Pareto Multi-Task Learning

Xi Lin[1], Hui-Ling Zhen[1], Zhenhua Li[2], Qingfu Zhang[1], Sam Kwong[1]

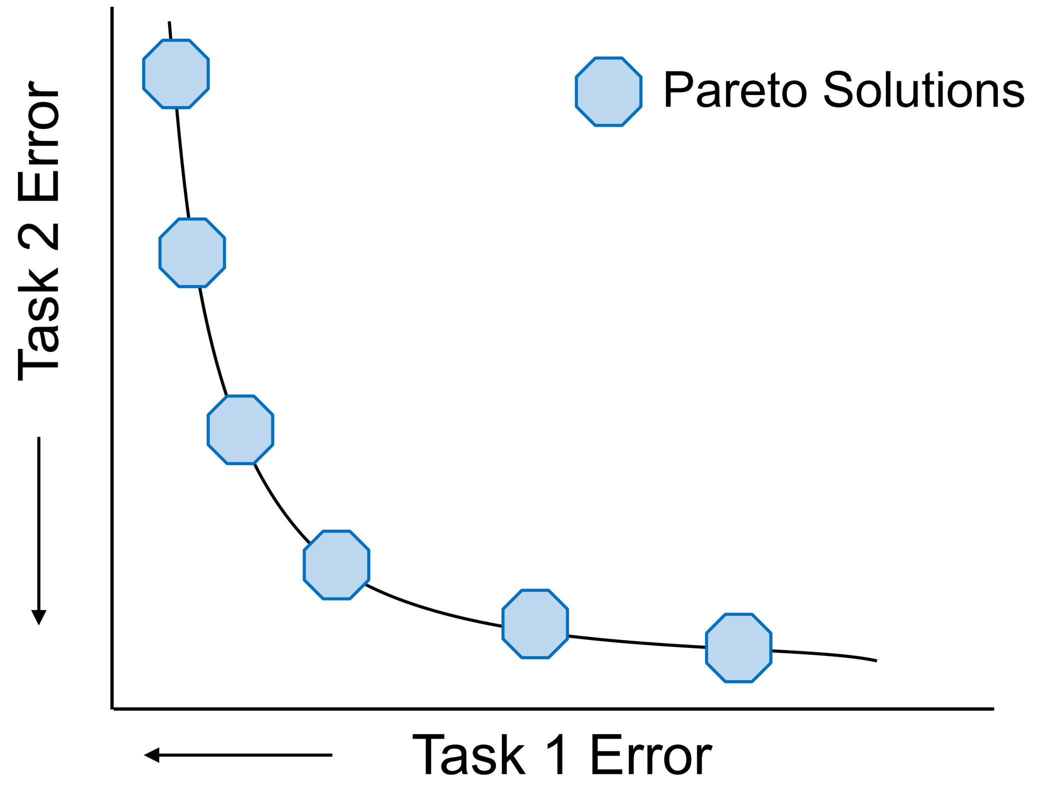
[1]City University of Hong Kong, [2]Nanjing University of Aeronautics and Astronautics

xi.lin@my.cityu.edu.hk

Goal and Motivation

Goal:

To generate a set of widely distributed Pareto solutions with **different optimal trade-offs among tasks** for a multi-task learning (MTL) problem.



Then MTL practitioners can easily select their solutions with preferred trade-offs.

Motivation:

A MTL problem with m tasks is a **multi-objective optimization problem**:

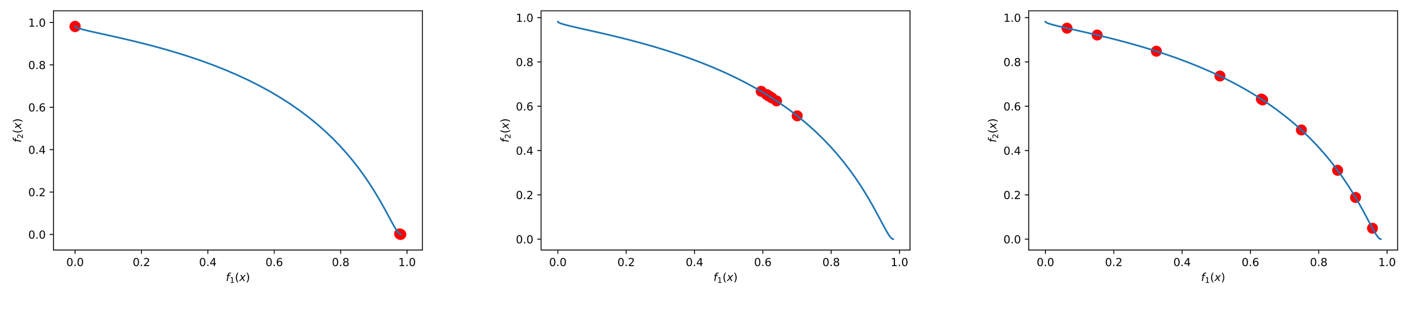
$$\min_\theta \mathcal{L}(\theta) = (\mathcal{L}_1(\theta), \mathcal{L}_2(\theta), \cdots, \mathcal{L}_m(\theta))^{\mathrm{T}}. \tag{1}$$

