KYUSHU UNIVERSITY

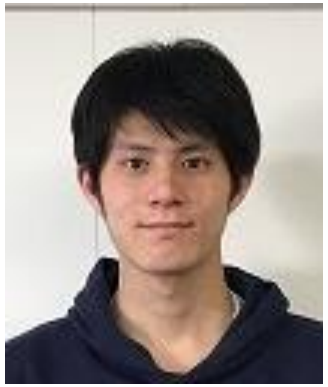
FUJITSU

Advanced Information and Communication Technology for Innovation ACT-i

ACT-X

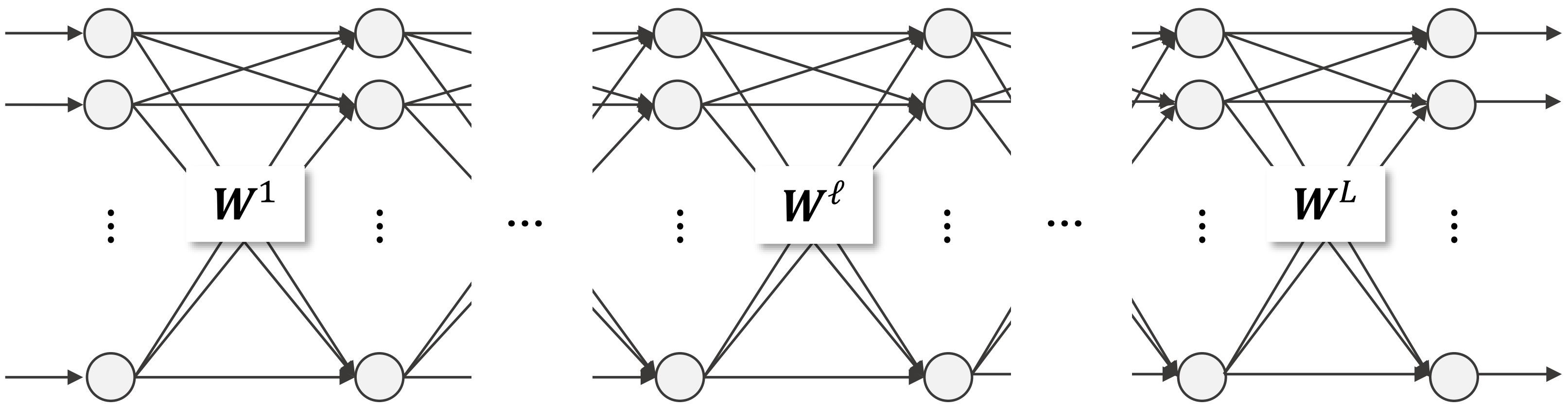
ICASSP 2021 TORONTO Canada
June 6-11, 2021
Metro Toronto Convention Centre

Layer-wise Interpretation of Deep Neural Networks Using Identity Initialization



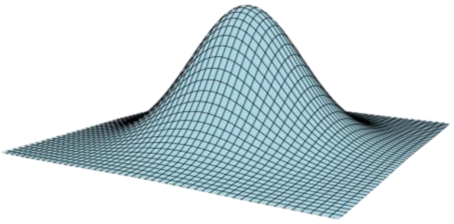

Shohei Kubota (Kyushu Univ.)
Hideaki Hayashi (Kyushu Univ.)
Tomohiro Hayase (Fujitsu Lab.)
Seiichi Uchida (Kyushu Univ.)

Can identity initialized deep neural nets be trained?



