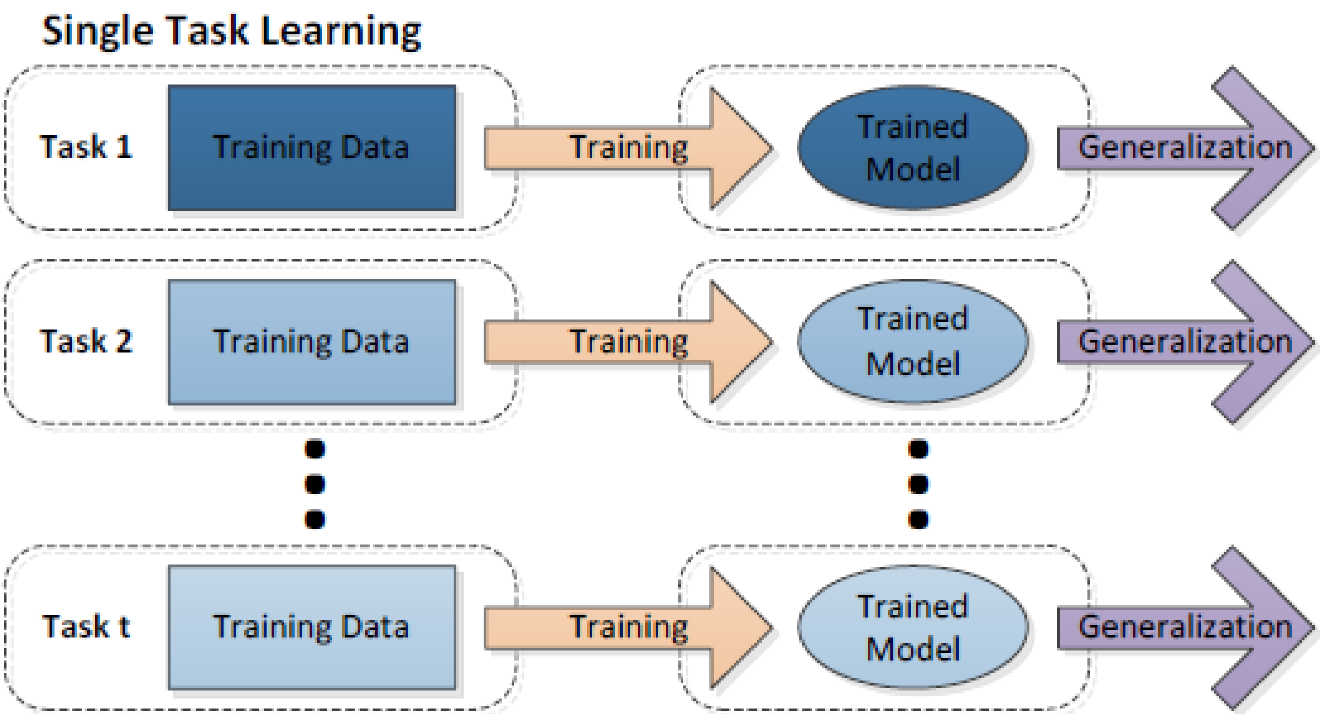
多任务学习及其在图像分类中的应用

报告的主要内容

- 几个常见的多任务学习模型
- 基于多任务学习的自回归分类模型
- 未来可以挖掘的潜力

- 几个常见的多任务学习模型
- 基于多任务学习的自回归分类模型
- 未来可以挖掘的潜力

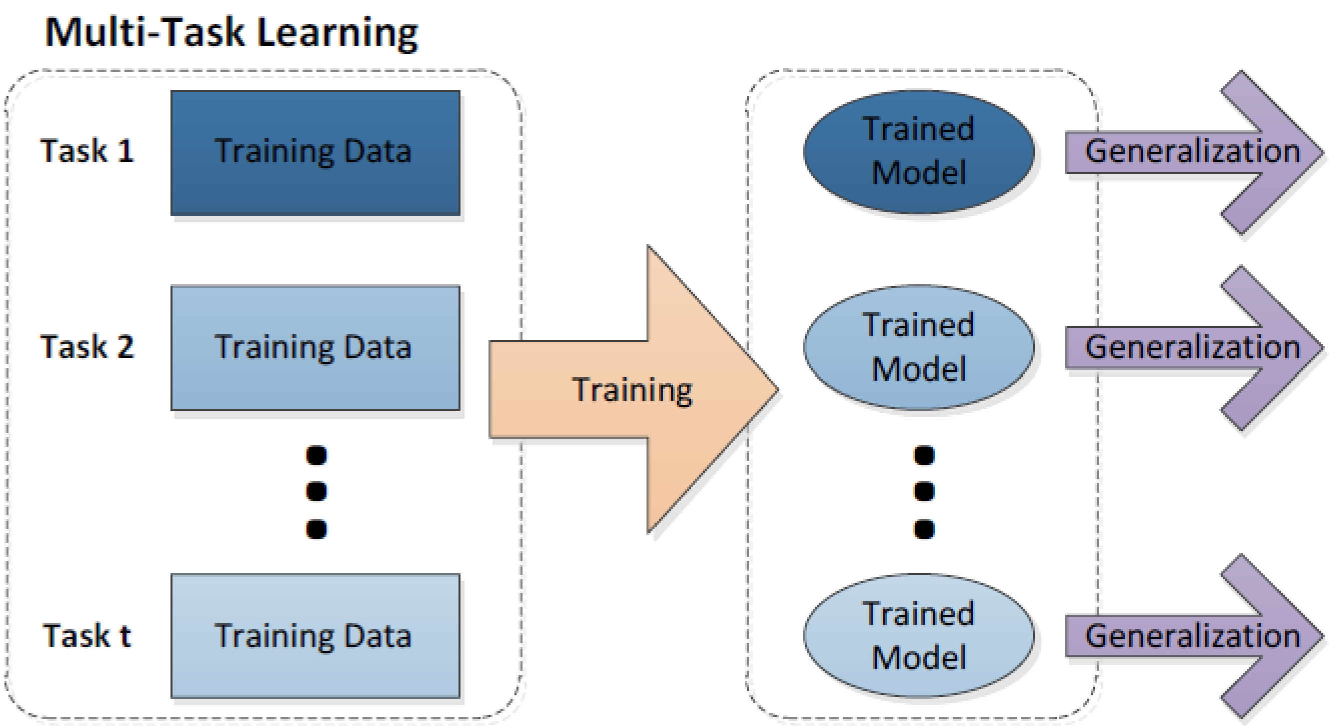
- 单任务学习



Single Task Learning
Task 1
Training Data
Training
Trained Model
Generalization
Task 2
Training Data
Training
Trained Model
Generalization
Task t
Training Data
Training
Trained Model
Generalization

Brief Introduction to Multi-task Learning

- 多任务学习



Brief Introduction to Multi-task Learning

- 多任务学习的构建原则
  - 建模任务之间的相关性；
  - 同时对多个任务的模型参数进行联合学习，挖掘其中的共享信息；
  - 考虑任务之间的差异性，增强模型的适应能力；