Traditional linear scalarization:

$$\min_\theta \mathcal{L}(\theta) = \sum_{i=1}^{m} \boldsymbol{w}_i \mathcal{L}_i(\theta). \tag{2}$$

where $\boldsymbol{w}_i$ is manually chosen by MTL practitioners.

However:

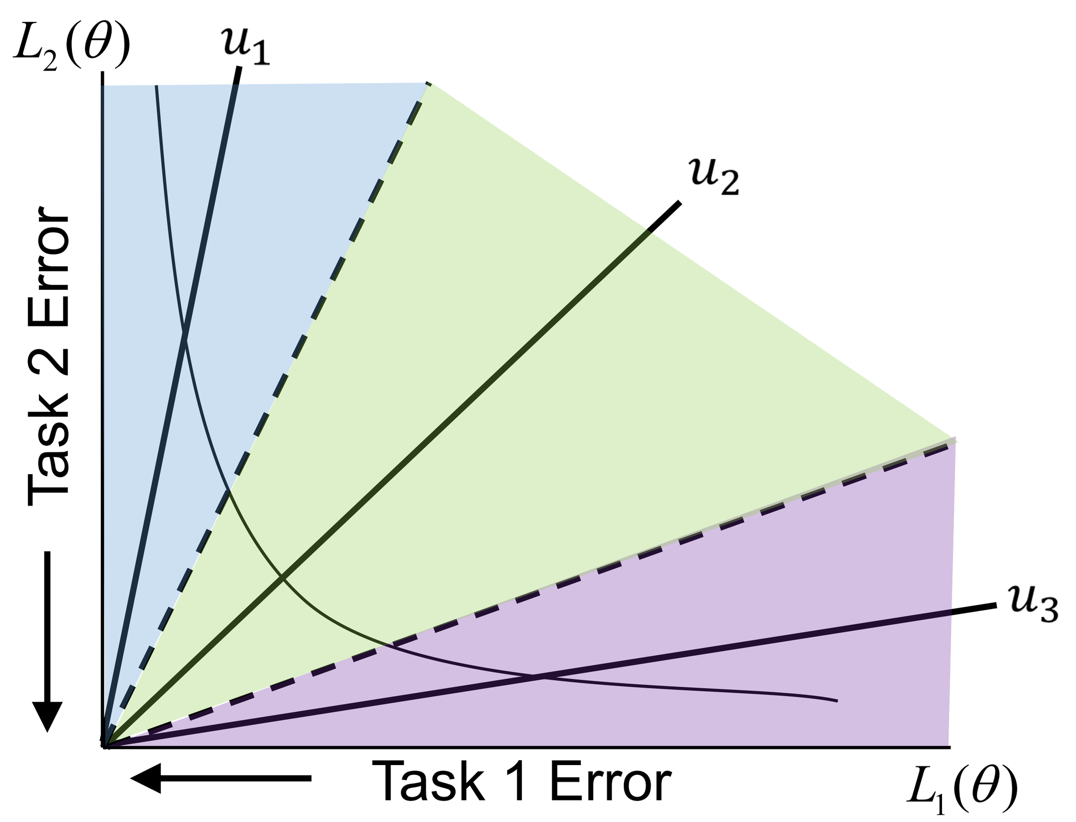


(a) Random Linear Scalarization (b) Adaptive Weight Methods (c) Pareto MTL (Ours)

- **Tasks would be conflicted** with each other when the model capacity is limited and fixed, and there is **no single best solution** for all tasks.
- **Exhaustive weights search** is usually needed for finding the solution with particular trade-off that satisfies the practitioner's need.
- Linear scalarization with fixed weight **can not handle concave Pareto front**.
- Current adaptive-weight methods **can not make different trade-offs** among tasks (GradNorm, Uncertainty, MOO-MTL).

Pareto Multi-Task Learning

Idea and Main Contribution:



- **Decompose** a MTL problem into multiple constrained subproblems with different trade-off preferences.
- **Reformulate** each constrained subproblem as a linear scalarization with dynamically adaptive weights.
- **Solve** all constrained subproblems and obtain a set of Pareto solutions with different trade-offs for the MTL.

