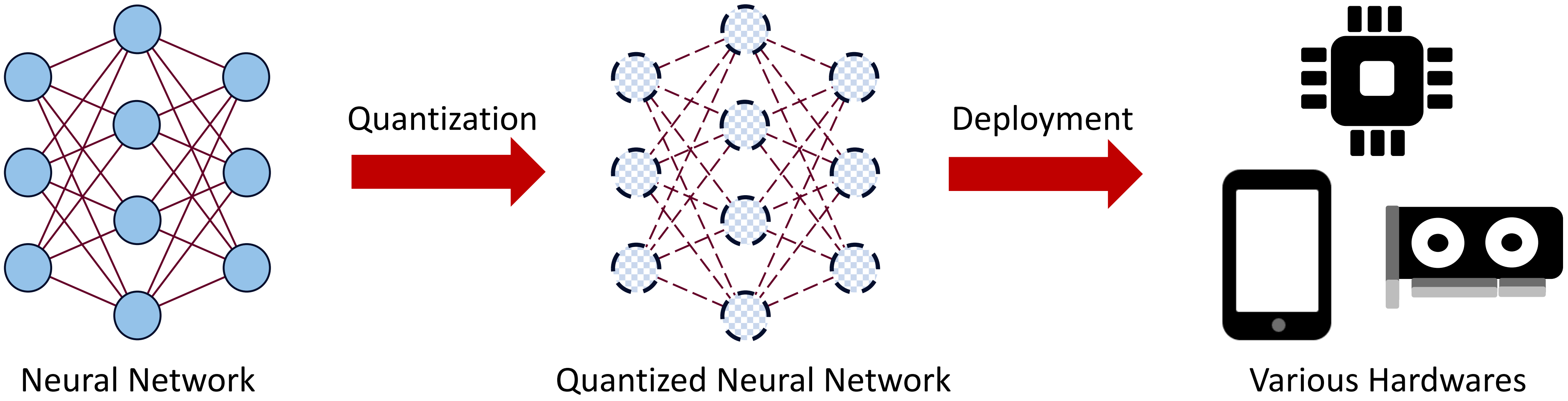
Merge-Friendly Post-Training Quantization for Mutli-Target Domain Adaptation

Juncheol Shin, Minsang Seok, Seonggon Kim, Eunhyeok Park
Pohang University of Science and Technology

Introduction

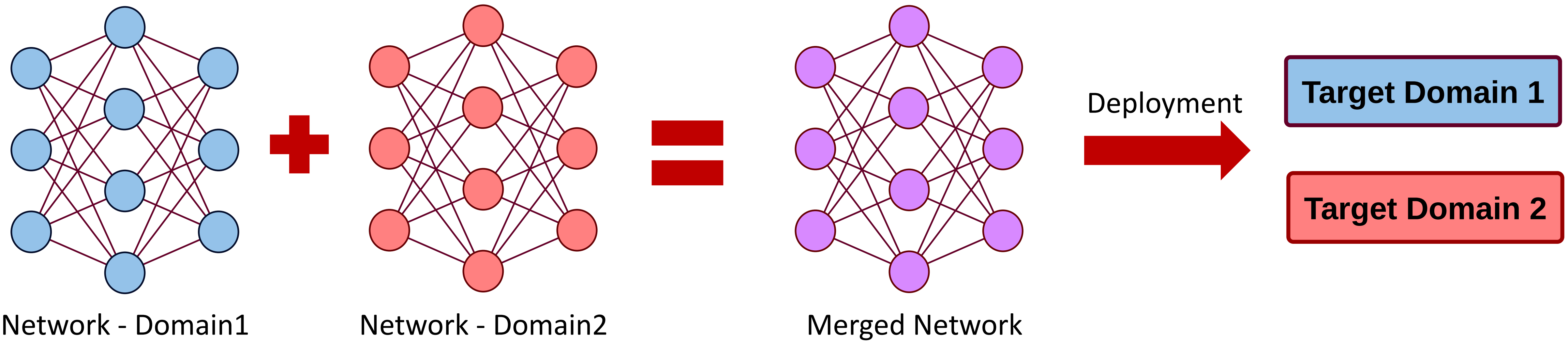
- Quantization
  - One of the most widely adopted optimization techniques
  - Activations and weights are stored in a **low-precision domain**
    - Reduced **memory usage** & **computational requirements**