Standard initialization

- Random matrix

$W^\ell \sim$ 

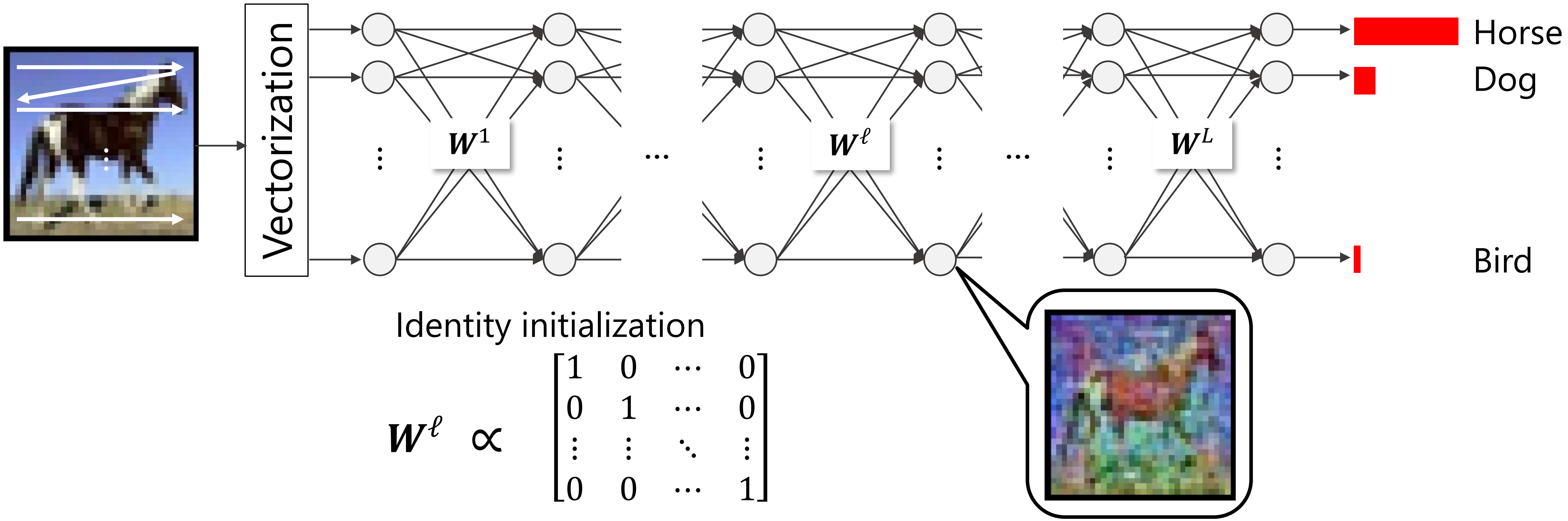
Initialization in this study

- Identity matrix

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

Interpretability of the internal process

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

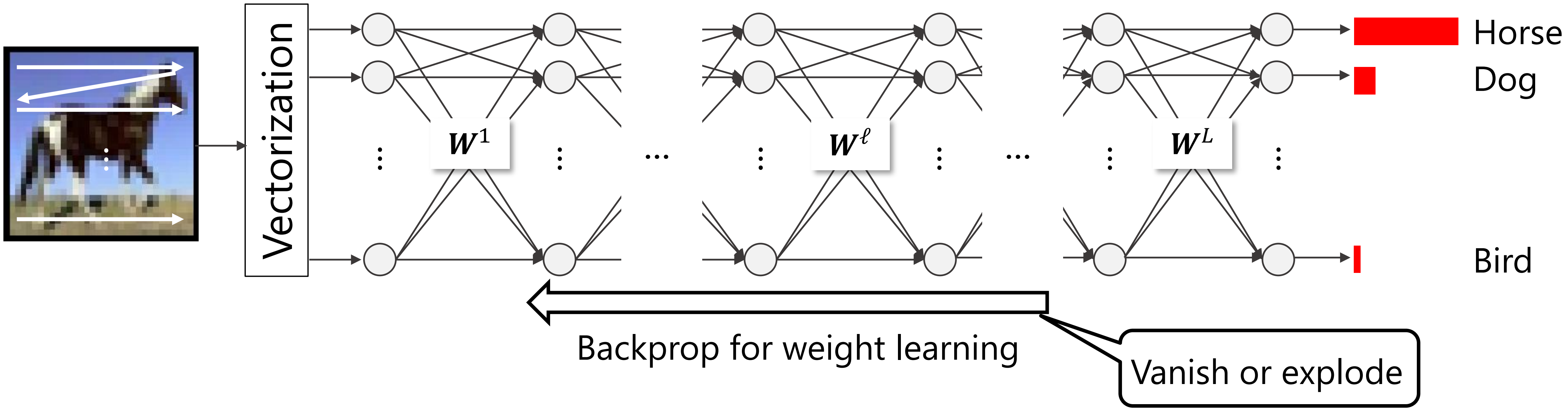


Identity initialization

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

Gradient vanishing and exploding

