Saliency-Regularized Deep Multi-Task Learning

Guangji Bai  
Emory University  
Atlanta, GA  
guangji.bai@emory.edu

Liang Zhao∗  
Emory University  
Atlanta, GA  
liang.zhao@emory.edu

ABSTRACT
Multi-task learning (MTL) is a framework that enforces multiple learning tasks to share their knowledge to improve their generalization abilities. While shallow multi-task learning can learn task relations, it can only handle pre-defined features. Modern deep multi-task learning can jointly learn latent features and task sharing, but they are obscure in task relation. Also, they pre-define which layers and neurons should share across tasks and cannot learn adaptively. To address these challenges, this paper proposes a new multi-task learning framework that jointly learns latent features and explicit task relations by complementing the strength of existing shallow and deep multitask learning scenarios. Specifically, we propose to model the task relation as the similarity between tasks’ input gradients, with a theoretical analysis of their equivalence. In addition, we innovatively propose a multi-task learning objective that explicitly learns task relations by a new regularizer. Theoretical analysis shows that the generalizability error has been reduced thanks to the proposed regularizer. Extensive experiments on several multi-task learning and image classification benchmarks demonstrate the proposed method’s effectiveness, efficiency as well as reasonableness in the learned task relation patterns.

CCS CONCEPTS
• Computing methodologies → Multi-task learning; Regularization.

KEYWORDS
Multi-task Learning, Task Relation, Saliency Detection

ACM Reference Format:

1 INTRODUCTION
Multi-task learning (MTL, [9]) is an important research domain based on the idea that the performance of one task can be improved using related tasks as inductive bias. While traditional shallow MTL methods can fit the models for individual tasks and learn task relations, they do not focus on generating features from scratch and instead rely on pre-defined and explicit features [51]. More recently, deep representation learning empowers MTL to go “deep” by equipping it with the capacity to generate features while fitting the tasks’ predictive models. Deep MTL is usually categorized according to the ways of correlating tasks’ models into two major types: hard-parameter sharing and soft-parameter sharing. Hard-parameter sharing methods [28, 39, 53] essentially hard-code which part of neurons or layers to share and which part does not for different tasks instead of doing it adaptively. Moreover, they usually share the layers for representation learning (e.g., convolutional layers) but not those for decision making (e.g., fully-connected layers for classification). On the other hand, soft-parameter sharing methods [12, 33, 48] do not require to hard-code the sharing pattern but instead build individual models for each task and “softly” regularize the relatedness among them. Hence, soft-parameter sharing has better flexibility in learning the task relation, while may not be efficient since its model parameters increase linearly with the number of tasks. Hard-parameter sharing, by contrast, is more “concise” but requires pre-define which parts are shared or not.

Therefore, although MTL is a long-lasting research domain, it remains a highly challenging and open domain that requires significantly more efforts to address challenges such as the trade-off between model flexibility and conciseness of hard- and soft-parameter sharing mentioned above. Although more recently, there have come a few attempts trying to alleviate the dilemma, such as those regularizing task relationships in task-specific layers in hard-parameter sharing to achieve knowledge transfer in unshared layers [28] and those adaptively learning which part to share or not by methods like branching [18, 29] or Neural Architecture Search [44], the research frontiers still suffer from several critical bottlenecks, including (1) Difficulty in regularizing deep non-linear functions of different tasks. Adaptively learning task relation requires regularizing different tasks’ predictive functions, which, however, are much harder to achieve for nonlinear-nonparametric functions since it requires regularizing in the whole continuous domain of input. To work around it, existing works [28, 43] typically resort to a reduced problem which is to regularize the neural network parameters. Notice that this reduction deviates from the original problem and is over-restricted. For example, first, two neural networks with different permutations of latent neurons can represent the same function [8]. Moreover, even if they have different architectures, they can still possibly represent the same function [23]. This gap deteriorates the model’s generalizability and effectiveness. (2) Lack of interpretability in joint feature generation and task relation learning. Despite incapability of generating features, shallow
MTL enjoys good interpretability since they learn explicit task correlations via how the hand-crafted features are utilized. However, in deep MTL, the generated features do not have explicit meaning and how the black-box models relate to each other is highly obscure. It is imperative yet challenging to increase the interpretability of both generated features and task relation. (3) Difficulty in theoretical analysis. While there are fruitful theoretical analyses on shallow MTL, such as on generalization error [4–6] and conditions for regularized MTL algorithms to satisfy representor theorems [2, 3], similar analyses meet strong hurdles to be extended to deep MTL due to the difficulty in reasoning about neural networks whose feature space is given by layer-wise embeddings [47]. It is crucial to enhance the theoretical analyses on the model capacity and theoretical relation among different deep MTL models.

This paper proposes a new Saliency-Regularized Deep Multi-task Learning (SRDML) framework to solve the challenges mentioned above. First, we reconsider the feature weights in traditional linear multitasking as the input gradient and then generalize the feature learning into the non-linear situation by borrowing the notion of saliency detection. Second, we recast the task relation problem as the similarity among saliency regions across tasks so as to regularize and infer the task relation. Third, to validate our hypothesis, we have given a theoretical analysis of their equivalency. Meanwhile, we also provide theoretical analysis on how the proposed regularization helps reduce the generalization error. Finally, we demonstrate our model’s effectiveness and efficiency on synthetic and multiple large-scale real-world datasets under comparison with various baselines.

2 RELATED WORK

Multi-task learning (MTL). Readers may refer to [11, 51] for a more comprehensive survey on MTL. Before the popularity of deep learning, traditional MTL usually focuses on hand-crafted features and can be generally divided into two categories: 1). multi-task feature learning, which aims to learn a shared/similar feature selection, latent space, or model parameters [1, 14]. 2). multi-task relation learning, which aims to quantify task relatedness via task clustering [19, 22] or task co-variance [16, 52]. However, they rely on hand-crafted features and the separation of feature generation and task learning may result in sub-optimal performance.

