Adaptive Adversarial Multi-task Representation Learning

Yuren Mao [1] **Weiwei Liu** [2] **Xuemin Lin** [1]

Abstract

Adversarial Multi-task Representation Learning (AMTRL) methods are capable of boosting the performance of Multi-task Representation Learning (MTRL) models. However, the theoretical mechanism behind AMTRL has been only minimally investigated. Accordingly, to fill this gap, we study the generalization error bound of AMTRL through the lens of Lagrangian duality. Based on this duality, we propose a novel adaptive AMTRL algorithm that improves the performance of the original AMTRL methods. We further conduct extensive experiments to back up our theoretical analysis and validate the superiority of our proposed algorithm.

1. Introduction

Multi-task Representation Learning (MTRL), which is an influential line of research on Multi-task Learning, learns related tasks simultaneously by sharing a common representation. Compared with learning each task independently, MTRL typically has a lower computational cost and better prediction performance. It has achieved great success in various applications ranging from computer vision (Kendall et al., 2018) to natural language processing (Collobert & Weston, 2008).

Recently, adversarial MTRL (AMTRL) methods (Liu et al., 2017; Chen et al., 2018a; Shi et al., 2018; Yu et al., 2018; Liu et al., 2018; Yadav et al., 2018) have been widely utilized in a range of applications. AMTRL methods improve the performance of original MTRL models by adding an extra adversarial module, i.e., a task discriminator in the representation space. Unfortunately, the theoretical mechanism behind AMTRL methods is still not well understood.

The findings of this paper suggest that AMTRL methods restrict the hypothesis class by enforcing all the tasks to share an identical distribution in the representation space. The identical distribution restriction provides further inductive bias and tightens the task-averaged generalization error bound for MTRL. Based on this restriction, we formulate AMTRL as a constrained optimization problem and propose to solve the problem using the augmented Lagrangian method.

To quantitatively measure how likely the tasks share an identical distribution in the representation space, we propose a pairwise relatedness metric for AMTRL. Based on this metric, a weight adaption strategy is proposed in order to accelerate the convergence of the adversarial module. Combining the weight adaption strategy and the augmented Lagrangian method, we present the adaptive AMTRL method.

This paper conducts experiments on two popular multi-task learning applications: sentiment analysis and topic classification. Experimental results verify our theoretical analysis and validate that the proposed algorithm outperforms several state-of-the-art methods.

2. Related Works

Adaptive weighting scalarization, which linearly scalarizes the tasks with adaptive weight assignment, is a typical MTRL method. Various adaptive weighting strategies (Kendall et al., 2018; Chen et al., 2018b; Sener & Koltun, 2018; Lin et al., 2019; Mao et al., 2020) have been proposed to balance the regularization between tasks and improve the performance of original MTRL. By contrast, existing AMTRL methods, for example (Liu et al., 2017; Chen et al., 2018a), only adopts the naïve uniform scalarization. In this paper, we propose a adaptive weighting strategy for AMTRL based on the augmented Lagrangian (Hestenes, 1969) and a novel task relatedness metric. The task relatedness metric is proposed based on the representation similarity. Comparing with the typical representation-similarity-based task relatedness metric (Kriegeskorte et al., 2008; McClure & Kriegeskorte, 2016; Dwivedi & Roig, 2019), the proposed task relatedness metric computes the representation similarity with the output of the adversarial module and does not require extra computation of correlation coefficients, which is more efficient for AMTRL.

[2]School of Computer Science, Wuhan University, China. [1]School of Computer Science and Engineering, University of New South Wales, Australia. Correspondence to: Weiwei Liu <liuweiwei863@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

3. Preliminaries

Consider a multi-task representation learning problem with T tasks over an input space $\mathcal{X}$ and a collection of task spaces $\{\mathcal{Y}\}_{t=1}^T$. We define the hypothesis class of the problem as $\mathcal{H}$ and $\mathcal{H} = \{\mathcal{F}\}_{t=1}^T \circ \mathcal{G}$. $\mathcal{G} = \{g : \mathcal{X} \to \mathbb{R}^K\}$ is the set of representation functions (i.e. the representation hypothesis class). K is the dimension of the representation space. $\{\mathcal{F}\}_{t=1}^T = \{f^t : \mathbb{R}^K \to \mathcal{Y}\}_{t=1}^T$ is a set of predictors (i.e. the prediction hypothesis class) and f^t is ρ-Lipschitz for all $t \in \{1, ..., T\}$. g is used across different tasks, while f^t is task-specific. $\mathcal{H} = \{h = \{f^t(g(\cdot))\}_{t=1}^T : \mathcal{X} \to \{\mathcal{Y}\}_{t=1}^T\}$. Learning $\mathcal{H}$ is based on the data observed for all the tasks. Without loss of generality, we assume that each task has n samples. The data takes the form of a multi-sample $S = \{S_t\}_{t=1}^T$ with $S_t = (\overline{X}_t, \overline{Y}_t)$ and $(\overline{X}_t, \overline{Y}_t) = \{x_i^t, y_i^t\}_{i=1}^n \sim \mathcal{D}_t^n$. $\mathcal{D}_t$ is a probability distribution over $\mathcal{X} \times \mathcal{Y}$. After representation mapping, $(g(\overline{X}_t), \overline{Y}_t) \sim \mu_t^n$ where μ_t is a distribution over $\mathbb{R}^K$.

