A Principled Approach for Learning Task Similarity in Multitask Learning

Changjian Shui†, Mahdieh Abbasi†, Louis-Émile Robitaille†,
Boyu Wang‡, Christian Gagné†

IJCAI 2019

† UNIVERSITÉ LAVAL ‡ Penn UNIVERSITY of PENNSYLVANIA

Multitask learning (MTL)

- MTL: learning a set of *related* tasks by using shared knowledge;

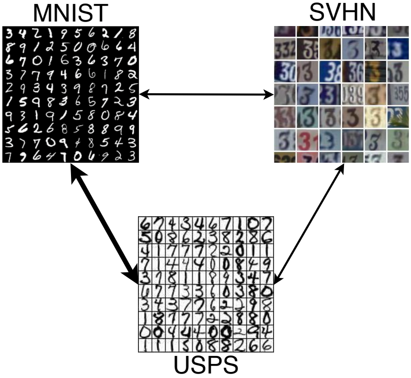
Multitask learning (MTL)

- MTL: learning a set of *related* tasks by using shared knowledge;
- Shared knowledge can improve the performance compared with learning individual tasks independently;

Multitask learning (MTL)

- MTL: learning a set of *related* tasks by using shared knowledge;
- Shared knowledge can improve the performance compared with learning individual tasks independently;
- How to express the shared knowledge?

Using task similarity as shared knowledge



- Intuition: *Tasks that are alike should be treated alike*

Our Contributions

- Theoretically prove the benefits of considering the task similarity: controlling the generalization error;

Our Contributions

- Theoretically prove the benefits of considering the task similarity: controlling the generalization error;
- A new training algorithm on deep neural network, based on the theoretical results
  - Developed for two task similarity metrics:
    - $\mathcal{H}$-divergence;
    - Wasserstein distance.

Problem setup

- Find T hypothesis $\{h_t\}_{t=1}^{T}$ from $\{\hat{\mathcal{D}}_t := (x_i, y_i)_{i=1}^{m}\}_{t=1}^{T}$;

Problem setup

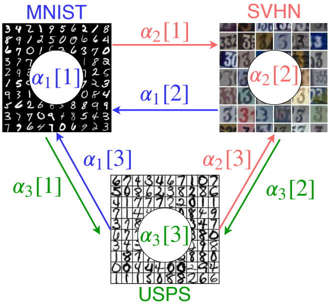
- Find T hypothesis $\{h_t\}_{t=1}^T$ from $\{\hat{\mathcal{D}}_t := (x_i, y_i)_{i=1}^m\}_{t=1}^T$;
- Generalization error:
  $\frac{1}{T}\sum_{t=1}^T R_t(h_t)$ with $R_t(h_t) = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\ell(h_t(\mathbf{x}), y)$;

Problem setup

- Find T hypothesis $\{h_t\}_{t=1}^{T}$ from $\{\hat{\mathcal{D}}_t := (x_i, y_i)_{i=1}^{m}\}_{t=1}^{T}$;
- Generalization error:
  $\frac{1}{T}\sum_{t=1}^{T} R_t(h_t)$ with $R_t(h_t) = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\ell(h_t(\mathbf{x}), y)$;
- *Relation coefficients*: $\{\boldsymbol{\alpha}_t\}_{t=1}^{T}$, each $\boldsymbol{\alpha}_t$ is T simplex;

Problem setup

- Find T hypothesis $\{h_t\}_{t=1}^T$ from $\{\hat{\mathcal{D}}_t := (x_i, y_i)_{i=1}^m\}_{t=1}^T$;
- Generalization error:
  $\frac{1}{T}\sum_{t=1}^T R_t(h_t)$ with $R_t(h_t) = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\ell(h_t(\mathbf{x}), y)$;
- *Relation coefficients*: $\{\boldsymbol{\alpha}_t\}_{t=1}^T$, each $\boldsymbol{\alpha}_t$ is T simplex;
- Empirical weighted loss for each task t:
  $\hat{R}_{\boldsymbol{\alpha}_t}(h) = \sum_{i=1}^T \boldsymbol{\alpha}_t[i]\hat{R}_i(h)$, $\hat{R}_i(h) = \frac{1}{m}\sum_{(x,y)\sim\hat{D}_i}\ell(h(x), y)$



Theoretical Result

Theorem (Wasserstein-1 distance, informal)

Supposing transport cost $c(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2$, with high probability $\geq 1-\delta$, we have:

$$\frac{1}{T}\sum_{t=1}^{T} R_t(h_t) \leq \underbrace{\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{\boldsymbol{\alpha}_t}(h_t)}_{\textit{Weighted empirical loss}} + \underbrace{C_1\sum_{t=1}^{T}\|\boldsymbol{\alpha}_t\|_2}_{\textit{Coefficient regularization}}$$

$$+\underbrace{\frac{2K}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_t[i]W_1(\hat{D}_t,\hat{D}_i)}_{\textit{Empirical distribution distance}} + \underbrace{C_2 + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_t[i]\lambda_{t,i}}_{\textit{Complexity \& optimal expected loss}}$$

C_1 and C_2 are constants related with Lipschtiz constant K, pseudo-dim d, m, T and δ.