More recently, MTL takes advantage of the advancement of deep neural networks which can directly take raw, complex data such as images, audio, and texts and learn in an end-to-end manner. Deep MTL integrates feature generation and task learning and simultaneously learns both of them [7]. In this domain, hard-parameter sharing [34, 53] requires to hard-code which part of the network is shared and which is not. Existing work usually shares the lower-level layers for representation learning (e.g., convolutions) while make higher-level layers (e.g., those for classification) separated across tasks. Some existing works extend hard-parameter sharing by considering Neural Architecture Search [13] like [29, 44], and [18], soft-parameter sharing based method has better flexibility where each task has its own models and regularization is used to enforce task relatedness by aligning their model parameters [12, 33, 48]. To achieve both hard-parameter sharing’s conciseness and efficiency and soft-parameter sharing’s flexibility, some recent work [28, 43] shares the representation learning layers while exploits task relations in task-specific layers.

Saliency detection. Saliency detection is to identify the most important and informative part of input features. It has been applied to various domains including CV [17, 37], NLP [24, 35], etc. The salience map approach is exemplified by [50] to test a network with portions of the input occluded to create a map showing which parts of the data actually have an influence on the network output. In [40], a salience map can be created by directly computing the input gradient. Since such derivatives can miss important aspects of the information that flows through a network, a number of other approaches have been designed to propagate quantities other than gradients through the network. In CV domain, Class Activation Mapping (CAM, [55]) modifies image classification CNN architectures by replacing fully-connected layers with convolutional layers and global average pooling [25], thus achieving class-specific feature maps. Grad-CAM [38] generalizes CAM by visualizing the linear combination of the last convolutional layer’s feature map activations and label-specific weights, which are calculated by the gradient of prediction score w.r.t the feature map activations. Grad-CAM invokes different versions of backpropagation and/or activation, which results in aesthetically pleasing, heuristic explanations of image saliency. While there exist some other saliency-based methods along this research line including Guided Propagation [42], Deconvolutional Network [50], etc., they are designed only for specific architectures like ReLU Network for Guided Propagation.

3 PROPOSED METHOD

In this section, we introduce our proposed Saliency-regularized Deep Multi-task Learning (SRDML) method. We first review the pros and cons for each MTL method and describe our main motivation, then formally introduce our model and its objective function.

3.1 Problem Formulation

Consider a multi-task learning problem with T tasks such that a dataset \( \{X, Y_1, Y_2, \ldots, Y_T\} \) is given with i.i.d training samples \( X_t = \{x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)}\}, Y_t = \{y_1^{(t)}, y_2^{(t)}, \ldots, y_n^{(t)}\} \), where \( n \) is the sample size and \( (x_i^{(t)}, y_i^{(t)}) \) is a pair of input and label such that \( x_i^{(t)} \in X \) and \( y_i^{(t)} \in Y \), \( i = 1, 2, \ldots, n \) and \( t = 1, 2, \ldots, T \).

Mathematically, consider a predictor \( g \) which factorizes as \( g = f \circ h \), where “\( \circ \)” stands for functional composition. The function \( h : X \rightarrow R^K \) is called the feature or representation extraction part and is shared for all tasks, while \( f : R^K \rightarrow R \) is a function defined on \( R^K \), a predictor specialized to each task at hand. K

<table>
<thead>
<tr>
<th>Model</th>
<th>Transparency in task relation</th>
<th>Feature generation</th>
<th>Conciseness and efficiency</th>
<th>Flexibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shallow MTL</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Hard-parameter sharing</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Soft-parameter sharing</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SRDML</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 1: Comparison over different MTL methods.
We propose a simple framework that can innovatively achieve all weights in shallow MTL as input gradient. 1) work: very commonly seen in deep MTL problems such as multi-task functional towards the space of explore an innovative alternative so that we can more easily work due to their gap with the actual function. For example, neural net-tic features and learns task relation to correlate different tasks with interpretability. This goal poses significant challenges to existing work: 1) Directly regularizing the prediction function of different tasks is extremely hard. Existing work considered a reduced problem by regularizing the feature weights of different \( f \) which is over-restricted. 2) How to learn interpretable task relations with deep/implicit features is still unclear. 3) Theoretical analysis is rare in deep MTL due to the non-linear and non-parametric functions of \( h \) and \( f \). To jointly solve these challenges, we reconsider the feature weights in shallow MTL as input gradient, i.e., \( \partial f(x)/\partial x \), and generalize the feature learning into the deep network by considering the saliency detection methods.

### 3.2 Motivations

We propose a simple framework that can innovatively achieve all the goals, as shown in Figure 1.

To achieve model conciseness and efficiency as well as task relatedness flexibility, we share the representation learning across different tasks. This is based on important neuro-inspirations: human sensory organs and retina are the same for all different tasks (meaning the convolution layers are shared). On the other hand, the working memory will leverage the long-term memory for each task and related tasks will have related memory (i.e., model) and their relatedness can be considered as the similarities of activation patterns for different tasks, namely the similarity among the saliency maps for different tasks.

Then, the next question is how to regularize the relation among different tasks, namely how to regularize the (dis)similarity of the predictive functions of different tasks. As mentioned above, it is problematic to directly regularize the neuron network parameters due to their gap with the actual function. For example, neural networks with different architecture or neuron permutation could represent the same function. Therefore, this motivates us to explore an innovative alternative so that we can more easily work towards the space of functional. Specifically, we propose to regularize first-order derivatives with respect to the input of different tasks. This new strategy has two crucial merits: First, it is mathematically equivalent to directly regularize the function without the gap in existing works mentioned above. Second, it also finds inspiration from the saliency map domain and comes with strong interpretability in how tasks correlate.