The loss function for task t is defined as $l^t : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ and assumed to be 1-Lipschitz. We define the true risk of a hypothesis $f^t \circ g$ for task t as $\mathcal{L}_{\mathcal{D}_t}(f^t \circ g) = \mathbb{E}_{(x^t, y^t) \sim \mathcal{D}_t}[l^t(f^t(g(x^t)), y^t)]$ and the task-averaged generalization error as $\mathcal{L}_{\mathcal{D}}(h) = \frac{1}{T}\sum_{t=1}^T \mathcal{L}_{\mathcal{D}_t}(f^t \circ g)$. Correspondingly, the empirical loss of the task t is defined as $\mathcal{L}_{S_t}(f^t \circ g) = \frac{1}{n}\sum_{i=1}^n l^t(f^t(g(x_i^t)), y_i^t)$ and the empirical task-averaged error is defined as $\mathcal{L}_S(h) = \frac{1}{T}\sum_{t=1}^T \mathcal{L}_{S_t}(f^t \circ g)$. We also denote the transpose of the vector/matrix by superscript $'$, the logarithms to base 2 by log.

Multi-task Representation Learning. Multi-task Representation Learning (MTRL) learns multiple tasks jointly by sharing representation across tasks. This representation is typically produced using a representation map that has the same parameters for each task. For example, in deep neural networks, the common representation is obtained by sharing hidden layers. The original MTRL module in Figure 1 shows a deep MTRL network model utilizing a hard parameter sharing strategy (Ruder, 2017). With the Empirical Risk Minimization (ERM) paradigm, MTRL is defined to minimize the task-averaged empirical error (1) (Maurer et al., 2016).

$$\min_{g, f^1, ..., f^T} \frac{1}{T}\sum_{t=1}^T \hat{\mathcal{L}}^t(g, f^t). \tag{1}$$

Theorem 1 (Maurer et al., 2016; Ando & Zhang, 2005) presents an upper bound for the task-averaged generalization error of MTRL.

Theorem 1. *For $0 < \delta < 1$, with probability at least $1 - \delta$ in S we have that*

$$\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h) \leq \frac{c_1 L\, G_a(\mathcal{G}(\overline{X}))}{nT} + \frac{c_2 Q sup_{g\in\mathcal{G}}\|g(\overline{X})\|}{n\sqrt{T}} + \sqrt{\frac{9ln(2/\delta)}{2nT}} \tag{2}$$

where c_1 and c_2 are universal constants. $G(\mathcal{G}(\overline{X}))$ is the Gaussian average defined in (3)

$$G_a(\mathcal{G}(\overline{X})) = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \sum_{k,t,i} \gamma_{kti} g_k(x_i^t) \mid x_i^t\right], \tag{3}$$

where γ_{kti} denote independent standard normal variables. $sup_{g\in\mathcal{G}}\|g(\overline{X})\|$ can be computed by (4)

$$sup_{g\in\mathcal{G}}\|g(\overline{X})\| = \sup_{g\in\mathcal{G}} \sqrt{\sum_{k,t,i} g_k(x_i^t)^2}. \tag{4}$$

Q is the quantity

$$Q \equiv \sup_{y \neq y^* \in \mathbb{R}^{Kn}} \frac{1}{\|y - y^*\|} \mathbb{E} \sup_{f\in\mathcal{F}} \sum_{i=1}^n \gamma_i (f(y_i) - f(y_i^*)), \tag{5}$$

where γ_i are independent standard normal variables.

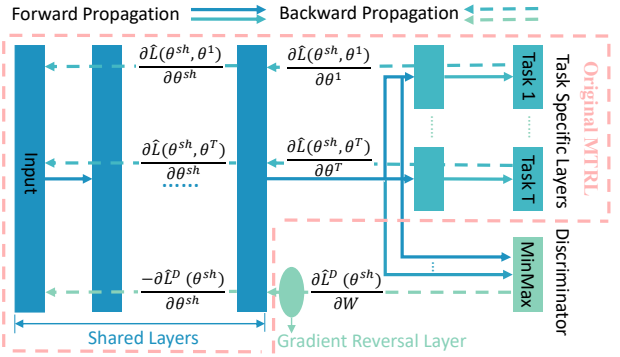


Figure 1. A deep adversarial MTRL Network model.

Adversarial Multi-task Representation Learning. Adversarial MTRL (AMTRL) adds an extra task discriminator to the original MTRL model shown in Figure 1. For each training sample, the discriminator can recognize which task the sample belongs to. The loss functions of existing adversarial MTRL methods (Liu et al., 2017; Chen et al., 2018a; Shi et al., 2018; Yu et al., 2018; Liu et al., 2018; Yadav et al., 2018) have a common part

$$\min_h L(h, \lambda) = \mathcal{L}_S(h) + \lambda \mathcal{L}^{adv}, \tag{6}$$

where λ is a hyper parameter and the adversarial term $\mathcal{L}^{adv}$ has the form

$$\mathcal{L}^{adv} = \max_{\Phi} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n e_t \Phi(g(x_i^t)). \tag{7}$$

$\Phi(\cdot) : \mathbb{R}^K \to [0,1]^T$ is a task discriminator that estimates which task the sample belongs to. e_t is the vector with all components equal to 0, except the t-th, which is 1.

(6) minimizes the task-averaged empirical risk and enforces the representation of each task to share an identical distribution ($\mu_1 = \mu_2, ..., = \mu_T$). When all tasks have an identical distribution in the representation space, $\mathcal{L}_{adv} = c$ where c is a discriminator-depended constant. For the widely used softmax function-based discriminator, where $\Phi(g(x_n^t)) = softmax(W' g(x_n^t) + b)$ and $W \in \mathbb{R}^{K \times T}$, $c = \frac{1}{T}$. Without loss of generality, we can set $\mathcal{L}^{adv} := \mathcal{L}^{adv} - c$.

