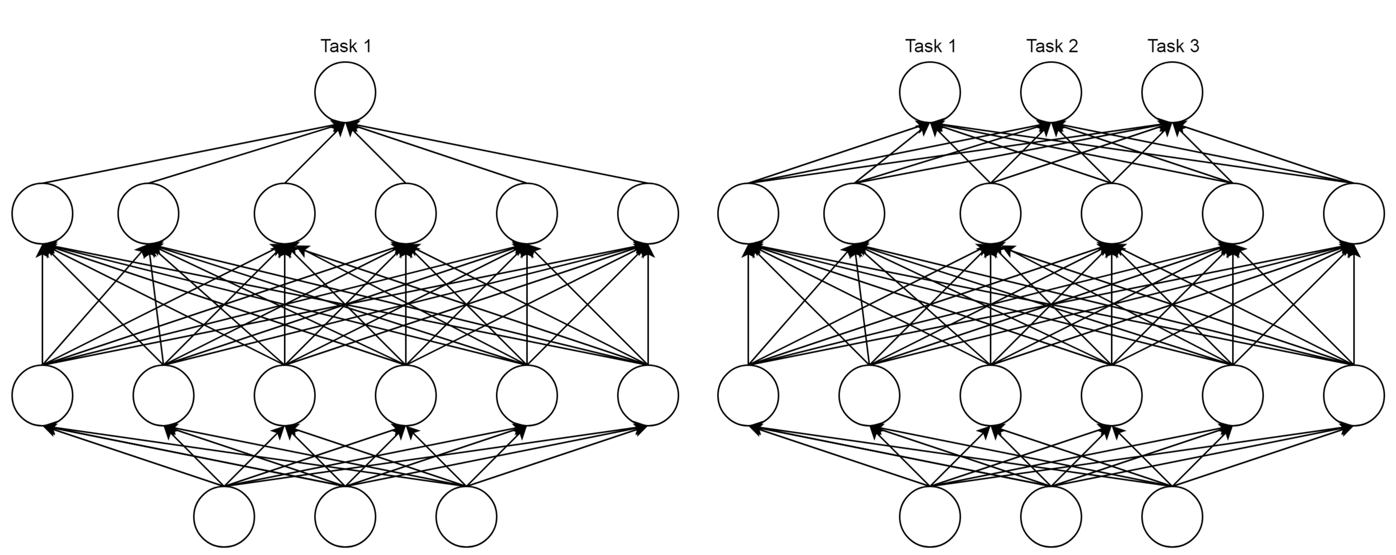
Loss-Balanced Task Weighting to Reduce Negative Transfer in Multi-Task Learning

Shengchao Liu[1,2], Yingyu Liang[1], Anthony Gitter[1,2,3]

Highlights

- Negative transfer: the multi-task performance is worse than a single-task model.
- Demonstrate the presence of negative transfer in a chemistry dataset.
- Present a preliminary algorithm to reduce negative transfer: learning task-specific weights.

Background



(a) Single-Task Learning (b) Multi-Task Learning

Figure 1: Figure 1(a) Single-Task Learning (STL) has T models for T tasks. Figure 1(b) Multi-Task Learning (MTL) has one shared model for T tasks.

- MTL is widely adopted due to better performance when averaged across all tasks.
- For some individual tasks, STL can be better than MTL.

[1]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI;
[2]Morgridge Institute for Research, Madison, WI;
[3]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI

Related Work

- Fine-Tuning
  1. Adopt knowledge.
  2. Pre-trained model.
- GradNorm [1]
  1. Larger loss dominates the training.
  2. A global and static task weight.
- Reinforced Multi-Task Learning (RMTL) [2]
  1. A dynamic task weight.
  2. Cosine similarity of the gradients.

Methods: Loss-Balanced Task Weighting

Loss-Balanced Task Weighting (LBTW) is a combination of ideas from GradNorm and RMTL.