POSTECH

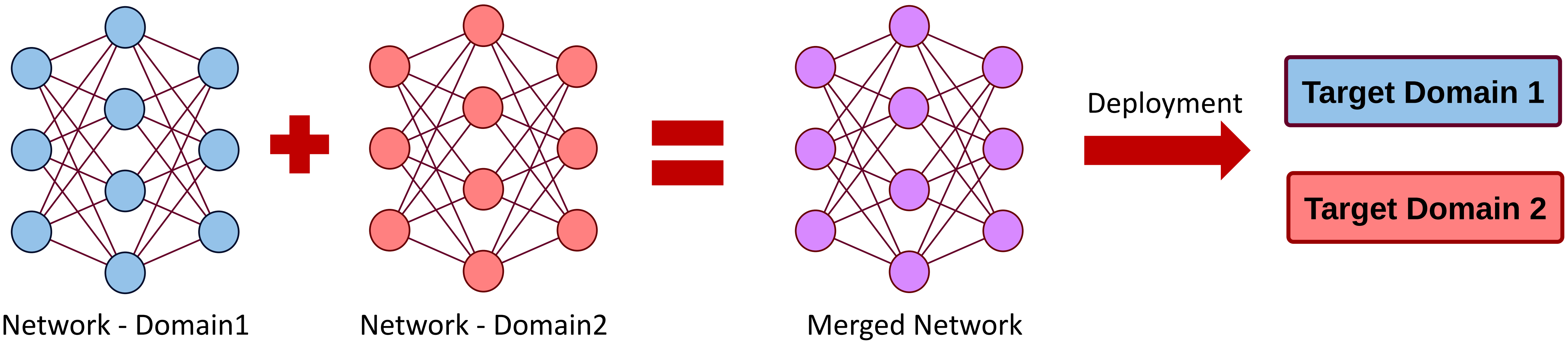
Introduction

- Model Merging
  - Emerging technique to generate model for multiple tasks
  - Recent study revealed even simple weight averaging outperforms other methods in MTDA
    - Shed light to real-time adaptive AI via model merging in edge devices



