Saliency-Regularized Deep Multi-Task Learning

Guangji Bai, Liang Zhao

Corresponding author: liang.zhao@emory.edu

Paper

Code

EMORY UNIVERSITY

Summary

We propose a new multi-task learning (MTL) framework that complements the strength of both shallow and deep multi-task learning scenarios. We propose to model the task relation as the similarity between tasks' input gradients and derive a new regularizer. We proved that the generalizability error has been reduced thanks to the proposed regularizer. Our method achieves state-of-the-art performance on several real-world multi-task learning benchmarks.

Challenges and Motivation

Existing works in deep multi-task learning suffer from the following challenges:

1. Difficulty in regularizing deep non-linear functions of different tasks.
2. Lack of interpretability in joint feature generation and task relation learning.
3. Difficulty in theoretical analyses.

Key motivation

Shallow multi-task learning does NOT suffer from any challenges above:

1. There exists a one-on-one mapping between functions and parameters for the linear model.
2. Linear models are known for great transparency and interpretability.
3. There already exist fruitful theoretical analyses over shallow multi-task learning, e.g., generalization bound., conditions for representer theorems.

Q: Can we achieve the merits of shallow MTL under the deep MTL setting?

Our Solution

We reconsider the feature weights in linear MTL as the **input gradien**t and generalize the feature learning into the non-linear situation by borrowing the notion of **saliency**.

THEOREM 1. *Define $\mathcal{F} := \{f \in \mathbf{C}^1 : f(0) = 0\}$, where $\mathbf{C}^k$ is the family of functions with k^{th}-order continuous derivatives for any non-negative integer k. Given $f_1, f_2 \in \mathcal{F}$, we have:*

$$f_1 = f_2 \quad \textbf{\textit{if and only if}} \quad f_1'(x) = f_2'(x),\ \forall x \in \mathcal{X}$$

Key insights 1:

The theorem above guarantees that, regularizing task functions by the input gradient is equivalent to directly regularizing in the **functional** space.

Key insights 2:

Similar tasks tend to have similar saliency map. Also, saliency is a form of input gradient, i.e., the derivative of the prediction w.r.t. input feature maps.

Proposed Objective Function

$$\min_{h, f_1, \cdots, f_T, \xi} \sum_{t=1}^{T} \mathcal{L}_t(f_t(h(\mathbf{X})), \mathrm{Y}_t), \ \text{s.t.}$$

$$\forall\, i, j,\ dist(\nabla_A f_i, \nabla_A f_j) \le \xi_{ij}\ , \ \sum_{1 \le i < j \le T} \xi_{ij} \le \alpha$$

- f_t Task-specific layers for task t
- h Shared representation layers
- A Feature map from the last layer of h
- $\nabla_A f_i$ Gradient of task i's prediction w.r.t A
- ξ_{ij} Slack variable
- $dist(\,)$ Some distance measure, e.g., ℓ_1, ℓ_2

By Lagrangian method,

$$\min_{h, f_1, \cdots, f_T, \omega} \sum_{t=1}^{T} \mathcal{L}_t(f_t(h(\mathbf{X})), \mathrm{Y}_t) + \lambda \cdot \sum_{1 \le i < j \le T} \omega_{ij} \cdot dist(\nabla_A f_i, \nabla_A f_j)$$

$$\text{s.t.} \ \forall\, i, j,\ \omega_{ij} \ge 0 \text{ and } \sum_{1 \le i < j \le T} \omega_{ij} \ge \beta$$

where $\{\omega_{ij}\}_{1 \le i < j \le T}$ is a set of learnable parameters to explicitly model task relations.

Theoretical Contribution

THEOREM 2 (GENERALIZATION ERROR). *Let $\delta > 0$ and $\mu_1, \mu_2, \ldots, \mu_T$ be the probability measure on $\mathcal{X} \times \mathbb{R}$. With probability of at least $1 - \delta$ in the draw of $\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \sim \prod_{t=1}^{T} \mu_t^n$, we have:*

$$\mathcal{E}(\hat{h}, \hat{\boldsymbol{f}}) - \mathcal{E}(h^*, \boldsymbol{f}^*) \le c_1 L \frac{G(\mathcal{H}(\boldsymbol{X}))}{nT} + c_2 B \frac{\sqrt{\lambda_{min}^{-1}} \sup_h \|h(\boldsymbol{X})\|}{n\sqrt{nT}} + \sqrt{\frac{8 \ln(4/\delta)}{nT}}$$