1. A dynamic task weight.
2. Task-specific loss is informative for balancing.
3. Loss ratio determines task weights.

Algorithm 1: Loss-Balanced Task Weighting

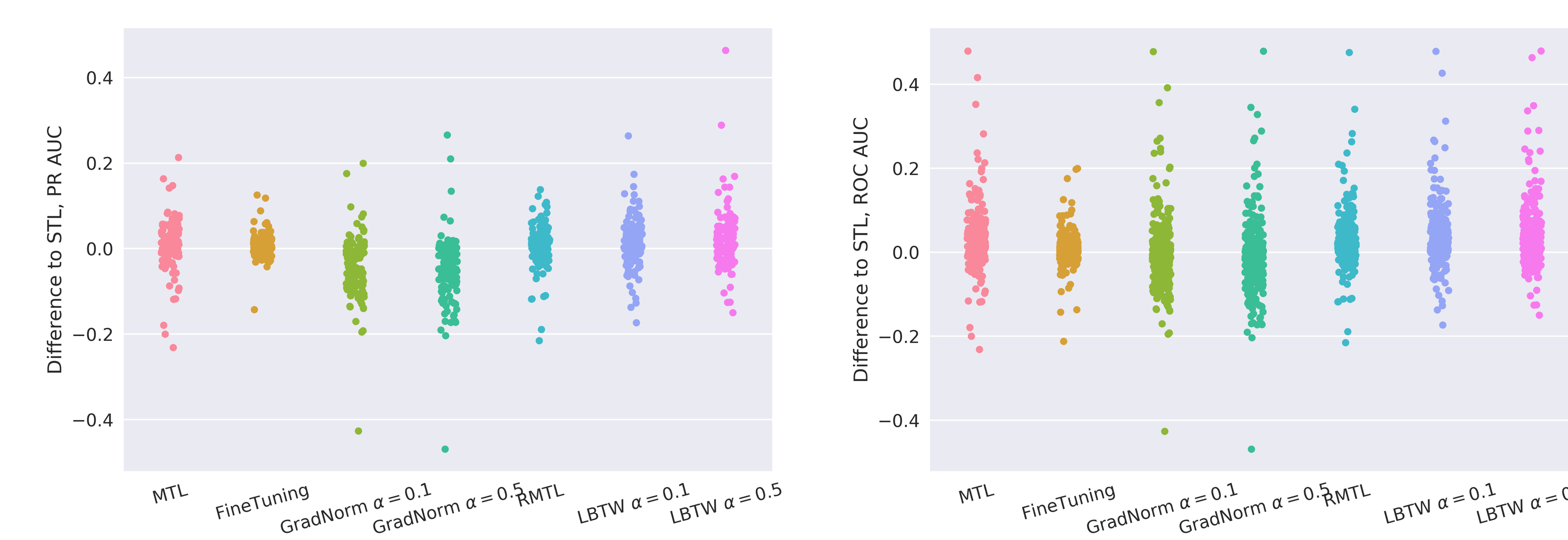
Given T tasks and parameter α.
Initialize neural network weights W.
for each epoch i **do**
  for each batch of data B **do**
    Get the loss on each task $\ell_B \in \mathbb{R}^T$.
    Store the first batch loss as $\ell_{(0,i)} \in \mathbb{R}^T$.
    for each task t **do**
      Set the task weight $w_t = \left(\frac{\ell_{(B,t)}}{\ell_{(0,i,t)}}\right)^\alpha$.
      Update weighted loss $\ell_{(B,t)} = \ell_{(B,t)} \times w_t$.
    end for
    Update W with respect to ℓ_B.
  end for
end for

Implementation is available on GitHub [3].

Experiments

Table 1: Mean PR and ROC AUC on all 128 PubChem BioAssay tasks and the number of tasks with negative transfer.

Evaluation Metric	STL	MTL	FineTuning	GradNorm ($\alpha = 0.1$)	GradNorm ($\alpha = 0.5$)	RMTL	LBTW ($\alpha = 0.1$)	LBTW ($\alpha = 0.5$)
Mean PR AUC	0.232	0.241	0.239	0.189	0.181	0.238	0.247	0.253
# Negative Transfer	-	48	46	98	103	47	45	42
Mean ROC AUC	0.799	0.857	0.806	0.840	0.833	0.852	0.859	0.863
# Negative Transfer	-	13	50	44	46	11	13	11



(a) AUC[PR] (b) AUC[ROC]

Figure 2: Distribution of the change in PR and ROC AUC relative to STL for 128 tasks. Values below 0 indicate tasks with negative transfer.

Conclusions

- No method eliminates negative transfer or even reduces it substantially.
- LBTW has the best overall performance and the fewest tasks with negative transfer.

References

[1] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *International Conference on Machine Learning*, pp. 793–802, 2018.

[2] S. Liu, "Exploration on deep drug discovery: Representation and learning," *Master's Thesis*, vol. TR1854, 2018.

[3] S. Liu, "Loss-balanced task weighting." `https://github.com/chao1224/Loss-Balanced-Task-Weighting`, 2018.