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



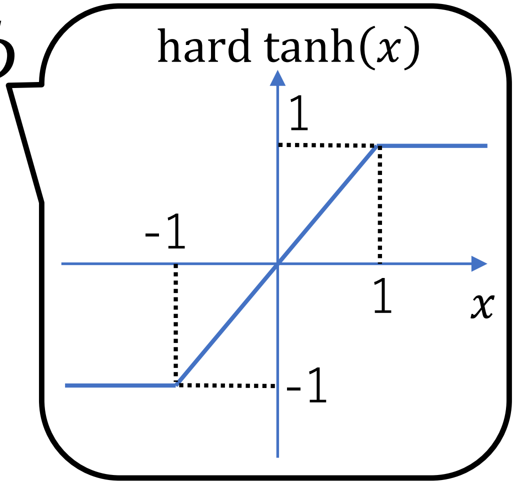
Investigate the potential of identity initialization

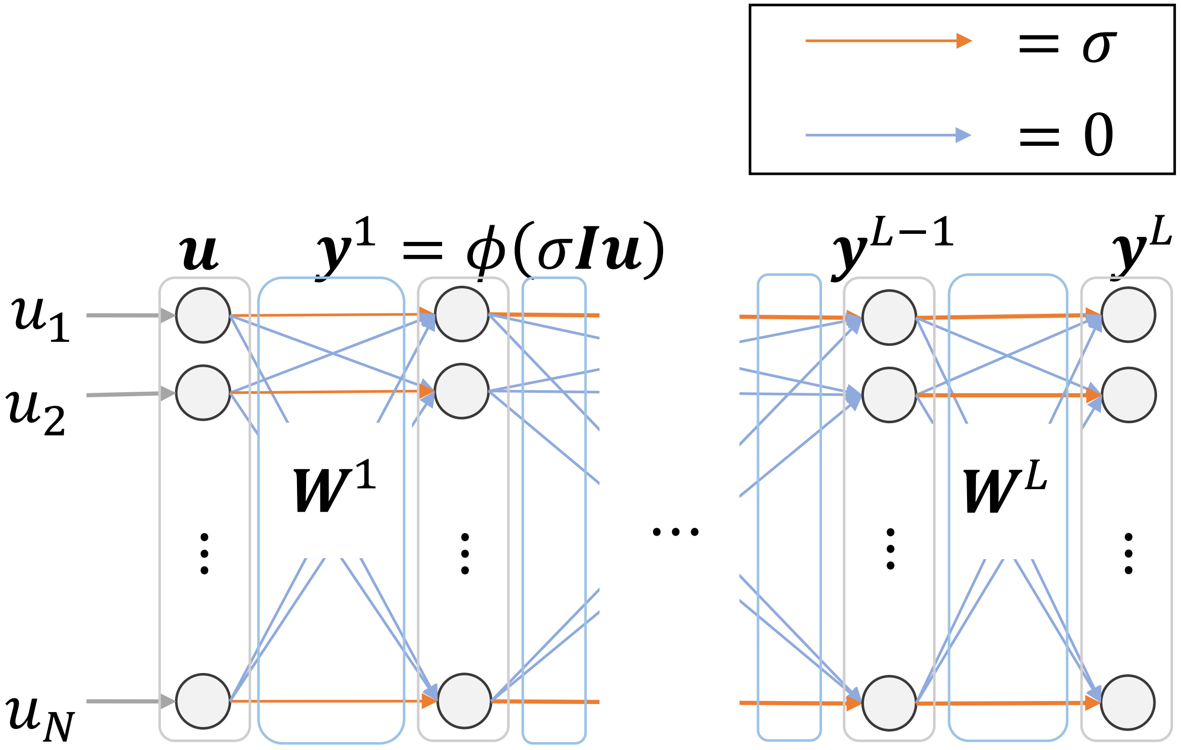
- **Theory:**
  We show a condition under which the identity-initialized deep multilayer perceptron (MLP) prevents gradient vanishing/exploding.

- **Application:**
  We propose an interpretable MLP structure using identity initialization.

