Towards Understanding the Importance of Shortcut Connections in Residual Networks

Tianyi Liu[1], Minshen Chen[1], Mo Zhou[2], Simon S. Du[3], Enlu Zhou[1], Tuo Zhao[1]
[1] Georgia Institute of Technology [2] Duke University
[3] Institute for Advanced Study

Background

Success of Deep Neural Networks (DNNs):

- Speech and image recognition;
- Nature Language Processing;
- Recommendation Systems.

Among different types of networks, ResNet is a Milestone!

- Shortcut connections: skip layers in the forward step of an input.
- Success over CNNs: He et al.(2016a), He et al.(2016b), Srivastava et al.(2015), Huang et al.(2017).
- Our Empirical Observation:

# of Layers	≤ 30	≥ 30
CNN	Good	Bad
RNN	Good	Good

- **Shortcut connections helps training.**

Existing Results:

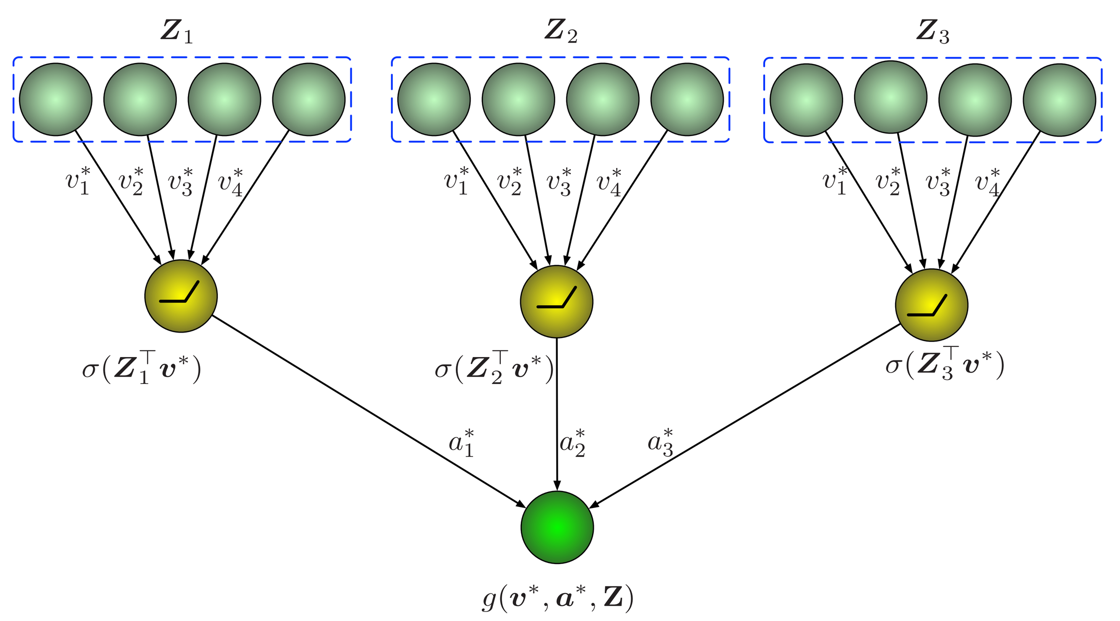
- Empirical: Veit et al. (2016), Balduzzi et al. (2017), Li et al. (2018).
- Hardt and Ma (2016): Linear ResNet has no spurious optima.
- Li and Yuan (2017): Two-layer ResNet with only one unknown layer has no spurious local optima and saddle points.

Question: How does the Shortcut Connection help training in the presence of bad optima?

- We Study: Two-Layer Nonoverlapping Convolutional NNs:
  1. A non-trivial spurious local optimum;
  2. GD gets trapped with constant probability ($\frac{1}{4} \sim \frac{3}{4}$);

A non-trival example provides new insights!