**Key Merit 1:** Regularizing task functions without theoretical gap. Specifically, Theorem 1 below tells us that enforcing multiple tasks to have similar input gradients is equivalent to encouraging those tasks themselves to be similar.

**Theorem 1.** Define \( \mathcal{F} := \{ f \in \mathbb{C} : f(0) = 0 \} \), where \( \mathbb{C}^k \) 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:

\[
\text{\( f_1 \equiv f_2 \) if and only if \( f_1'(x) = f_2'(x), \forall x \in \mathbb{X} \) (1)}
\]

**Proof.** Please refer to the appendix for the formal proof. □

Our analysis above allows us to regularize the prediction functions of different tasks in the functional space instead of parameter space. The assumption over function family \( \mathcal{F} := \{ f \in \mathbb{C} : f(0) = 0 \} \) is reasonable in practice since an all-zero input \( x \) simply corresponds to a "black" picture, and for any task we assume a black picture contains no useful information and should be classified as the negative sample (i.e., ground-truth label should be 0).

**Key Merit 2:** Inspiration from saliency map and enhancement of interpretability. Evaluating task relation with derivative similarity has justification from saliency perspective. A saliency is a derivative of the prediction score w.r.t. input features, and it denotes the semantic features that influence the prediction most. In addition, similar tasks tend to have similar saliency, while dissimilar tasks tend to have dissimilar saliency. As shown in Figure 2, we enforce higher-level semantic features as saliency.

Many previous work have asserted that deeper representations in a CNN capture higher-level visual constructs [7, 31]. Furthermore, convolutional layers naturally retain spatial information which is lost in fully connected layers, so we expect the last convolutional layers to have the best compromise between high-level semantics and detailed spatial information. By following a recent work called Grad-CAM [38], we use the gradient information flowing into the last convolutional layer of the CNN to capture the saliency map to each neuron for a particular task or class of interests.

### 3.3 Objective Function

We first give a formal definition of saliency. For example, in computer vision, given an input image \( I \), a classification ConvNet \( f \)
where $W = \sum_{1 \leq i < j \leq T} \omega_{ij}$. Thanks to our normalization trick, the overall objective of SRDML is differentiable and can be trained in an end-to-end manner. We use standard gradient descent (e.g., Adam [20]) to solve Eq. (4), where we aim to learn multiple tasks and the task relationship simultaneously. Although the normalization trick introduced in Eq. (4) no longer guarantees that the hard constraint of the lower bound of all $\omega_{ij}$ can be strictly satisfied, our empirical results show that our normalization trick works well in practice and SRDML can capture reasonable task relationship by optimizing Eq. (4) with finetuned hyperparameters.

A general overview of our SRDML architecture can be found in Figure 3. First, the input image is fed into a shared feature extractor which in our case is implemented by a sequence of convolutional layers. Right after the feature extraction process, we obtain a set of flattened feature map (shown as the blue bar in Figure 3) which contains high-level semantic information with respect to the original image [38]. On top of the feature map, each task-specific head will first calculate the saliency map with respect to its own prediction. Based on the saliency map for all the tasks, the task similarity can be calculated via some distance measure. Note that our overall framework is differentiable and can be trained in an end-to-end manner.

4 THEORETICAL ANALYSIS

In this section, we present the theoretical analyses of our SRDML model. First, we prove that our proposed regularizer can help reduce the generalization error. Second, we formally analyze the relation between SRDML and other MTL methods. We put all formal proofs in the appendix due to the limited space.

4.1 Generalization Error Bound

Here we show the generalization bound of our model. The major contribution is that we theoretically proved that our proposed regularization term can help reduce the generalization error.

For simpler notation, define

$$\mathcal{F}_{e(\alpha)} := \{ f \in F^T : \forall 1 \leq i < j \leq T, x \in X, \quad \text{dist}(\nabla f_i, \nabla f_j) \leq \epsilon_{ij}, \sum_{1 \leq i < j \leq T} \epsilon_{ij} \leq \alpha \}$$

where $f = (f_1, f_2, \cdots, f_T)$ is the vectorization of each task’s function, and $\{\epsilon_{ij}\}_{1 \leq i < j \leq T}$ is a set of global slack variables. Hence, the optimization problem of Eq. (2) can be simplified as

$$\min_{h, f_1, \cdots, f_T, \alpha} \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} L_t(f_i(h(x)^t)), y^{(t)}$$

Before introducing the theorem, we make the following standard assumptions over the loss function:

**Assumption 1 ([32]).** The loss function $L$ has values in $[0, 1]$ and has Lipschitz constant 1 in the first argument, i.e.: (1) $L(y, y') \in [0, 1]$ (2) $L(y, y') \leq y, \forall y'$.

Different Lipschitz constants can be absorbed in the scaling of the predictors and different ranges than $[0, 1]$ can be handled by a simple scaling of our results.

**Definition 1 (Expected risk, Empirical risk).** Given any set of function $h, f_1, \cdots, f_T$, we denote the expected risk as:
Given the data \( Z = (X, Y) \), the empirical risk is defined as:
\[
\hat{E}(h, f_i, \cdots, f_T | Z) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(f_i(h(x_i)), y_i^{(t)})
\]

**Definition 2 (Global optimal solution, Optimized solution).** Denote \((h^*, f^*)\) as the global optimal solution of the expected risk:
\[
(h^*, f^*) := \arg \min_{h \in \mathcal{H}, f \in \mathcal{F}_{\tau(a)}} E(h, f, \cdots, f_T)
\]
and denote \((\hat{h}, \hat{f})\) as the optimized solution by minimizing the empirical risk:
\[
(\hat{h}, \hat{f}) := \arg \min_{h \in \mathcal{H}, f \in \mathcal{F}_{\tau(a)}} \hat{E}(h, f_i, \cdots, f_T | Z)
\]

The following theorem provides theoretical guarantee of our proposed regularizer’s effectiveness.

