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

Shengchao Liu,[1,2] **Yingyu Liang,**[1] **Anthony Gitter**[1,2,3]
[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
{shengchao, yliang}@cs.wisc.edu, gitter@biostat.wisc.edu

Abstract

In settings with related prediction tasks, integrated multi-task learning models can often improve performance relative to independent single-task models. However, even when the average task performance improves, individual tasks may experience **negative transfer** in which the multi-task model's predictions are worse than the single-task model's. We show the prevalence of negative transfer in a computational chemistry case study with 128 tasks and introduce a framework that provides a foundation for reducing negative transfer in multi-task models. Our Loss-Balanced Task Weighting approach dynamically updates task weights during model training to control the influence of individual tasks.

Introduction

Multi-task learning aims to exploit information from related tasks to improve the generalization performance of all the tasks jointly. Deep learning-based multi-task learning has been successfully applied in chemical screening, genomics, object detection, natural language processing, and other domains. Shared hidden layers in a neural network can transfer knowledge among related tasks, which may reduce overfitting and improve learned latent representations, especially when the task-specific training data is limited. However, when the tasks considered are not sufficiently related, the multi-task setting can be detrimental to performance.

Although the performance may improve on average over all tasks in the multi-task setting, for some specific tasks, the multi-task performance can be worse than a single-task model. This decrease in performance is known as **negative transfer**. Negative transfer occurs naturally in real scenarios and is especially problematic if a subset of tasks is of primary interest and the others are used only to improve the representation learning. We have two conjectures for why negative transfer may happen. (1) All tasks are diverse and unrelated to each other; there is no suitable common latent representation so multi-task learning produces poor representations. (2) One group of related tasks dominates the training process. The performance for those tasks improves as more related tasks are added, but tasks outside the dominant group suffer.

Despite abundant approaches for multi-task learning, few methods aim to improve average task performance while simultaneously minimizing negative transfer. Our contributions are (1) demonstrating the presence of negative transfer in a chemistry dataset and (2) proposing a preliminary algorithm intended to reduce negative transfer by learning task-specific weights. On our computational chemistry case study, this algorithm has the best average performance and fewest tasks with negative transfer.

Methods

Our goal is to design a multi-task learning framework that reduces negative transfer while still improving the average task performance. We consider five neural network-based transfer learning algorithms and a single-task baseline.

Single-Task Learning (STL) and **Multi-Task Learning (MTL)** are fully-connected neural networks, where MTL has shared hidden layers and separate outputs for T tasks in the last layer. **Fine Tuning** is another transfer learning strategy that first trains a MTL model on $T-1$ tasks and then initializes a STL model for the final task with the MTL weights. This strategy transfers the latent representation from the larger multi-task dataset to the single-task dataset so as to alleviate data insufficiency.

Another approach is to apply a task-specific weight vector. **Reinforced Multi-Task Learning (RMTL)** (Liu 2018) uses cosine similarity of the gradients between $T-1$ tasks and a single emphasized task as task weights. **GradNorm** (Chen et al. 2018) assumes that tasks with larger loss dominate the training and therefore learns a balanced global task weight. In GradNorm the task weight is static because it is identical among all batches, whereas RMTL assumes the task weight should be dynamic. Dynamic means that the task weight differs given different inputs. We propose that the challenging tasks can be identified by their loss, and the loss dynamically changes for different batches of data. Therefore, we introduce **Loss-Balanced Task Weighting (LBTW)**, which combines and expands upon ideas from RMTL and GradNorm.

LBTW follows the RMTL framework with dynamic task weights. However, LBTW assumes that the task-specific loss is informative for balancing different tasks. For each task and batch, LBTW considers the loss ratio between the current loss and the initial loss, which is a proxy for how

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

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

well the model has trained for that task. Poorly trained tasks have ratios close to 1 and contribute more to the overall loss and gradient. A hyperparameter α balances the influence of the task-specific weights, and LBTW approaches standard MTL as α goes to 0 (Algorithm 1). Implementation details and model hyperparameters are provided in the Appendix.

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

Results

We test LBTW on the PubChem BioAssay chemistry dataset. It includes 128 tasks and approximately 440,000 chemicals, where each task is a binary classification problem on whether the chemical affects a biological target. The data processing follows (Liu et al. 2018).

We compute the difference in predictive performance for each transfer learning model versus the STL baseline using PR AUC (Figure 1 and Table 1) or ROC AUC (Appendix) for evaluation. In this domain, there are far more inactive than active chemicals, so PR AUC is more meaningful than ROC AUC (Liu et al. 2018). All five approaches improve mean ROC AUC, but on average the GradNorm PR AUC is worse than STL (Table 1). No method eliminates negative transfer or even reduces it substantially for PR AUC. However, LBTW has the best mean PR AUC overall and the fewest tasks with negative transfer, slightly fewer than MTL, Fine Tuning, and RMTL.

Conclusion

Although the preliminary version of LBTW provides only a minor reduction in the number of tasks with negative transfer, the LBTW framework provides flexibility to tune task weights. Currently, LBTW uses uniform task weights when making predictions, but future versions could apply gradient-based task weighting during prediction as well. In

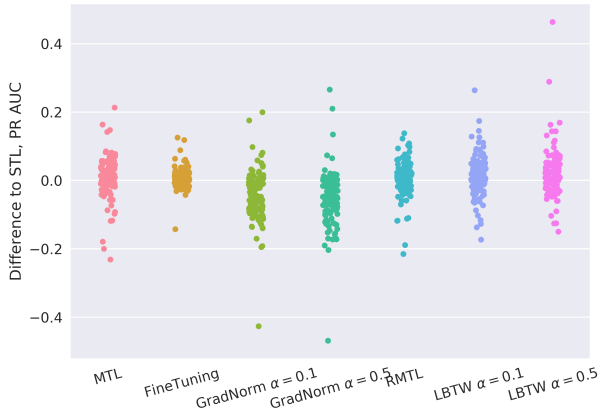

Figure 1: Distribution of the change in PR AUC relative to STL for 128 tasks. Values below 0 indicate tasks with negative transfer. Details and ROC AUC results are in the Appendix.

settings where there is no shared optimal latent representation for all tasks, there will be natural tradeoffs between average performance and instances of negative transfer. LBTW provides a platform to continue to explore and tune that tradeoff.

Acknowledgements

The authors acknowledge computing resources from the University of Wisconsin-Madison Center for High Throughput Computing and GPU hardware from NVIDIA. This work was also supported in part by FA9550-18-1-0166, the Morgridge Institute for Research, and the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. The authors thank Spencer S. Ericksen for transfer learning discussions.

References

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

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

Liu, S.; Alnammi, M.; Ericksen, S. S.; Voter, A. F.; Ananiev, G. E.; Keck, J. L.; Hoffmann, F. M.; Wildman, S. A.; and Gitter, A. 2018. Practical model selection for prospective virtual screening. *bioRxiv* 337956.