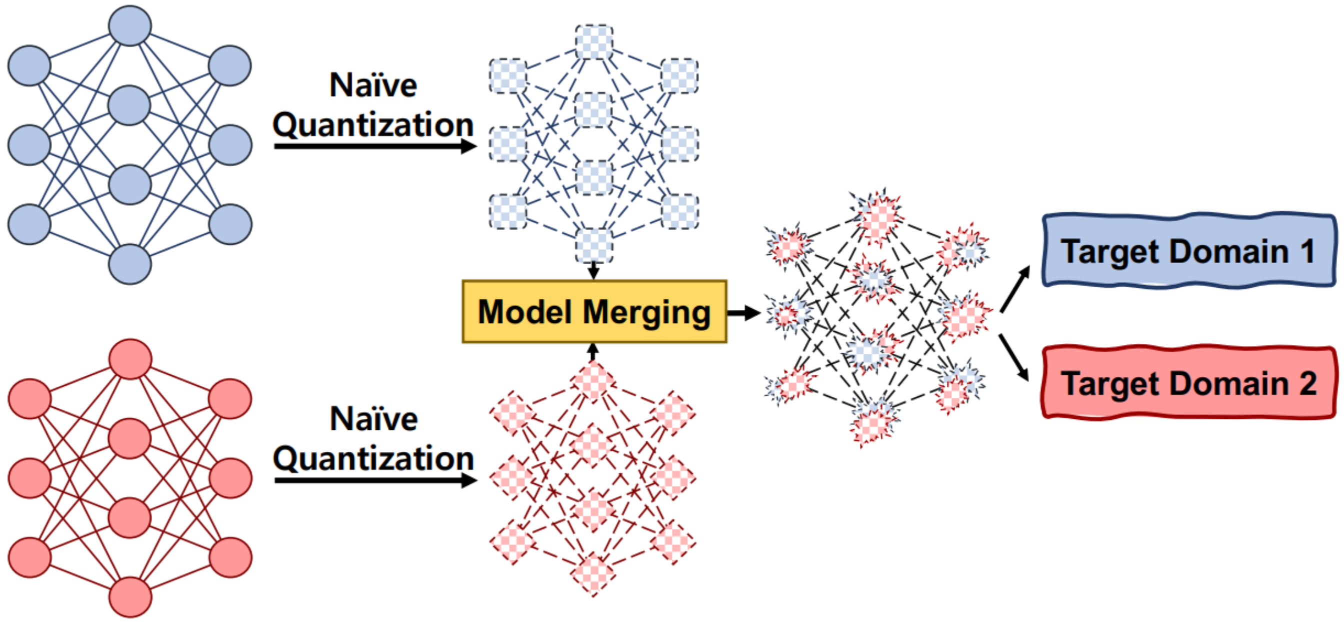
Introduction

- Model Merging
  - Emerging technique to generate model for multiple tasks
  - Recent study revealed even simple weight averaging outperforms other methods in MTDA
    - Shed light to real-time adaptive AI via model merging in edge devices
      - **+ Quantization**?



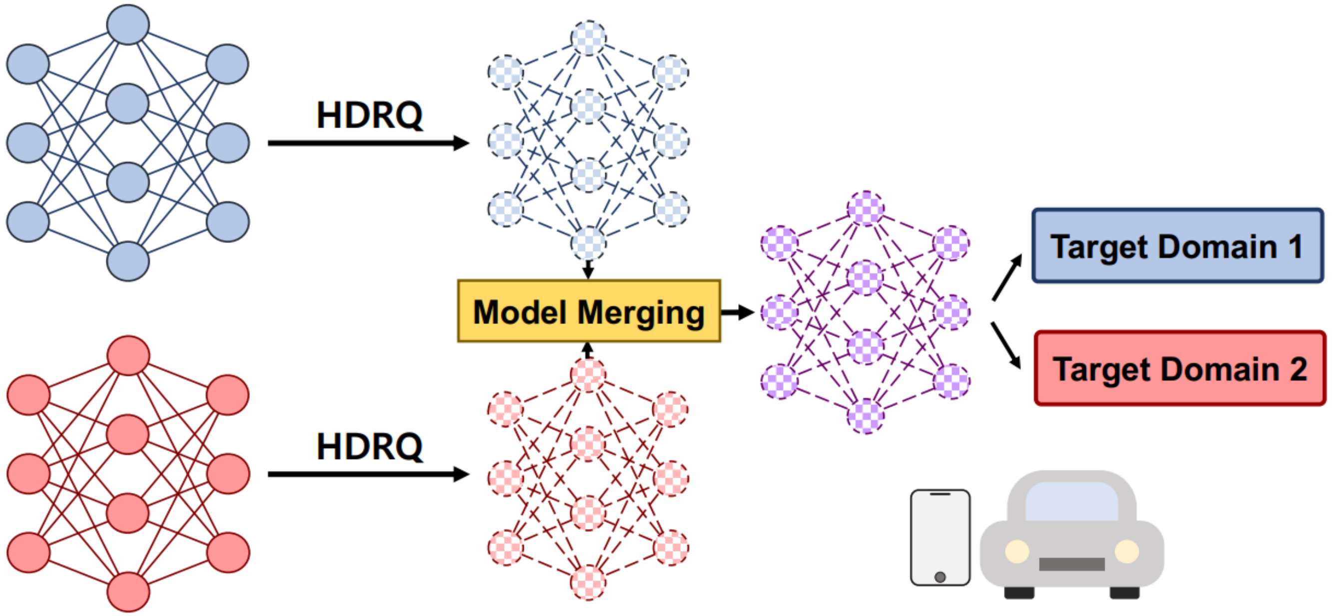
Motivation

- Quantization + Model Merging ?
  - Discretization that is not well aligned with the merging
    - Suboptimal and degraded performance with naïve quantization
  - Little attention has been given to the interplay



Motivation

- **HDRQ** : **H**essian and **D**istance **R**egulariziang **Q**uantization
  - **Theoretical analysis** of quantization's impact on model merging
  - Propose **regularization techniques** for merge-friendly quantization
  - **Noise-sampling-based rounding** to handle ambiguity problem



Analysis

- Error Barrier
  - Quantifies the degree of interpolation-induced performance degradation
    - θ_1 and θ_2 denotes converged weights for each domain
    - θ_λ denotes interpolated weight
      - $\theta_\lambda = (1-\lambda)\theta_1 + \lambda\theta_2,\ \lambda \in [0,1]$

Should be minimized!

$$\max_{\lambda \in [0,1]} \left[L(\theta_\lambda) - \frac{1}{2}(L(\theta_1) + L(\theta_2))\right]$$

