Neuron Campaign for Initialization Guided by Information Bottleneck Theory

Haitao Mao[1,2], Xu Chen[1,3], Qiang Fu[1], Lun Du[1], Shi Han[1], Domei Zhang[1]

1. Microsoft Research Asia
2. University of Electronic Science and Technology of China
3. Peking University

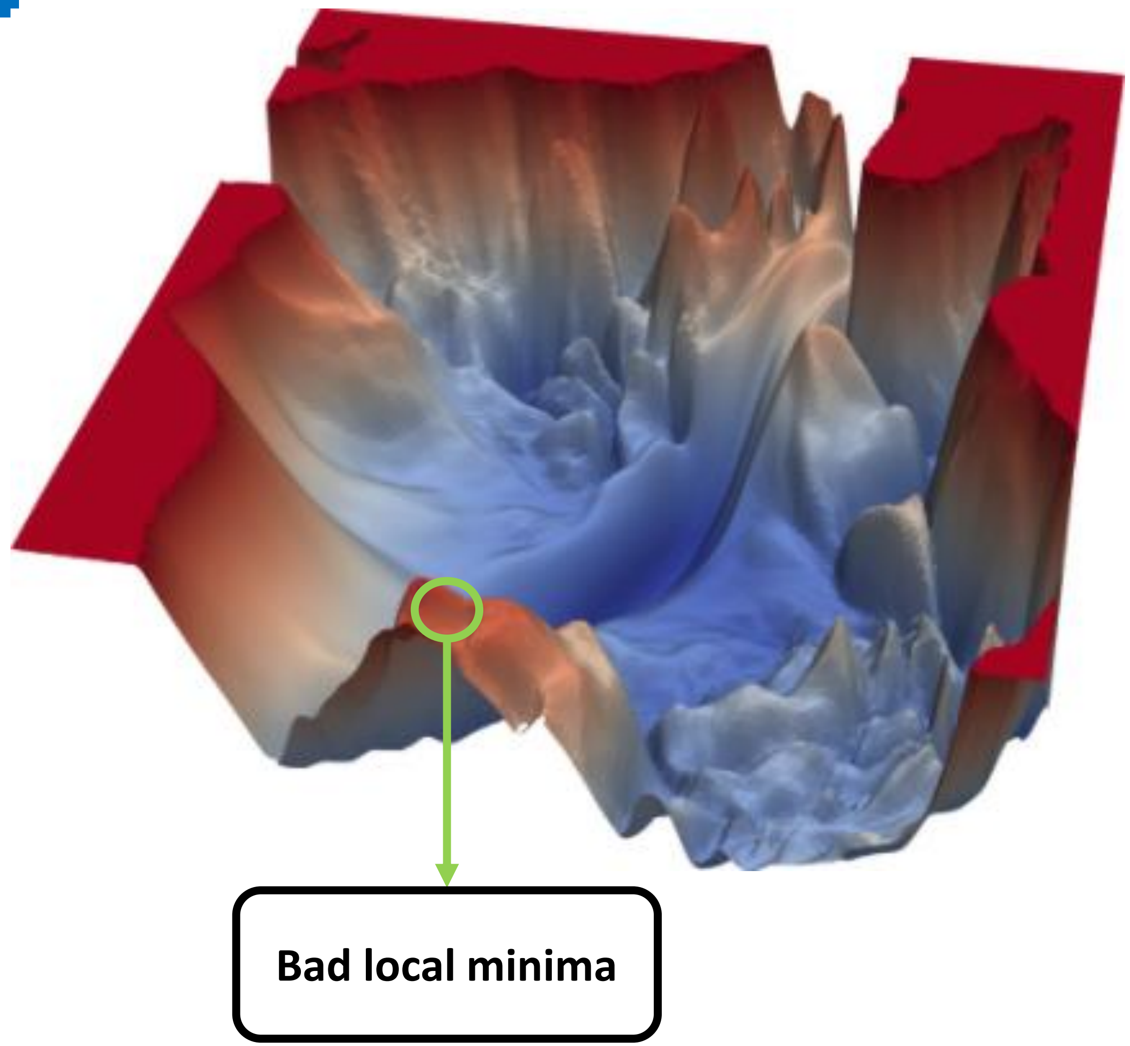
Contents

- Background
- Related work & Limitation
- Approach
- Evaluation
- Conclusion & Future work

Microsoft

Background

Background



- Training a DNN is to find a good local minima.
- A bad initialization may lead to stuck in a bad local minima.