Identity initialization

- **MLP of L layers with width N and $\boldsymbol{W}^\ell \in \mathbb{R}^{N\times N}$**
- **Initialization: $\boldsymbol{W}^\ell = \sigma \boldsymbol{I}$**
- **Activation ϕ**





$\boldsymbol{W}^\ell = \sigma \boldsymbol{I}$

- **Input u follows $N(0, q^0)$**

Condition to prevent gradient vanishing/exploding

Dynamical isometry

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

Output

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

Input

Weight of the first layer

Derivative of the activation function at the L-th layer

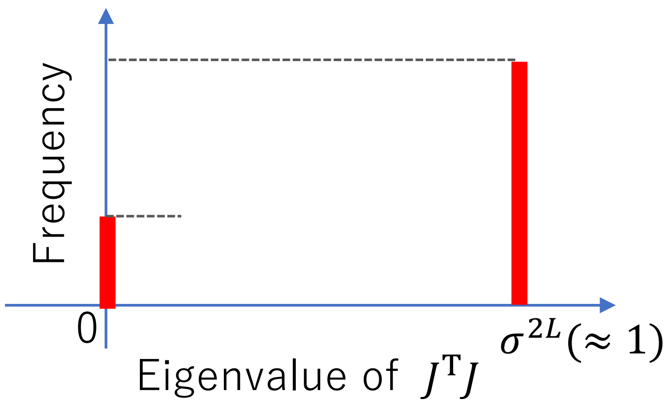
Theoretical result

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

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

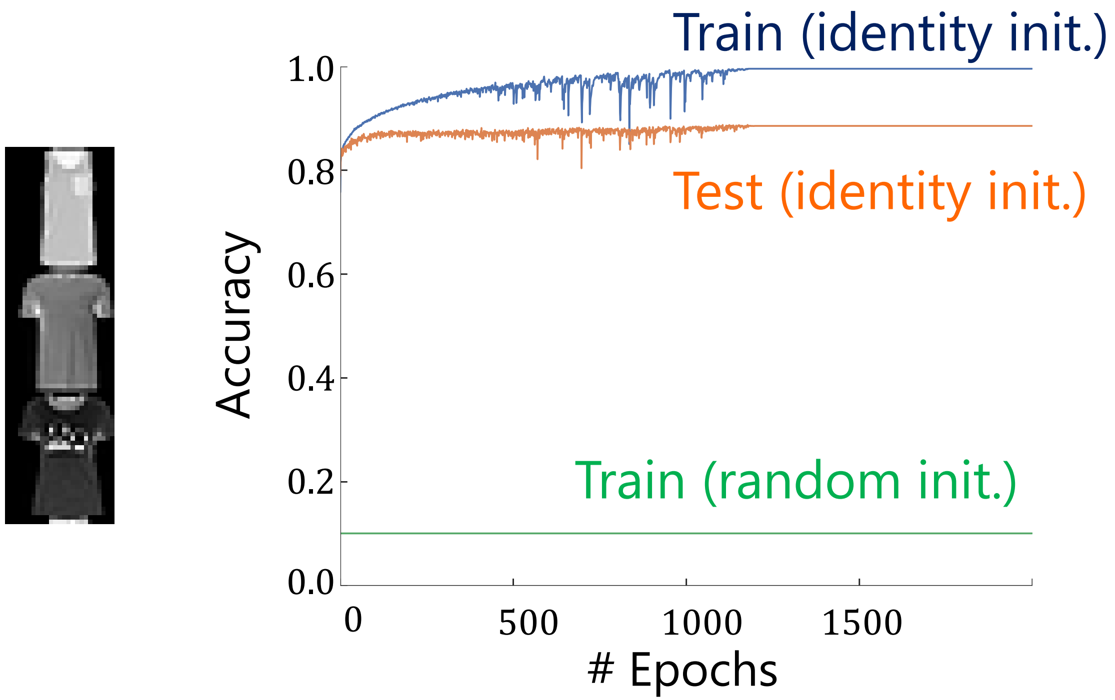
α^{ℓ}: Constant depending on σ (constant) and q^0 (input variance)