4. Proposed Methods

4.1. Task-averaged Generalization Error Bound

Assuming the representation of each task shares an identical distribution, Corollary 1 outlines the task-averaged generalization error bound for AMTRL.

Corollary 1. *Assume $\mu_1 = \mu_2, ..., = \mu_T$. For $0 < \delta < 1$, with probability at least $1 - \delta$ in $\overline{S}$ we have that*

$$\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h) \leq \frac{c_1 \rho \, G_a(\mathcal{G}^*(\overline{X}_1))}{n} + \frac{c_2 Q sup_{g \in \mathcal{G}^*} \|g(\overline{X}_1)\|}{\sqrt{n}} + \sqrt{\frac{9ln(2/\delta)}{2nT}} \tag{8}$$

where c_1 and c_2 are universal constants, while $\mathcal{G}* = \{g \in \mathcal{G} : \mu_1 = \mu_2 =, ..., \mu_T\}$. $G(\mathcal{G}^*(\overline{X}_1))$ is the Gaussian average of task 1 defined in (9)

$$G_a(\mathcal{G}^*(\overline{X}_1)) = \mathbb{E}\left[\sup_{g \in \mathcal{G}^*} \sum_{k,i} \gamma_{ki} g_k(x_i^1) \mid x_i^1\right], \tag{9}$$

where γ_{ki} are independent standard normal variables. $sup_{g \in \mathcal{G}^*} \|g(\overline{X}_1)\|$ can be computed by (10):

$$sup_{g \in \mathcal{G}^*} \|g(\overline{X}_1)\| = \sup_{g \in \mathcal{G}^*} \sqrt{\sum_{k,i} g_k(x_i^1)^2}. \tag{10}$$

Q is the quantity

$$Q \equiv \sup_{y \neq y' \in \mathbb{R}^{Kn}} \frac{1}{\|y - y'\|} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \gamma_i (f(y_i) - f(y_i')), \tag{11}$$

where γ_i denote independent standard normal variables.

Proof. For $\mu_1 = \mu_2, ..., = \mu_T$,

$$G_a(\mathcal{G}^*(\overline{X})) = \mathbb{E}\left[\sup_{g \in \mathcal{G}^*} \sum_{k,t,i} \gamma_{kti} g_k(x_i^t) \mid x_i^t\right] = T\mathbb{E}\left[\sup_{g \in \mathcal{G}^*} \sum_{k,i} \gamma_{ki} g_k(x_i^1) \mid x_i^1\right] = T G_a(\mathcal{G}^*(\overline{X}_1)). \tag{12}$$

$$sup_{g \in \mathcal{G}^*} \|g(\overline{X})\| = \sqrt{T} sup_{g \in \mathcal{G}^*} \|g(\overline{X}_1)\|. \tag{13}$$

By combining (12) and (13) with Theorem 1, we conclude our proof. □

Remarks:

- The first term of the bound, which can be interpreted as the cost of estimating the representation g, is typically of order $\frac{1}{n}$. Moreover, the second term, which corresponds to the cost of estimating task-specific predictors, is typically of order $\frac{1}{\sqrt{n}}$. The last term contains the confidence parameter. According to Theorem 3 in (Maurer, 2014), c_1, c_2 are rather large; the last term typically makes only a small contribution.
- From the property of the Gaussian average, $TG_a(\mathcal{G}^*(\overline{X}_1)) \leq G_a(\mathcal{G}(\overline{X}))$ for $G^* \subseteq \mathcal{G}$. Furthermore, we have $\sqrt{T} sup_{g \in \mathcal{G}^*} \|g(\overline{X}_1)\| \leq sup_{g \in \mathcal{G}} \|g(\overline{X})\|$. The generalization error bound for AMTRL is tighter than that for MTRL.
- In AMTRL, the number of tasks has little to do with the generalization error bound.

4.2. Task Relatedness in Representation Space

The above analysis shows that the similarity of distributions between tasks in the representation space determines the performance of AMTRL. The similarity is a data-dependent between-task relatedness. This paper proposes a novel relatedness metric for AMTRL based on the task discriminator to quantitatively measure the similarity. Based on the metric, we are able to visualize the relatedness between tasks during training.

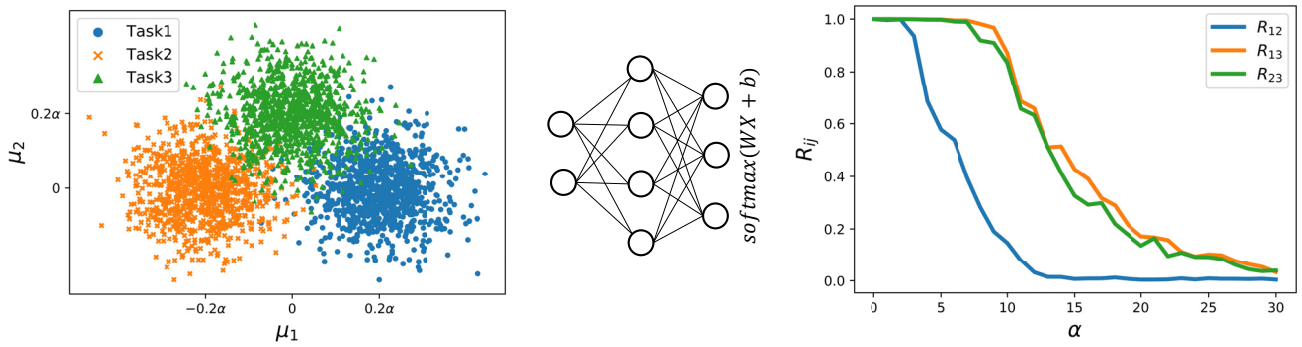
Assume that the discriminator $\Phi(\cdot)$ is the Bayes optimal classifer. We propose to measure the relatedness between task i and task j as follows:

$$R_{ij} = \frac{\Phi_j(g(x^i)) + \Phi_i(g(x^j))}{\Phi_i(g(x^i)) + \Phi_j(g(x^j))}, \tag{14}$$

where x^i and x^j are sampled from $\mathcal{D}_i$ and $\mathcal{D}_j$ respectively, $g(x^i) \sim \mu_i$ and $g(x^j) \sim \mu_j$. $\Phi_i(\cdot)$, $\Phi_j(\cdot)$ represent the probability that $\Phi(\cdot)$ classify the input into tasks i, j respectively. $R_{ij} \in [0,1]$ reflects the similarity between μ_i and μ_j. R_{ij} is equal to 1 when μ_i is the same as μ_j and equals to 0 when μ_i and μ_j are totally different.

In the Empirical Risk Minimization (ERM) setting, we approximate R_{ij} with (15), as follows:

$$R_{ij} = \min\{\frac{\sum_{n=1}^{N} e_j \Phi(g(x_n^i)) + e_i \Phi(g(x_n^j))}{\sum_{n=1}^{N} e_i \Phi(g(x_n^i)) + e_j \Phi(g(x_n^j))}, 1\}, \tag{15}$$



(a) Three 2-D Gaussian distributions. (b) Discriminator. (c) Relatedness change curve.

Figure 2. Performance of the proposed relatedness measure R_{ij} across three two-dimensional Gaussian distributions. (a) Illustration of three tasks with 2-D Gaussian distributions over their representation space. A total of 3000 samples are used in this case. The mean of the Gaussian distributions corresponding to task 1, 2 and 3 are $[0.2\alpha, 0]$, $[-0.2\alpha, 0]$, $[0, 0.2\alpha]$ respectively, and all of them have the same variance-covariance matrix $\sum = I_T$, where I is the $T \times T$ identity matrix. (b) Discriminator constructed using a two-layers fully connected network ending with a softmax function. (c) Illustration of relatedness R_{ij} between tasks decreases as α increases.

where e_t is the vector with all components equal to 0, except the t-th, which is 1.

Figure 2 presents the performance of the proposed relatedness metric in a two-dimensional Gaussian distribution case. It verifies that the metric is sensitive to the variation of the similarity between distributions.

We then propose a relatedness matrix R, where

$$R = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1T} \\ R_{21} & R_{22} & \cdots & R_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ R_{T1} & R_{T2} & \cdots & R_{TT} \end{bmatrix}. \tag{16}$$

4.3. Adaptive Adversarial MTRL

Motivated by considering the task relatedness and duality, we present an adaptive AMTRL algorithm with an novel weighting strategy in §4.3.1 and optimize it with the augmented Lagrangian method in §4.3.2.

4.3.1. Weight Adaptation

Based on the relatedness matrix, we propose a weighting strategy designed to accelerate the convergence of the adversarial module for AMTRL models. Let $\mathbf{w} = (w_1, w_2, ..., w_T)'$ and $\mathbf{1} = (1, 1, ..., 1)$ be a T-dimension vector with all components being 1. The weighting strategy is used in formulating the empirical loss of the proposed adaptive AMTRL method (17).

$$\mathcal{L}_S(h) = \frac{1}{T} \sum_{t=1}^{T} w_t \mathcal{L}_{S_t}(f^t \circ g), \tag{17}$$

where

$$\mathbf{w} = \frac{1}{\mathbf{1}R\mathbf{1}'} \mathbf{1}R\,. \tag{18}$$

Tasks that have a closer relationship with other tasks in the representation space have a larger weight. This has an intuitive interpretation: that is, the weighting strategy motivates tasks to be more similar in the representation space, which meets the constraint of AMTRL. The experimental result in Section 5.2.1 verifies this intuition.

4.3.2. Augmented Lagrangian

(6) can be regard as the Lagrangian dual function of the following equality-constrained optimization problem (Problem 1).

Problem 1.

$$\begin{aligned} \min_{h} \quad & \mathcal{L}_S(h) \\ s.t. \quad & \mathcal{L}^{adv} = 0, \end{aligned}$$

In existing adversarial MTRL works, λ is manually tuned; this process is highly time-consuming and makes it almost impossible to achieve the optimal Lagrange multiplier. As a result, an adaptive method that can choose λ automatically is desired. Moreover, an MTL Problem like Problem 1 is usually non-convex, such that the solution obtained from the Lagrangian duality is in fact not optimal due to the duality gap (Rockafellar, 1974; Hager, 1987).

Accordingly, we propose an Augmented Lagrangian-based Algorithm to dynamically tune λ and reduce the duality gap. The basic idea behind augmented Lagrangian involves augmenting the ordinary Lagrangian with a penalty term, which usually has a quadratic form. Combining the proposed weighting strategy with the augmented Lagrangian

Algorithm 1 Adaptive Adversarial MTRL

Input: S
Initialize λ_0, r_0, R^0.
for $q = 0$ **to** N **do**
  $\mathbf{w}^q = \frac{1}{IR_q I'} IR_q$
  Train the AMTRL model with loss (19)
  Update R^{q+1} using (15) with $\Phi_q(\cdot)$
  if $\lambda_{q+1} > 0$ **then**
    Update Lagrange multipliers using (20) to obtain λ_{q+1}
  else
    $\lambda_{q+1} = \lambda_q$
  end if
  Choose new penalty parameter $r_{q+1} > r_q$
end for

method, the optimization objective of our adaptive AMTRL method is given in (19).

$$\min_h \frac{1}{T} \sum_{t=1}^{T} w_t \mathcal{L}_{S_t}(f^t \circ g) + \lambda \mathcal{L}^{adv} + \frac{r}{2} \mathcal{L}^{adv^2}, \quad (19)$$

where λ is the Lagrangian multiplier, while r is the the penalty parameter with $r > 0$. As r increases, the gap between the value of the primal problem and the value of the dual problem decreases.