**Theorem 2 (Generalization Error (Informal)).** Let \( \delta > 0 \) and \( \mu_1, \mu_2, \ldots, \mu_T \) be the probability measure on \( X \times \mathbb{R} \). With probability of at least \( 1 - \delta \) in the draw of \( Z = (X, Y) \sim \prod_{t=1}^{T} \mu_t \), we have:
\[
E(\hat{h}, \hat{f}) - E(h^*, f^*) \leq c_1 L \frac{G(H(X))}{nT} + c_2 B \frac{\sqrt{\lambda_{\min} \sup_{k} \|h(x)^2\|}}{n \sqrt{nT}} + \sqrt{\frac{8 \ln(4/\delta)}{nT}}
\]

where \( c_1, c_2 \) are universal constants, \( G(H(X)) \) is the Gaussian average defined as \( G(H(X)) = \mathbb{E}[\sup_{h \in \mathcal{H}} K \sum_{k \in T} t^{2} K \{k \} h(x)^2] \), where \( \{ t \} \) is i.i.d standard normal variables. \( L \) is the Laplacian matrix of graph with \( T \) vertices and edge-weights \( \{ \omega_{ij} \} \) with \( \lambda_{\min} \) is its smallest non-zero eigenvalue. \( B \) is any positive value that satisfies the condition \( \sum_{i,j=1}^{T} \omega_{ij} \cdot \text{dist}^2(\nabla \lambda_i, \nabla \lambda_j) \leq B^2 \).

Some remarks over Theorem 2: 1) The first term of the bound can be interpreted as the cost of estimating the shared representation learning function \( h \in \mathcal{H} \). This term is typically of order \( \frac{1}{nT} \). The last term contains the confidence parameter. According to [32], the constant \( c_1 \) and \( c_2 \) are pretty large, so the last term typically makes limited contribution in the bound. 2) The second or middle term contains the cost of estimating task-specific predictors \( f \in \mathcal{F}_{\tau(a)} \), and this term is typically of order \( \frac{\sqrt{\lambda_{\min} \|h(x)^2\|}}{\sqrt{nT}} \). Here the positive constant \( B \) provides important insights into the relationship between our proposed regularizer and the error bound. The smaller our regularization term becomes, the smaller values \( B \) could take and in turn reduces the second term in the bound. In general, our generalization error result bounds the gap between the test error of the model trained from finite samples and that trained from infinite data, namely the theoretically optimal model/solution. In other words, Theorem 2 provides theoretical guarantee for our performance on actual test set.

### 4.2 Relation with Other MTL Frameworks

In this section, we mathematically elucidate the relation and difference between our proposed SRDML and other MTL methods, i.e., shallow MTL and deep MTL.

**Natural generalization of shallow MTL.** Following [51], traditional multi-task learning methods (i.e., linear model based MTL) can be generally classified into two categories: **multi-task feature learning** and **multi-task relation learning**, with objective function \( \min_{W_1, b, b} L(W, b) + \lambda_1/2 \cdot tr(W^T \Theta^{-1} W) \) and \( \min_{W, b, \Sigma} L(W, b) + \lambda_2/2 \cdot \text{tr}(W^T \Sigma^{-1} W) \), where \( \Theta \) and \( \Sigma \) models the covariance between different features and tasks, respectively. For any regularization-based shallow MTL defined as above, it can be formulated as a **special case** under the general framework of SRDML, with identity feature extraction function \( h \), linear task-specific function \( f \) and the corresponding regularizer on the input gradients.

**Relation with deep MTL.** Define two hyperparameters: 1) The coefficient of regularizer in SRDML, \( \lambda_1 \), and 2) the number of layers \( T \) before which the model is shared cross tasks. When \( \lambda \) equals 0 and \( T \) is greater than 1 and less than \( L \) (total number of layers), SRDML degenerates to hard-parameter sharing. On the other hand, when \( T \) equals to 1 and \( \lambda \) is greater than 0, our SRDML is equivalent to soft-parameter sharing. Hence, both hard-parameter sharing and soft-parameter sharing can be formally formulated as special cases of our proposed SRDML framework.

### 5 EXPERIMENTS

In this section, we validate SRDML on both synthetic and real-world datasets against state-of-the-art baselines, on various aspects including performance, sensitivity, qualitative analysis and ablation study. The experiments were performed on a 64-bit machine with 4-core Intel Xeon W-2123 @ 3.60GHz, 32GB memory and NVIDIA Quadro RTX 5000.

#### 5.1 Experimental Settings

**Controlled Synthetic Dataset.** We first check the validity of SRDML on a controlled regression synthetic dataset. We generate \( T \) tasks \( T = 12 \) and for each task \( i \) we generate \( m \) samples \( m = 100 \). The input data \( X_i \in \mathbb{R}^{m \times d} \) is for each task \( i \) generated from \( X_i \sim \mathcal{N}(\eta_i, I) \) with mean vector \( \eta_i \) and identity covariance matrix \( I \). Next, we generate feature weight \( W \) by the following steps: 1) Generate two base feature weights. As shown in Figure 4a, the first base feature weight (on the LHS column) corresponds to \( w_1 = (1; 0)^T \) and the second base feature weight (on the RHS column) corresponds to \( w_2 = (0; 1)^T \), where \( 1 \) and \( 0 \) each denotes a 10-dimensional all-one and all-zero vector respectively. In this way, \( w_1 \) and \( w_2 \) will simulate two different regions in the input \( X \) once the regions zeroed out by \( w \) will not be helpful in corresponding tasks. 2) Generate task specific feature weight. Based on \( w_1 \) and \( w_2 \), we further consider creating different levels of saliency by multiplying the base feature weights by some magnitude parameter. Here we select 3 different magnitude parameters to create different levels of saliency for each base feature weight, and for each level of saliency we create two tasks which are basically twin tasks.

