Achievement-based Training Progress Balancing for Multi-Task Learning

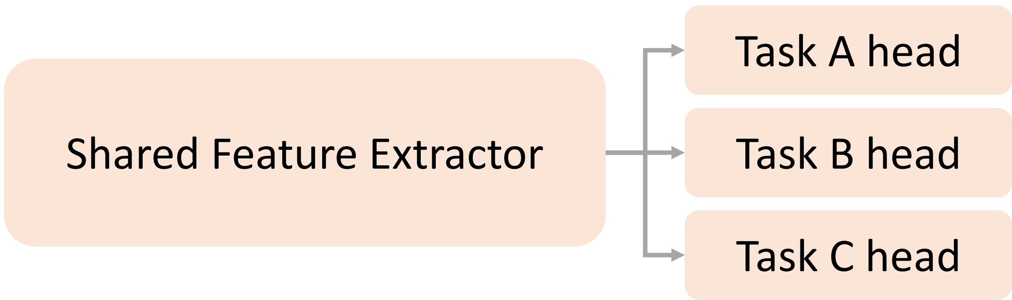
ICCV23 PARIS

Samsung Research

Hayoung Yun and Hanjoo Cho*, Samsung Research

Motivation

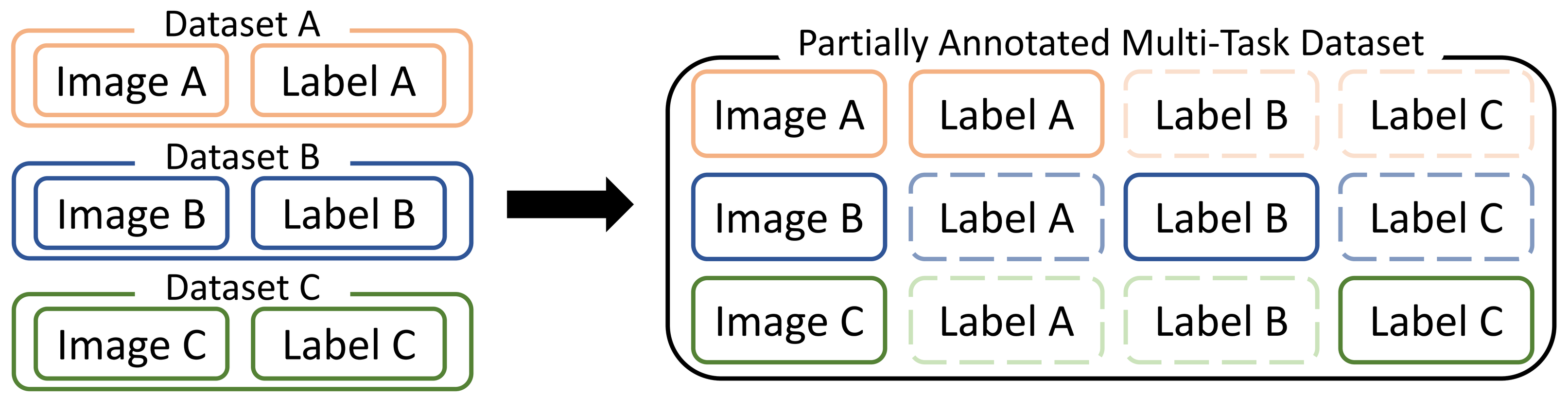
- Vision neural networks generally consist of a feature extractor and a prediction head
  - Most computations are concentrated on the feature extractor
  - Multi-task learning allows the feature extractor to be shared among different tasks
    - Accelerating processing greatly
    - Enabling the learning of more general representation



- Multi-task learning faces two major challenges
  - The high cost of annotating labels of all tasks for plenty of images
  - Balancing the training progress of various tasks with different nature

Partially Annotated Multi-Task Dataset

- To address high cost of annotation, we propose constructing a large scale multi-task dataset by merging task-specific datasets
  - The multi-task dataset is partially annotated because its images are labeled only for the tasks from which they originated



  - The disparity in the number of labels for individual tasks may exacerbate the imbalance in training process among tasks

Previous Work for Balancing Training Progress

- Scale-based methods [RLW, DWA, GLS]
  - Multi-task losses are generally formulated as the weighted sum of task losses

$$L_{total} = \sum_{t=1}^{N_T} w_t L_t$$

L_{total}: total loss
w_t: task weight of task t
L_t: task loss of task t

  - Adjusting task weights to control training progress based on the scale of the task losses
- Gradient-based methods
  - **Magnitude of Gradients**: Modulate task weights to balance the magnitudes of task gradients at the last shared layer [GradNorm, IMTL-G, IMTL]
  - **Directional Conflict**: Directly manipulate task gradients, without designing a multi-task loss, to resolve the directional conflict among task gradients [MGDA, PCGrad, CAGrad]
- Accuracy-based methods [DTP]
  - Control task weights based on the current validation accuracy of each task

