Quantifying Task Priority for Multi-Task Optimization

Wooseong Jeong, and Kuk-Jin Yoon
{stk14570, kjyoon}@kaist.ac.kr ; Visual Intelligence Lab., KAIST, Korea

Background

- **Multi-Task Learning** (MTL) is learning paradigm that handle multiple tasks in a single shared model.
- Optimization for MTL aims to mitigate **negative transfer** among tasks and finds **Pareto-optimal solutions**.
- Previous works directly modify task-specific gradients to address task conflicts.

Motivation

- We define the relative importance of tasks in the shared parameters as task priority.
- Divide shared parameters into subsets based on the task priority and use them for multi-task optimization.

Definition 3 (Task priority). *Assume that the task losses $\mathcal{L}_i$ for $i = 1, 2, ..., \mathcal{K}$ are differentiable. Consider $\mathcal{X}^t$ as the input data at time t. We initiate with shared parameters Θ_s^t and task-specific parameters Θ_i^t with sufficiently small learning rate $\eta > 0$. A subset of shared parameters at time t is denoted as θ^t, such that $\theta^t \subset \Theta_s^t$. For any task $\tau_i \in \mathcal{T}$, the task's gradient for θ^t is as follows:*

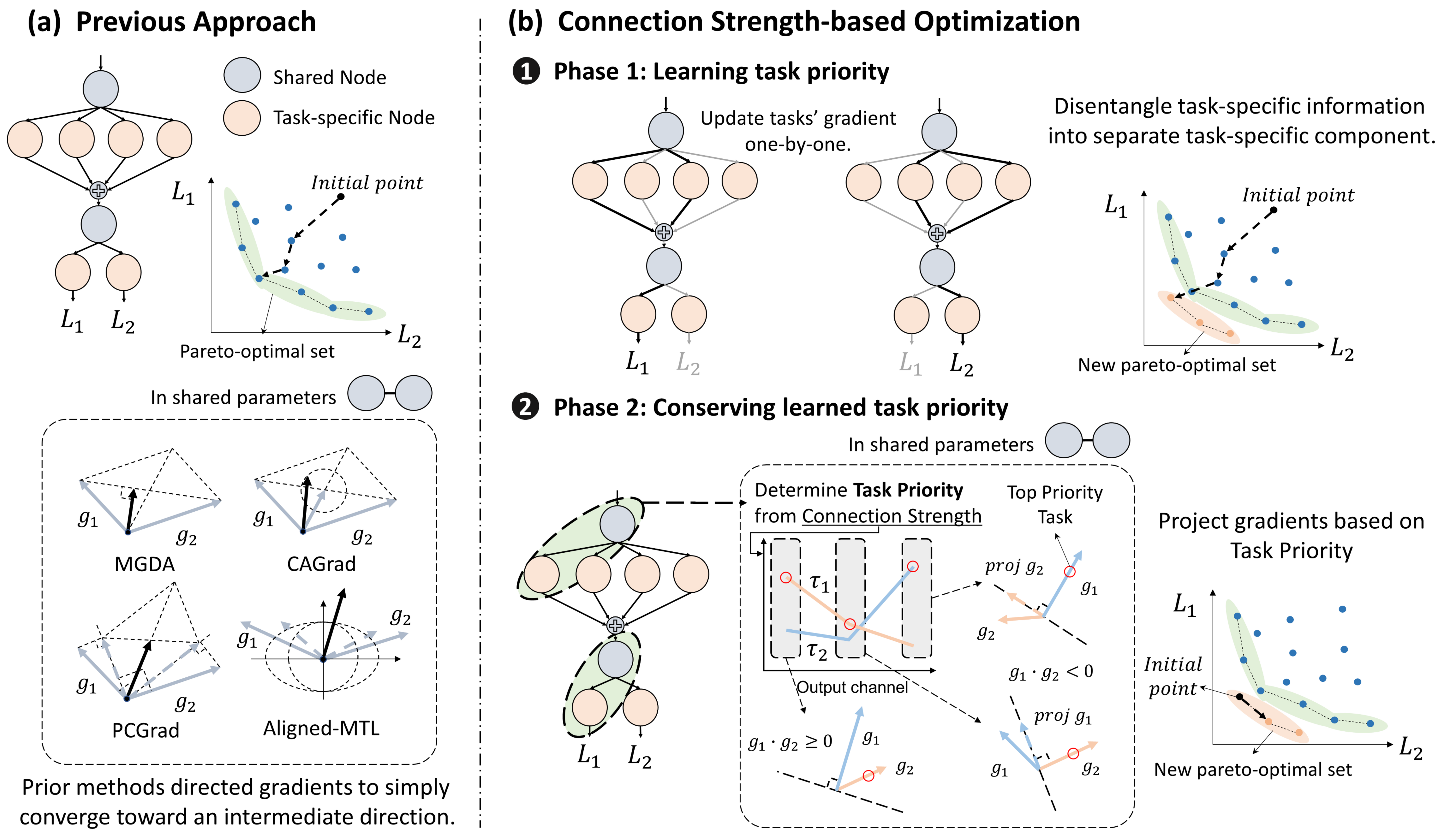
$$g_i = \nabla_{\theta^t} \mathcal{L}_i(\mathcal{X}^t, \tilde{\Theta}_s^t, \theta^t, \Theta_i^t)$$