Brief Introduction to Multi-task Learning

- 多任务学习的两种主要方式
  - 基于参数的共享

    例如：神经网络隐层节点的共享

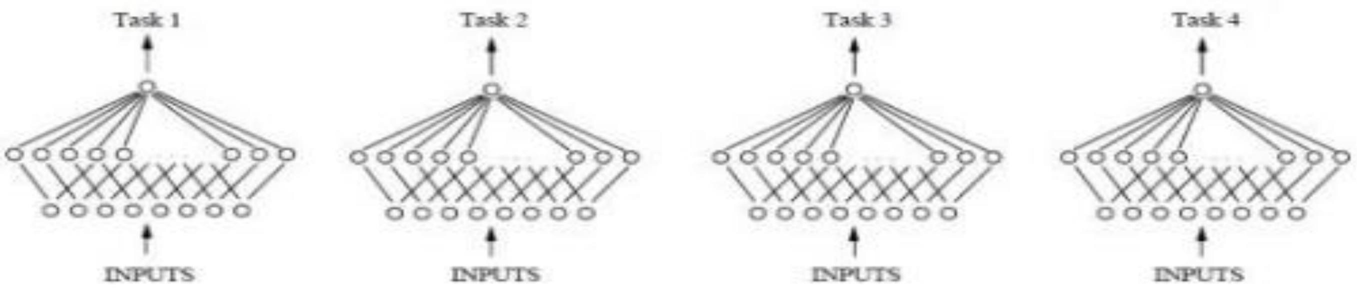
  - 基于正则化约束的共享

    例如：均值约束、联合特征学习等

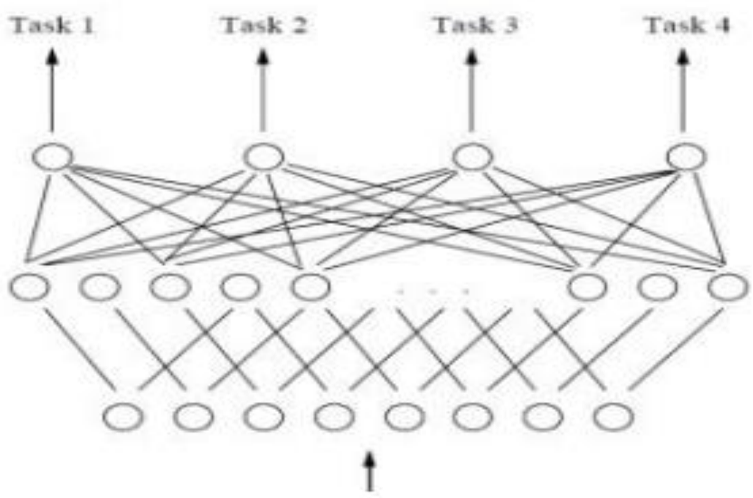
Brief Introduction to Multi-task Learning

- 参数共享-神经网络节点共享

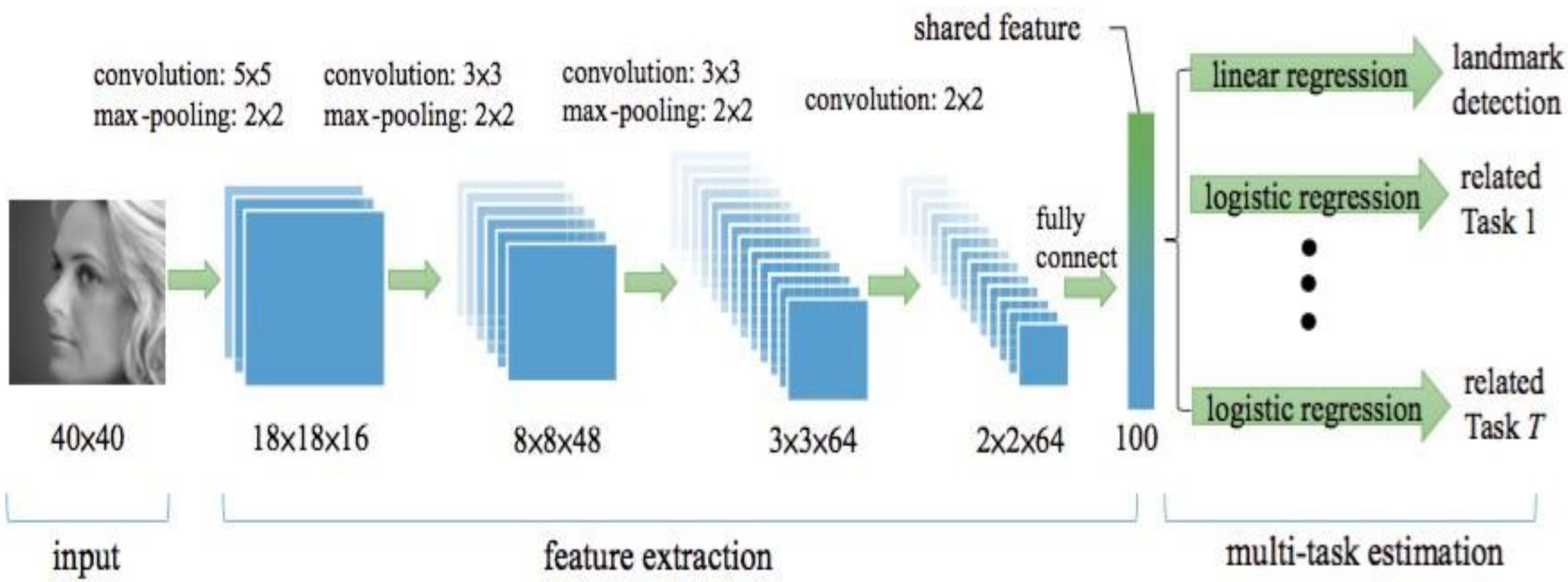
单任务学习



多任务学习



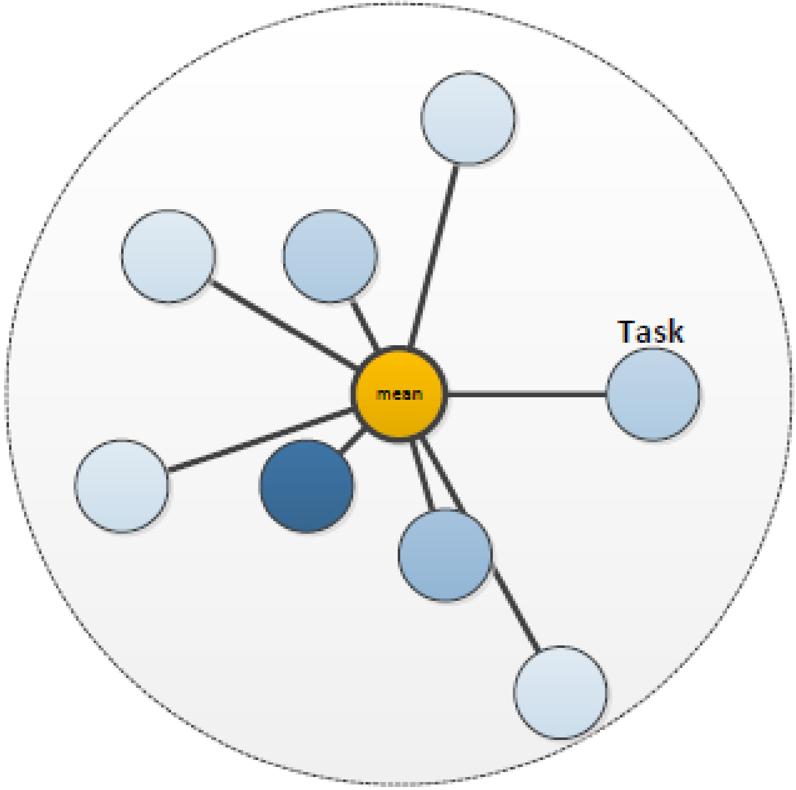
- 基于CNN的实例



Z Zhang, P Luo, CL Chen, X Tang. **Facial Landmark Detection by Deep Multi-task Learning**, **ECCV, 2014**