Analysis

- Error Barrier
  - Quantifies the degree of interpolation-induced performance degradation
    - θ_1 and θ_2 denotes converged weights for each domain
    - θ_λ denotes interpolated weight
      - $\theta_\lambda = (1-\lambda)\theta_1 + \lambda\theta_2,\ \lambda \in [0,1]$

Should be minimized!

$$\max_{\lambda\in[0,1]} \left[L(\theta_\lambda) - \frac{1}{2}(L(\theta_1) + L(\theta_2)\right]$$

- + Quantization
  - Error induced by quantization can be approximated as additive uniform noise
    - $\epsilon_1 \sim U\left[-\frac{s_1}{2}, \frac{s_1}{2}\right]$ and $\epsilon_2 \sim U\left[-\frac{s_2}{2}, \frac{s_2}{2}\right]$ with quantization step sizes s_1 and s_2

$$\max_{\lambda\in[0,1]} \left[L(\theta_\lambda + \epsilon_\lambda) - \frac{1}{2}(L(\theta_1 + \epsilon_1) + L(\theta_2 + \epsilon_2)\right]$$

Analysis

- Error Barrier + Quantization
  - Applying a second-order Taylor expansion, we obtain:

$$\max_{\lambda\in[0,1]}[L(\theta_\lambda) - \frac{1}{2}(L(\theta_1) + L(\theta_2)] +$$
$$\max_{\lambda\in[0,1]}[\epsilon_\lambda \cdot \nabla_\theta L(\theta_\lambda) + \frac{1}{2}\epsilon_\lambda^T \cdot \nabla_\theta^2 L(\theta_\lambda) \cdot \epsilon_\lambda - \frac{1}{2}(\epsilon_1 \cdot \nabla_\theta L(\theta_1) + \frac{1}{2}\epsilon_1^T \cdot \nabla_\theta^2 L(\theta_1) \cdot \epsilon_1 +$$
$$\epsilon_2 \cdot \nabla_\theta L(\theta_2) + \frac{1}{2}\epsilon_2^T \cdot \nabla_\theta^2 L(\theta_2) \cdot \epsilon_2)]$$

Analysis

- Error Barrier + Quantization
  - Applying a second-order Taylor expansion, we obtain:

Assuming zero error barrier for simplicity

$$\max_{\lambda\in[0,1]}[L(\theta_\lambda) - \frac{1}{2}(L(\theta_1) + L(\theta_2)] +$$

$$\max_{\lambda\in[0,1]}[\epsilon_\lambda \cdot \nabla_\theta L(\theta_\lambda) + \frac{1}{2}\epsilon_\lambda^T \cdot \nabla_\theta^2 L(\theta_\lambda)\cdot\epsilon_\lambda - \frac{1}{2}(\epsilon_1 \cdot \nabla_\theta L(\theta_1) + \frac{1}{2}\epsilon_1^T \cdot \nabla_\theta^2 L(\theta_1)\cdot\epsilon_1 +$$

$$\epsilon_2 \cdot \nabla_\theta L(\theta_2) + \frac{1}{2}\epsilon_2^T \cdot \nabla_\theta^2 L(\theta_2)\cdot\epsilon_2)]$$

First-order term of converged point

First-order term of converged point

Analysis

- Error Barrier + Quantization
  - Applying a second-order Taylor expansion, we obtain:

$$\max_{\lambda\in[0,1]}[\epsilon_\lambda \cdot \nabla_\theta L(\theta_\lambda) + \frac{1}{2}\epsilon_\lambda^T \cdot \nabla_\theta^2 L(\theta_\lambda) \cdot \epsilon_\lambda - \frac{1}{4}(\epsilon_1^T \cdot \nabla_\theta^2 L(\theta_1) \cdot \epsilon_1 + \epsilon_2^T \cdot \nabla_\theta^2 L(\theta_2) \cdot \epsilon_2)]$$

Analysis

- Error Barrier + Quantization
  - Applying a second-order Taylor expansion, we obtain:

$$\max_{\lambda \in [0,1]} \left[\epsilon_\lambda \cdot \nabla_\theta L(\theta_\lambda) + \frac{1}{2}\epsilon_\lambda^T \cdot \nabla_\theta^2 L(\theta_\lambda) \cdot \epsilon_\lambda - \frac{1}{4}\left(\epsilon_1^T \cdot \nabla_\theta^2 L(\theta_1) \cdot \epsilon_1 + \epsilon_2^T \cdot \nabla_\theta^2 L(\theta_2) \cdot \epsilon_2\right)\right]$$