Related work & Limitation

Traditional Initialization strategy

- Random Initialization

$$W \sim N\left(0, 0.01^2\right)$$

- Gradient Exposure and vanish
  - Forward

$$y = W_3 * W_2 * W_1 * x$$

  - Backward

$$\nabla W_1 = \boxed{\frac{\partial Loss}{\partial f_3}} * \boxed{\frac{\partial f_3}{\partial z_3} * \frac{\partial f_2}{\partial z_2} * \frac{\partial f_1}{\partial z_1}} * \boxed{W_3 * W_2} * X$$

Variance scaling Initialization strategy

- Xavier Initialization (for linear and sigmoid activation function)

$$W \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}, \frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}\right]$$

- He initialization (for Relu activation function)

$$W \sim N\left(0, \sqrt{\frac{2}{n_j}}\right)$$

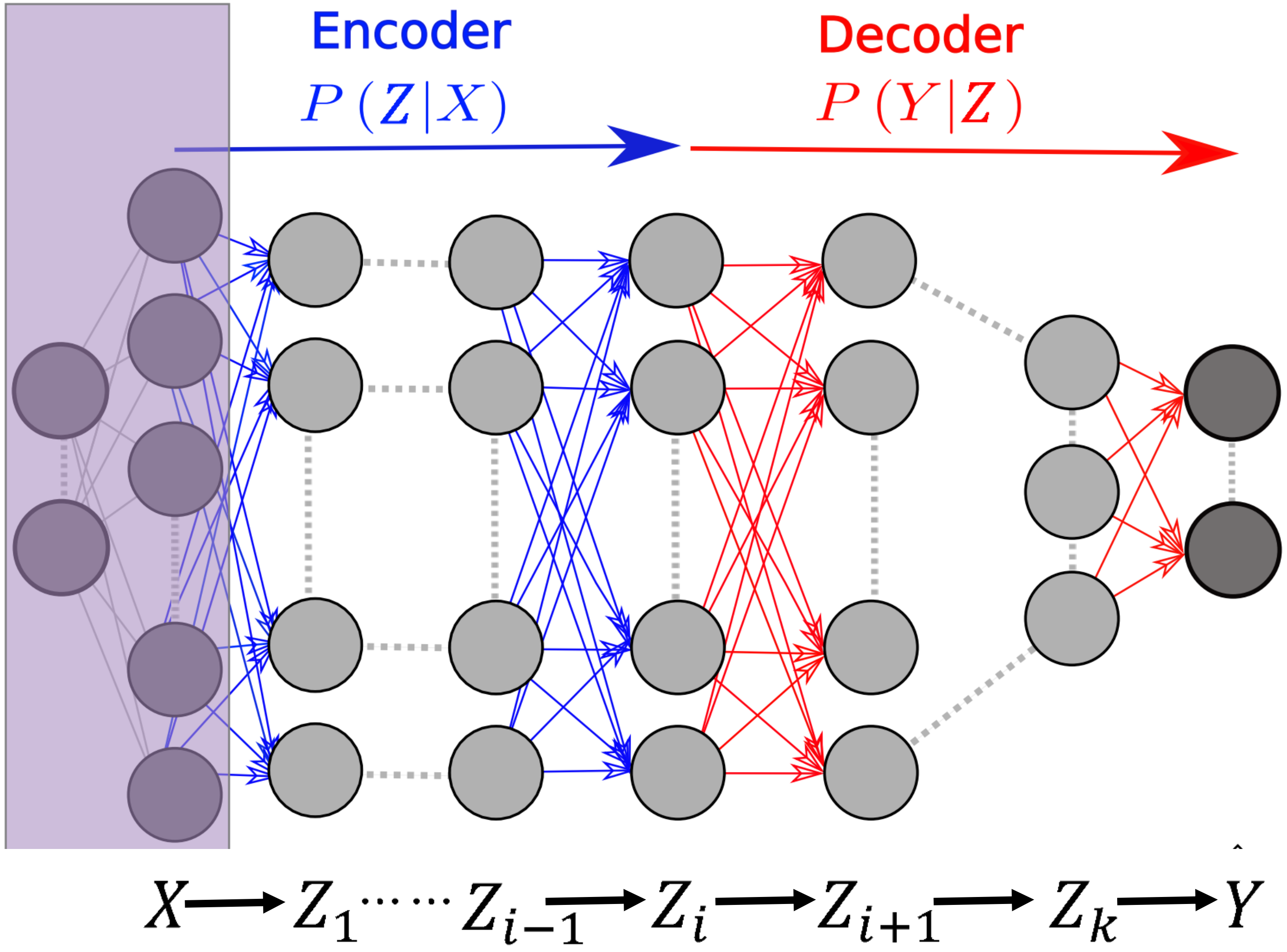
- LSUV (Layer Sequential Unit-Variance initialization)

$$\mathrm{W}_L = \mathrm{W}_L / \sqrt{var(Z_i)}$$

Approach

- **Identify the desire initialization strategy leading to better generalization**
- **Two criteria guided by the Information Bottleneck Theory**
- **Neuron Campaign Initialization algorithm**