Teacher Network



- Two-layer Nonoverlapping CNNs:

$$f(\boldsymbol{w}^*, \boldsymbol{a}^*, \boldsymbol{Z}) = \sum_{j=1}^{k} a_j^* \sigma(\boldsymbol{Z}_j^\top \boldsymbol{w}^*),$$

– $\|\boldsymbol{w}^*\|_2 = 1$, $\boldsymbol{w} \in \mathbb{R}^p$, $\boldsymbol{a} \in \mathbb{R}^k$, $\sigma(\cdot) = \max\{\cdot, 0\}$.
– $\mathbf{Z} = [\boldsymbol{Z}_1, ..., \boldsymbol{Z}_k]$ with $\boldsymbol{Z}_j$'s i.i.d. $N(\mathbf{0}, \mathbf{I})$,

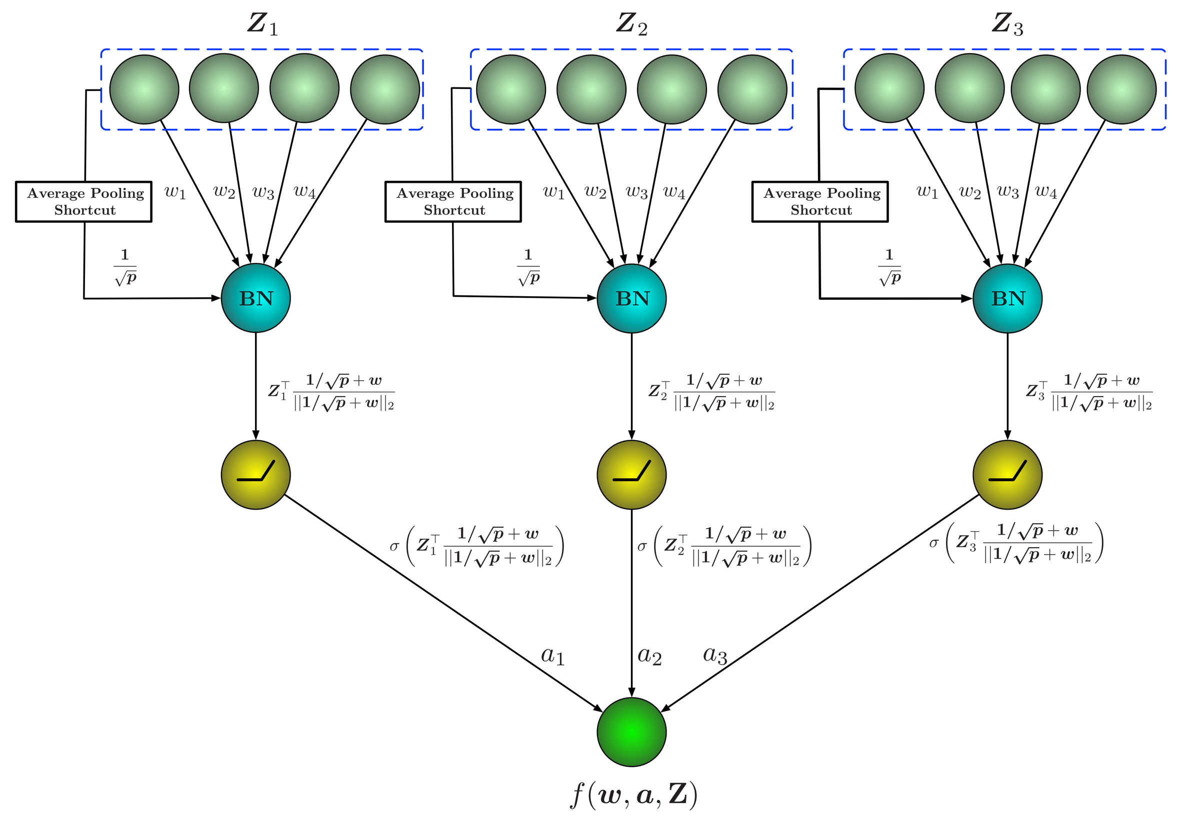
Student Network

- Student Network with shortcut connection:

$$h'(\boldsymbol{w}, \boldsymbol{a}, \boldsymbol{Z}) = \sum_{j=1}^{k} a_j \sigma\left(\boldsymbol{Z}_j^\top \left(\frac{\mathbb{1}}{\sqrt{p}} + w\right)\right)$$