where $\tilde{\Theta}_s^t$ represents the parameters that are part of Θ_s^t but not in θ^t. For two distinct tasks $\tau_m, \tau_n \in \mathcal{T}$, if τ_m holds priority over τ_n in θ^t, then the following inequality holds:

$$\sum_{i=1}^{\mathcal{K}} w_i \mathcal{L}_i(\tilde{\Theta}_s^t, \theta^t - \eta g_m, \Theta_i^t) \leq \sum_{i=1}^{\mathcal{K}} w_i \mathcal{L}_i(\tilde{\Theta}_s^t, \theta^t - \eta g_n, \Theta_i^t)$$

Proposed Method

- To measure **task priority**, we establish the **connections** in the network and assess their **strength**.
- Proposed novel multi-task optimization for MTL termed **connection strength-based optimization**.
- Proposed optimization breaks down the optimization process into two phases.
- In Phase 1, the network is instructed to capture task-specific details **by learning the task priority**.
- In Phase 2, task priority within the shared parameters is determined and project gradients to **preserve the priority**.



Experimental Results

Table 1. Comparisons with previous optimization (NYUD-v2)

Tasks	Depth		SemSeg			Surface Normal					
Method	Distance (Lower Better)		(%) (Higher Better)			Angle Distance (Lower Better)		Within t degree (%) (Higher Better)			MTP
	rmse	abs_rel	mIoU	PAcc	mAcc	mean	median	11.25	22.5	30	$\triangle_m \uparrow$(%)
Independent	0.667	0.186	33.18	65.04	45.07	20.75	14.04	41.32	68.26	78.04	+ 0.00
GD	0.594	0.150	38.67	69.16	51.12	20.52	13.46	42.63	69.00	78.42	+ 9.53
MGDA [36]	0.603	0.159	38.89	69.39	51.53	20.58	13.56	42.28	68.79	78.33	+ 9.21
PCGrad [45]	0.596	0.149	38.61	69.30	51.51	20.50	13.54	42.56	69.14	78.55	+ 9.40
CAGrad [26]	0.595	0.153	38.80	68.95	50.78	20.38	13.53	42.89	69.33	78.71	+ 9.84
Aligned-MTL [37]	0.592	0.150	39.02	68.98	51.83	20.40	13.57	42.83	69.26	78.69	+ 10.17
Ours	**0.565**	**0.148**	**41.10**	**70.37**	**53.74**	**19.54**	**12.45**	**46.11**	**71.54**	**80.12**	**+ 15.00**

Table 2. Comparisons with previous optimization (PACSAL-Context)