- **We can prevent gradient vanishing by setting σ and q^0 appropreately**

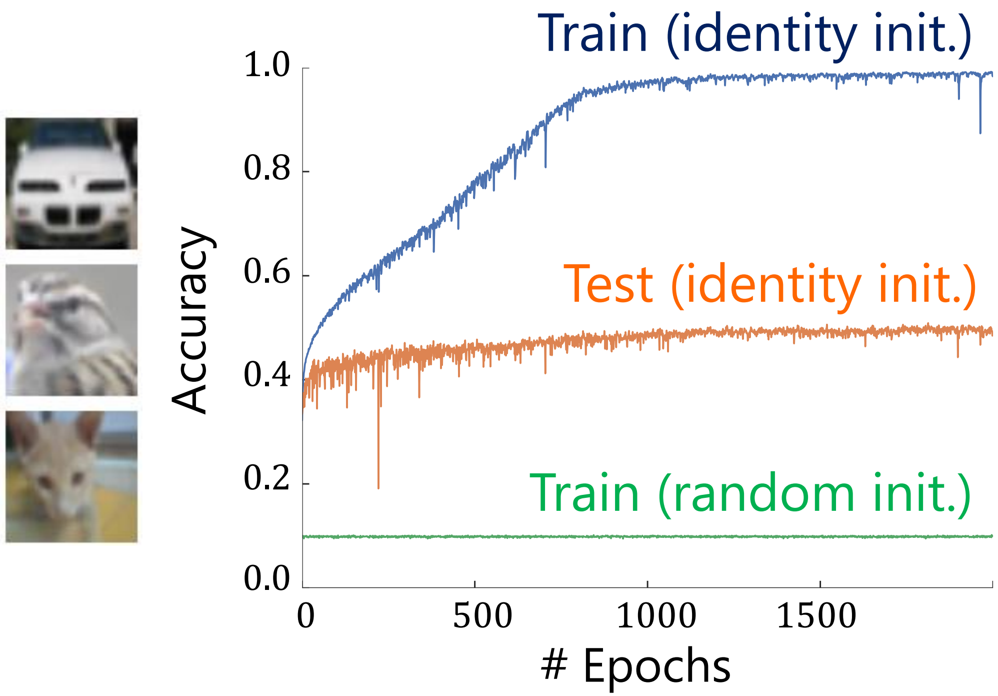


Identity-initialized networks can be trained

- **MLP with 100 layers**
- **Dataset: Fashion-MNIST and CIFAR-10 (10 classes)**



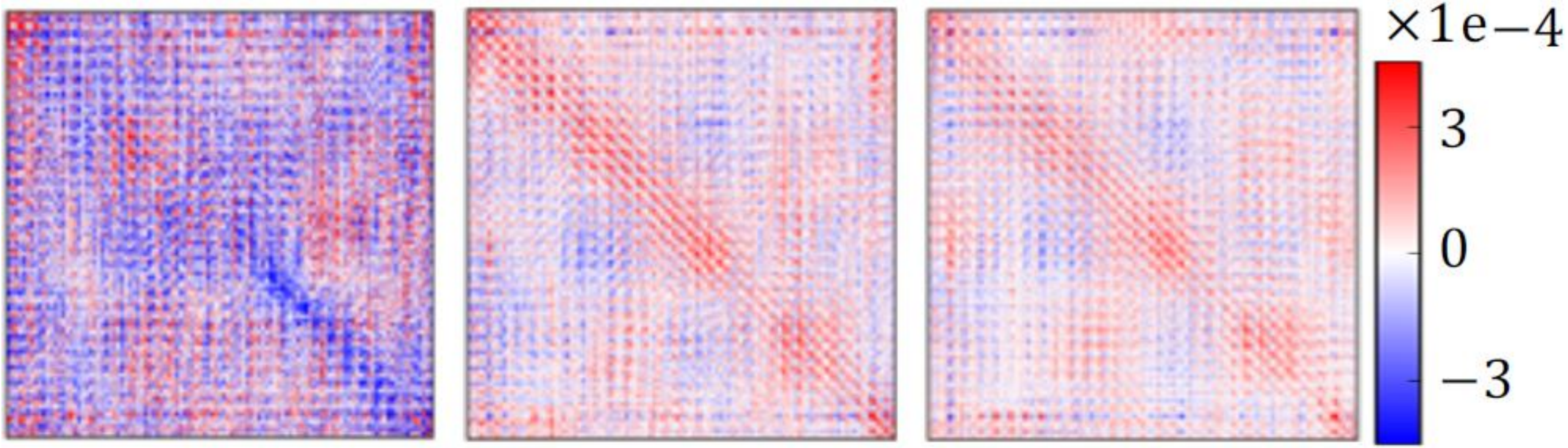
(a) Fashion-MNIST



(b) CIFAR-10

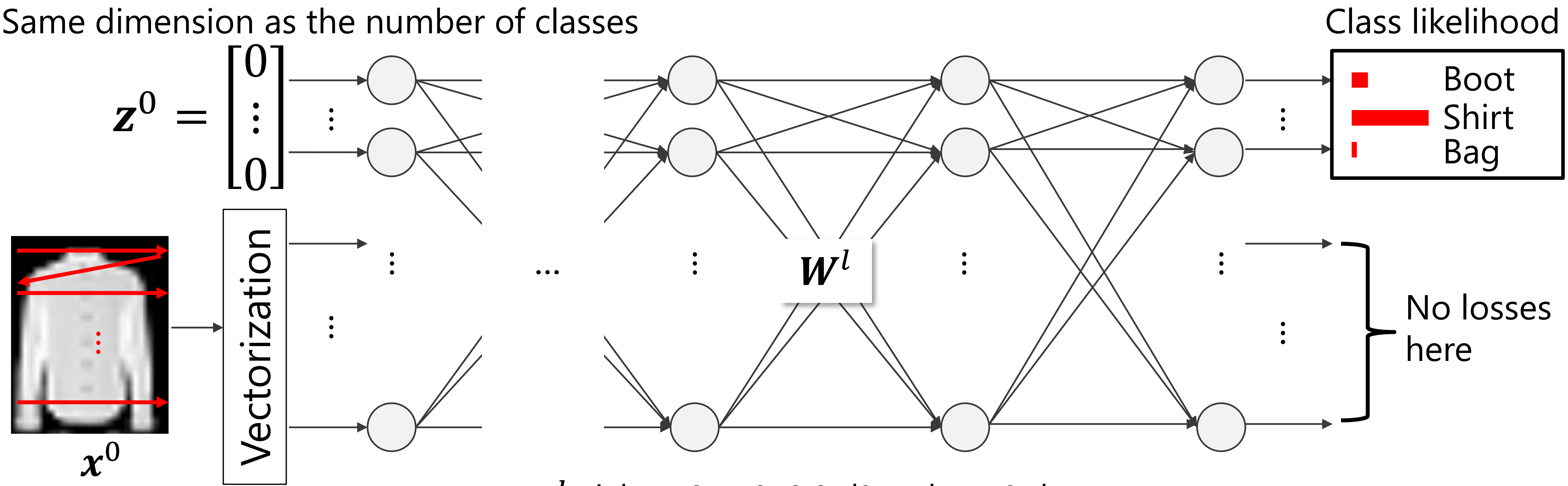
Close to the identity matrix