- Normalization to achieve identifiability:

$$h(\boldsymbol{w}, \boldsymbol{a}, \boldsymbol{Z}) = \sum_{j=1}^{k} a_j \sigma\left(\boldsymbol{Z}_j^\top \frac{\mathbb{1}/\sqrt{p} + \boldsymbol{w}}{\|\mathbb{1}/\sqrt{p} + \boldsymbol{w}\|_2}\right).$$



- Nonconvex Optimization:

$$(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{a}}) = \underset{\boldsymbol{w}, \boldsymbol{a}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{w}, \boldsymbol{a}),$$

where $\mathcal{L}(\boldsymbol{w}, \boldsymbol{a}) = \mathbb{E}_{\mathbf{Z}}(f(\boldsymbol{v}^*, \boldsymbol{a}^*, \mathbf{Z}) - h(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}))^2$.

- $(\boldsymbol{w}, \boldsymbol{a})$ is a global optimum , if

$$\mathbb{1}/\sqrt{p} + \boldsymbol{w} = \alpha \boldsymbol{v}^* \text{ and } \boldsymbol{a} = \boldsymbol{a}^*.$$

- $(\boldsymbol{w}, \boldsymbol{a})$ is a **spurious local optimum**, if

$$\overline{\boldsymbol{w}} = -\boldsymbol{w}^*,\ \overline{\boldsymbol{a}} = (\mathbf{1}\mathbf{1}^\top + (\pi - 1)\mathbf{I})^{-1}(\mathbf{1}\mathbf{1}^\top - \mathbf{I})\boldsymbol{a}^*.$$

Gradient Descent with Normalization

- Initialization: $\boldsymbol{a}_0 \in \mathbb{B}_0\left(|\mathbf{1}^\top \boldsymbol{a}^*|/\sqrt{k}\right)$ and $\boldsymbol{w}_0 = \mathbf{0}$.
- At the t-th iteration, we update $\boldsymbol{w}$ and $\boldsymbol{a}$ by

$$\widetilde{\boldsymbol{w}}_{t+1} = \boldsymbol{w}_t - \eta_{\boldsymbol{w}} \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{a}_t),$$
$$\boldsymbol{w}_{t+1} = \frac{\mathbb{1}/\sqrt{p} + \widetilde{\boldsymbol{w}}_{t+1}}{\|\mathbb{1}/\sqrt{p} + \widetilde{\boldsymbol{w}}_{t+1}\|_2} - \frac{\mathbb{1}}{\sqrt{p}},$$
$$\boldsymbol{a}_{t+1} = \boldsymbol{a}_t - \eta_{\boldsymbol{a}} \nabla_{\boldsymbol{a}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{a}_t).$$

where $\mathcal{L}(\boldsymbol{w}, \boldsymbol{a}) = \mathbb{E}_{\mathbf{Z}}(f(\boldsymbol{v}^*, \boldsymbol{a}^*, \mathbf{Z}) - h(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}))^2$.

- Normalization ensures

$$\mathrm{Var}\left(\boldsymbol{Z}_j^\top(\mathbb{1}/\sqrt{p} + \boldsymbol{w}_{t+1})\right) = 1,$$

$\Longleftrightarrow$ a population version of the batch normalization.

Skip-Layer Prior

Assumption. *There exists a w^* with $\|w^*\|_2 \leq 1$, such that $v^* = w^* + \mathbb{1}/\sqrt{p}$.*

- Supported by Existing Results:
  - Li et al. (2016) and Yu et al. (2018): The weight has a small and vanishing magnitude.
  - Hardt and Ma (2016): For linear ResNet, the norm of the weight in each layer scales as $O(1/D)$ with D being the depth.
  - Bartlett et al. (2018): The norm of the weight of order $O(\log D/D)$ is sufficient to express differentiable functions.