  - To minimize total error barrier,
    1. Minimze red term
    2. Maximize blue term

POSTECH

Analysis

- Error Barrier + Quantization
  - Applying a second-order Taylor expansion, we obtain:

$$\max_{\lambda\in[0,1]}\left[\epsilon_\lambda \cdot \nabla_\theta L(\theta_\lambda) + \frac{1}{2}\epsilon_\lambda^T \cdot \nabla_\theta^2 L(\theta_\lambda) \cdot \epsilon_\lambda - \frac{1}{4}(\epsilon_1^T \cdot \nabla_\theta^2 L(\theta_1) \cdot \epsilon_1 + \epsilon_2^T \cdot \nabla_\theta^2 L(\theta_2) \cdot \epsilon_2)\right]$$

  - To minimize total error barrier,
    1. Minimze red term
    2. ~~Maximize blue term~~ : Increased Hessian → Degraded robustness and quality

POSTECH

Analysis

- Error Barrier + Quantization
  - Applying a second-order Taylor expansion, we obtain:

$$\max_{\lambda\in[0,1]}\left[\epsilon_\lambda \cdot \nabla_\theta L(\theta_\lambda) + \frac{1}{2}\epsilon_\lambda^T \cdot \nabla_\theta^2 L(\theta_\lambda)\cdot \epsilon_\lambda - \frac{1}{4}\left(\epsilon_1^T \cdot \nabla_\theta^2 L(\theta_1)\cdot \epsilon_1 + \epsilon_2^T \cdot \nabla_\theta^2 L(\theta_2)\cdot \epsilon_2\right)\right]$$

  - To minimize total error barrier,
    1. Minimze red term : How?
    2. ~~Maximize blue term~~ : Increased Hessian → Degraded robustness and quality

POSTECH

Analysis

- Error Barrier + Quantization
  - Assuming Hessian of loss L is M-Lipschitz continuous between θ_1 and θ_2,

$$\left| \boxed{\nabla_\theta^2 L(\theta_\lambda)} - \frac{\nabla_\theta^2 L(\theta_1) + \nabla_\theta^2 L(\theta_2)}{2} \right| \leq \frac{M\|\theta_2 - \theta_1\|}{2}$$

  - Hessian at merged point can be effectively regularized by,

Analysis

- Error Barrier + Quantization
  - Assuming Hessian of loss L is M-Lipschitz continuous between θ_1 and θ_2,

$$\left| \nabla_\theta^2 L(\theta_\lambda) - \frac{\nabla_\theta^2 L(\theta_1) + \nabla_\theta^2 L(\theta_2)}{2} \right| \leq \frac{M \|\theta_2 - \theta_1\|}{2}$$

  - Hessian at merged point can be effectively regularized by,
    - **Controlling Hessians** at the θ_1 and θ_2
    - **Minimizing Distance** between θ_1 and θ_2
    - This leads to **minimization of first-order term, as it also becomes lipschitz continuous**

Analysis

- Error Barrier + Quantization
  - Assuming Hessian of loss L is M-Lipschitz continuous between θ_1 and θ_2,

$$\left| \nabla_\theta^2 L(\theta_\lambda) - \frac{\nabla_\theta^2 L(\theta_1) + \nabla_\theta^2 L(\theta_2)}{2} \right| \le \frac{M \|\theta_2 - \theta_1\|}{2}$$

  - Hessian at merged point can be effectively regularized by,
    - **Controlling Hessians** at the θ_1 and θ_2
    - **Minimizing Distance** between θ_1 and θ_2
    - This leads to **minimization of first-order term, as it also becomes lipschitz continuous**
  - Domain Adaptation Case ($\nabla L_1(\theta_2) \neq 0, \quad \nabla L_2(\theta_1) \neq 0$)?
    - Able to derive same conclusion

$$\max_{\lambda \in [0,1]} \left[(\epsilon_\lambda + k \cdot \epsilon_2) \cdot \nabla_\theta L_1(\theta_\lambda) + \frac{1}{2} \epsilon_\lambda^T \cdot \nabla_\theta^2 L_1(\theta_\lambda) \cdot \epsilon_\lambda - \frac{1}{4} (\epsilon_1^T \cdot \nabla_\theta^2 L(\theta_1) \cdot \epsilon_1 + \epsilon_2^T \cdot \nabla_\theta^2 L(\theta_2) \cdot \epsilon_2) \right]$$