- 基于正则化约束的共享
  - 均值约束共享

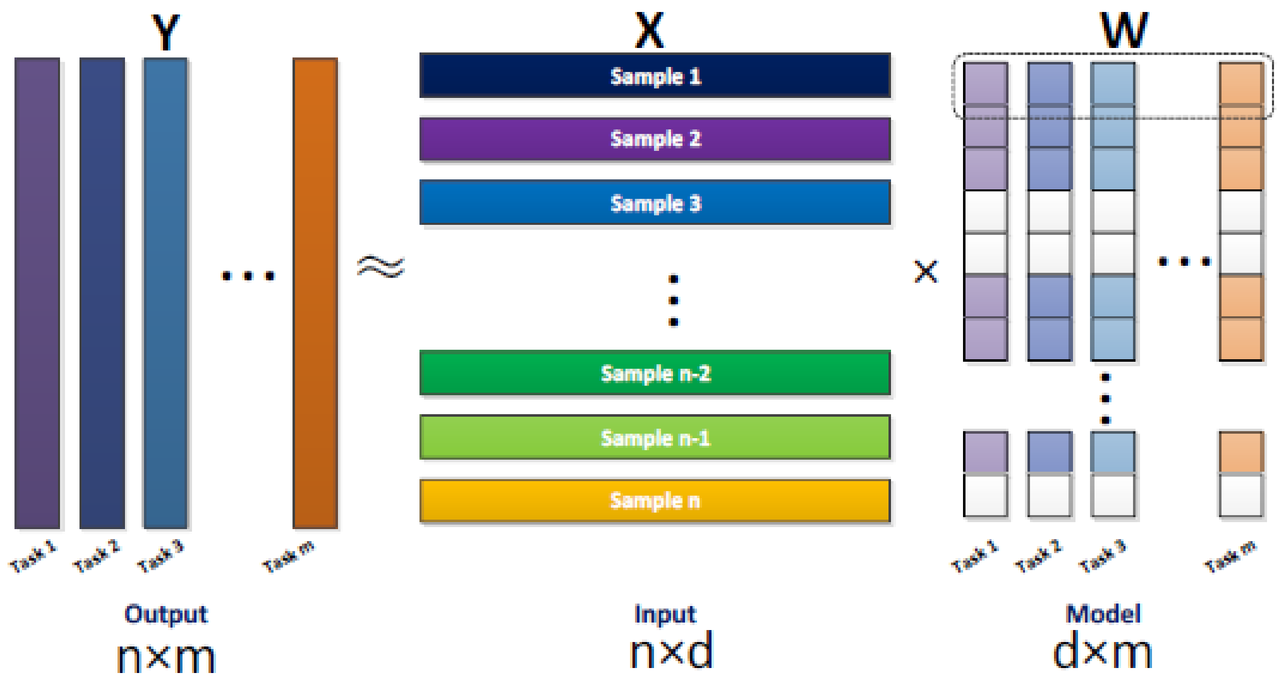
$$\min_{W} \frac{1}{2}\|XW - Y\|_F^2 + \lambda \sum_{i=1}^{m} \left\| W_i - \frac{1}{m}\sum_{s=1}^{m} W_s \right\|_2^2$$



Brief Introduction to Multi-task Learning

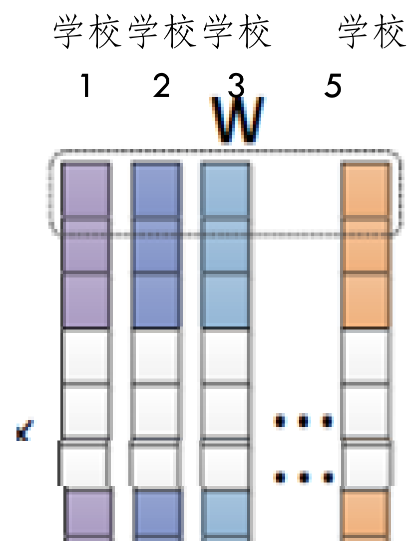
- 联合特征约束多任务学习

$$\min_{W} \frac{1}{2} \|XW - Y\|_F^2 + \lambda \|W\|_{1,q}$$

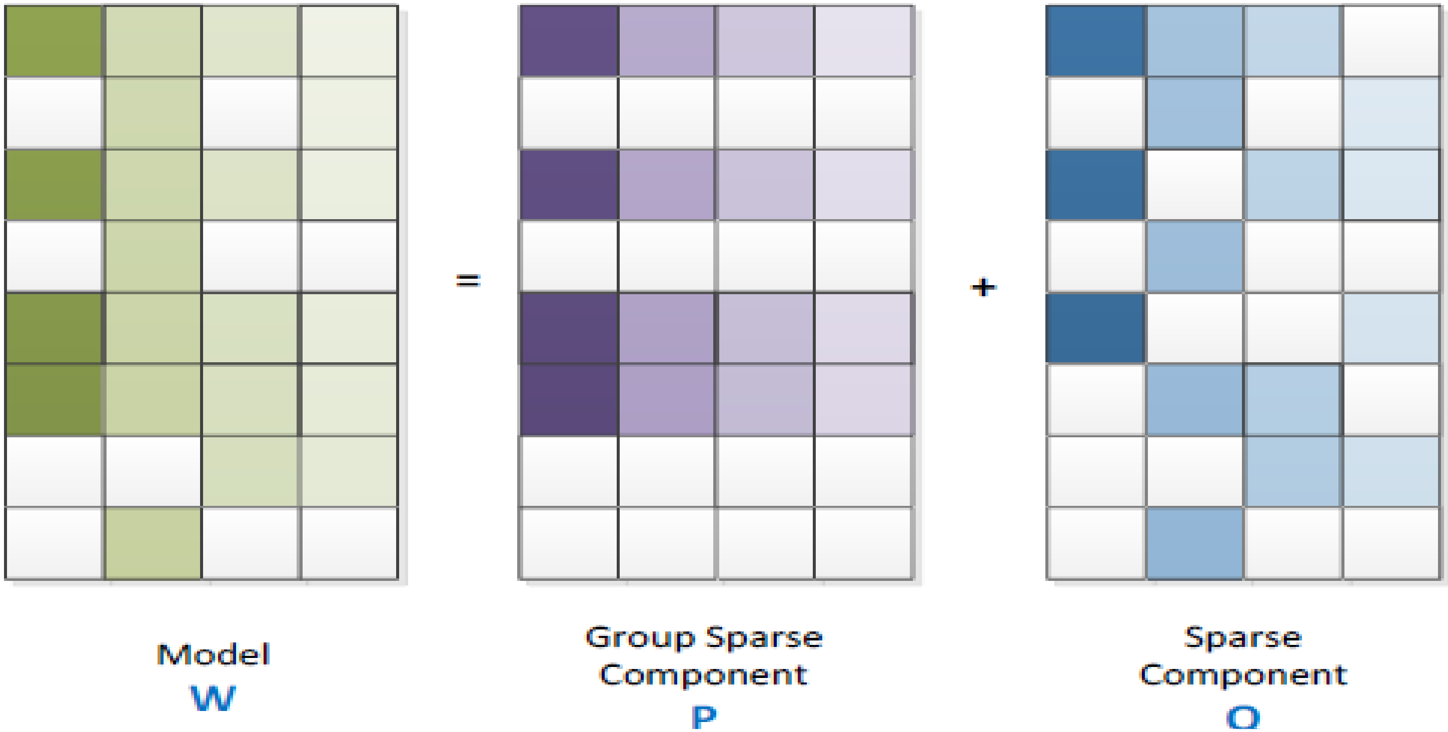


Brief Introduction to Multi-task Learning

$$
\begin{array}{c}
\begin{array}{ccccccc}
\text{上次成绩} & \text{上次排名} & \text{睡觉时间} & \text{学号} & \text{姓名} & \text{身高} & \text{父母教育背景}
\end{array} \\
Y = \begin{bmatrix}
87 & 12 & 22.8 & 76 & \text{张三} & 1.72 & \text{博士} \\
93 & 5 & 20.5 & 105 & \text{李四} & 1.65 & \text{本科} \\
80 & 35 & 21.2 & 35 & \text{王五} & 1.60 & \text{硕士} \\
52 & 102 & 19 & 57 & \text{赵六} & 1.83 & \text{初中} \\
65 & 78 & 20 & 82 & \text{宋七} & 1.78 & \text{高中}
\end{bmatrix}
\end{array}
$$



