Fixed-Weight Difference Target Propagation

Tatsukichi Shibuya[1], Nakamasa Inoue[1],
Rei Kawakami[1], Ikuro Sato[1,2]

[1]Tokyo Institute of Technology, Japan [2]DENSO IT Laboratory, Inc., Japan

東京工業大学
Tokyo Institute of Technology

Recognition and Learning Algorithm Laboratory

DENSO IT LAB

1. Background and Contributions

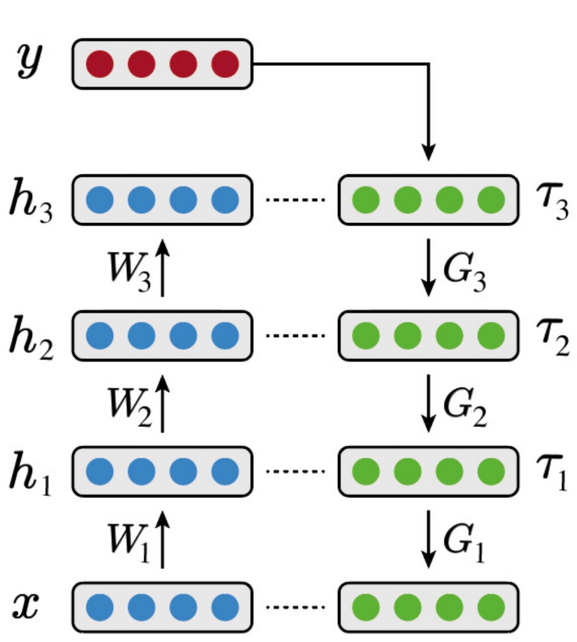
Criticisms of BP about its inconsistencies with neuroscience
➡ **Difference Target Propagation** [Lee+, 2015] was proposed

Problems:
- Low generalization performance
- High computational cost
- Hyperparameter instability

We propose **Fixed-Weight Difference Target Propagation**
- Generalization performance comparable to SOTA
- Significantly reduced computational cost
- Improved stability with different hyperparameters

2. Difference Target Propagation [Lee+, 2015]

Feedback network propagates targets of each activation



Forward Network

$h_l = f_l(h_{l-1}) = \sigma_l(W_l h_{l-1})$

Update weight W_l to reduce layer-wise loss:

$\|h_l - \tau_l\|^2$

Feedback Network

$\tau_{l-1} = h_{l-1} + g_l(\tau_l) - g_l(h_l)$

Update weight G_l by forming layer-wise autoencoder with forward network

3. Fixed-Weight Difference Target Propagation

Feedback network with fixed-weights:

$$\tilde{g}_l(\tau_l) = \sigma_l(B_l \tau_l)$$

B_l: Fixed-weight initialized with uniform distribution $U(-0.01, 0.01)$

Forward Network

$h_l = f_l(h_{l-1}) = \sigma_l(W_l h_{l-1})$

Update weight W_l to reduce layer-wise loss:

$\|h_l - \tau_l\|^2$

Feedback Network

$\tau_{l-1} = h_{l-1} + \tilde{g}_l(\tau_l) - \tilde{g}_l(h_l)$

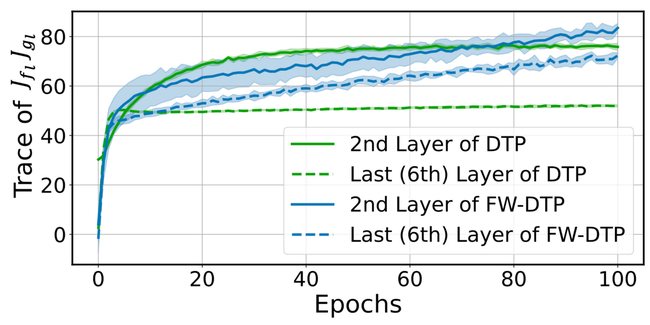
Weight B_l is initialized with random and full-rank matrix and *fixed in training process*

Analysis 1. (Condition of Jacobian)

- Forward weights self-align s.t. the forward and feedback Jacobian satisfies

$$\mathrm{tr}(J_{f_l} J_{g_l}) \geq 0$$

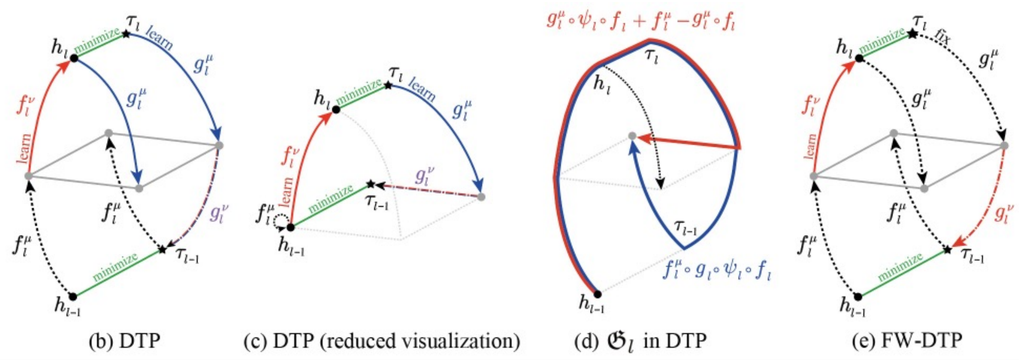
J_f: Jacobian of f
$\mathrm{tr}(X)$: trace of X