Method

- Noise-based hessian regularization
  - Simulates quantization error by introducing additive sampled noise, ϵ
    - Quantized weight : $\widehat{w} = clamp\left(\left\lfloor\frac{w}{\Delta}\right\rceil, -2^{b-1}, 2^{b-1} - 1\right) \cdot \Delta$
      - Δ , b denotes step size and bit-width, respectively
    - ϵ is sampled from $w - \widehat{w}$
      - Quantization noise follows uniform distribution, $U[-\frac{\Delta}{2}, \frac{\Delta}{2}]$

$$\widehat{w}_{HDRQ} = w + \epsilon$$

Method

- Noise-based hessian regularization
  - Simulates quantization error by introducing additive sampled noise, ϵ
    - Quantized weight : $\hat{w} = clamp\left(\left\lfloor \frac{w}{\Delta} \right\rceil, -2^{b-1}, 2^{b-1} - 1\right) \cdot \Delta$
      - Δ , b denotes step size and bit-width, respectively
    - ϵ is sampled from $w - \hat{w}$
      - Quantization noise follows uniform distribution, $U[-\frac{\Delta}{2}, \frac{\Delta}{2}]$

$$\hat{w}_{HDRQ} = w + \epsilon$$

  - Inherently regularizes Hessian as follows:

$$E[L(\hat{w})] \approx E[\hat{w}_{HDRQ}] = E[w + \epsilon]$$

First-order term of converged point

$$\approx E[L(w) + \epsilon \cdot \nabla_w L(w) + \frac{1}{2}\epsilon^T \cdot \nabla_w^2 L(w) \cdot \epsilon]$$

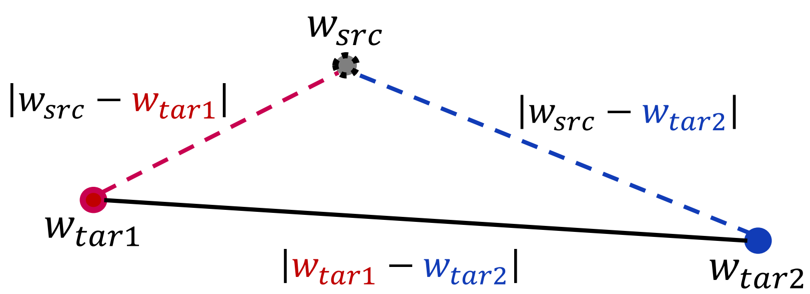
$$\approx E[L(w) + \frac{1}{2}\epsilon^T \cdot \nabla_w^2 L(w) \cdot \epsilon]$$

POSTECH

Method

- Weight distance regularization
  - Regularize upper bound derived from triangular inequality
    - Without prior information about target domains and weights

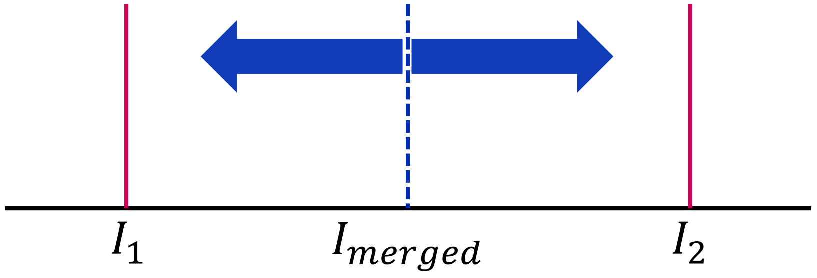
$$|w_{tar1} - w_{tar2}| \leq |w_{src} - w_{tar1}| + |w_{src} - w_{tar2}|$$



  - Access to source weights?
    - Generally models pretrained from source data are adapted and deployed
      - Provider must maintain source weight

Method

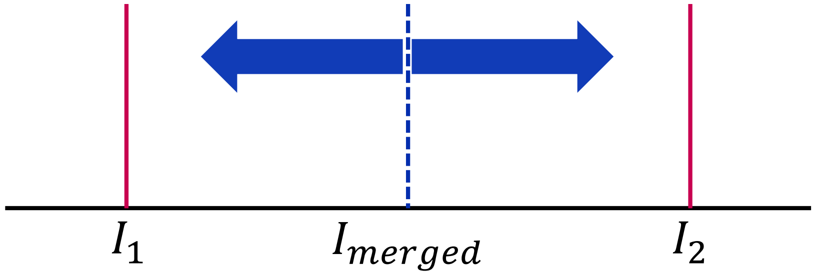
- Handling Ambiguity in Rounding Policy
  - Consider two quantized values being merged,
    - I_1 and I_2 : Integer representations
    - Δ_1 and Δ_2 : Step sizes
  - If sum of I_1 and I_2 is an odd number, ambiguity in the rounding direction arises



