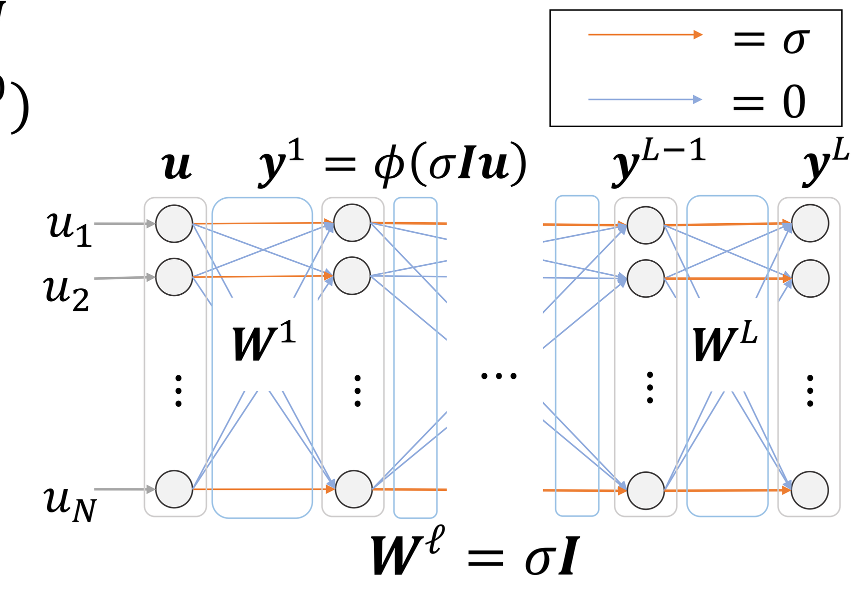
- **Theory:**
  We show a condition under which the identity-initialized deep multilayer perceptron (MLP) prevents gradient vanishing/exploding.
- **Application:**
  We propose an interpretable MLP structure using identity initialization.

## Identity initialization

- **MLP of $L$ layers with width $N$ and $W^\ell \in \mathbb{R}^{N\times N}$**
- **Initialization: $W^\ell = \sigma I$**
- **Input $u$ follows $N(0, q^0)$**
- **Activation $\phi$**

$\phi(x) = \text{hard}\tanh(x)$




## Condition to prevent gradient vanishing

- **Dynamical isometry: Singular values of Jacobian $J$ (or eigenvalues of $J^{\mathrm{T}}J$) concentrate around 1**

$$J = \frac{\partial y^L}{\partial u} = D^1 W^1 D^2 W^2 \cdots D^L W^L$$

Output: $y^L$; Input: $u$; $W^1$: Weight of the first layer; $D^L$: Derivative of the activation function at the $L$-th layer