Similar theoretical results with $\mathcal{H}$-divergence.

Key factors from the bounds

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{\alpha_t}(h_t)}_{\text{Weighted empirical loss}} + \underbrace{\frac{2K}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_t[i]W_1(\hat{D}_t,\hat{D}_i)}_{\text{Empirical distribution distance}} + \underbrace{C_1\sum_{t=1}^{T}\|\boldsymbol{\alpha}_t\|_2}_{\text{Coefficient regularization}} + \underbrace{C_2 + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_t[i]\lambda_{t,i}}_{\text{Complexity \& optimal expected loss}}$$

- According to the theoretical results, we should:
  1. Minimize the weighted prediction loss for each task, $\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{\alpha_t}(h_t)$;

Key factors from the bounds

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{\alpha_t}(h_t)}_{\text{Weighted empirical loss}} + \underbrace{\frac{2K}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\alpha_t[i]W_1(\hat{D}_t,\hat{D}_i)}_{\text{Empirical distribution distance}} + \underbrace{C_1\sum_{t=1}^{T}\|\boldsymbol{\alpha}_t\|_2}_{\text{Coefficient regularization}} + \underbrace{C_2 + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_t[i]\lambda_{t,i}}_{\text{Complexity \& optimal expected loss}}$$

- According to the theoretical results, we should:
    1. Minimize the weighted prediction loss for each task, $\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{\boldsymbol{\alpha}_t}(h_t)$;
    2. Minimize the weighted pairwise-distribution divergence, $\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_{t,i}d(\hat{\mathcal{D}}_t,\hat{\mathcal{D}}_i)$;

Key factors from the bounds

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{\boldsymbol{\alpha}_t}(h_t)}_{\text{Weighted empirical loss}} + \underbrace{\frac{2K}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_t[i]W_1(\hat{D}_t,\hat{D}_i)}_{\text{Empirical distribution distance}} + \underbrace{C_1\sum_{t=1}^{T}\|\boldsymbol{\alpha}_t\|_2}_{\text{Coefficient regularization}} + \underbrace{C_2 + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_t[i]\lambda_{t,i}}_{\text{Complexity \& optimal expected loss}}$$

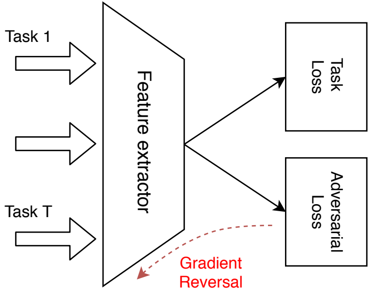
- According to the theoretical results, we should:
  1. Minimize the weighted prediction loss for each task, $\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{\boldsymbol{\alpha}_t}(h_t)$;
  2. Minimize the weighted pairwise-distribution divergence, $\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_{t,i}d(\hat{\mathcal{D}}_t,\hat{\mathcal{D}}_i)$;
  3. Control the relation coefficient $\{\boldsymbol{\alpha}_t\}_{t=1}^{T}$ (regularization term) $\sum_{t=1}^{T}\|\boldsymbol{\alpha}_t\|_2$.

Key factors from the bounds

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{\boldsymbol{\alpha}_t}(h_t)}_{\text{Weighted empirical loss}} + \underbrace{\frac{2K}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_t[i]W_1(\hat{D}_t,\hat{D}_i)}_{\text{Empirical distribution distance}} + \underbrace{C_1\sum_{t=1}^{T}\|\boldsymbol{\alpha}_t\|_2}_{\text{Coefficient regularization}} + \underbrace{C_2 + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_t[i]\lambda_{t,i}}_{\text{Complexity \& optimal expected loss}}$$

- According to the theoretical results, we should:
  1. Minimize the weighted prediction loss for each task, $\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{\boldsymbol{\alpha}_t}(h_t)$;
  2. Minimize the weighted pairwise-distribution divergence, $\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_{t,i}d(\hat{\mathcal{D}}_t,\hat{\mathcal{D}}_i)$;
  3. Control the relation coefficient $\{\boldsymbol{\alpha}_t\}_{t=1}^{T}$ (regularization term) $\sum_{t=1}^{T}\|\boldsymbol{\alpha}_t\|_2$.
- Underlying assumptions: optimal expected loss $\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\boldsymbol{\alpha}_{t,i}\lambda_{t,i}$ is *much smaller* than the empirical term.