Pareto MTL Decomposition:

Pareto MTL decomposes a MTL problem with a set of unit preference vectors $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_K\}$ into K subproblems:

$$\min_\theta \mathcal{L}(\theta) = (\mathcal{L}_1(\theta), \cdots, \mathcal{L}_m(\theta))^{\mathrm{T}}, s.t. \mathcal{L}(\theta) \in \Omega_k, \tag{3}$$

where Ω_k is the attracted subregion with smallest acute angle to $\boldsymbol{u}_k$:

$$\Omega_k = \{\boldsymbol{v} \in R_+^m | \boldsymbol{u}_j^T \boldsymbol{v} \le \boldsymbol{u}_k^T \boldsymbol{v}, \forall j = 1, ..., K\} \tag{4}$$

and it can be further simplified as multiple constraints:

$$\mathcal{G}_j(\theta_t) = (\boldsymbol{u}_j - \boldsymbol{u}_k)^T \mathcal{L}(\theta_t) \le 0, \forall j = 1, ..., K. \tag{5}$$

Solving Each Subproblem:

$\theta_{t+1} = \theta_t + \eta d_t$ with a valid direction to minimize all losses and constraints:

$$\begin{aligned} (d_t, \alpha_t) = &\arg \min_{d \in R^n, \alpha \in R} \alpha + \frac{1}{2}||d||^2 \\ s.t.\ & \nabla \mathcal{L}_i(\theta_t)^T d \le \alpha, i = 1, ..., m. \\ & \nabla \mathcal{G}_j(\theta_t)^T d \le \alpha, j \in I_\epsilon(\theta_t), \end{aligned} \tag{6}$$

where $I_\epsilon(\theta) = \{j \in I | \mathcal{G}_j(\theta) \ge -\epsilon\}$ is the index set of activated constraints.

Pareto MTL as Linear Adaptive Scalarization

Dual Problem:

$$\begin{aligned} \max_{\lambda_i, \beta_j} \ & -\frac{1}{2}||\sum_{i=1}^{m} \lambda_i \nabla \mathcal{L}_i(\theta_t) + \sum_{j \in I_\epsilon(\boldsymbol{x})} \beta_i \nabla \mathcal{G}_j(\theta_t)||^2 \\ s.t.\ & \sum_{i=1}^{m} \lambda_i + \sum_{j \in I_\epsilon(\theta)} \beta_j = 1, \lambda_i \ge 0, \beta_j \ge 0, \forall i = 1, ..., m, \forall j \in I_\epsilon(\theta). \end{aligned} \tag{7}$$

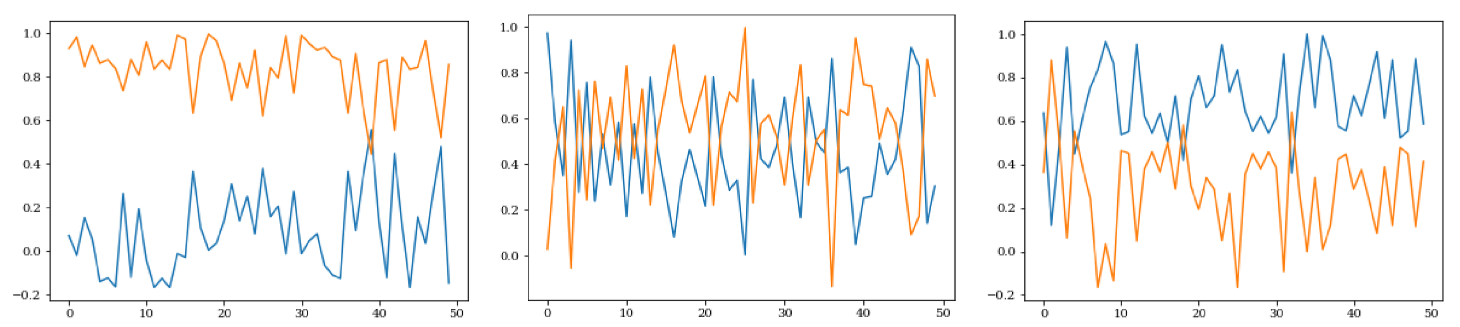
Pareto MTL as Adaptive Linear Scalarization:

$$\mathcal{L}(\theta_t) = \sum_{i=1}^{m} \alpha_i \mathcal{L}_i(\theta_t), \quad \text{where } \alpha_i = \lambda_i + \sum_{j \in I_\epsilon(\theta)} \beta_j (\boldsymbol{u}_{ji} - \boldsymbol{u}_{ki}), \tag{8}$$