Based on the typical augmented Lagrangian algorithmic framework, λ_k is updated as follows:

$$\lambda_{q+1} = \lambda_q - r_q \mathcal{L}^{adv}, \quad (20)$$

with r_q increasing linearly. The specific procedure of the algorithm is shown in Algorithm 1.

The adaptive AMTRL algorithm is shown in Algorithm 1.

5. Experiments

In this section, we perform experimental studies on sentiment analysis and topic classification in order to evaluate the performance of our proposed method and verify our theoretical analysis respectively. The implementation is based on PyTorch (Paszke et al., 2019). The code can be found in the Supplementary Materials.

5.1. Experimental Setup

5.1.1. Datasets

Sentiment Analysis [1]. We evaluated our algorithm on product reviews from Amazon. The dataset (Blitzer et al., 2007) contains product reviews from 14 domains, including books, DVDs, electronics, kitchen appliances, etc. We consider each domain as a binary classification task. Reviews with ratings > 3 were labeled positive, while those with ratings < 3 were labeled negative, reviews with rating $= 3$ are discarded, as the sentiments were ambiguous and difficult to predict. The training/testing/validation partition is randomly split into 70% training, 10% testing and 20% validation.

Table 1. Data Allocation for Topic Classification Tasks.

Tasks	Newsgroups
COMP	OS.MS-WINDOWS.MISC, SYS.MAC.HARDWARE, GRAPHICS, WINDOWS.X
REC	SPORT.BASEBALL, SPORT.HOCKEY AUTOS, MOTORCYCLES
SCI	CRYPT, ELECTRONICS, MED, SPACE
TALK	POLITICS.MIDEAST, RELIGION.MISC, POLITICS.MISC, POLITICS.GUNS

Topic Classification [2]. We select 16 newsgroups from the 20 Newsgroup dataset, which is a collection of approximately 20,000 newsgroup documents and partitioned (nearly) evenly across 20 different newsgroups, and formulate them into four 4-class classification tasks (shown in Table 1) to evaluate the performance of our algorithm on topic classification. The training/testing/validation partition is randomly split into 60% training, 20% testing and 20% validation.

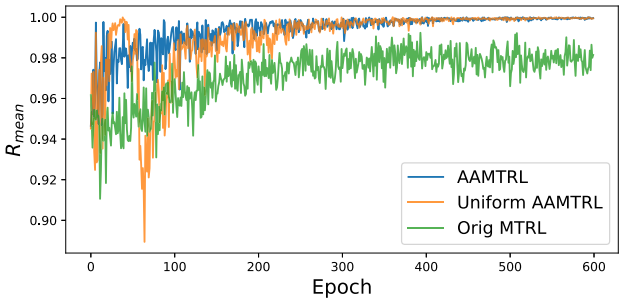
5.1.2. Network Model

We implement our adaptive AMTRL algorithm on the most prevalent deep multi-task representation learning network model (i.e. hard parameter sharing network model (Caruana, 1997)). As shown in Figure 1, all tasks have task-specific output layers and share the representation extraction layers in the model.

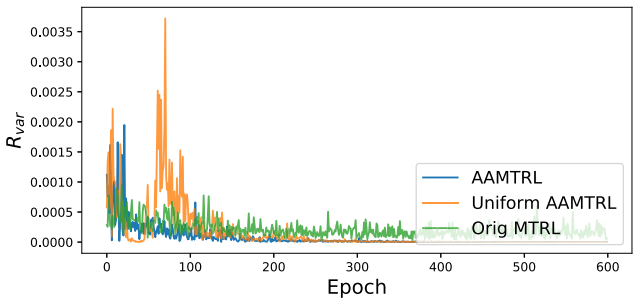
The shared representation extraction layers are typically built with a feature extraction structure such as Convolutional Neural Networks (CNN) or Recurrent Neural Network (RNN), and the task-specific output layers are typically formulated using fully connected layers. In our experiments, either TextCNN (Kim, 2014) or BiLSTM (Hochreiter & Schmidhuber, 1997) is used to build the shared representation extraction layers. The TextCNN module is structured with three parallel convolutional layers with kernel sizes of 3, 5, and 7 respectively. The BiLSTM module is structured with two bi-directional hidden layers with size 32. The extracted feature representations are then concatenated and classified using the task-specific output module, which has one fully connected layer.

[1] https://www.cs.jhu.edu/~mdredze/datasets/sentiment/

[2] http://qwone.com/~jason/20Newsgroups/

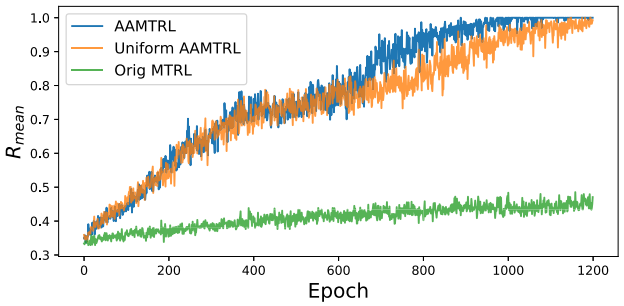


(a) R_{mean} changes in the training process.