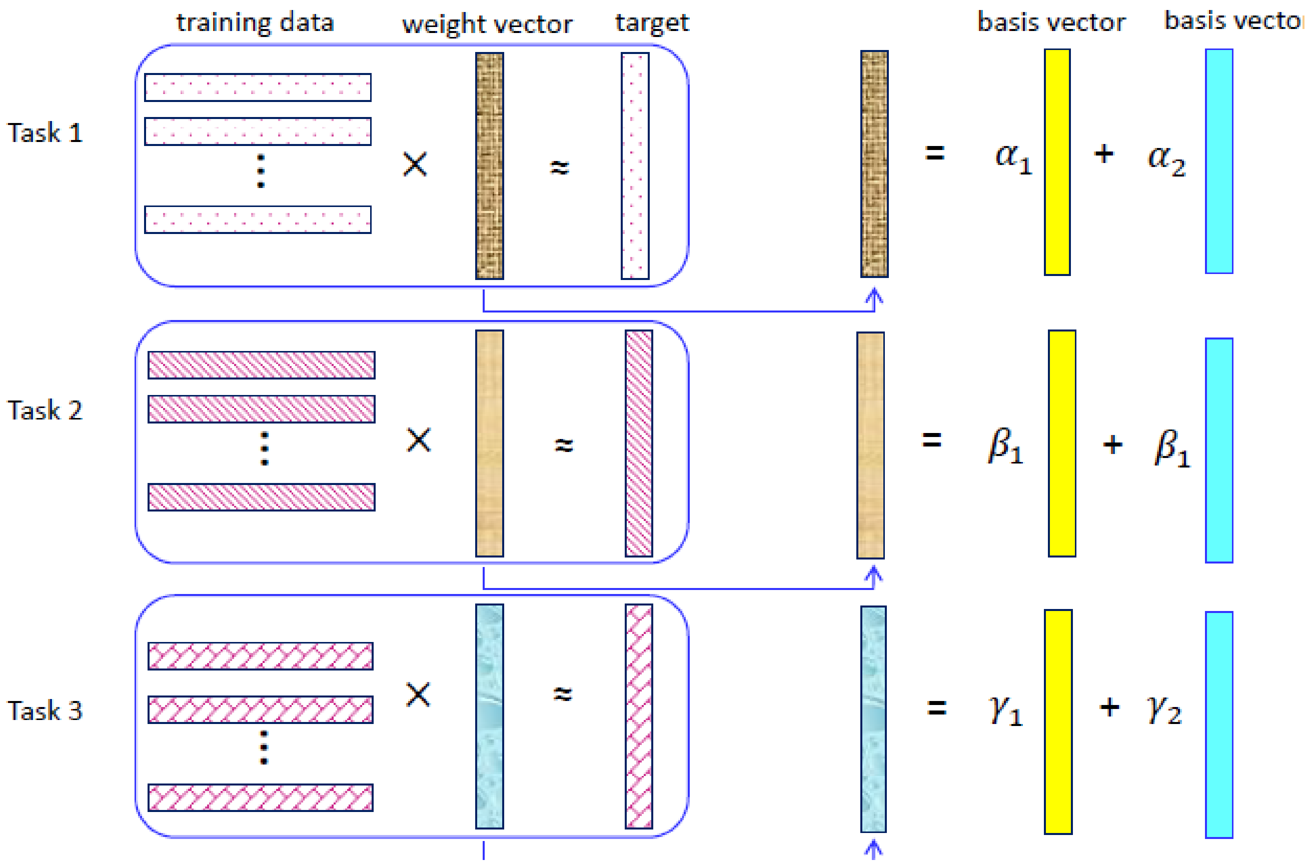
- Dirty Model 多任务学习



$$\min_{P,Q} \|Y - X(P+Q)\|_F^2 + \lambda_1 \|P\|_{1,q} + \lambda_2 \|Q\|_1$$

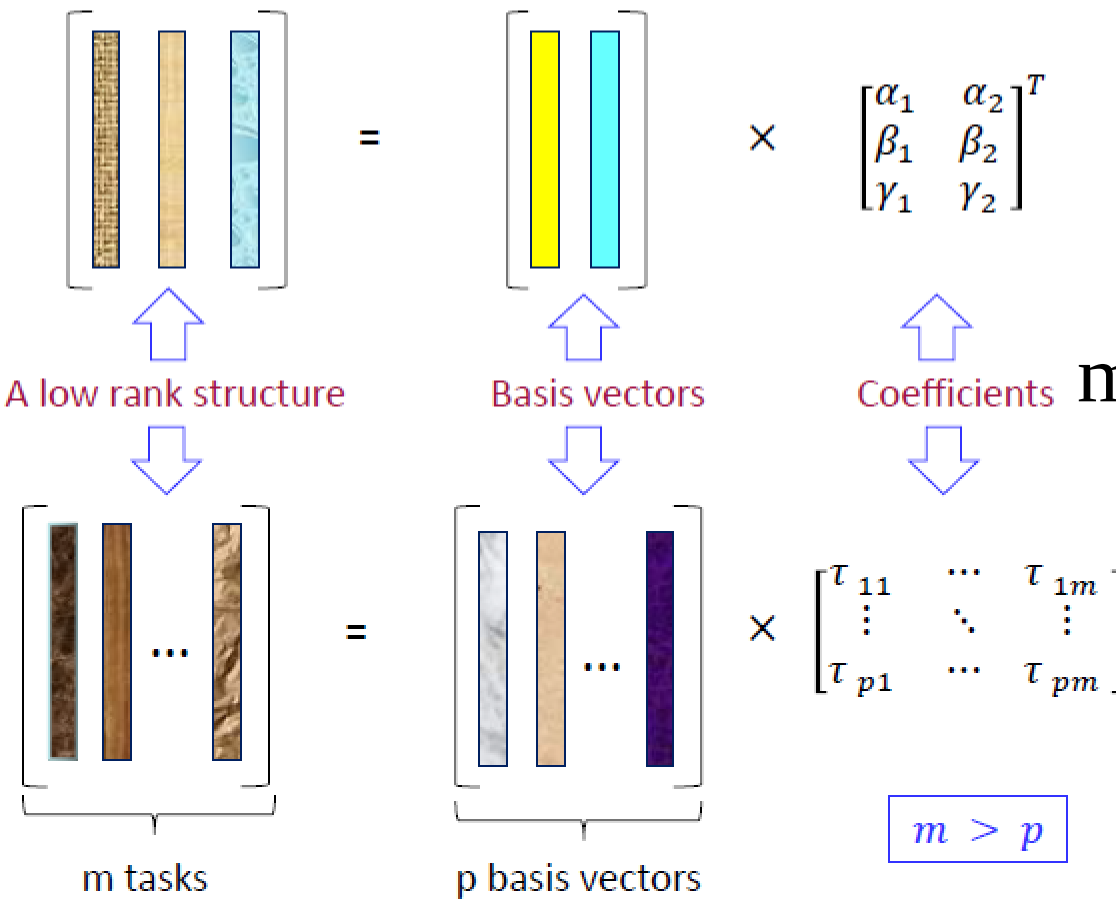
Brief Introduction to Multi-task Learning

- 低秩约束的多任务学习



多任务学习及其在图像分类中的应用

Brief Introduction to Multi-task Learning



$$\begin{bmatrix}\alpha_1 & \alpha_2\\ \beta_1 & \beta_2\\ \gamma_1 & \gamma_2\end{bmatrix}^T$$

$$\begin{bmatrix}\tau_{11} & \cdots & \tau_{1m}\\ \vdots & \ddots & \vdots\\ \tau_{p1} & \cdots & \tau_{pm}\end{bmatrix}$$

$$\min_{W} \|Y - XW\|_F^2 + \|W\|_*$$

Brief Introduction to Multi-task Learning

- 存在的问题
  - 认为任务之间的相关性仅与模型参数有关;
  - 忽略了样本原始特征的差异性;