Proposed Multi-Task Loss

- Achievement-based task weight
  - Define the *potential* of task accuracy as the accuracy of a single-task model
  - Assess the *achievement* by the ratio of the current accuracy to its potential

$$w_t = \left(1 - \frac{acc_t}{p_t}\right)^{\gamma}$$

acc_t: current accuracy of task t
p_t: single-task accuracy of task t
γ: focusing factor

  - Encourage tasks with low achievements while slowing down tasks converged early
- Weighted geometric mean
  - The multi-task losses formulated as the weighted sum can be easily dominated by the largest one if significant scale differences exist among task losses
  - The geometric mean equitably reflects the variations in all losses, regardless of their magnitude, preventing any single task from dominating the overall loss

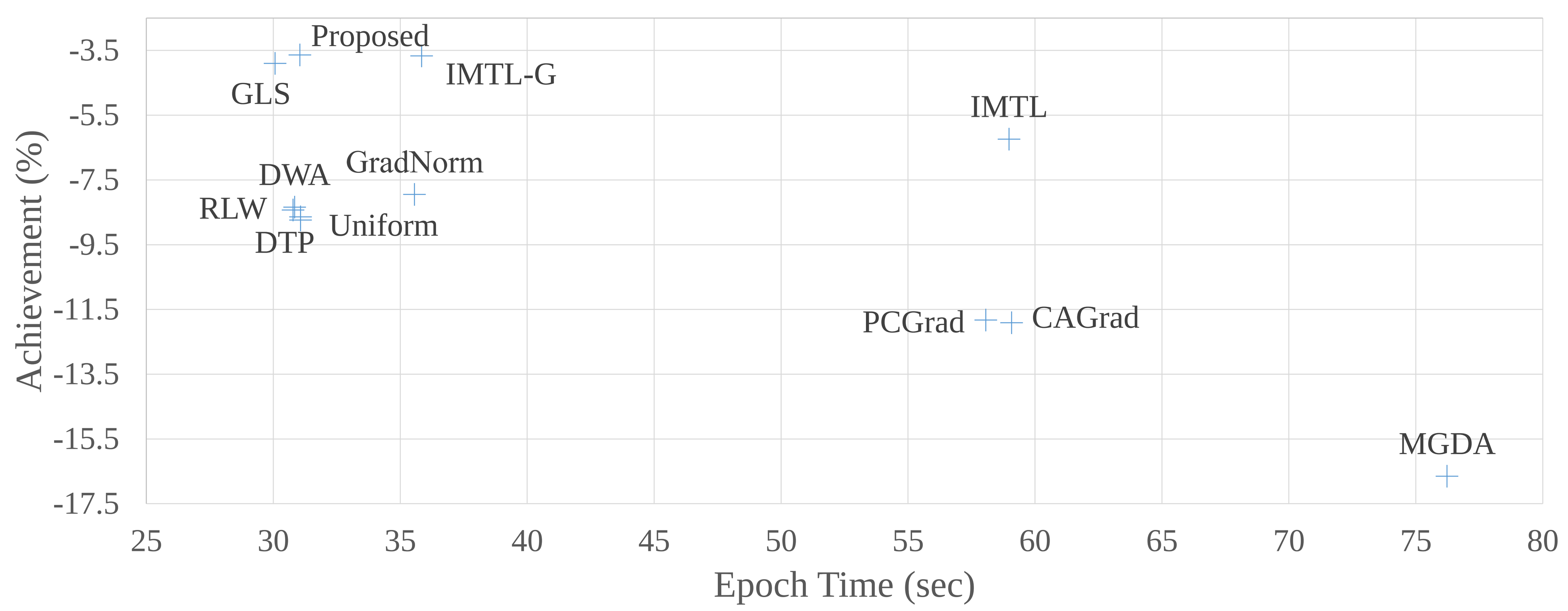
$$L_{total} = \prod_{t=1}^{N_T} L_t^{w_t}$$

L_{total}: total loss
w_t: task weight of task t
L_t: task loss of task t

Experimental Results

- Comparison on the NYU v2 multi-task dataset (795 training images)
  - Configuration
    - Tasks: semantic segmentation, depth estimation, and surface normal
    - Network: Dilated ResNet50 based DeepLabV3 architecture

Comparison of accuracy achievement and training time on the NYU v2 dataset



    - The proposed multi-task loss achieved similar multi-task accuracy to the state-of-the-art loss (IMTL-G), **without incurring training overhead**
  - Ablation Study
    - Whereas DTP considered current accuracy only, the proposed weight considered the *achievement*, thereby improving multi-task accuracy
    - The weighted geometric mean effectively prevented any single task from dominating the loss, resulting in the improvement of multi-task accuracy

	Δacc
DTP	-8.74%
+achievement-based task weight	-6.11%
+ weighted geometric mean	-3.64%

Experimental Results (Cont.)