POSTECH

Method

- Handling Ambiguity in Rounding Policy
  - Consider two quantized values being merged,
    - I_1 and I_2 : Integer representations
    - Δ_1 and Δ_2 : Step sizes
  - If sum of I_1 and I_2 is an odd number, ambiguity in the rounding direction arises



  - Merging in floating point domain?
    - Again degenerates when $\Delta_1 \approx \Delta_2$

$$I_{merged} = \left\lfloor \frac{I_1 \cdot \Delta_1 + I_2 \cdot \Delta_2}{\Delta_1 + \Delta_2} \right\rceil \approx \left\lfloor \frac{I_1 \cdot \Delta_1 + I_2 \cdot \Delta_1}{2 \cdot \Delta_1} \right\rceil \approx \left\lfloor \frac{I_1 + I_2}{2} \right\rceil$$

Method

- Handling Ambiguity in Rounding Policy
  - Our Solution : Employ noise sampling
    - Maintains same quantized representation while mitigating ambiguity
    - $\epsilon \sim U[-\frac{\Delta}{2}, \frac{\Delta}{2}]$

$$I_{merged} = \left\lfloor \frac{(I_1 \cdot \Delta_1 + \epsilon_1) + (I_2 \cdot \Delta_2 + \epsilon_2)}{\Delta_1 + \Delta_2} \right\rceil$$

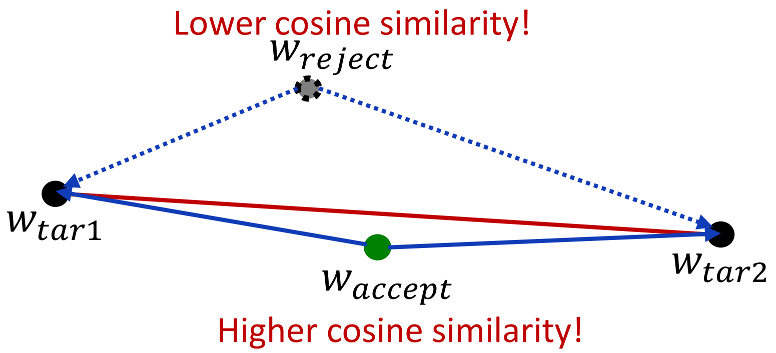
POSTECH

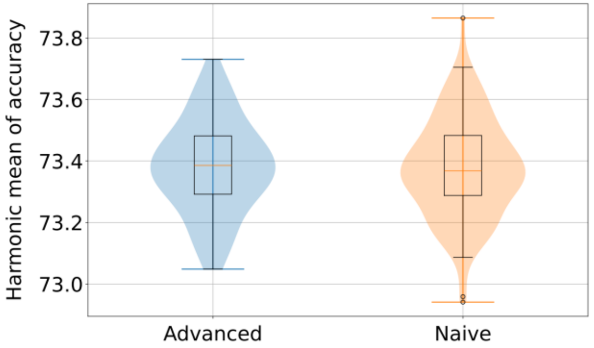
Method

- Handling Ambiguity in Rounding Policy
  - Our Solution : Employ noise sampling
    - Maintains same quantized representation while mitigating ambiguity
    - $\epsilon \sim U[-\frac{\Delta}{2}, \frac{\Delta}{2}]$

$$I_{merged} = \left\lfloor \frac{(I_1 \cdot \Delta_1 + \epsilon_1) + (I_2 \cdot \Delta_2 + \epsilon_2)}{\Delta_1 + \Delta_2} \right\rceil$$

  - Stabilize merge result?
    - Utilizing noise, it is important to have **stable results with low variance**
    - Highest Cosine similiary between original interpolation vector and new vectors