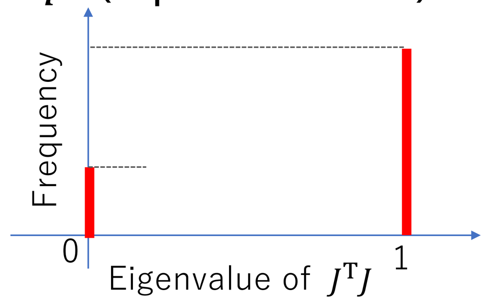
## Theoretical result

- **The eigenvalue distribution is a Bernoulli function**

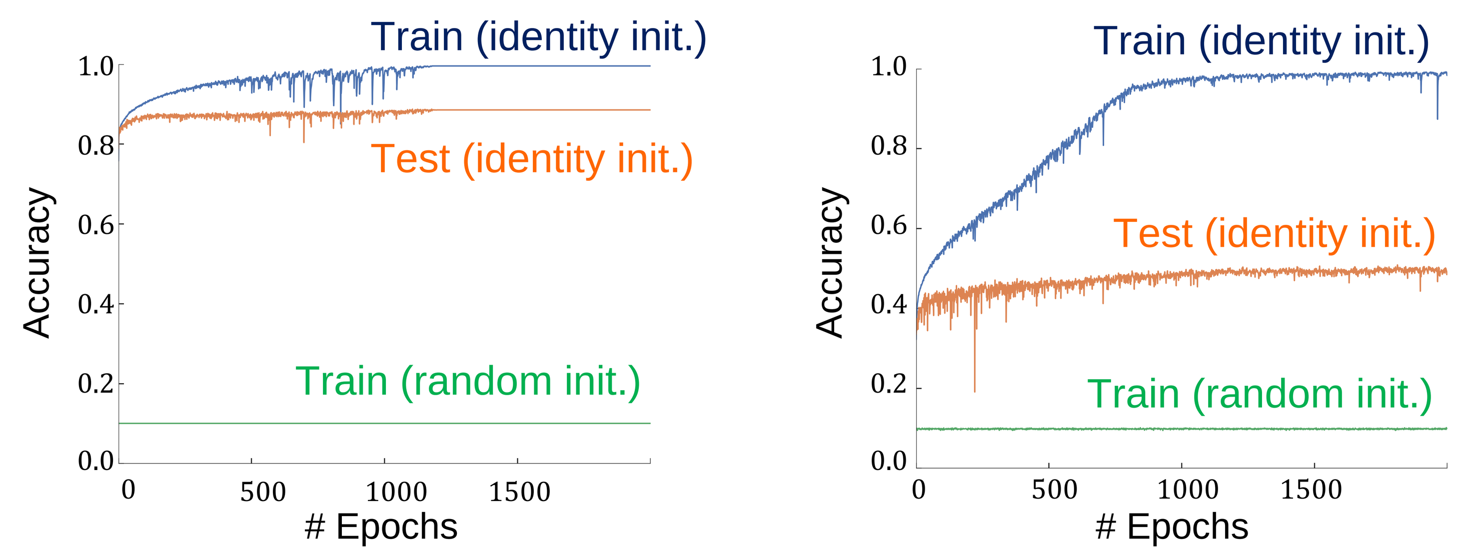
$$\mu_{JJ^{\mathrm{T}}}(\lambda) = \begin{cases} \alpha^L\,\delta(\lambda - \sigma^{2L}) - (1 - \alpha^L)\,\delta(\lambda) & (\sigma > 1) \\ \alpha^1 \delta(\lambda - \sigma^{2L}) - (1 - \alpha^1)\,\delta(\lambda) & (\sigma \leq 1) \end{cases}$$

$\alpha^\ell$: Constant depending on $\sigma$ (constant) and $q^0$ (input variance)

- **We can prevent gradient vanishing by setting $\sigma$ and $q^0$ approprately**