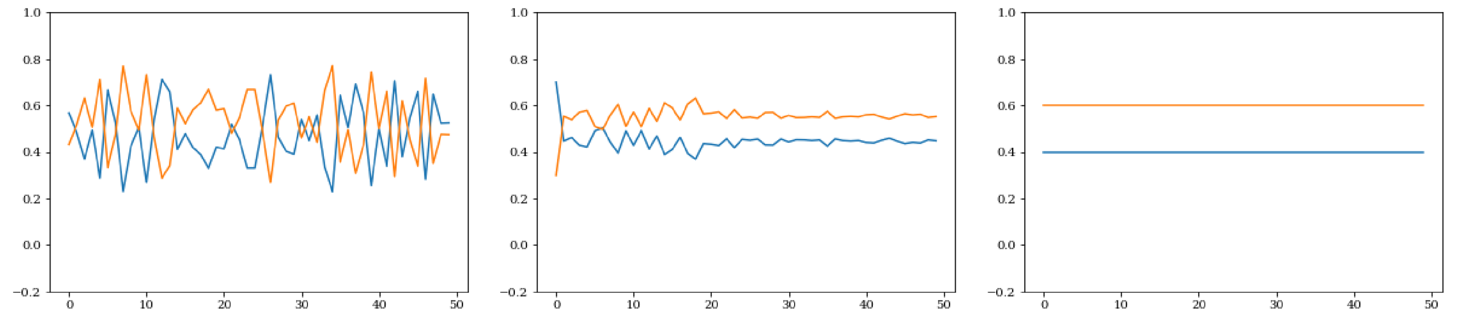
where λ_i and β_j are obtained by solving dual problem (7) with assigned reference vector $\boldsymbol{u}_k$ at each step.

Pareto MTL obtains a set of restricted Pareto optimal solutions with different trade-offs among tasks by solving all subproblems.

Adaptive Weights for Different Algorithms:



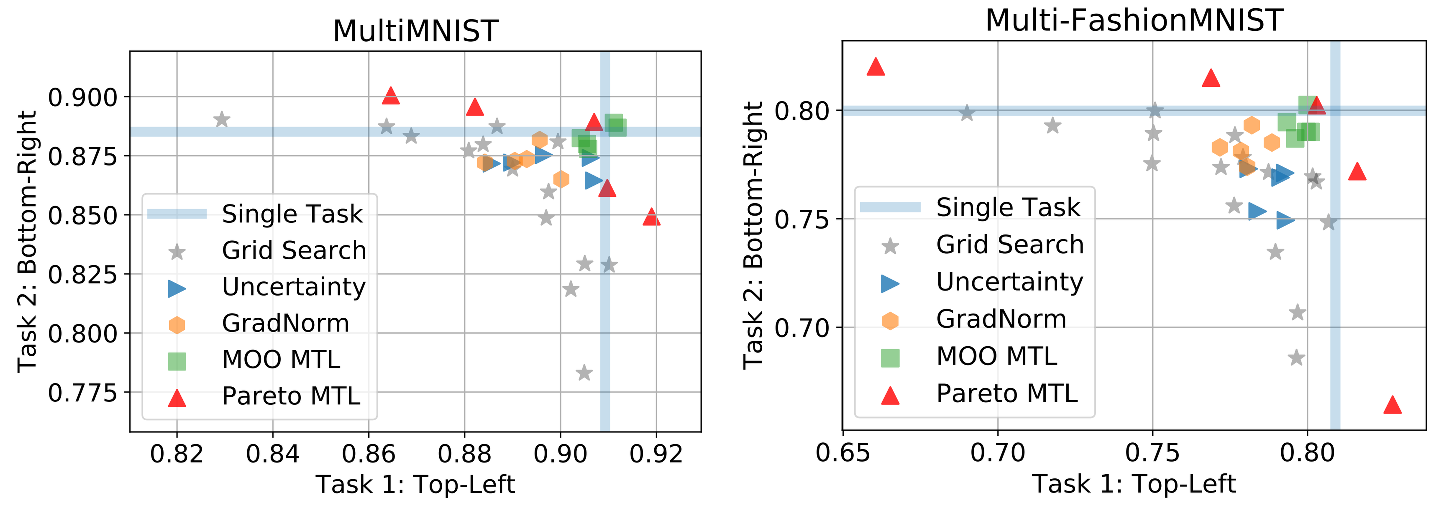
ParetoMTL (1,0) ParetoMTL ($\sqrt{2}/2$, $\sqrt{2}/2$) ParetoMTL (0,1)