Convergence Analysis

Partial Dissipativity Condition: Given any $\delta \geq 0$ and $c \geq 0$,

$$C1: \ \langle -\nabla_w \mathcal{L}(w, a), w^* - w\rangle \geq c\|w - w^*\|_2^2 - \delta;$$
$$C2: \ \langle -\nabla_a \mathcal{L}(w, a), a^* - a\rangle \geq c\|a - a^*\|_2^2 - \delta;$$

- **Stage I: Avoid the spurious local optimum:**
  - C1 holds $\Longleftrightarrow$ Improvement of a.
  - C2 does not hold, but w will not move far away!

Theorem 1. *Initialize with arbitrary $a_0 \in \mathbb{B}_0\left(\frac{|\mathbb{1}^\top a^*|}{\sqrt{k}}\right)$ and $w_0 = 0$. We choose step sizes*

$$\eta_a = \frac{\pi}{20(k + \pi - 1)^2} = O\left(\frac{1}{k^2}\right),\ \eta_w = C\|a^*\|_2^2 \eta_a^2 = \widetilde{O}(\eta_a^2)$$

for some constant $C > 0$. Then, we have

$$\phi_t \leq \frac{5\pi}{12} \quad \text{and} \quad 0 \leq m \leq a_t^\top a^* \leq M, \qquad (1)$$

for all $t \in [T_1, T]$, where $0 < m < M$ are some constants and

$$T_1 = \widetilde{O}\left(\frac{1}{\eta_a}\right),\quad T = O\left(\frac{1}{\eta_a^2}\right).$$

- **Stage II: Converging to Global Optima:**
  - C1, C2 jointly hold $\Longleftrightarrow$ Convergence!

Theorem 2. *Given the output (1), for any $\delta > 0$, choose*

$$\eta_a = \eta_w = \eta = \min\left\{\frac{m}{2M^2}, \frac{5\pi^2}{4(k + \pi - 1)^2}\right\} = \widetilde{O}\left(\frac{1}{k^2}\right),$$

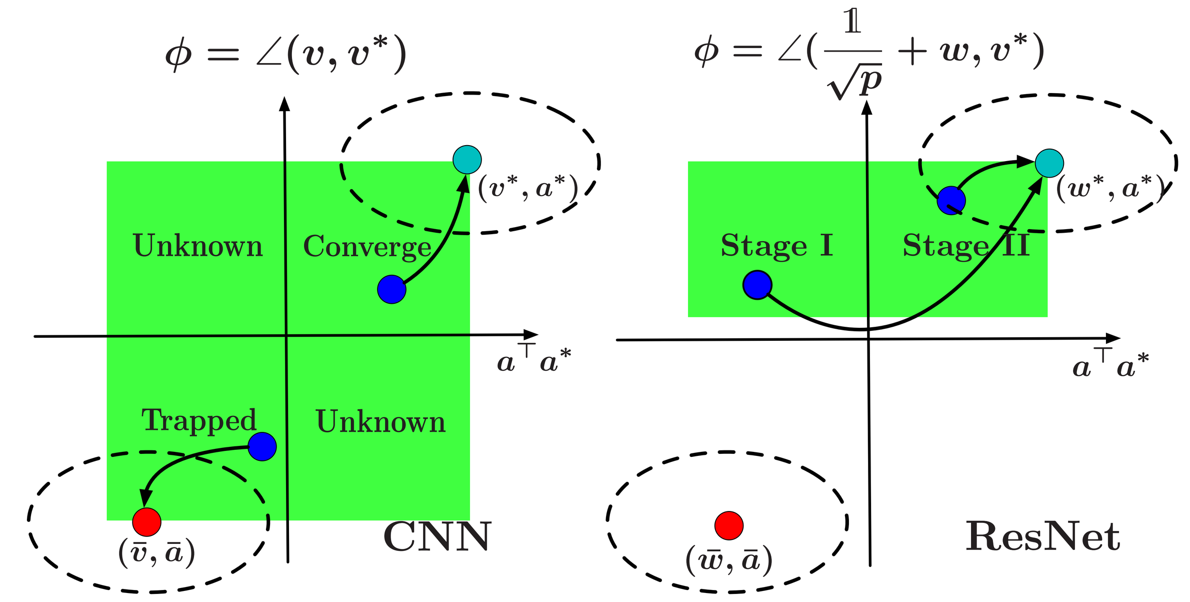
then we have

$$\|w_t - w^*\|_2^2 \leq \delta \quad \text{and} \quad \|a_t - a^*\|_2^2 \leq 5\delta$$