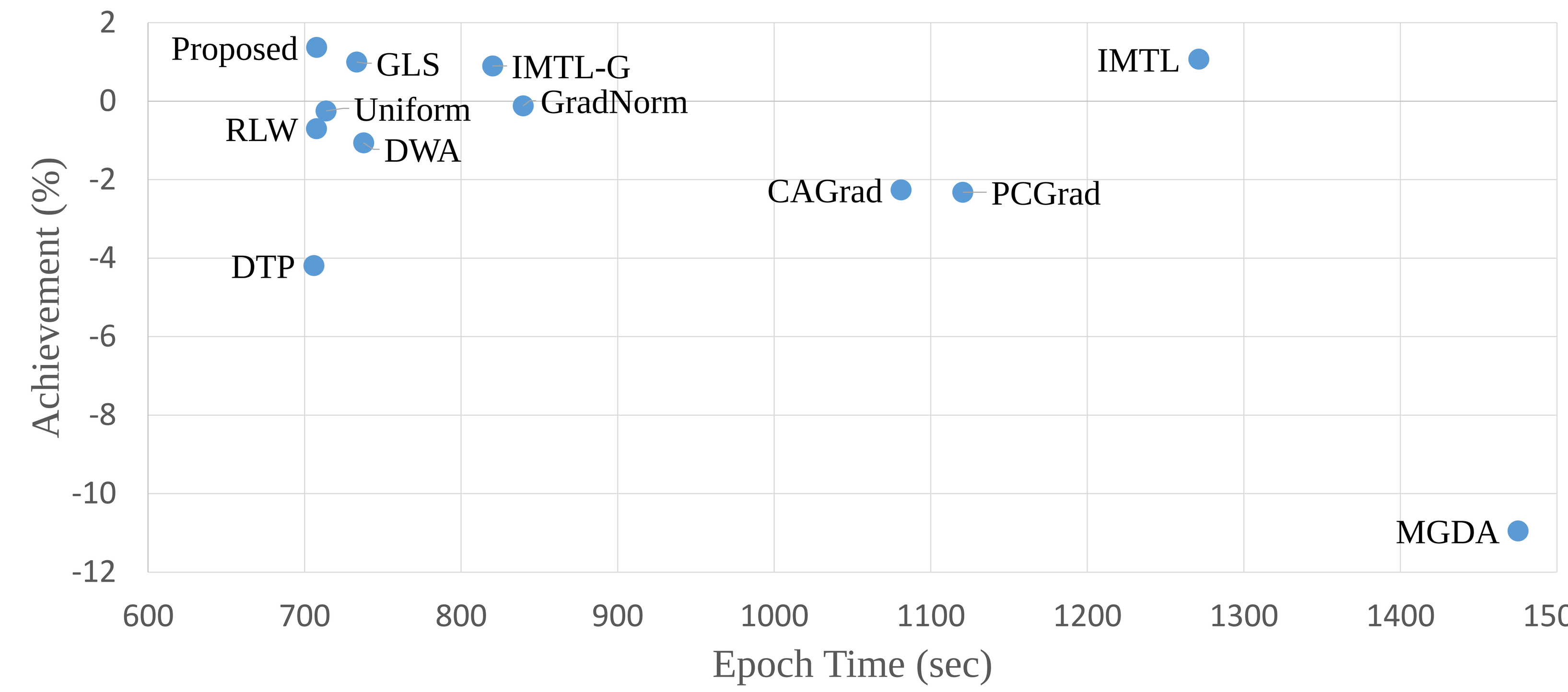
- Effectiveness of the achievement-based weight and weighted geometric mean
  - The proposed task weight also enhanced multi-task accuracy of optimization-based methods resolving gradient conflicts [PCGrad and CAGrad]
  - The weighted geometric mean improved the multi-task accuracy of scale-based [RLW, DWA] and accuracy-based [DTP] methods

	PCGrad	CAGrad
baseline	-11.83%	-11.91%
w/ proposed weight	-8.73%	-8.98%

	RLW	DWA	DTP
arithmetic mean	-8.43%	-8.34%	-8.74%
geometric mean	-5.59%	-4.60%	-4.81%

- Comparison on the PASCAL VOC + NYU dataset (39,446 training images)
  - Configuration
    - PASCAL VOC: object detection (15,215 images) and segmentation (10,477 images)
    - NYU depth: depth estimation (24,231 images)
    - Networks: EfficientNetV2-S based EfficientDet architecture

Comparison of achievement and training time on the PASCAL VOC and NYUD dataset



  - The proposed one outperformed all benchmarks on the partially annotated dataset because **the *achievement* was not disturbed by the imbalance in task labels**

Conclusion

- We addressed the high cost of annotating labels for all tasks by constructing a large scale partially annotated multi-task dataset by integrating task-specific datasets
  - The disparity in the number of task labels may escalate the imbalance in training progress among tasks
- We proposed a novel multi-task loss to balance the training progress of various tasks with different natures
  - We assessed training progress based on the accuracy achievement, successfully balancing the progress of various tasks with different difficulty
  - We employed a weighted geometric mean to capture the variations of task losses regardless of their magnitude, effectively preventing any task from dominating it
- The proposed loss achieved the best multi-task accuracy on both conventional multi-task dataset and partially annotated dataset