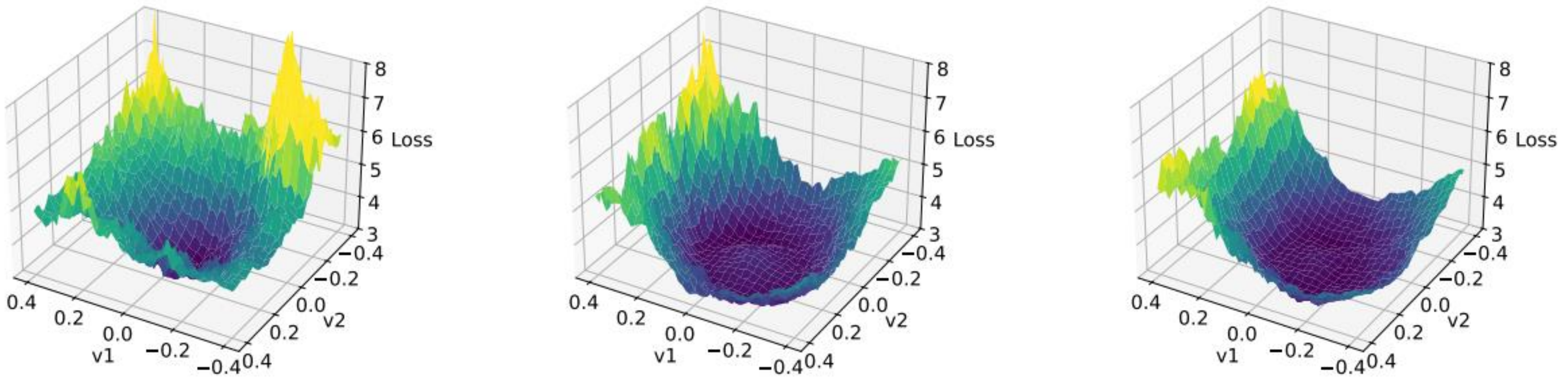




POSTECH

Experimental Results

- Landscape Visualization
  - HDRQ guides the network to smoother loss surface
    - Direct injection of noise to weights effectively handles local lumps



(a) BRECQ (b) QDrop (c) HDRQ(Ours)

Experimental Results

- Semantic Segmentation
  - Quantization results on each target domain are comparable
  - However when merged, HDRQ outperforms other methods by large margin

Method	Bit(W/A)	Domain	mIOU
FP	32/32	G → C	61.69
		G → I	52.06
BRECQ	4/4	G → C	53.67
		G → I	45.50
Qdrop	4/4	G → C	58.92
		G → I	49.44
HDRQ	4/4	G → C	58.23
		G → I	48.68

Quantization result on each domain

Method	Bit(W/A)	Metric	mIOU
FP	32/32	C	58.12
		I	53.50
		H	55.71
BRECQ	4/4	C	29.21
		I	35.34
		H	31.95
Qdrop	4/4	C	39.91
		I	43.37
		H	41.54
HDRQ	4/4	C	44.44
		I	47.17
		H	**45.75**

Merging Results

Experimental Results

- Image Classification (Merging 3 networks)
  - HDRQ outperforms other methods, especially when weights are quantized into low bit
    - **Bold** indicates best one
    - Red indicates accuracy gain of over > 1% compared to the second-best

Domain	FP	Methods	W4A8	W4A4	W3A3
R→A,C,P	67.68	BRECQ	64.15	60.95	43.66
		QDrop	64.85	66.26	62.99
		HDRQ	**66.74**	**66.41**	**64.70**
A→R,C,P	68.80	BRECQ	66.06	62.53	48.04
		QDrop	66.83	66.04	64.22
		HDRQ	**67.80**	**67.58**	**65.29**
C→R,A,P	75.07	BRECQ	73.22	71.31	56.16
		QDrop	73.81	73.25	71.01
		HDRQ	**74.26**	**73.58**	**71.63**
P→R,A,C	65.25	BRECQ	**64.09**	61.92	45.09
		QDrop	62.52	**63.22**	61.24
		HDRQ	63.93	63.19	**61.55**

Experimental Results

- Incremental Ablation Study
  - Office-Home dataset, W3A3 precision, R→A,C,P setting
  - Noise-based quantization scheme yields performance gain of 1.22%
  - Further incorporating weight distance regularization gives additional 0.49% gain

Method	Accuracy
Baseline	62.99 (73.14|58.69|83.62)
+ Noise-based Quantization	64.21 (+1.22%) (73.42|58.65|83.67)
+ Distance Regularization	64.70 (+0.49%) (72.81|58.72|83.33)

POSTECH