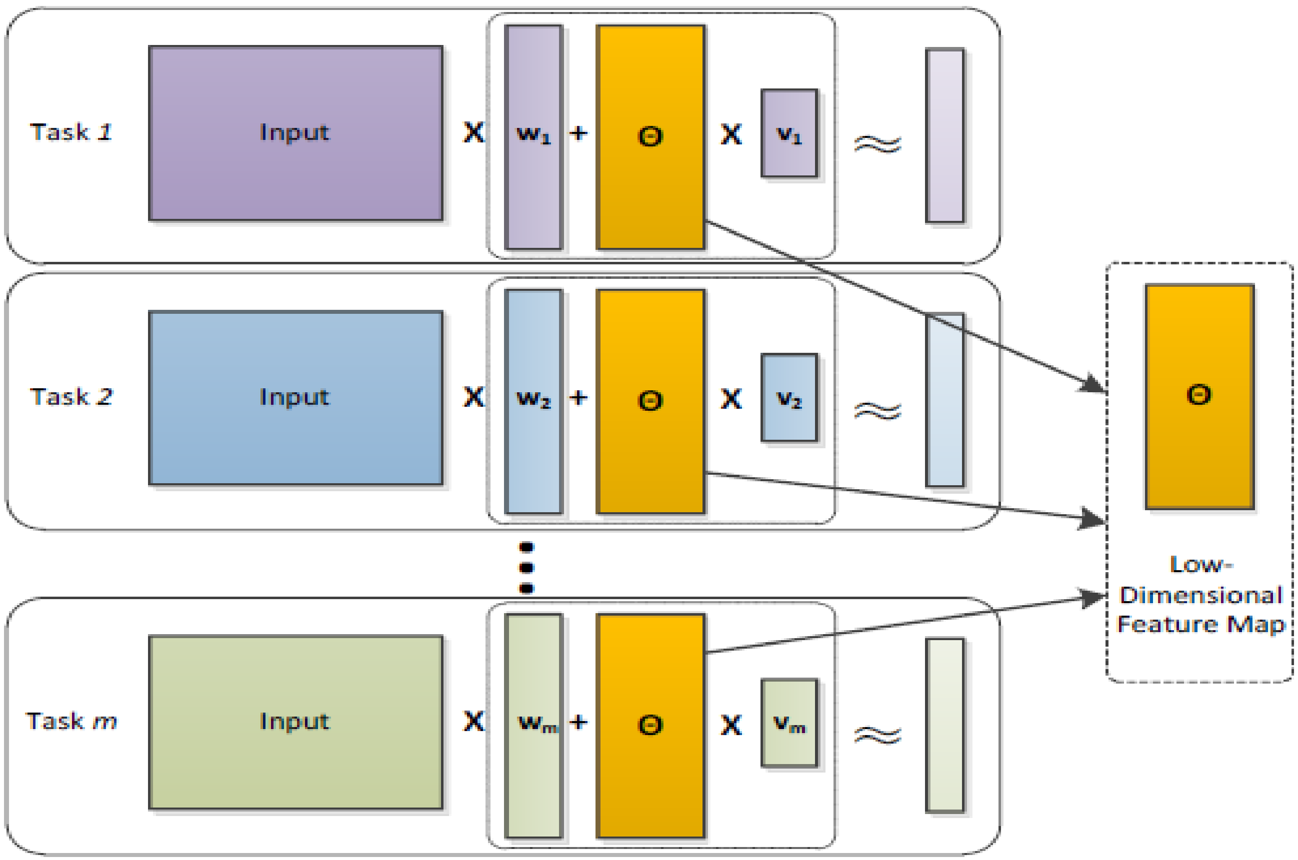
- 交替结构优化(Alternating Structure Optimization, ASO)的多任务学习

$$\mathbf{u}_i = \mathbf{w}_i + \boldsymbol{\theta}\mathbf{v}_i$$

$$\mathbf{u}_i^T x = (\mathbf{w}_i^T + \mathbf{v}_i^T \theta^T) x$$

$$\mathbf{U} = \mathbf{W} + \boldsymbol{\theta}\mathbf{V}$$

Brief Introduction to Multi-task Learning



基于多任务学习的图像分类

Brief Introduction to Multi-task Learning

- Loss function:

$$\mathcal{L}_i(X_i(\Theta v_i + w_i), y_i) = \|X_i(\Theta v_i + w_i) - y_i\|^2$$

- ASO:

$$\min_{\Theta,\{v_i,w_i\}} \sum_{i=1}^{m} \{\mathcal{L}_i(X_i(\Theta v_i + w_i), y_i) + \alpha\|w_i\|^2\}$$

$$\text{subject to} \quad \Theta^T\Theta = I$$

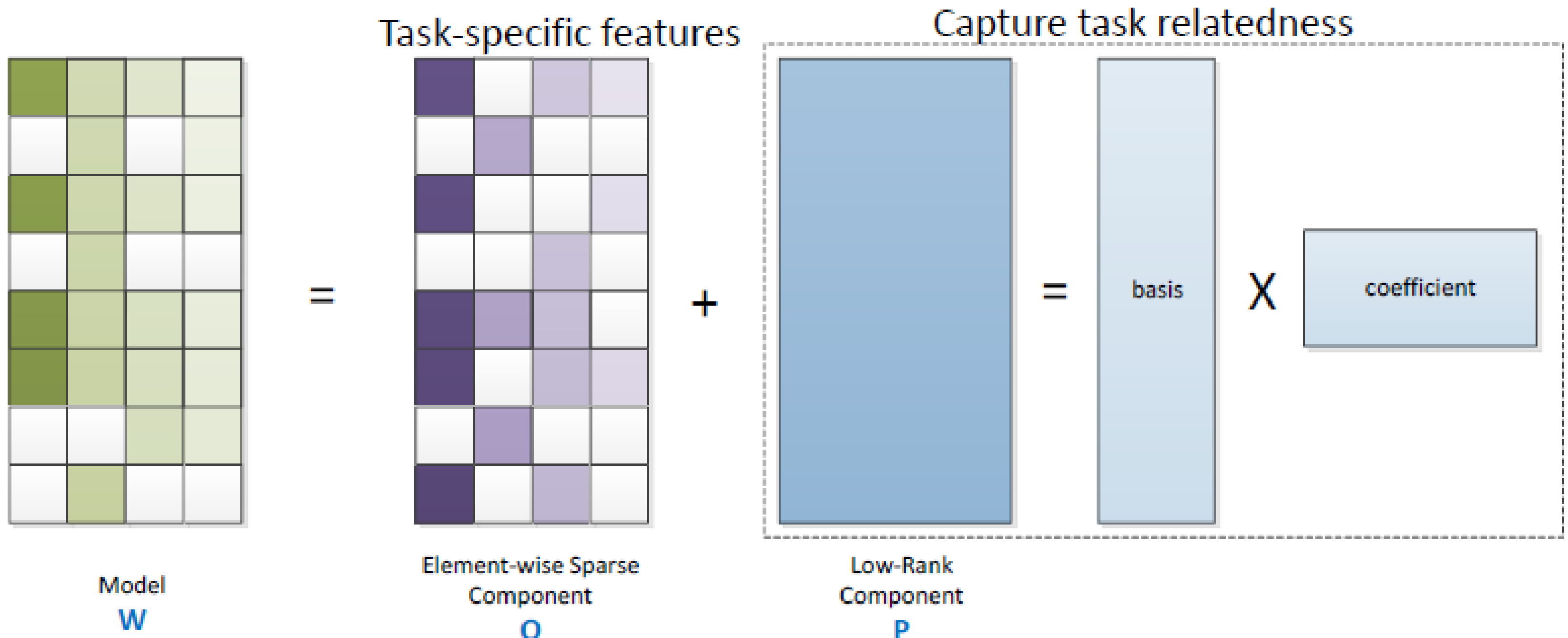
- iASO

$$\min_{\Theta,\{v_i,w_i\}} \sum_{i=1}^{m} \{\mathcal{L}_i(X_i(\Theta v_i + w_i), y_i) + \alpha\|\Theta v_i + w_i\|^2 + \beta\|w_i\|^2\}$$