| Table 1: Attributes summary in CelebA and COCO. |
|---------------------------------|-----------|---------|---------|
| Tid  | CelebA | COCO | Tid  | CelebA | COCO |
| 1    | 13      | Young  | 11    | PaleSkin | coach |
| 2    | 16      | Mouth | 12    | Sideburns | bed   |
| 3    | BlackHair | dog  | 13    | Smiling | dining table |
| 4    | BrownHair | horse | 14    | WavyHair | laptop |
| 5    | Chubby | 15    | WearingLipstick | tv   |
| 6    | Mustache | 16    | Young | cell phone |
| 7    | 17    | bottle | 8    | HeavyMakeup | motorcycle |
| 9    | Mustache | 18    | cup   | 9    | MouthSlightlyOpen | bicycle |
| 10   |         | 19    |       | Mustache | chair |
| 11   |         |       |       |         | bowl |
For example, in Figure 4a, task 1 and task 2 are twin tasks which share the same level of saliency, and the lightest blue color means they are generated by the lowest magnitude parameter. We denote each generated task-specific feature weight as $w_i$, $i \in \{1, 2, \cdots, T\}$. The aforementioned logistics are basically symmetric for $w_1$ and $w_2$. 3) Add noise and create labels. We first inject some noise into each task’s feature weight by randomly flipping the sign of the value in some positions of each $w_i$. The proportion of the flipped positions is controlled to guarantee the overall pattern can be well kept. Then, we generate the label for each task by $Y_i = X_i \cdot w_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 0.1 \cdot I)$ is random normal noise.

**Real-world Dataset.** We evaluate the proposed method on 3 real-world benchmarks with varying number of tasks and difficulty, including: multi-task version of CIFAR-10 [21] (CIFAR-MTL), a modified version of CelebA [27] and a modified version of MS-COCO [26]. To follow our model’s assumption, all tasks are image classification ones. For CIFAR-MTL, we follow existing work [36] creating one task for each of the 10 classes in original CIFAR-10 dataset. There are 10 binary classification tasks with 2k training samples and 1k testing samples per task. CelebA has 200 thousand images of celebrity faces and each image is labeled with 40 facial attributes. We follow existing work [54] selecting 16 attributes more related to face appearance and ignore attributes around decoration such as eyeglasses and hat for our experiments. We randomly selected 30k training samples and include whole validation and test set. For MS-COCO we select 19 types of objects and remove those with too sparse labels. We include all images that contain at least two of the 19 types of objects and randomly split them into training and testing set by half. All results are reported on the test set. For hyper-parameter tuning of our method, without further specification, we applied grid search on the range of $\{10^{-3}, 5 \cdot 10^{-3}, \cdots, 0.5, 1\}$ for the regularization coefficient.

**Comparison Methods** We compare SRDML with various existing methods, including two baselines, three shallow and five deep state-of-the-art MTL methods:

- **Shallow MTL Methods:**
  1. **Lasso [46]** is an $\ell_1$-norm regularized method which introduce sparsity into the model to reduce model complexity and feature learning, and that the parameter controlling the sparsity is shared among all tasks. 2. **Joint Feature Learning (L21) [14]** assumes the tasks share a set of common features that represent the relatedness of multiple tasks. 3. **Robust Multi-task Learning (RMTL) [10]** method assumes that some tasks are more relevant than others. It assumes that the model $W$ can be decomposed into a low rank structure $L$ that captures task-relatedness and a group-sparse structure $S$ that detects outliers.

- **Deep MTL Methods:**
  1. **Multilinear Relationship Networks (MRNs)** places a tensor normal prior on task-specific layers of the deep multi-task learning model [28]. 2. **Multi-gate Mixture-of-Experts (MMoE) [30]** adapt the Mixture-ofExperts (MoE) structure to multi-task learning by sharing the expert submodels across all tasks, while also having a gating network. 3. **Progressive layer extraction (PLE) [45]** separates shared components and task-specific components explicitly and adopts a progressive routing mechanism to extract and separate deeper semantic knowledge gradually, improving efficiency of joint representation learning and information routing across tasks in a general setup. 4. **Multi-task Learning as Multi-Objective Optimization (MGDA-UB) [39]** considers multi-task learning from optimization perspective by using Pareto optimality and Multiple Gradient Descent Algorithm. 5. **Gradient Surgery for Multi-task Learning (PCGrad) [49]** aims to solve the problem of gradient interference by gradient surgery, which is basically by gradient projection to make sure the gradients of different tasks have direction smaller than 90°.