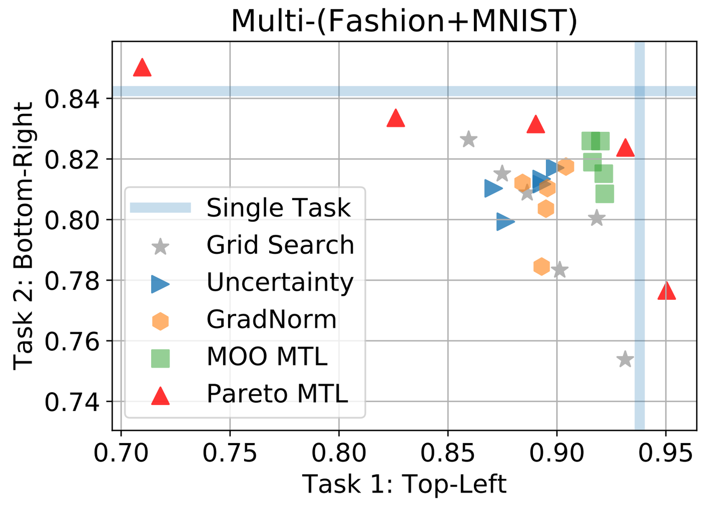
MOO-MTL GradNorm Fixed Weight

Experimental Results

MultiMNIST, MultiFashionMNIST and Multi-(Fashion+MNIST):

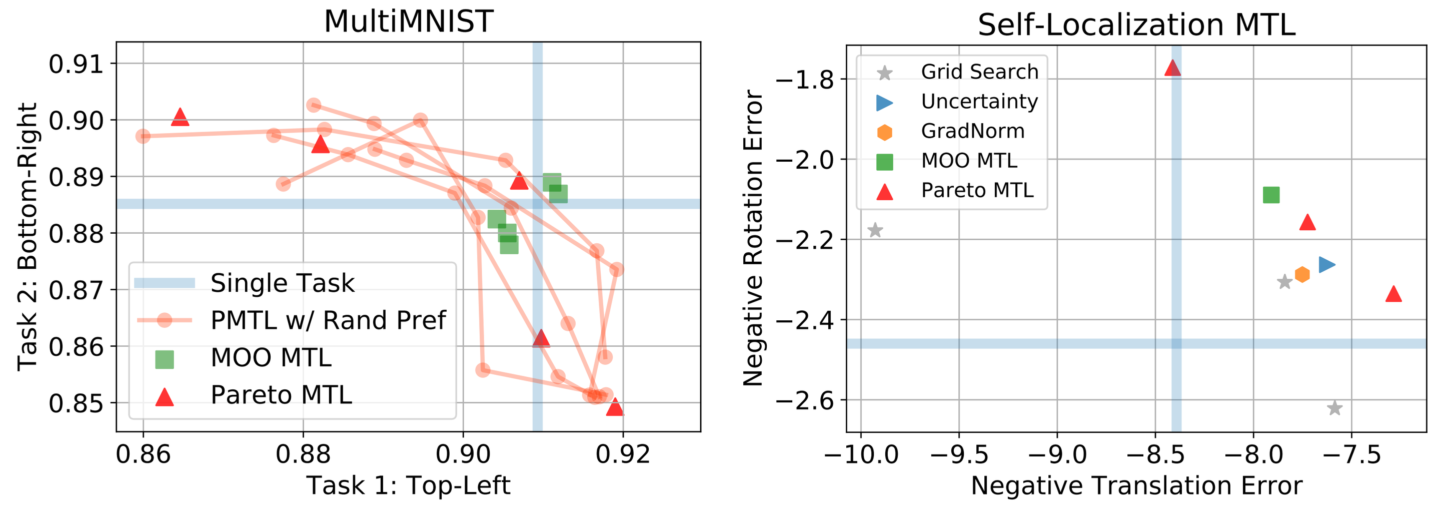


MultiMNIST MultiFashionMNIST



Multi-(Fashion+MNIST)

More details, discussions, and experimental results can be founded in the paper and supplementary material :



Pareto MTL with Random Preferences Self-Driving Car: Localization

Conclusion

Summary:

We proposed a novel Pareto Multi-Task Learning (Pareto MTL) algorithm to generate a set of well-distributed Pareto solutions with different trade-offs among tasks for a given multi-task learning (MTL) problem.

Code & Paper:

Key Reference:

Uncertainty: Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses forscene geometry and semantics. CVPR 2018

GradNorm: Chen et al. Gradnorm: Gradient normal-ization for adaptive loss balancing in deep multitask networks. ICML 2018

MOO MTL: Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. NeurIPS 2018