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] Dongmei Zhang[1]

[1] Microsoft Research Asia, China,
[2] University of Electronic Science and Technology of China, China
[3] Peking University, China,
⋆ These authors contribute equally in this work.

1. Introduction

Initialization plays a critical role in the training of deep neural networks (DNN). Existing initialization strategies mainly focus on stabilizing the training process to mitigate gradient vanish/explosion problems. However, these initialization methods are lacking in consideration about how to enhance generalization ability. The **Information Bottleneck (IB) theory** is a well-known understanding framework to provide an explanation about the generalization of DNN. Guided by the insights provided by IB theory, we design two criteria for better initializing DNN. And we further design a neuron campaign initialization algorithm to efficiently select a good initialization for a neural network on a given dataset. The experiments on MNIST dataset show that our method can lead to a better generalization performance with faster convergence.

2. motivation & target

- The purpose of DNN's training procedure: **find a good local minima** near where the loss landscape is flat.
- Effect of initialization strategy: inappropriate initialization may leads to **stuck in a bad local minima**.
- A proper initialization: help to faster convergence to the desired solution with good generalization

3. Challenges

This work try to address the following challenges:

- How to enhance the generalization ability during the initialization phase?
- What is the relationship between model's primary discriminative ability and the generalization ability?
- How to promote primary discriminative ability of an initial model?

4. Information Bottleneck Theory

The information bottleneck (IB) theory provides insights for DNN's generalization ability from: (1) **Input information maintenance**: in the early training phase, DNN increases mutual information between input X and the latent representation Z as $\max I(X, Z)$, while keeps compressing the input information by minimization of $I(X, Z)$. (2) **Target-related information enhancement**: the target-related information keeps growing during the entire training procedure through maximizing the mutual information between the model target Y and latent representation Z as $\max I(Z, Y)$.

7. Conclusion

In this work, we explore the initialization guided with the Information Bottleneck Theory and propose IBCI with efficient neuron campaign algorithm. Two practical criteria, **Input information maintenance** and **Target-related information enhancement** are proposed to enhance the model's primary discrininative ability while initializing. Comparison with widely-used initialization strategies (Xavier and He) demonstrate that our IBCI leads to better generalization ability with faster convergence.

5. Proposed Method: IBCI

1.IB based Initialization Principle

The two principles guided by IB can be formulated as follows:

$$\max_{W} \sum_{d=1}^{D} \alpha_i I(\mathbf{X}; \mathbf{Z}_d) + (1 - \alpha_i) I(\mathbf{Z}_d; \mathbf{Y}) \qquad (1)$$
$$\text{s.t.} \quad \forall j, \ \left\| \mathbf{W}^{(i)}_{\cdot,j} \right\|_2 < \varepsilon, \quad \mathbf{W} = \{\mathbf{W^{(i)}} | i = 1, 2, \ldots, D\},$$

where α_i controlling the importance between input information maintenance and target-related information enhancement.

2. Mutual Information Simplification

The input information maintenance criterion can be simplified as

$$tr(\Sigma_i), \qquad (2)$$

where Σ_i is the covariance matrix of Z_i.

The target-related maintenance criterion is approximated as:

$$tr\left(\hat{\mathbf{H}}\hat{\mathbf{H}}^T\right) - \frac{1}{N}\sum_{j=0}^{N} tr\left(\hat{\mathbf{Z}}_i^T \Pi^j \hat{\mathbf{Z}}_i\right) \qquad (3)$$

More details of the notations can be found in our paper. Intuitively, the first term denotes the inter-class variance and the second term represents the intra-class variance.

3. Neuron Campaign

We utilize a greedy algorithm to find the optimal neuron subset from a large set of randomly initialized neurons and then integrate them as the initialization weight.

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 $\mathbf{W}'$

6. Experiments

We conduct experiments on MNIST compared with widely-used initialization strategies, Xavier, He and LSUV. Three different MLPs with ReLU as the activation function are selected as the basic model. The results are organized on Table 1. We also study the effect of **T**arget **I**nformation **E**nhancement (TIE) and **I**nput **I**nformation. **M**aintenance (IIE) on Table 2.

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)**

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

Strategy	Layers		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)

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