$$\text{subject to} \quad \Theta^T\Theta = I$$

Brief Introduction to Multi-task Learning

- Incoherent Low Rank and Sparse Structure



$$\min_{P,Q} \sum_{i=1}^{m} \mathcal{L}_i(X_i(P_i + Q_i), y_i) + \lambda \|Q\|_1$$

$$\text{subject to} \quad \|P\|_* \leqslant \eta$$

- 多任务学习的构建原则
  - 如何建模任务之间的相关性；
  - 如何提取任务间的共享信息进行联合学习；
  - 如何体现任务之间的差异性；

- 几个常见的多任务学习模型
- 基于多任务学习的自回归分类模型
- 未来可以挖掘的潜力

Our work

传统多任务学习框架特点：

$$\min_{P,Q} \sum_{i=1}^{m} \mathcal{L}_i(X_i(P_i + Q_i), y_i) + \lambda \|Q\|_1$$

$$\text{subject to} \quad \|P\|_* \leqslant \eta$$

1. 能够用于解决多种机器学习问题：

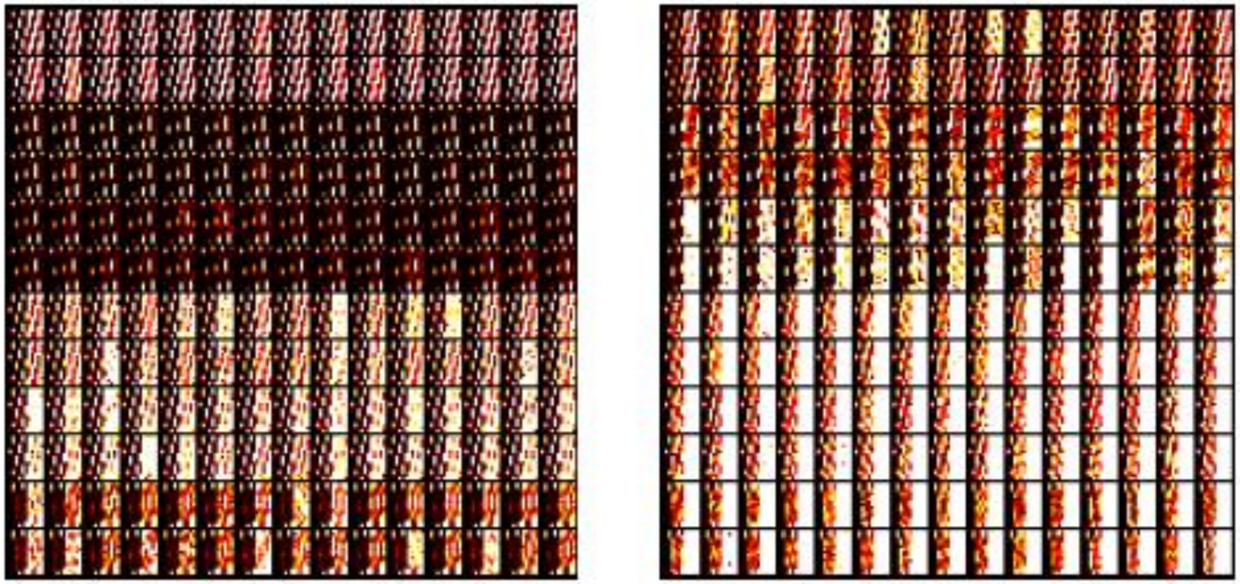
   (1) 股票价格预测　(2)学生成绩预测 …

2. 实际上是一个特征提取的过程，还需训练额外的分类器；

Our work

- 我们的工作思路：
  - 针对图像的结构，寻找任务之间的相关性来设计一种有效的多任务学习框架
  - 一次性的进行特征提取和分类器训练，将两个独立的过程统一到一个框架中；
  - 如兼顾的考虑共享特征与差异性，并给出适当的数值求解算法

Our work



(a) Gaussian Channel (b) Rayleigh Channel

https://cyclostationary.blog

基于多任务学习的图像分类

Our work

- 提出的损失函数：

$$L\left(x_n^i, W_S, W_i\right) = \left\| A\left(x_n\right) - \left(W_s + W_i\right) \Theta \left(P_i x_n\right) \right\|_2^2$$

$$\text{令：}\quad X = \left[X_1 \quad X_2 \quad \cdots \quad X_C \right] ,\quad X_i \in R^{d*n_i}$$

$$L\left(X, W_S, W_i\right) = \sum_i \sum_{n=1}^{n_i} L\left(x_n^i, W_S, W_i\right) = \sum_{i=1,2,\ldots C} \left\| A\left(X_i\right) - \left(W_S + W_i\right) \Theta \left(P_i X_i\right) \right\|_F^2$$

Our work

- The proposed objective function

$$\min_{\{W_i,P_i\}_{i=1}^{C},W_S} \sum_{i=1}^{C} \left(\left\| \mathcal{A}\left(X_i\right) - \left(W_S + W_i\right)\Theta\left(P_i X_i\right) \right\|_F^2 + \varphi\left(P_i\right) + \lambda \left\| W_i \right\|_1 \right) + \lambda_S \left\| W_S \right\|_F^2$$

$$\varphi\left(P_i\right) = \gamma \| P_i \|_F^2 + \theta \| P_i \bar{X}_i \|_F^2 \quad , \qquad \Theta_1\left(t\right) = \frac{t}{1+e^{-t}}$$

$$\Theta_2\left(t\right) = sign\left(t\right) Shrink_{\tau}\left(t\right)$$

Our work

- Numerical Solution

$$\min_{\{W_i,P_i,\Phi_i\},W_S} \sum_{i=1}^{C} \left(\|\mathcal{A}(X_i) - (W_S + W_i)\,\Phi_i\|_F^2 + \varphi(P_i) + \lambda \|W_i\|_1 \right) + \lambda_S \|W_S\|_F^2$$

$$s.t. \quad \Phi_i = \Theta(P_i X_i),\ i = 1, 2, \ldots, C$$

Our work

- Solve the following minimization problems iteratively:

① $$\min_{W_S} \sum_{i=1}^{C} \|\mathcal{A}(X_i) - (W_S + W_i)\,\Phi_i\|_F^2 + \lambda_S \|W_S\|_F^2$$

② $$\min_{W_i} \|\mathcal{A}(X_i) - (W_S + W_i)\,\Phi_i\|_F^2 + \lambda_i \|W_i\|_1$$

③ $$\min_{P_i,\Phi_i} \|\mathcal{A}(X_i) - (W_S + W_i)\,\Phi_i\|_F^2 + \varphi(P_i)$$
$$s.t. \quad \Phi_i = \Theta(P_i X_i)$$

Our work

- Minimization - 1

$$\min_{W_S} \sum_{i=1}^{C} \|\mathcal{A}(X_i) - (W_S + W_i)\,\Phi_i\|_F^2 + \lambda_S \|W_S\|_F^2$$

$$\frac{\partial J_{WS}}{\partial W_S} = 0 \quad \Rightarrow$$

$$W_S^* = \left(\sum_{i=1}^{C} (\mathcal{A}(X_i) - W_i\Phi_i)\,\Phi_i^T\right)\left(\sum_{i=1}^{C} \Phi_i\Phi_i^T + \lambda_S I\right)^{-1}$$

Our work

- Minimization - 2

$$\min_{W_i} \|\mathcal{A}(X_i) - (W_S + W_i)\Phi_i\|_F^2 + \lambda_i \|W_i\|_1$$