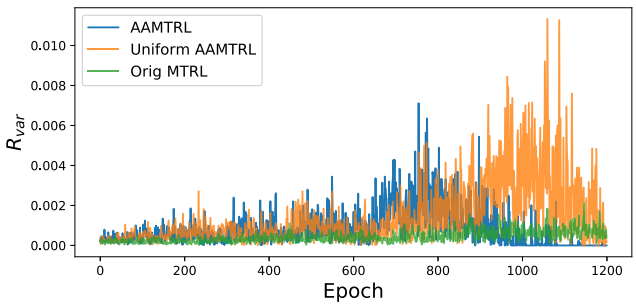


(b) R_{var} changes in the training process.

Figure 3. Evolution of relatedness between tasks during training for sentiment analysis. (a) presents the change in R_{mean} for the original MTRL (Orig MTRL), AAMTRL without the weighting strategy (Uniform AAMTRL) and AAMTRL respectively. (b) presents the change in R_{var} for Orig MTRL, Uniform AAMTRL and AAMTRL respectively.



(a) R_{mean} changes in the training process.



(b) R_{var} changes in the training process.

Figure 4. Evolution of relatedness between tasks during training for topic classification. (a) presents the change in R_{mean} for the original MTRL (Orig MTRL), AAMTRL without the weighting strategy (Uniform AAMTRL) and AAMTRL respectively. (b) presents the change in R_{var} for Orig MTRL, Uniform AAMTRL and AAMTRL respectively.

The adversarial module is built with one fully connected layer, the output size of which is equal to the number of tasks. It is noteworthy that the adversarial module connects to the shared layers via a gradient reversal layer (Ganin & Lempitsky, 2015). This gradient reversal layer multiplies the gradient by -1 during the backpropagation, which optimizes the adversarial loss function (7).

5.1.3. Training Parameters

We train the deep AAMTRL network model with Algorithm 1 settings $\lambda_0 = 1$, $r_0 = 10$ and $r_{k+1} = r_k + 2$; here, R_0 is a matrix of ones. We use the Adam optimizer (Kingma & Ba, 2015) and train 600 epochs for sentiment analysis and 1200 epochs for topic classification, The batch size is 256 for both sentiment analysis and topic classification. We use dropout with probability of 0.5 for all task-specific output modules. For all experiments, we search over the set $\{1e{-}4, 5e{-}4, 1e{-}3, 5e{-}3, 1e{-}2, 5e{-}2\}$ of learning rates and choose the model with the highest validation accuracy.

5.2. Results and Analysis

5.2.1. Relatedness Evolution

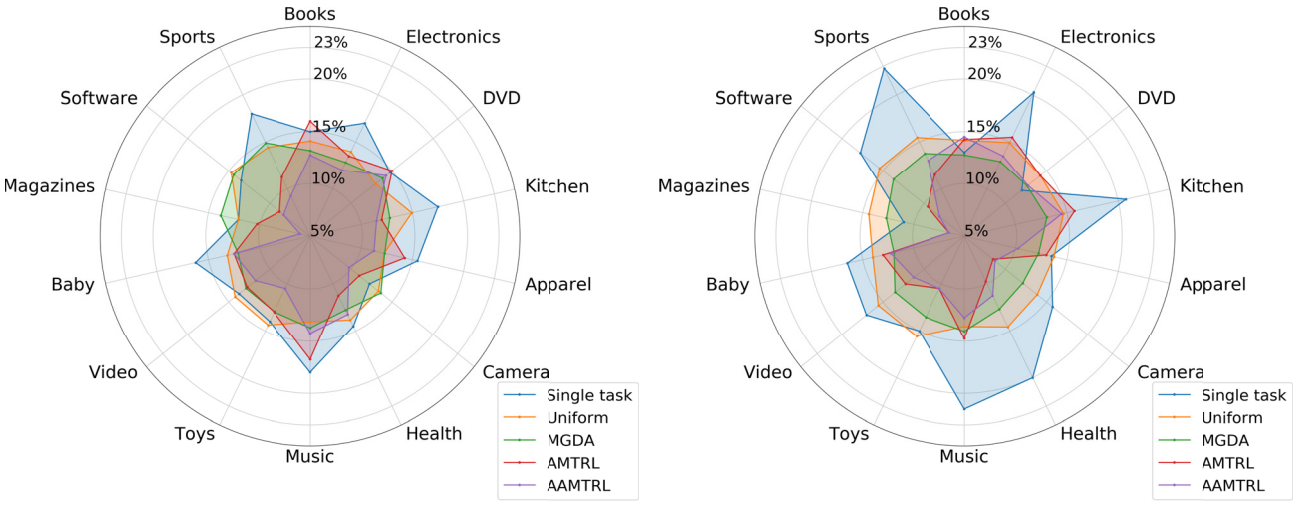
To evaluate the performance of the adversarial module for AAMTRL, we record the change in the relatedness matrix during training. In this experiment, the text CNN module is used to extract representation.

The relatedness matrix is summarized by the mean and variance of $\{R_1, R_2, ..., R_T\}$, where R_t for $t \in \{1, ..., T\}$ is defined in (21). Let R_{mean}, R_{var} be the mean and the variance respectively. The results for sentiment analysis and topic classification are shown in Fig. 3 and Fig.4 respectively.