Information Bottleneck Theory and measurement



- Input information maintenance:
  $$I(X; Z_i)$$
- Target-related information enhancement
  $$I(Z_i; Y)$$
- Criterion:
  $$\alpha I(X; Z_i) + (1 - \alpha) I(Z_i; Y)$$
- The front layer should focus more on input information maintenance

Criteria Simplification with High Efficiency

- $I(X; Z_i)$ input information maintenance criterion

$$tr(\Sigma_i)$$

  where Σ_i is the covariance matrix of Z_i

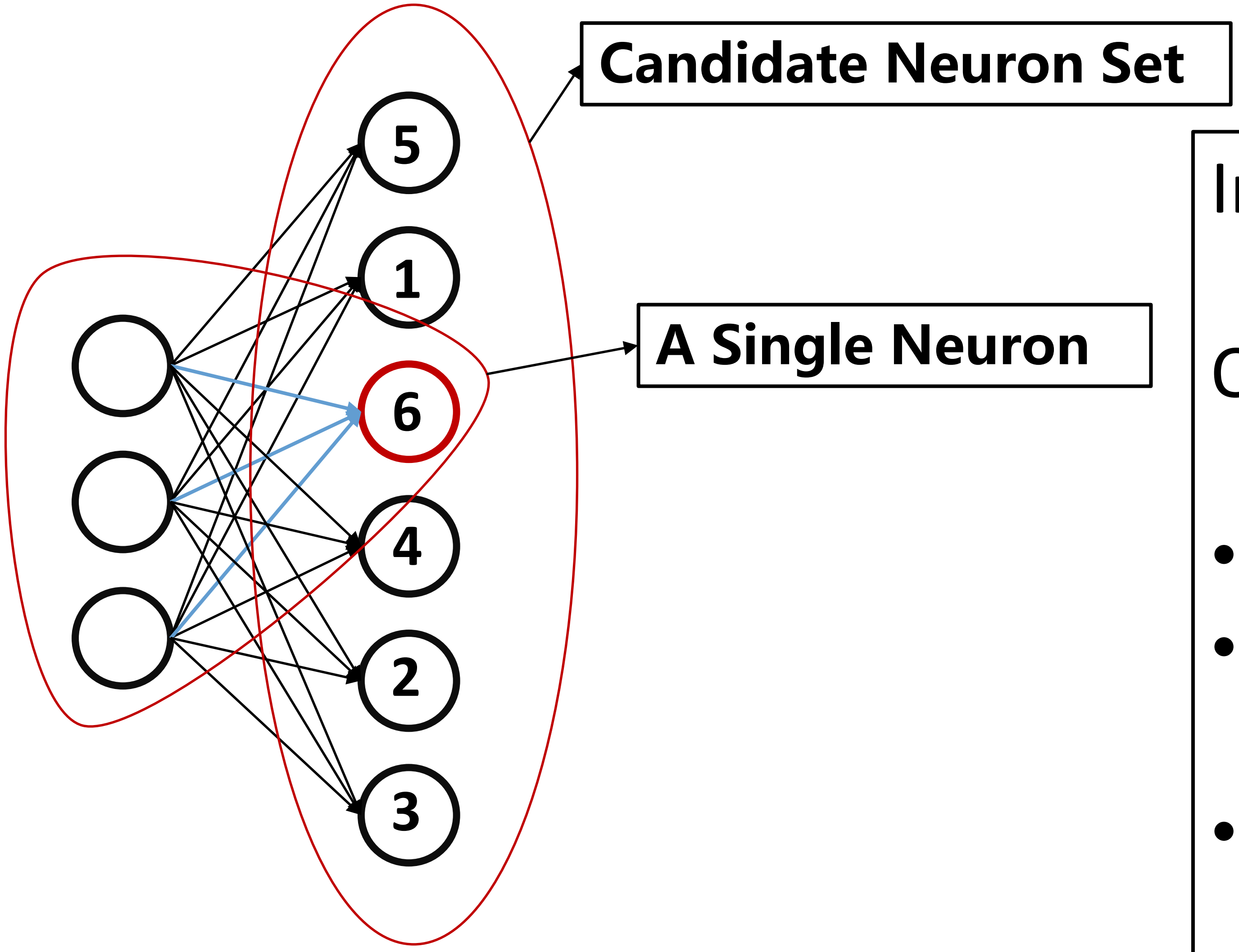
- $I(Z_i; Y)$ target-related maintenance criterion

$$tr\left(\widehat{H}\widehat{H}^T\right) - \frac{1}{N}\sum_{j=0}^{N} tr\left(\hat{Z}_i^T \Pi^j \hat{Z}_i\right)$$