**Implementation Details.** All shallow MTL methods are implemented according to standard package MALSAR [56]. Deep MTL methods and our SRDML are built based on VGG-16 [41], which is a very popular architecture in computer vision. The convolutional layers are followed by one fully connected layer with 128 hidden neurons and one classification layer for our SRDML. Each model is trained by Adam [20]. For PCGrad, due to the fact that...
Table 2: Performance (%) on real-world large-scale multi-task learning datasets. Our proposed SRMTL outperforms most comparison methods on all three datasets. Bold and underline score are for the best and second best methods, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>CIFAR-MTL</th>
<th>CelebA</th>
<th>COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>AUC</td>
<td>Precision</td>
</tr>
<tr>
<td>STL (Deep)</td>
<td>94.35</td>
<td>66.69</td>
<td>74.32</td>
</tr>
<tr>
<td>Hard-Share</td>
<td>94.70</td>
<td>95.56</td>
<td>76.30</td>
</tr>
<tr>
<td>Lasso</td>
<td>91.48</td>
<td>86.64</td>
<td>68.90</td>
</tr>
<tr>
<td>L21</td>
<td>91.50</td>
<td>87.58</td>
<td>68.01</td>
</tr>
<tr>
<td>RMTL</td>
<td>92.28</td>
<td>85.65</td>
<td>61.34</td>
</tr>
<tr>
<td>MRN</td>
<td>94.51</td>
<td>96.67</td>
<td>79.94</td>
</tr>
<tr>
<td>MMDA-UB</td>
<td>93.53</td>
<td>93.17</td>
<td>73.42</td>
</tr>
<tr>
<td>PLE</td>
<td>94.01</td>
<td>93.32</td>
<td>75.26</td>
</tr>
<tr>
<td>MMoE</td>
<td>90.74</td>
<td>84.38</td>
<td>57.80</td>
</tr>
<tr>
<td>PCGrad</td>
<td>95.11</td>
<td>96.69</td>
<td>79.03</td>
</tr>
<tr>
<td>SRDML</td>
<td>95.82</td>
<td>96.43</td>
<td>81.22</td>
</tr>
<tr>
<td>SRDML (w/. PCGrad)</td>
<td>96.03</td>
<td>96.72</td>
<td>82.59</td>
</tr>
</tbody>
</table>

it is a gradient surgery method which is model-agnostic and can be applied onto any deep MTL method, we report its performance by combining it with the best baseline on each real-world dataset (i.e., Hard-Share on CIFAR-MTL, MGD-A-UB on CelebA, MRN on COCO). In addition, we also consider applying PCGrad onto our own method SRDML, resulting in two versions of our method, namely SRDML and SRDML with PCGrad.

5.2 Experimental Results

Effectiveness on controlled synthetic dataset. The empirical results on the regression synthetic dataset demonstrate that our model can generally outperform single task learning and is capable to capture the ground-truth task relations. Quantitative evaluation in Figure 4c shows that SRDML can outperform single task learning in general, which can be attributed to the effective knowledge sharing between task-specific layers. In addition, the task relationship pattern (i.e., \( w_{ij} \) in (4)) learned by SRDML as shown in Figure 4b is accurate and reasonable, since tasks belong to different bases are well-separated and meanwhile each pair of twin tasks shows very strong correlation (corresponds to those yellow boxes). Within each base, different pairs of twin tasks also show relatively strong relationship due to the fact that they share the same base and only differ in the magnitude.

Sensitivity analysis. The sensitivity of hyperparameter \( \lambda \) in SRDML on synthetic dataset is shown in Figure 4d. As can be seen, the optimal value for \( \lambda \) is around 0.5 measured by RMSE. The general “U” shape is potentially reasonable because as \( \lambda \) goes to infinity the regularization term would dominate the overall objective function while too small \( \lambda \) will reduce the functionality of the regularizer and finally degenerate to single task learning.

Effectiveness on real-world datasets.

• CIFAR-MTL: Table 2 shows the performance results of our proposed SRDML and other baselines on CIFAR-MTL dataset. We can make the following observations from the results. 1) Deep multi-task learning models generally outperform shallow ones by a great margin, which confirms the importance of learning the deep representation features as well as the shared policy of feature extraction part which allows knowledge transfer across tasks. 2) Our proposed SRDML outperforms baselines in majority of metrics and achieves comparable performance in rest. 3) Combining with PCGrad can further improve the performance of SRDML due to the mitigated negative transfer by gradient surgery of PCGrad.

• CelebA: In this case, we tackle a larger and more challenging benchmark, where we tailored the dataset to contain 16 binary classification tasks with each one corresponding to a certain human’s facial feature. As shown in Table 2, our model outperforms all comparison methods in majority of metrics, which is attributed to the potential fact that the salient regions in some tasks are close to those in the related tasks. For example, there are two tasks to classify whether a celebrity’s beard is goatee or mustache, respectively. For both tasks the salient regions are highly overlapped around the mouth area (as can be seen in Section “Saliency map visualization” in appendix) so enforcing similar input gradients around the mouth area could improve the knowledge transfer and achieve better performance.

• COCO: To evaluate our model under various settings, we consider COCO which contains different types of objects like human, animals, vehicles, furniture, etc, and each type object has varying rate of occurrence. In Table 2, we report the task-average classification error with lower values indicating better performance. As shown in Table 2, our proposed SRDML outperforms all the baselines by a great margin. This experiment also validates the effectiveness of our model when the number of tasks is relatively large and the image context is complicated. Moreover, MMoE and PLE perform generally not quite well probably due to the fact that these two approaches are designed for multi-task learning under recommender system scenario, which is not similar to that in multi-task image classification, e.g., the number of tasks in our case is much larger and hence more challenging.

Qualitative analysis. Here we demonstrate that SRDML can learn reasonable task relations on challenging real-world dataset, by visualizing the task weight matrix (i.e., \( w_{ij} \) in (4)). As shown in Figure 5, many highlighted task relations are intuitive. In CelebA, our proposed SRDML successfully learned the similarity of tasks sharing the same/similar regions around face, lie ‘Arched Eyebrow’ and ‘Bags Under Eyes’; ‘Black Hair’, ‘Brown Hair’ and ‘Wavy Hair’;
"Goatee", "Sideburns" and "mustache", etc. On the other hand, our model can also learn reasonable task similarities in COCO, including "cat" and "dog"; "car", "bus" and "bicycle"; "couch" and "bed", etc. We also conduct qualitative analysis experiment on the saliency map generated by our proposed SRDML on similar or related tasks. Please refer to the appendix for the detail.

**Adaptive regularizer on contradicting tasks.** In this section, we conducted another sensitivity analysis when all tasks compete (we generate such synthetic dataset by following similar procedure introduced in Section 5.1), and the results in Table 3 demonstrate the efficacy of our regularization term which can adaptively decrease the task-similarity weight to zero and avoid competition.