$$R_t = \frac{1}{T}\sum_{k=0}^{T} R_{tk}. \tag{21}$$

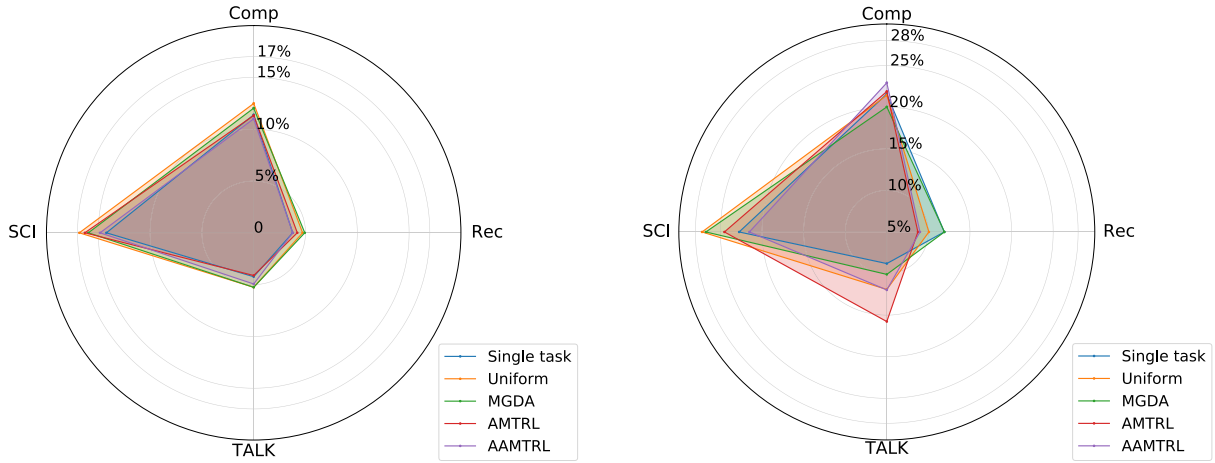
The results show the following:

- The proposed AAMTRL is able to enforce the tasks



(a) text-CNN. (b) BiLSTM.

Figure 5. Radar chart of the error rate for each task in sentiment analysis. (a) shows the results for MTRL models with text CNN-based representation extraction layers. (b) shows the results for MTRL models with BiLSTM-based representation extraction layers.



(a) text-CNN. (b) BiLSTM.

Figure 6. Radar chart of the error rate for each task in topic classification. (a) shows the results for MTRL models with text CNN-based representation extraction layers. (b) shows the results for MTRL models with BiLSTM-based representation extraction layers

to share an identical distribution in the representation space.

- The weighting strategy can accelerate and smooth the convergence process of the adversarial module during training.
- The tasks in sentiment analysis initially have a much closer relationship than those in topic classification.

5.2.2. Classification Accuracy

We compare our proposed methods with two baselines — (i) **Single Task**, which solves tasks independently, and (ii) **Uniform Scaling**, which minimizes a uniformly weighted sum of loss functions—as well as two state-of-the-art methods: (i) **MGDA** , which uses the MGDA-UB method proposed by (Sener & Koltun, 2018). (ii) **Adversarial MTRL**, which uses the original adversarial MTL framework proposed by (Liu et al., 2017).

We report the error rate of each task for sentiment analysis and topic classification in Figure 5 and Figure 6 respectively. The exact results can be referred to in the supplementary materials. The results shows the following:

- The proposed AAMTRL outperforms the state-of-the-

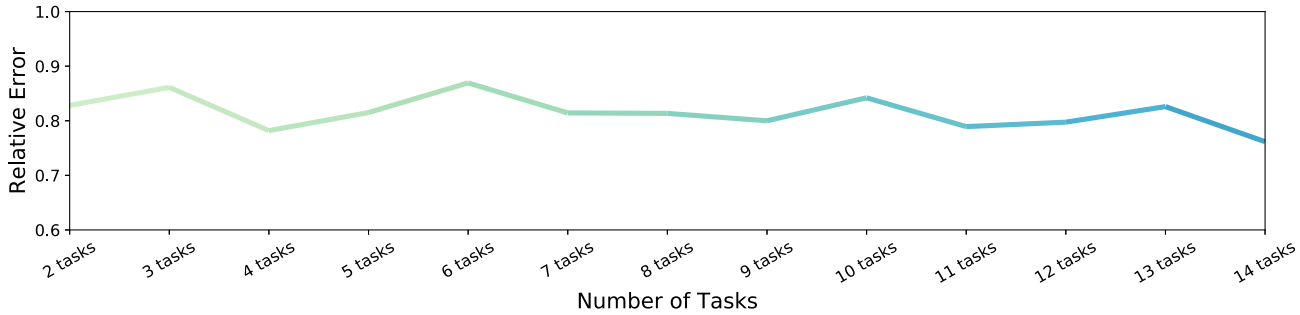


Figure 7. Change of the relative task-averaged risk along the number of tasks.

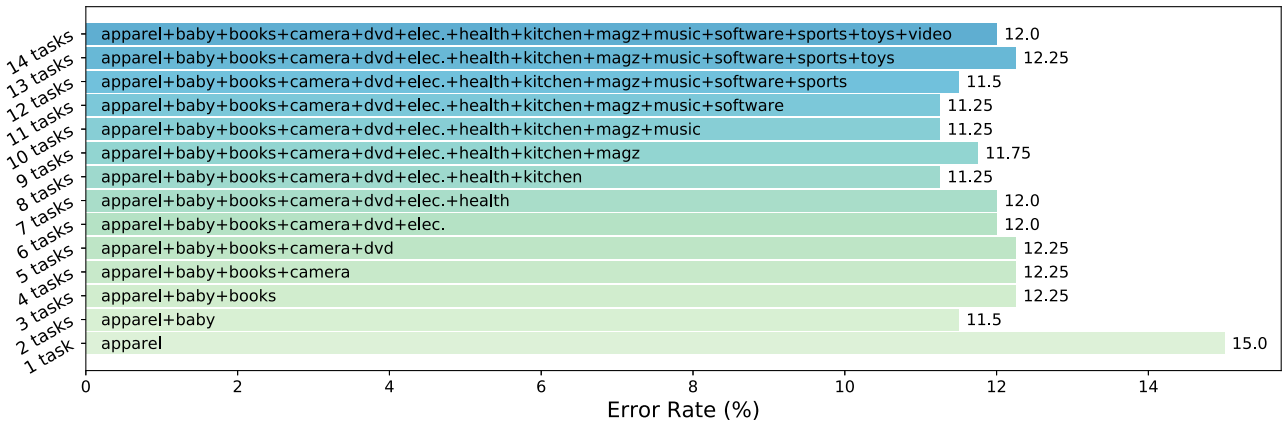


Figure 8. variety of the test error for task (Appeal) according to learning with different tasks.

art methods on sentiment analysis and achieves similar performance for topic classification.