Suppose：

$$f(W_i) = \left\| A(X_i) - (W_s + W_i)\Phi_i \right\|_F^2$$

$$\hat{W}_i^{(k+1)} = \arg\min \left\| W_i - \left(W_i^{(k)} - \nabla f\left(W_i^{(k)}\right)\delta_i \right) \right\|_F^2 + \lambda \left\| W_i \right\|_1$$

Our work

- Minimization - 3

$$\min_{P_i,\Phi_i} \left\|\mathcal{A}\left(X_i\right)-\left(W_S+W_i\right)\Phi_i\right\|_F^2+\varphi\left(P_i\right)$$
$$s.t. \quad \Phi_i=\Theta\left(P_i X_i\right)$$

$$\min_{P_i,\Phi_i} \left\|\mathcal{A}\left(X_i\right)-\left(W_S+W_i\right)\Phi_i\right\|_F^2+\varphi\left(P_i\right)$$
$$s.t. \quad \Phi_i=\Theta\left(Q_i\right), Q_i=P_i X_i$$

Our work

- The augmented Lagrangian function:

$$\begin{aligned}\min_{P_i,\Phi_i,Q_i,U_1,U_2} & \left\|\mathcal{A}\left(X_i\right)-\left(W_S+W_i\right)\Phi_i\right\|_F^2+\varphi\left(P_i\right)\\ & +\frac{\rho}{2}\left\|\Phi_i-\Theta\left(Q_i\right)\right\|_F^2+\frac{\rho}{2}\left\|Q_i-P_iX_i\right\|_F^2\\ & +\left\langle U_1,\Phi_i-\Theta\left(Q_i\right)\right\rangle+\left\langle U_2,Q_i-P_iX_i\right\rangle\end{aligned}$$

Our work

- With the ADMM, the minimization can be solved by the following three subproblems:

$$\Phi_i^{(t+1)} = \arg\min_{\Phi_i} \|\mathcal{A}(X_i) - (W_S + W_i)\,\Phi_i\|_F^2 + \frac{\rho}{2}\|\Phi_i - \Theta\left(Q_i^{(t)}\right) + \frac{1}{\rho}U_1^{(t)}\|_F^2$$

$$P_i^{(t+1)} = \arg\min_{P_i} \frac{\rho}{2}\|Q_i^{(t)} - P_i X_i + \frac{1}{\rho}U_2^{(t)}\|_F^2 + \gamma\|P_i\|_F^2 + \theta\|P_i \bar{X}_i\|_F^2$$

$$Q_i^{(t+1)} = \arg\min_{Q_i} \frac{\rho}{2}\|\Phi_i - \Theta(Q_i) + \frac{1}{\rho}U_1^{(t)}\|_F^2 + \frac{\rho}{2}\|Q_i - P_i X_i + \frac{1}{\rho}U_2^{(t)}\|_F^2$$

Our work

- Overall Algorithm

Input: Training data set $\{X_i\}_{i=1}^{C}$,self-projected operator $\mathcal{A}$ regularization parameters γ, θ, λ, λ_S ρ initial random $W_S^{(0)}$,$\{W_i^{(0)}, P_i^{(0)}, \Phi_i^{(0)}\}_{i=1}^{C}$ $U_1^{(0)} = U_2^{(0)} = \mathbf{0}$, iteration number T, initial $k = 0$.

Repeat

1. Calculate matrix $W_S^{(k+1)}$ with Eq.(10);
2. Calculate matrix $W_i^{(k+1)}$ for each task with Eq.(12);
3. **for** $t = 0$ to $T - 1$ **do**, $(i = 1, 2, \ldots, C)$

    Update $\Phi_i^{(t+1)}$ in subproblem in Eq.(16);

    Update $P_i^{(t+1)}$ in subproblem in Eq.(17);

    Update $Q_i^{(t+1)}$ in subproblem in Eq.(18);

    Update Lagrangian multipliers:

$$U_1^{(t+1)} = U_1^{(t)} + \rho\left(\Phi_i^{(t+1)} - \Theta\left(Q_i^{(t+1)}\right)\right)$$

$$U_2^{(t+1)} = U_2^{(t)} + \rho\left(Q_i^{(t+1)} - P_i^{(t+1)} X_i\right)$$

   end for

4. For each i, $\Phi_i^{(k+1)} \leftarrow \Phi^T$, $P_i^{(k+1)} \leftarrow P^T$, $Q_i^{(k+1)} \leftarrow Q^T$
5. $k \leftarrow k + 1$;

Until convergence

Output: W_S^*,$\{W_i^*, P_i^*\}_{i=1}^{C}$

Our work

- Classification Criterion

$$\hat{l}_t = \arg\min_i \left\| \mathcal{A}(x_t) - (W_S + W_i)\,\Theta\,(P_i x_t) \right\|_F^2$$

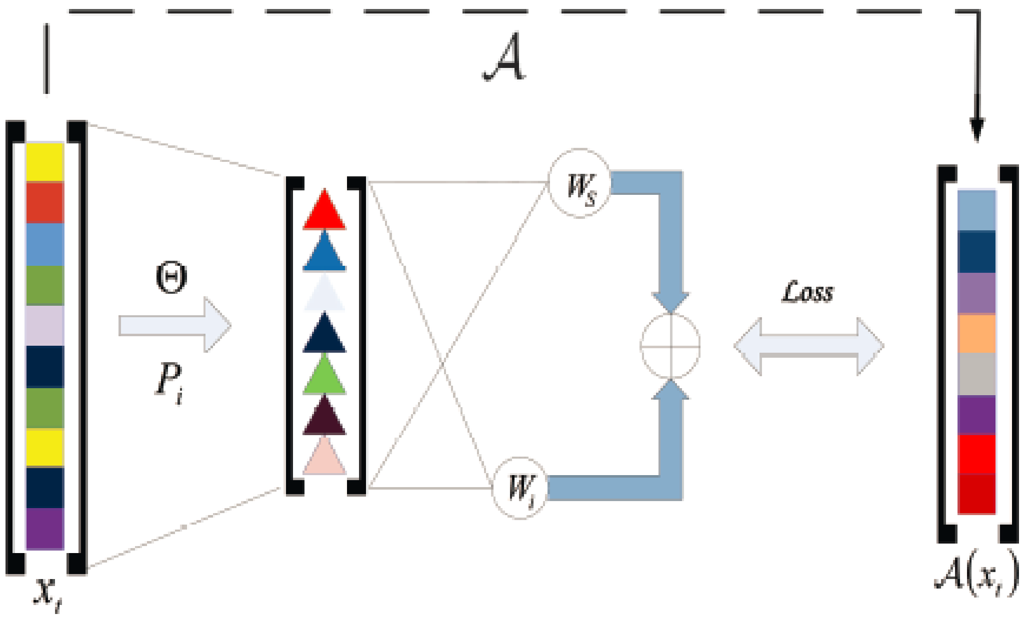
$$p\left(x_n \middle| y_n, \{W_i\}_{i=1}^{C}, W_s\right) = \sum_i y_{ni} \exp\left\{ -\left\| A(x_n) - (W_s + W_i)\, x_n \right\|_2^2 \right\}$$

$$-\log P\left(X \middle| Y, \{W_i\}_{i=1}^{C}, W_s\right) = -\log \prod_n p\left(x_n \middle| y_n\right)$$

$$= \sum_i \sum_n y_{ni} \left\| A(x_n) - (W_s + W_i)\, S(P_i x_n) \right\|_2^2$$