**Effect of normalization on input gradient** We add an experiment on SRDML with normalizing the input gradients and compare its results with our original method (e.g., without normalization) on ALL 3 real-world dataset. As shown in Table 4, adding normalization did not obviously change the performance in task-average classification error. The task-average classification error decreased by < 0.2% on CIFAR-MTL and increased by around 0.1% on CelebA and COCO. One explanation is, for similar tasks like "Black hair" and "Brown hair" in CelebA, we empirically observed that the magnitude for the gradients was close to each other, which might limit the point in applying gradient normalization in such case.

**Ablation study.** In this section, we present ablation study on the task relation learning part in the regularizer. Specifically, we remove the \{ω_{ij}\}_{1≤i<j≤T} in (3) and the coefficient for each term in the regularizer is just the hyperparameter \(\lambda\). We conduct experiments on all three real-world dataset to see the difference and the results are shown in Table 5.

### Table 3: Sensitivity analysis on regularizer coefficient when tasks are contradicting. Our regularizer coefficient can adaptively reduce to zero and avoid negative transfer.

<table>
<thead>
<tr>
<th>(\lambda)</th>
<th>1.0</th>
<th>0.1</th>
<th>0.01</th>
<th>0.001</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>2.726</td>
<td>1.550</td>
<td>1.405</td>
<td>1.393</td>
<td>1.392</td>
</tr>
<tr>
<td>MAE</td>
<td>2.198</td>
<td>1.260</td>
<td>1.127</td>
<td>1.127</td>
<td>1.126</td>
</tr>
</tbody>
</table>

### Table 4: Normalization of input gradient

<table>
<thead>
<tr>
<th></th>
<th>CIFAR-MTL</th>
<th>CelebA</th>
<th>MS-COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRDML w/o normalization</td>
<td>4.18</td>
<td>9.91</td>
<td>14.32</td>
</tr>
<tr>
<td>SRDML w/ normalization</td>
<td>4.02</td>
<td>10.03</td>
<td>14.41</td>
</tr>
</tbody>
</table>

### Table 5: Ablation study on adaptive regularizer (Accuracy)

<table>
<thead>
<tr>
<th></th>
<th>CIFAR-MTL</th>
<th>CelebA</th>
<th>MS-COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRDML (w/o regularizer)</td>
<td>94.92</td>
<td>89.74</td>
<td>85.18</td>
</tr>
<tr>
<td>SRDML (w/ regularizer)</td>
<td>95.82</td>
<td>90.15</td>
<td>85.68</td>
</tr>
</tbody>
</table>

### 6 CONCLUSION

Learning interpretable task relation is challenging in multi-task learning problem. In this paper, we proposed Saliency-regularized Deep Multi-task Learning (SRDML) framework which regularizes the input gradient of different tasks by saliency and achieves good task relation interpretability. Instead of regularizing parameters like existing work, we directly regularize in functional space which allows better expressiveness. Theoretical analyses show that our regularizer can help reduce the generalization error. Experiments on multiple synthetic and real-world datasets demonstrate the effectiveness and efficiency of our methods in various metrics, compared with several comparison methods and baselines. The reasonableness of the task relation learned by SRDML is also validated on different challenging real-world datasets.

### ACKNOWLEDGEMENT

This work was supported by the National Science Foundation (NSF) Grant No. 1755850, No. 1841520, No. 2007716, No. 2007976, No. 1942594, No. 1907805, a Jeffress Memorial Trust Award, Amazon Research Award, NVIDIA GPU Grant, and Design Knowledge Company (subcontract number: 10827.002.120.04).
A THEORETICAL PROOF

A.1 Proof of Theorem 1

Proof. Suppose $\mathcal{X} \subseteq \mathbb{R}^K$ is an open set and $f_1, f_2 : \mathcal{X} \to \mathbb{R}$, where both functions are differentiable and equal to zero at the origin. We establish the following: This direction is obvious, since two exactly the same functions will have the same gradient at any input point. Given $\nabla f_1(x) = \nabla f_2(x)$, we know that $f_1 = f_2$ (see section 3.1 in main paper). As mentioned in section 3.1 in main paper, we denote the feature extraction function $h$ and the only assumption here is that $H$ and $\mathcal{H}$ are open sets and $\mathcal{H}$ is an open set and $\mathcal{H}$ is an open set.

In general, we will use $\gamma$ to denote a generic vector of i.i.d standard normal variables, whose dimension will be clear in context. In addition, without further specification, we will use $K, T, n$ to denote the (flattened) dimension of the output space from the feature extraction function $h$, number of tasks, and number of training samples, respectively. We denote the representation class for task-specific function $f$ and representation extraction function $h$ as $\mathcal{F}$ and $\mathcal{H}$, respectively. Two hypothesis classes here can be very general, and the only assumption here is that $\forall f \in \mathcal{F}, h$ has Lipschitz constant at most $L$, for any positive $L$.

DEFINITION 3. Given a set $V \subseteq \mathbb{R}^n$, define the Gaussian average of $V$ as:

$$G(V) = \mathbb{E} \sup_{v \in V} \langle v ; \gamma \rangle = \mathbb{E} \sup_{v \in V} \sum_{i=1}^{n} Y_i \bar{u}_i$$

(12)

As mentioned in section 3.1 in main paper, we denote the feature representation learning part as function $h \in \mathcal{H}$. As we will see later, the complexity of representation class $\mathcal{H}$ is important in our proof for the error bound, so we define a measure of its complexity by Gaussian average.

DEFINITION 4. Given observed input data $X \subseteq \mathcal{X}^{Tn}$, define a random set $\mathcal{H}(X) \subseteq \mathbb{R}^{KTn}$ by

$$\mathcal{H}(X) = \{ (h_k(x^t_i)) : h \in \mathcal{H} \}. \quad (13)$$

The Gaussian average over $\mathcal{H}(X)$ can be defined accordingly as

$$G(\mathcal{H}(X)) = \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{K} \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{i=1}^{n} Y_i \bar{u}_i h_k(x^t_i) | x^t_i \right]$$

(14)

The following lemmas are useful in our proof later, and we introduce them here in advance.