Tasks	SemSeg		PartSeg	Saliency		Surface Normal					
Method	(Higher Better)		(Higher Better)	(Higher Better)		Angle Distance (Lower Better)		Within t degree (%) (Higher Better)			MTP
	mIoU	PAcc	mIoU	mIoU	maxF	mean	median	11.25	22.5	30	$\triangle_m \uparrow$(%)
Independent	60.30	89.88	60.56	67.05	78.98	14.76	11.92	47.61	81.02	90.65	+ 0.00
GD	62.17	90.27	61.15	67.99	79.60	14.70	11.81	47.55	80.97	90.56	+ 1.47
MGDA [36]	61.75	89.98	61.69	67.32	78.98	14.77	12.22	47.02	80.91	90.14	+ 1.15
PCGrad [45]	62.47	90.57	61.46	67.86	79.38	14.59	11.77	47.72	81.28	90.81	+ 1.86
CAGrad [26]	62.22	90.01	61.89	67.46	79.12	14.97	12.10	47.23	80.54	90.30	+ 1.14
Aligned-MTL [37]	62.43	90.51	62.05	67.94	79.57	14.76	11.86	47.44	80.78	90.46	+ 1.83
Ours	**63.86**	**90.65**	**63.05**	**68.30**	**79.26**	**14.33**	**11.45**	**49.08**	**81.86**	**91.05**	**+ 3.70**

Table 3. Ablation studies for phase mixing strategies.

Phase		Depth	Seg	Norm	MTP	Averaged
1	2	rmse	mIoU	mean	$\triangle_m \uparrow$	Loss
✓		0.581	40.36	19.55	+ 13.44	**0.5396**
	✓	0.597	39.23	20.39	+ 10.32	0.6519
$\checkmark_{seq}$	$\checkmark_{seq}$	0.574	40.38	19.56	+ 13.79	0.5788
$\checkmark_{mix}$	$\checkmark_{mix}$	**0.565**	**41.10**	**19.54**	**+ 15.50**	0.5942

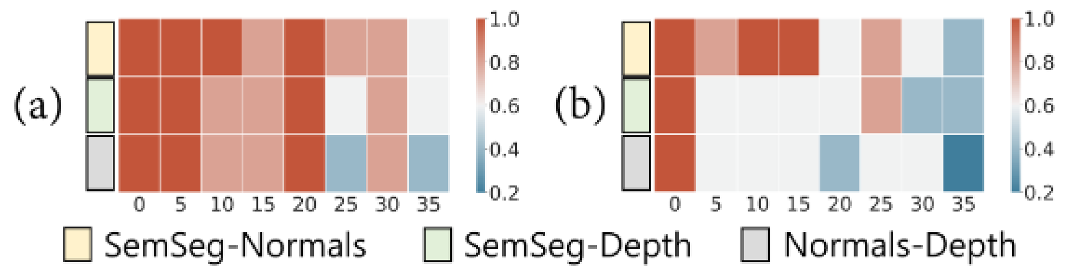


Figure 1. Correlation of loss trends across tasks during the epochs
(a) Phase 1, (b) Phase 2

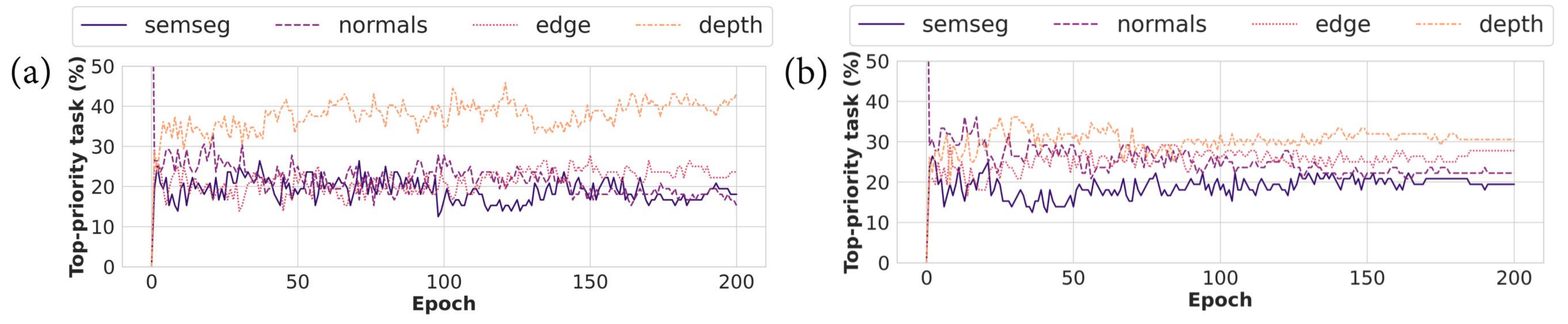


Figure 2. Visualization of the percentage of top priority task
(a) Phase 1, (b) Mixing Phase1 + Phase 2