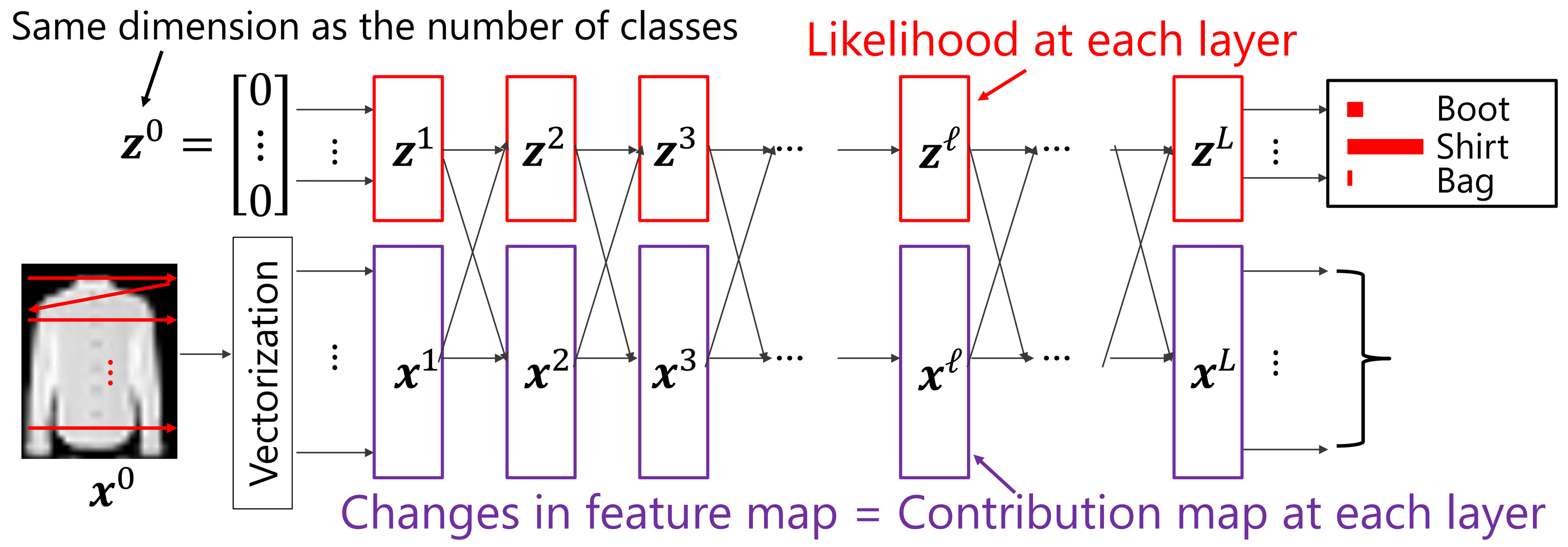



## Experimental Results

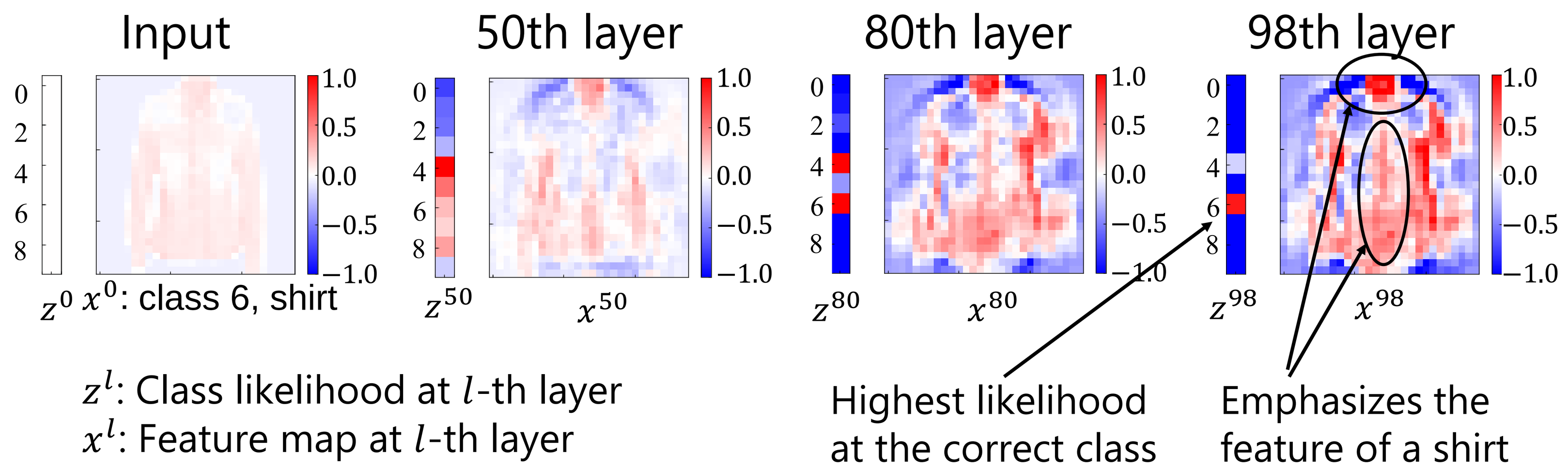



(a) Fashion-MNIST (b) CIFAR-10

## Network structure that enhances interpretability

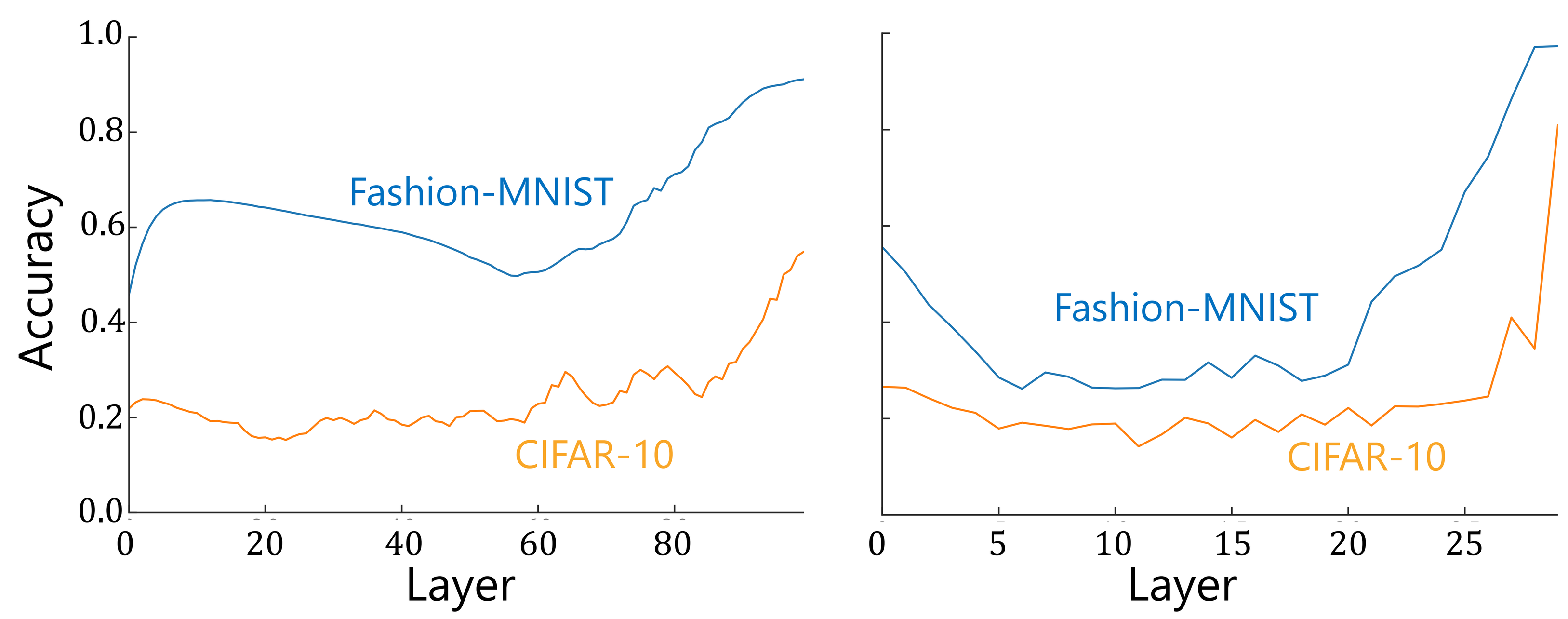



## Changes in feature map




$z^l$: Class likelihood at $l$-th layer
$x^l$: Feature map at $l$-th layer

## Classification accuracy for each layer




(a) 100 layer (b) 30 layer

## Conclusion

- **Investigate the potential of identity initialization**
  - Condition for preventing the vanishing/exploding gradients
  - Network structure enhancing interpretability
- **Future work**
  - Analysis of changes in feature values during learning
  - Application to other structures such as convolutional neural networks