LEMMA 1. $\forall A, C \in \mathbb{R}^{mxn}$ and $B \in \mathbb{R}^{nxm}$,

$$tr(ABA^TBC) = \sum_{i,j} B_{ij} \sum_{k=1}^{m} A_{ik} \bar{C}_{jk}.$$  \hspace{1cm} (15)

LEMMA 2. Suppose $X \subseteq \mathbb{R}^K$ is an open set, and two differentiable functions $f_1, f_2 : \mathcal{X} \to \mathbb{R}$. Then

$$\exists B > 0, s.t \parallel \nabla f_1(x) - \nabla f_2(x) \parallel < B \quad (16)$$

and

$$\lim_{\Delta x \to 0} \left| \frac{f_1(x + \Delta x) - f_1(x)}{\parallel \Delta x \parallel} - \frac{f_2(x + \Delta x) - f_2(x)}{\parallel \Delta x \parallel} \right| < B. \quad (17)$$

Given everything above, we can prove our Theorem 2 with Maurer et al. [32]. However, due to the limited space of appendix here, we decide to put the formal proof for our Theorem 2 along with the proof for everything above into the put link. Please refer to there for the formal proof. Thanks!

A.3 Proof of Section 4.2

Natural generalization of shallow MTL

Proof. Basically, when the feature extraction function $h$ is identity function and each task-specific function $f_t, t = 1, 2, \cdots, T$ are linear functions, we know for any input $x \in \mathcal{X}$,

$$h(x) = x, \nabla f_t(x) = \omega_t, \forall t \quad (18)$$

where $\omega_t$ is the model parameter of linear model $f_t$. Hence, denote $W = [\omega_1; \omega_2; \cdots; \omega_T]$ and take the $dist()$ function in (3) to be inner product, by Lemma 1 we have $\sum_{t=1}^{T} \omega_{tij} \cdot dist(\nabla f_t(x), \nabla f_t(x)) = \sum_{t=1}^{T} \omega_{tij} \cdot \langle \omega_t, \omega_t \rangle = tr(W^T \Omega W)$, where $\Omega = (\omega_i \omega_j)$. Let $\Omega$ to be either $\Theta^{-1}$ or $\Sigma^{-1}$ in as section 4.2 can finish the proof.

Relation with deep MTL

Proof. First, we define two hyperparameters:

- $\lambda$: The coefficient of our regularizer in SRDML
- $t$: The index of the layer before which the model is shared cross different tasks.

Case 1. If $\lambda = 0$ and $1 < t < L$, where $L$ (please differentiate this $L$ with that for Lipschitz constant) denotes the total number of layers, our SRDML has no regularization and is simply equivalent to hard-parameter sharing.

Case 2. If $\lambda > 0$ and $t = 1$, each layer in our SRDML is separate for different tasks and the regularization is posed on all the layers, which is equivalent to soft-parameter sharing.

B DETAILS ON SYNTHETIC DATASET GENERATION

- "What are base feature weights": Since we want to generate tasks with different level of similarity in our synthetic dataset, we achieved it by controlling the similarity in the feature weight (i.e., $w_t$) of different tasks. The base feature weights $w_1$ and $w_2$ are basically two vectors (with length equal to number of features) for generating the feature weight vectors for all the tasks. We call them "base" feature weight because they serve as the base vector or unit vector for generating all the tasks’ feature weights. In addition, $w_1$ and $w_2$ are orthogonal and each has length 1 in any dimension.
"How are base feature weights used": The base feature weights are used to generate each task’s feature weight in the following steps: 1) We choose which base the current task belongs to. In our setting, we chose the first half of tasks to belong to the first base (i.e., \( w_1 \)) and the second half of tasks to belong to the second base (i.e., \( w_2 \)). Since two bases are orthogonal, they can actually simulate two non-overlapping regions in pictures which means tasks from different bases should not be similar while those from the same base should be similar since, they focus on the same region. 2) Within each base, we multiply the base vector (i.e., \( w_1 \) and \( w_2 \)) by some positive integers to generate the actual feature weight for the tasks. For example, we multiply \( w_1 \) by integer 1, 2 and 3 to generate the feature weight vectors for the first half of tasks.

C ADDITIONAL QUALITATIVE ANALYSIS

We also conduct qualitative analysis experiment on the saliency map generated by our proposed SRDML on similar or related tasks. As can be seen in Figure 6 and Figure 7, our proposed SRDML can generate saliency map focusing on similar regions for related tasks. For example, the saliency map generated for “Black hair” and “Brown hair” both generally overlap around the hair region of the woman, and the saliency map generated for three types of beard all overlap around the mouth and beard region of the man. Notice that the quality of saliency itself is not the main focus of this paper, but instead we are more interested in the task relation induced by the saliency map similarity (i.e., saliency across tasks).

D EXTRA REMARKS ON THEOREM 2

In this section, we provide more remarks on our main theorem, namely Theorem 2, for better understanding.

- The equation above bounds the gap between the test error of the model trained from finite samples and that trained from infinite data, namely the theoretically optimal model/solution. In other words, Theorem 2 provides theoretical guarantee for our performance on actual test error.
- In Equation 2-9, we assume all tasks share the same set of \( X \) which is a very common case in Multi-task Learning on image dataset. Theorem 2 does not need different tasks to have different \( X(t) \), since \( \mu_1 = \mu_2 = \cdots = \mu_T = \mu \) is a special case of the version in Theorem 2. Our current assumption is actually a more general one and can handle the case in Equation 2-9.