- For topic classification, in which the tasks are not closely related (as shown in Figure 4 (a)), MTL strategies do not outperform single-task learning. This shows that the performance of MTL is dependent on the initial relatedness between tasks.

5.2.3. Influence of the Number of Tasks

In this section, we investigate the influence of the number of tasks on the task-averaged risk. We define a relative task-averaged risk with respect to single-task learning (STL) in (22).

$$er_{rel} = \frac{er_{MTL}}{\frac{1}{T}\sum_1^T er_{STL}^t}, \tag{22}$$

where er_{MTL} is the task-averaged test error of a MTL model, while er_{STL}^t is the test error of the STL model t. The MTL model and the STL models are the best-performing models generated from our experimental setting. The MTL model is trained using our AAMTRL algorithm.

We also carry out an experiment on sentiment analysis. In this experiment, the text CNN module is used to extract representation. Figure 7 presents the change in the relative task-averaged risk depending on the number of tasks. Figure 8 presents the variety of the test error for task (Appeal) according to learning with different tasks.

The results show the following:

- In AMTRL, an increase in task numbers does not decrease the task-averaged error.

- For a specific task in AMTRL, learning with more tasks does not guarantee better performance.

The results verify our analysis in Section 4.1.

6. Conclusion

While performance of AMTRL is attractive, the theoretical mechanism is unexplored. To fill this gap, we analyze the task-averaged generalization error bound for AMTRL. Based on the analysis, we propose a novel AMTRL method, named Adaptive AMTRL, that is designed to improve the performance of existing AMTRL methods. Numerical experiments support our theoretical results and demonstrate the effectiveness of our proposed approach.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grants 61976161.

References

Ando, R. K. and Zhang, T. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

Blitzer, J., Dredze, M., and Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

Caruana, R. Multitask learning. *Machine Learning*, 28(1): 41–75, 1997.

Chen, C., Yang, Y., Zhou, J., Li, X., and Bao, F. S. Cross-domain review helpfulness prediction based on convolutional neural networks with auxiliary domain discriminators. In *NAACL*, pp. 602–607, 2018a.

Chen, Z., Badrinarayanan, V., Lee, C., and Rabinovich, A. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, pp. 793–802, 2018b.

Collobert, R. and Weston, J. A unified architecture for natural language processing: deep neural networks with multitask learning. In *ICML*, pp. 160–167, 2008.

Dwivedi, K. and Roig, G. Representation similarity analysis for efficient task taxonomy & transfer learning. In *CVPR*, 2019.

Ganin, Y. and Lempitsky, V. S. Unsupervised domain adaptation by backpropagation. In *ICML*, pp. 1180–1189, 2015.

Hager, W. W. Dual techniques for constrained optimization. *Journal of Optimization Theory and Applications*, 55(1): 37–71, 1987.

Hestenes, M. R. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pp. 7482–7491, 2018.

Kim, Y. Convolutional neural networks for sentence classification. In *EMNLP*, pp. 1746–1751, 2014.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Lin, X., Zhen, H., Li, Z., Zhang, Q., and Kwong, S. Pareto multi-task learning. In *NeurIPS*, 2019.

Liu, P., Qiu, X., and Huang, X. Adversarial multi-task learning for text classification. In *ACL*, pp. 1–10, 2017.

Liu, Y., Wang, Z., Jin, H., and Wassell, I. J. Multi-task adversarial network for disentangled feature learning. In *CVPR*, pp. 3743–3751, 2018.

Mao, Y., Yun, S., Liu, W., and Du, B. Tchebycheff procedure for multi-task text classification. In *ACL*, 2020.

Maurer, A. A chain rule for the expected suprema of gaussian processes. In *ALT*, pp. 245–259, 2014.

Maurer, A., Pontil, M., and Romera-Paredes, B. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17:81:1–81:32, 2016.

McClure, P. and Kriegeskorte, N. Representational distance learning for deep neural networks. *Frontiers in Computational Neuroscience*, 10:131, 2016.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 2019.

Rockafellar, R. T. Augmented lagrange multiplier functions and duality in nonconvex programming. *SIAM Journal on Control*, 12(2):268–285, 1974.

Ruder, S. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098, 2017.

Sener, O. and Koltun, V. Multi-task learning as multi-objective optimization. In *NeurIPS*, pp. 525–536, 2018.

Shi, G., Feng, C., Huang, L., Zhang, B., Ji, H., Liao, L., and Huang, H. Genre separation network with adversarial training for cross-genre relation extraction. In *EMNLP*, pp. 1018–1023, 2018.

Yadav, S., Ekbal, A., Saha, S., Bhattacharyya, P., and Sheth, A. P. Multi-task learning framework for mining crowd intelligence towards clinical treatment. In *NAACL*, pp. 271–277, 2018.

Yu, J., Qiu, M., Jiang, J., Huang, J., Song, S., Chu, W., and Chen, H. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *WSDM*, pp. 682–690, 2018.