- **Difference of the weight matrix from the identity matrix after training**



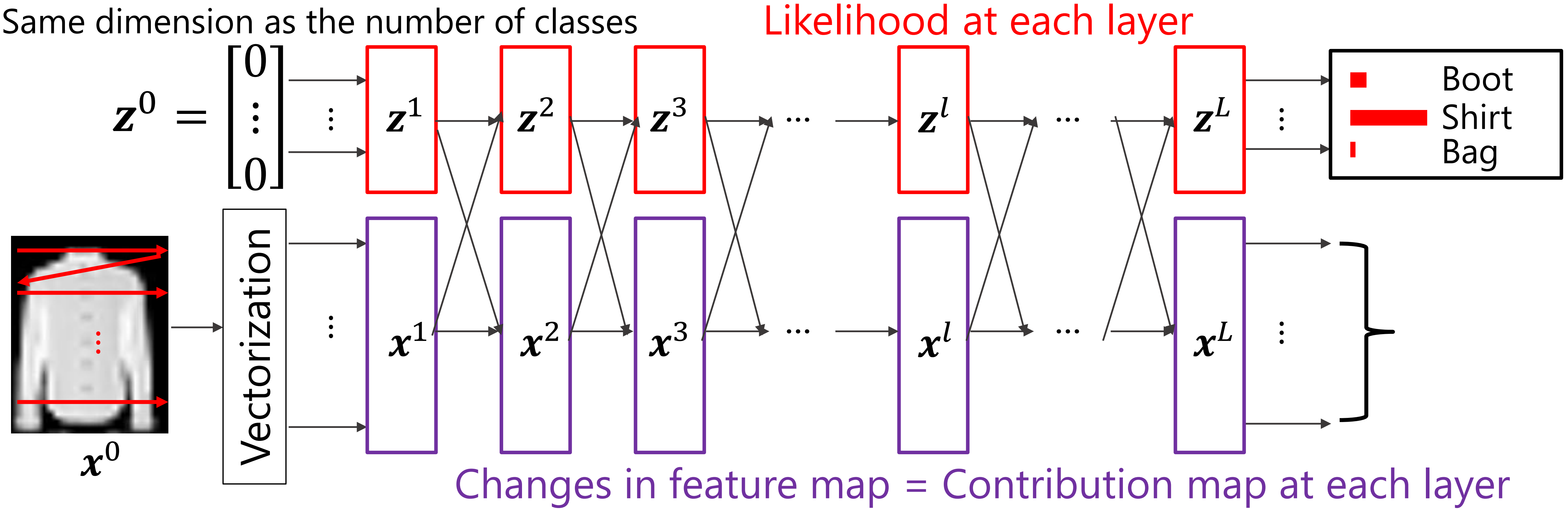
10th layer　　40th layer　　100th layer

Network structure that enhances interpretability

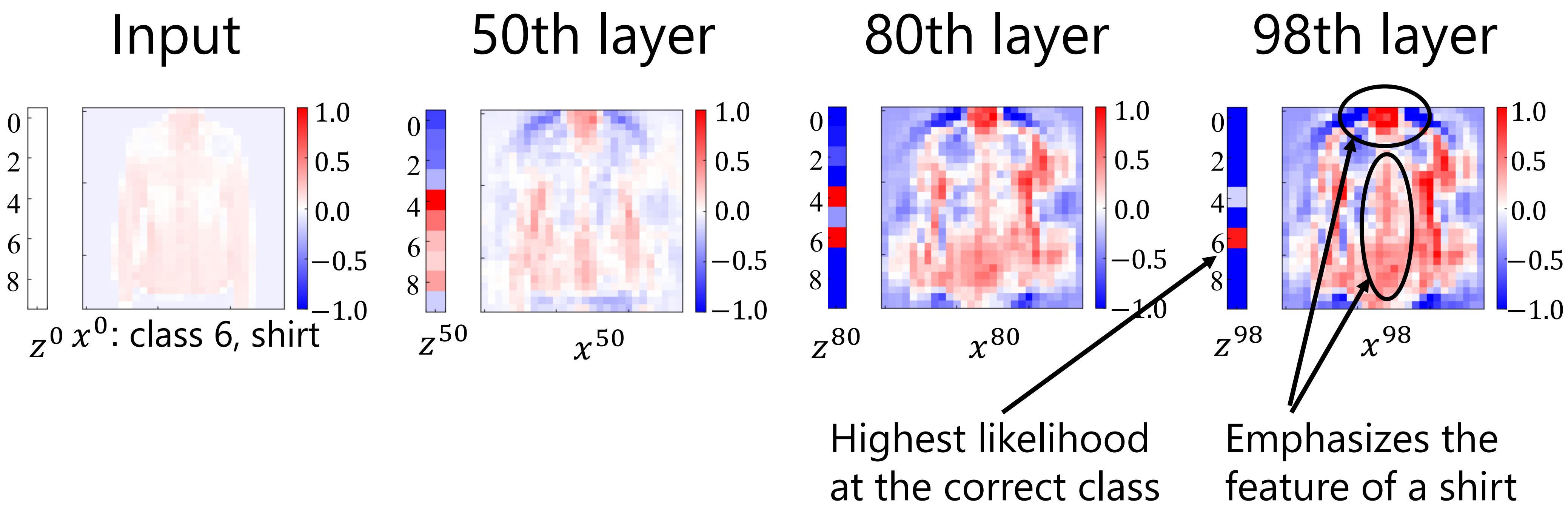


W^l: Identity-initialized weight

Network structure that enhances interpretability



Emphasize the areas important for classification



0: T-shirt
1: Trouser
2: Pullover
3: Dress
4: Coat
5: Sandal
6: Shirt
7: Sneaker
8: Bag
9: Ankle boot

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

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

Layer-wise Interpretation of Deep Neural Networks Using Identity Initialization

Shohei Kubota, Hideaki Hayashi, Tomohiro Hayase, and Seiichi Uchida

Poster Session

MLSP-42: Neural Network Pruning

Friday, 11 June from 11:30 to 12:15 (EDT)