for any $t \geq T_2 = \widetilde{O}\left(\frac{1}{\eta}\log\frac{1}{\delta}\right)$.

- Remark: Step Size Warm Up: $\eta_w^1 < \eta_w^2$.

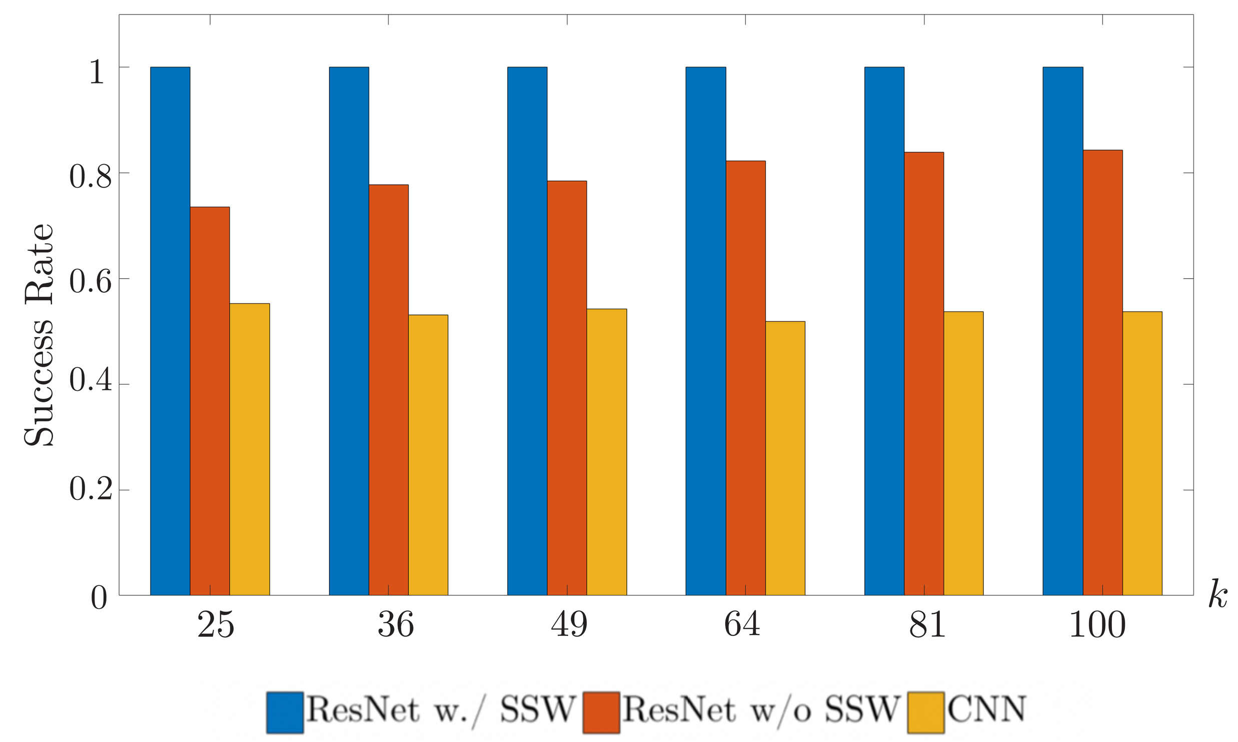
Comparison



Skip-layer prior helps avoid spurious local optima!