Training Adversarial MultiTask Neural Network (AMTNN)



- A new training algorithm based on the mentioned factors;
- Task loss: weighted empirical loss;
- Adversarial loss: empirical distribution distance.

- Two kinds of parameters:
  - Neural networks parameters: $\boldsymbol{\theta}^{f}$ (feature extractor); $\boldsymbol{\theta}_{\cdot}^{h}$ (predictor); $\boldsymbol{\theta}_{\cdot}^{d}$ (discriminator);

Alternative training strategy

- Two kinds of parameters:
  - Neural networks parameters: $\boldsymbol{\theta}^f$ (feature extractor); $\boldsymbol{\theta}_\cdot^h$ (predictor); $\boldsymbol{\theta}_\cdot^d$ (discriminator);
  - Relation coefficients: $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T$.

Alternative training strategy

- Two kinds of parameters:
  - Neural networks parameters: $\boldsymbol{\theta}^{f}$ (feature extractor); $\boldsymbol{\theta}^{h}_{\cdot}$ (predictor); $\boldsymbol{\theta}^{d}_{\cdot}$ (discriminator);
  - Relation coefficients: $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T$.
- Alternative updating:
  1. Given a fixed coefficients, training adversarial multitask neural network;

Alternative training strategy

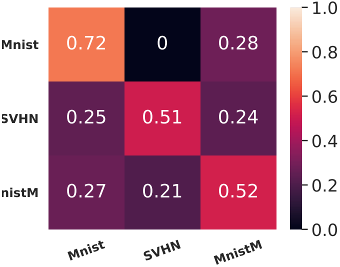
- Two kinds of parameters:
  - Neural networks parameters: $\boldsymbol{\theta}^{f}$ (feature extractor); $\boldsymbol{\theta}^{h}_{\cdot}$ (predictor); $\boldsymbol{\theta}^{d}_{\cdot}$ (discriminator);
  - Relation coefficients: $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T$.
- Alternative updating:
  1. Given a fixed coefficients, training adversarial multitask neural network;
  2. Given a fixed neural network, estimate $\{\boldsymbol{\alpha}_t\}_{t=1}^{T}$ through solving a convex optimization problem.

Empirical validation: digits recognition

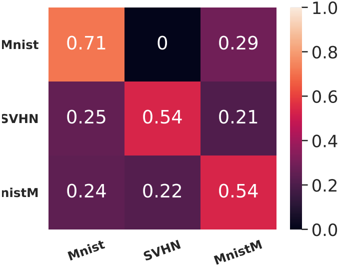
	3K				5K				8K			
Approach	**MNIST**	**MNIST_M**	**SVHN**	Average.	**MNIST**	**MNIST_M**	**SVHN**	Average	**MNIST**	**MNIST_M**	**SVHN**	Average
MTL_uni	93.23	76.85	57.20	75.76	97.41	77.72	67.86	81.00	97.73	83.05	71.19	83.99
MTL_weighted	89.09	73.69	68.63	77.13	91.43	74.07	73.81	79.77	92.01	76.69	73.77	80.82
MTL_disH	89.91	**81.13**	70.31	80.45	91.92	82.68	73.27	82.62	92.96	**85.04**	78.50	85.50
MTL_disW	96.77	80.38	68.40	81.85	95.47	**83.48**	72.66	83.87	98.09	84.13	74.37	85.53
AMTNN_H	**97.47**	77.87	71.26	82.20	**97.94**	76.28	76.06	83.43	**98.28**	82.75	76.63	85.89
AMTNN_W	97.20	80.70	**76.93**	**84.95**	97.67	82.50	**76.36**	**85.51**	98.01	82.53	**79.97**	**86.84**

- Improved performance ($\sim 1-2\%$), particularly on the SVHN ($\sim 4-6\%$);
- Similar results on the Amazon review dataset.

Robust and interpretable relation coefficient



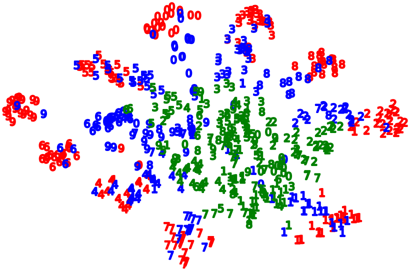
(a) Wasserstein



(b) $\mathcal{H}$-divergence

- Asymmetric relation coefficients
- For task MNIST, SVHN is not helpful;
- For task SVHN, MNIST is helpful.

Role of weighted sum



t-SNE plot of task MNIST.
Red: MNIST;
Blue: MNIST_M;
Green: SVHN.

Similar task naturally extends the decision boundary of the original task.

Thank You

Thanks for listening, for more information:

- Come and see the poster
- Paper link: `https://arxiv.org/abs/1903.09109`