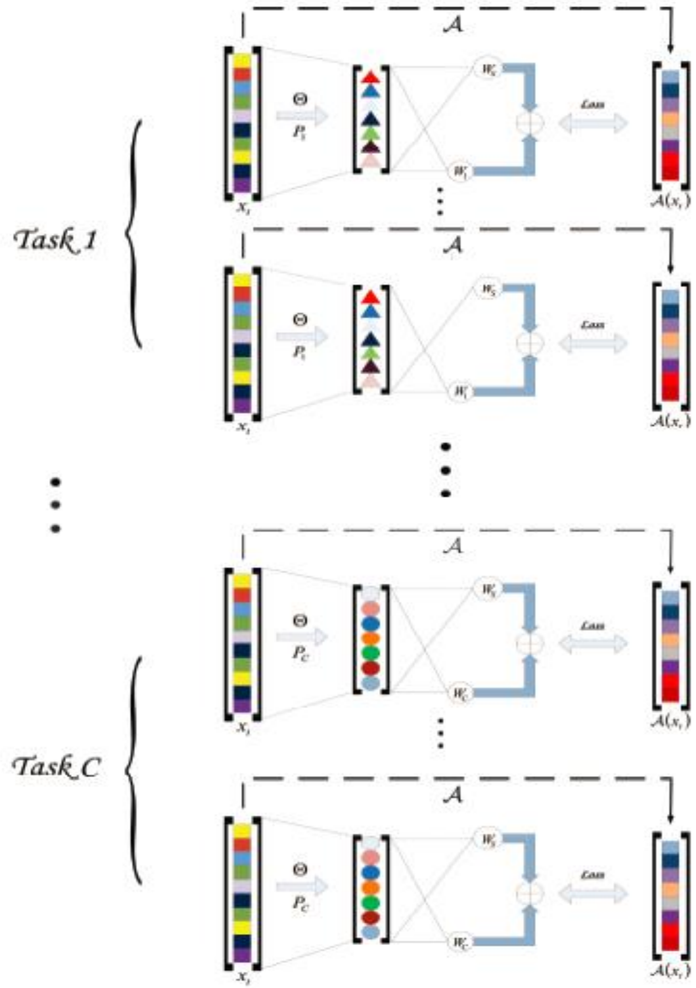
Our work

- Structure for each task



Our work

- Structure of our proposed model



Our work

- Experimental results

	Extended YaleB	WSU-HOS-G	WSU-HOS-R
SRC	95.42±0.47	92.87±1.56	79.85±2.79
CRC	97.23±0.35	96.42±0.66	83.62±0.93
ProCRC	93.25±1.28	95.87±1.93	88.89±0.62
NNGR	92.32±1.43	95.47±1.40	82.86±1.33
Ours with Θ_1	96.98±0.28	98.63±1.29	**90.63±1.23**
Ours with Θ_2	**97.08±0.48**	**98.79±0.23**	89.80±0.88

Our work

- Scene Classification

	Scene-15
SPM	79.95±0.27
LLC	79.81±0.35
LLCDC	80.30±0.62
LLCDCSIFT	**82.40±0.35**
Ours with Θ_1	82.23±0.49
Ours with Θ_2	81.51±0.54

Our work

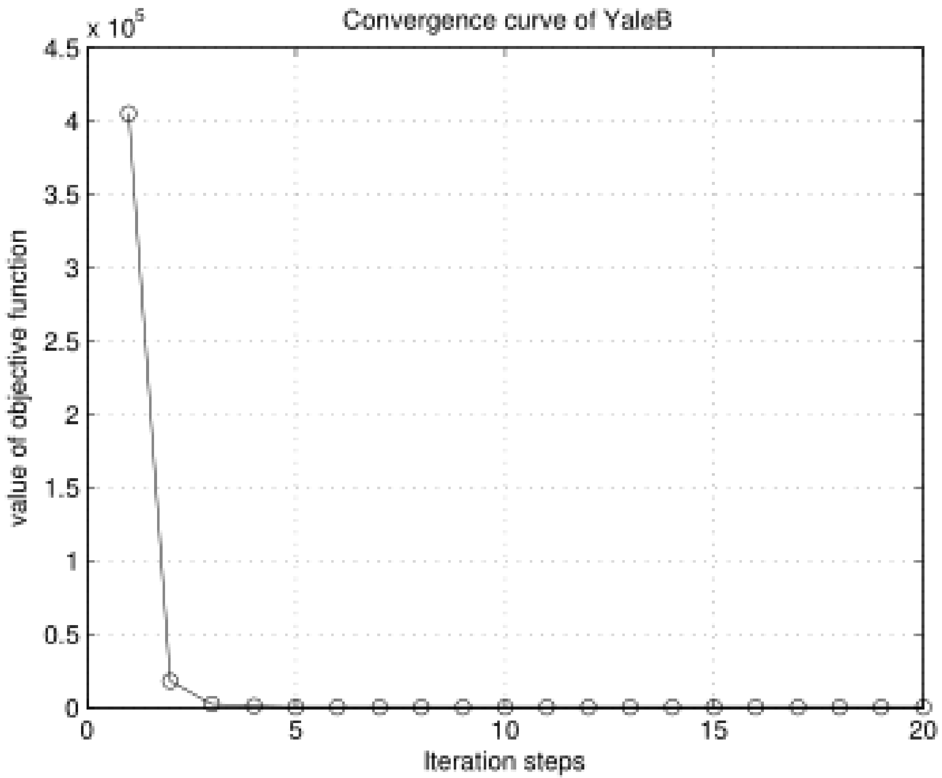
	suburb	coast	forest	highway	insidecity	mountain	opencountry	street	tallbuilding	office	bedroom	industrial	kitchen	livingroom	store
suburb	0.9929	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0071	0.0000
coast	0.0000	0.8692	0.0038	0.0192	0.0000	0.0077	0.0962	0.0038	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
forest	0.0000	0.0000	0.9549	0.0000	0.0000	0.0219	0.0132	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
highway	0.0000	0.0625	0.0000	0.8938	0.0000	0.0125	0.0250	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0063
insidecity	0.0048	0.0000	0.0000	0.0048	0.8173	0.0048	0.0096	0.0240	0.0337	0.0048	0.0000	0.0288	0.0192	0.0048	0.0433
mountain	0.0000	0.0219	0.0219	0.0036	0.0000	0.9015	0.0474	0.0000	0.0036	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
opencountry	0.0065	0.1226	0.0258	0.0032	0.0000	0.0387	0.7903	0.0129	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
street	0.0000	0.0000	0.0052	0.0104	0.0417	0.0052	0.0104	0.9010	0.0052	0.0000	0.0000	0.0208	0.0000	0.0000	0.0000
tallbuilding	0.0000	0.0000	0.0078	0.0000	0.0273	0.0156	0.0000	0.0039	0.9063	0.0000	0.0000	0.0195	0.0039	0.0000	0.0156
office	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9304	0.0087	0.0000	0.0435	0.0087	0.0087
bedroom	0.0000	0.0086	0.0000	0.0000	0.0000	0.0086	0.0000	0.0000	0.0000	0.0431	0.7672	0.0086	0.0345	0.1034	0.0259
industrial	0.0142	0.0474	0.0142	0.0047	0.0474	0.0521	0.0237	0.0332	0.0900	0.0095	0.0095	0.4882	0.0142	0.0095	0.1422
kitchen	0.0000	0.0000	0.0000	0.0000	0.0182	0.0091	0.0000	0.0000	0.0000	0.0273	0.0818	0.0091	0.7000	0.1000	0.0545
livingroom	0.0000	0.0000	0.0000	0.0000	0.0159	0.0053	0.0000	0.0053	0.0053	0.0476	0.1693	0.0212	0.1005	0.5344	0.0952
store	0.0000	0.0047	0.0186	0.0000	0.0326	0.0233	0.0000	0.0093	0.0047	0.0093	0.0140	0.0140	0.0326	0.0233	0.8140

Our work

- Visualization Result

Our work

- Convergence Analysis



- 几个常见的多任务学习模型
- 基于多任务学习的自回归分类模型
- 未来可以挖掘的潜力

- 未来可以尝试的工作：

（1）降维算子学习；

（2）非线性映射方式：使用核技术

（3）考虑引入属性特征应用于其它视觉任务；

Thank you for your attentions!