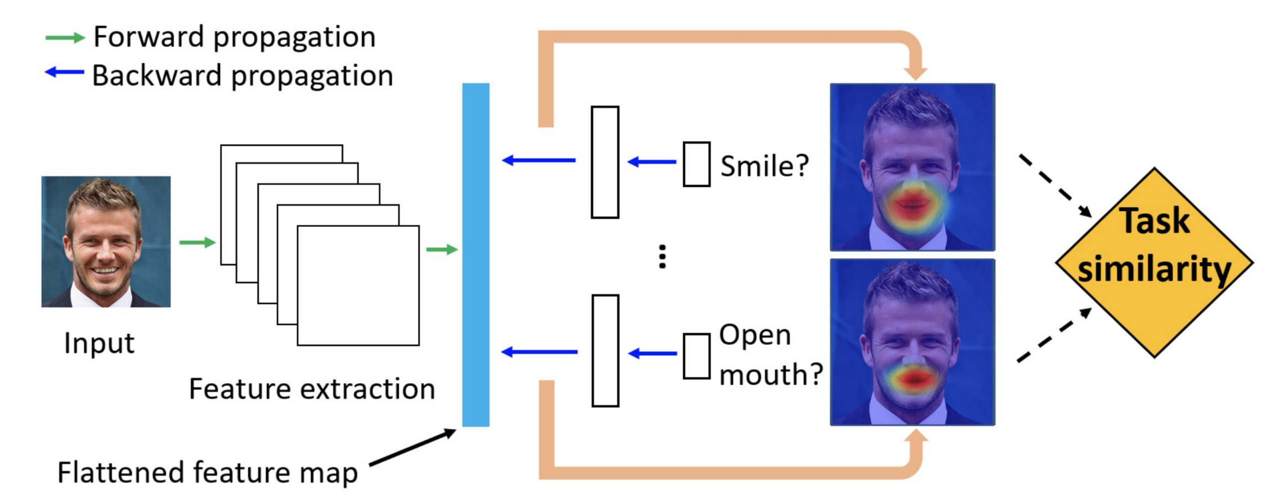
where $\sum_{i,j=1}^{T} \omega_{ij} \cdot dist^2(\nabla_A f_i, \nabla_A f_j) \le B^2$

Remark: By minimizing the proposed regularizer, B may take smaller value thus tightening the generalization error bound, i.e., smaller generalization error.

In our paper, we **also** proved that:

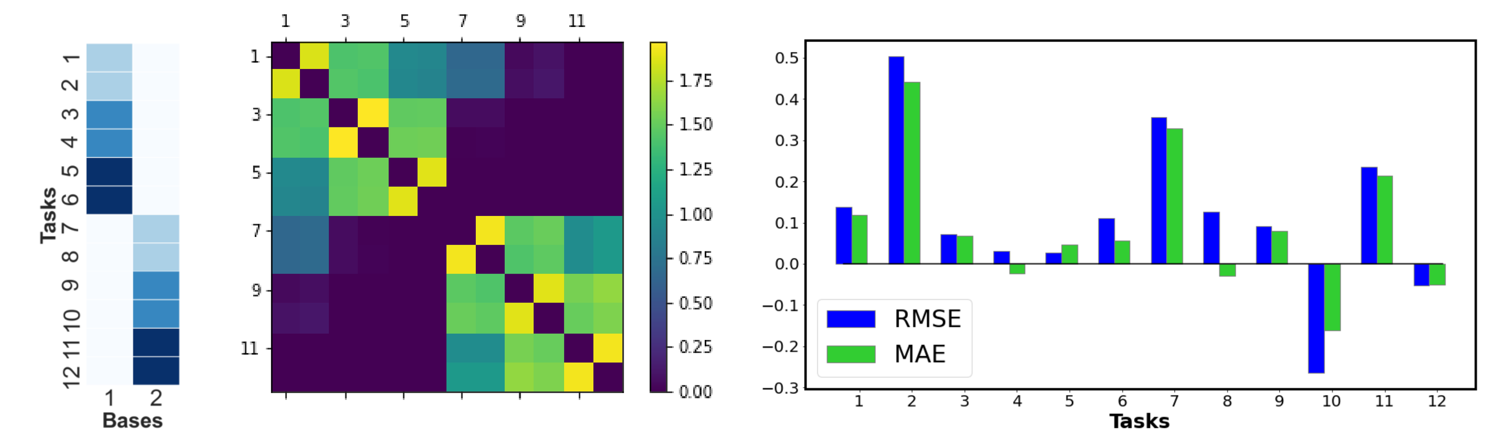
1. SRDML is a natural generalization of shallow multi-task learning
2. Hard/Soft-parameter sharing are special cases of SRDML

Framework Architecture



Experiments

Controlled synthetic dataset



Left to right: Ground-truth of each task's feature weights; Task relation learned by SRDML; Performance improvement of SRDML over single-task learning.

- Twin tasks (same weight, e.g., Tasks 1 & 2) show extremely strong similarity
- Tasks from the same base (same weight sign, e.g., Tasks 1 & 3) show strong similarity
- Tasks from different bases (opposite weight sign, e.g., Tasks 1 & 12) show very strong dissimilarity.

Real-world datasets

Model	CIFAR-MTL				CelebA			
	Accuracy	AUC	Precision	Recall	Accuracy	AUC	Precision	Recall
STL	92.65	66.20	71.32	69.83	86.83	90.96	70.53	60.39
Hard-Share	94.70	95.56	76.30	72.28	89.24	91.38	71.40	58.84
Lasso	91.48	86.64	68.90	24.74	76.55	66.69	37.38	36.62
L21	91.50	87.58	68.01	29.32	76.09	66.12	37.11	36.13
RMTL	92.28	85.65	61.54	28.15	75.52	66.99	37.48	36.74
MRN	94.51	96.67	79.94	76.95	89.35	91.54	71.51	64.64
MMoE	93.53	93.17	73.42	69.32	77.57	67.84	68.79	58.92
PLE	94.01	93.32	75.26	70.15	83.21	69.32	70.03	59.72
MGDA-UB	90.74	84.38	57.80	24.10	90.03	92.92	73.42	62.65
PCGrad	95.11	96.69	79.03	74.82	90.11	92.87	73.51	62.92
SRDML	95.82	96.43	81.22	75.93	90.15	92.95	73.87	64.91
SRDML (w/. PCGrad)	**96.03**	**96.72**	**82.59**	**77.01**	**90.26**	**93.01**	**73.93**	**65.30**

Our method (SRDML) outperforms both shallow and deep multi-task learning methods on CIFAR-MTL and CelebA benchmarks. Refer to our paper for more details.