Inter-class variance

Intra-class variance

Neuron Campaign Initialization algorithm



Input: weight with size [3, 6]

Output: weight with size [3, 3]

- Pre-Initialize a large neuron set
- Select neuron with well-designed criteria (based on IB)
- Combine neurons as initial weight

Algorithm details

Algorithm 1 Neuron Campaign initialization algorithm

Input: Candidate weight matrix $\mathbf{W}$

1: **for** t=1 to T **do**

2: Update generalized orthnormalization matrix at t steps: $\mathbf{A}_t = (\mathbf{A}_{t-1}, \mathbf{a}_t^T)^T$

3: Calculate the null space projection by $\mathbf{P}_t = \mathbf{P}_{t-1} - \mathbf{a}_t \mathbf{a}_t^T \mathbf{W}$

4: Select optimal neuron whose index is chosen by $i = \max_i s_i \frac{||\mathbf{p}_t^i||}{||\mathbf{W}_i||}$

5: Update $\mathbf{w}^* = \mathbf{W}_{\cdot,i}$

6: Normalize basis of the generalized orthnormalization matrix as $\mathbf{a}_{t+1} = \mathbf{p}_t^i / ||\mathbf{p}_t^i||$

7: **end for**

Output: Winning neurons formed weight matrix $\mathbf{W}'$

Ensure orthogonality of the selected neurons

Select the neuron with largest score

Evaluation

Experimental details

Table 1: minimal error rate and corresponding epoch comparison of IBCI with baseline methods on MNIST.

Strategy	Layers	Vanilla	LSUV	**IBCI**
Xavier	2	2.04 ± 0.03 (75)	2.05 ± 0.06 (51)	**1.93 ± 0.06 (60)**
	3	1.82 ± 0.05 (52)	1.80 ± 0.07 (63)	**1.71 ± 0.09 (36)**
	5	2.83 ± 0.16 (98)	3.13 ± 0.17 (69)	**2.53 ± 0.09 (78)**
He	2	2.03 ± 0.03 (65)	2.00 ± 0.04 (70)	**1.93 ± 0.07 (57)**
	3	1.83 ± 0.05 (54)	1.86 ± 0.07 (71)	**1.73 ± 0.04 (35)**
	5	2.76 ± 0.07 (80)	2.90 ± 0.12 (77)	**2.62 ± 0.08 (73)**

Hidden layer dimension setting

layers	Hidden Layer Dimension
2	784, 100, 10
3	784, 256, 100, 10
5	784, 32, 32, 32, 32, 10

Ablation Study

Table 2: minimal error rate and corresponding epoch comparison of IBCI with methods with only one criterion.

Strategy	Layers	IBCI	TIE	IIM
Xavier	2	**1.93 ± 0.06 (60)**	2.04 ± 0.07 (58)	2.07 ± 0.09 (84)
	3	**1.71 ± 0.09 (36)**	1.82 ± 0.03 (43)	1.82 ± 0.05 (52)
	5	**2.53 ± 0.09 (78)**	2.68 ± 0.05 (82)	2.57 ± 0.09 (84)
He	2	**1.93 ± 0.07 (57)**	2.07 ± 0.06 (59)	2.034 ± 0.09 (62)
	3	**1.73 ± 0.04 (35)**	1.83 ± 0.07 (42)	1.856 ± 0.05 (55)
	5	**2.62± 0.08 (73)**	2.89 ± 0.11 (74)	2.67 ± 0.12 (86)

Target Information Enhancement, i.e., IBCI without IIM

Input Information Maximization , i.e., IBCI without TIE

Conclusion & Future work

Conclusion & Future Work

Conclusion

- Introduce the Information Bottleneck Theory into practice use.
- Propose a novel and interesting neuron campaign initialization algorithm.

Future work

- Introduce to broader neural network architectures.
- Can we help to understand the recent popular initialization with pretrain?

Reference

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.
- Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 249–256
- Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. arXiv Prepreint arXiv:1703.00810 (2017)
- Dmytro Mishkin and Jiri Matas. 2015. All you need is a good init. 2017 IEEE Conference on Computer Vision and Pattern Recognition.

Thanks & QA