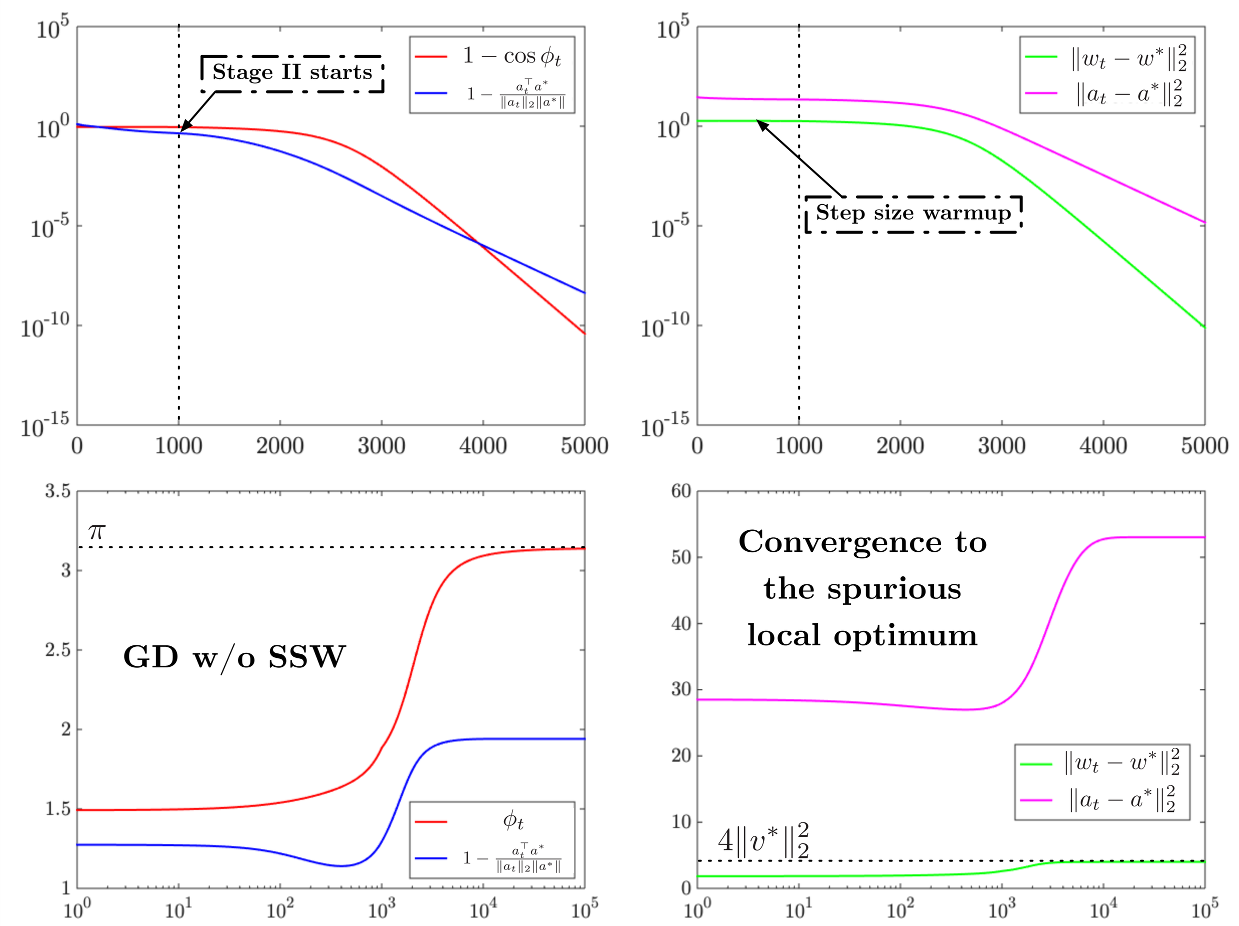
Experiments

- Success Rates with $p = 8$ and Varing k.



– The skip-layer connection improves the success rate.
– Step size warm-up makes ResNet even better .

- Empirical Convergence:



–1st Row: The algorithm has a phase transition.
–2nd Row: GD w/o SSW is trapped in the spurious local optimum.