It guarantees that the average global loss decreases

Analysis 2. (Exact form of feedback function)

- The feedback function in FW-DTP was cooperatively modified with updated forward weights



4. Experimental Results

1. Generalization Performance

Set-up:
- Model: Fully connected network (with 3~5 hidden layers)
- Datasets: Four image classification datasets

Methods	#Params	MNIST	F-MNIST	#Params	CIFAR-10	CIFAR-100
BP	0.5M	$1.85_{\pm 0.09}$	$10.42_{\pm 0.08}$	6.3M	$46.16_{\pm 1.15}$	$75.96_{\pm 0.52}$
TP	1.1M	$78.99_{\pm 2.04}$	–	13.0M	–	–
DTP	0.5M	$3.24_{\pm 0.15}$	$11.86_{\pm 0.14}$	6.3M	$52.17_{\pm 0.79}$	$77.89_{\pm 0.39}$
	1.1M	$2.77_{\pm 0.10}$	$11.77_{\pm 0.16}$	13.0M	$52.01_{\pm 0.80}$	$77.11_{\pm 0.20}$
DRL	0.5M	$3.13_{\pm 0.03}$	$12.75_{\pm 0.52}$	6.3M	$50.11_{\pm 0.67}$	$76.69_{\pm 0.30}$
	1.1M	$2.84_{\pm 0.09}$	$12.15_{\pm 0.25}$	13.0M	$\mathbf{48.79_{\pm 0.58}}$	$75.62_{\pm 0.35}$
L-DRL	0.5M	$3.14_{\pm 0.03}$	$12.45_{\pm 0.36}$	6.3M	$49.58_{\pm 0.33}$	$76.72_{\pm 0.26}$
	1.1M	$2.82_{\pm 0.10}$	$12.29_{\pm 0.46}$	13.0M	$49.84_{\pm 0.55}$	$\mathbf{75.62_{\pm 0.31}}$
FW-DTP	0.5M	$\mathbf{2.76_{\pm 0.10}}$	$\mathbf{11.76_{\pm 0.37}}$	6.3M	$48.97_{\pm 0.32}$	$76.76_{\pm 0.45}$

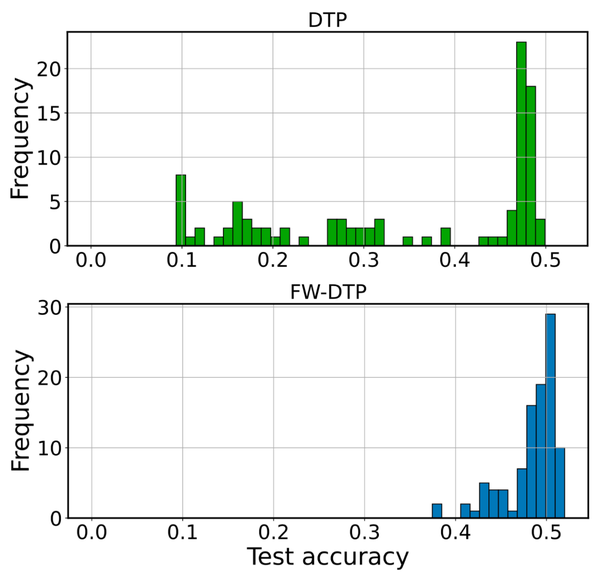
2. Computational cost & Hyperparameter sensitivity

Set-up:
- Cost: Training time per epoch
- Sensitivity: Test acc. with different hyperparameters

	Time[sec]	Ratio to FW-DTP
FW-DTP	$2.22_{\pm 0.02}$	$1.00_{\pm 0.00}$
DTP	$8.32_{\pm 0.36}$	$3.74_{\pm 0.17}$
DRL	$9.52_{\pm 0.08}$	$4.29_{\pm 0.05}$
L-DRL	$8.86_{\pm 0.08}$	$3.99_{\pm 0.05}$
BP	$0.76_{\pm 0.03}$	$0.34_{\pm 0.01}$

Computational cost (↑)

Hyperparameter sensitivity (→)



AAAI Association for the Advancement of Artificial Intelligence

Paper with supplementary materials (arXiv)

This work is supported by DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Tokyo Tech.).