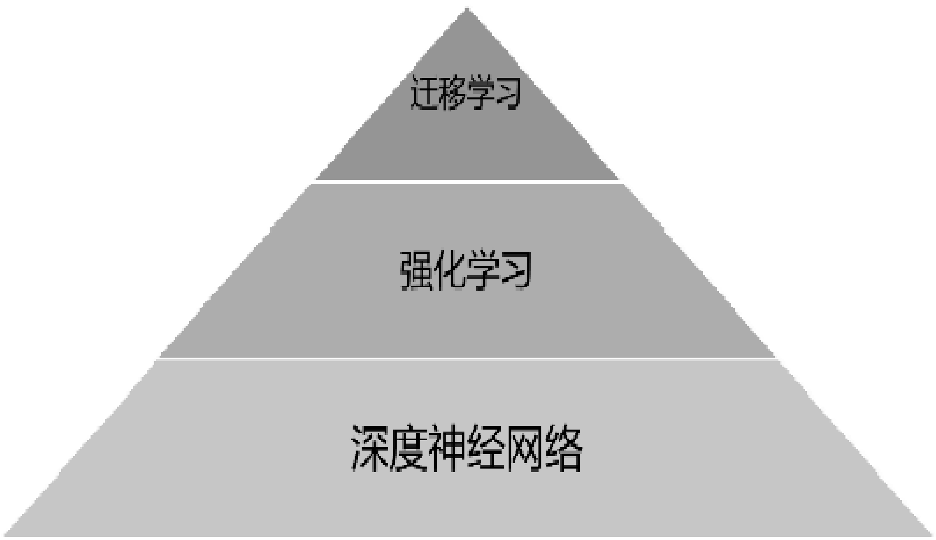
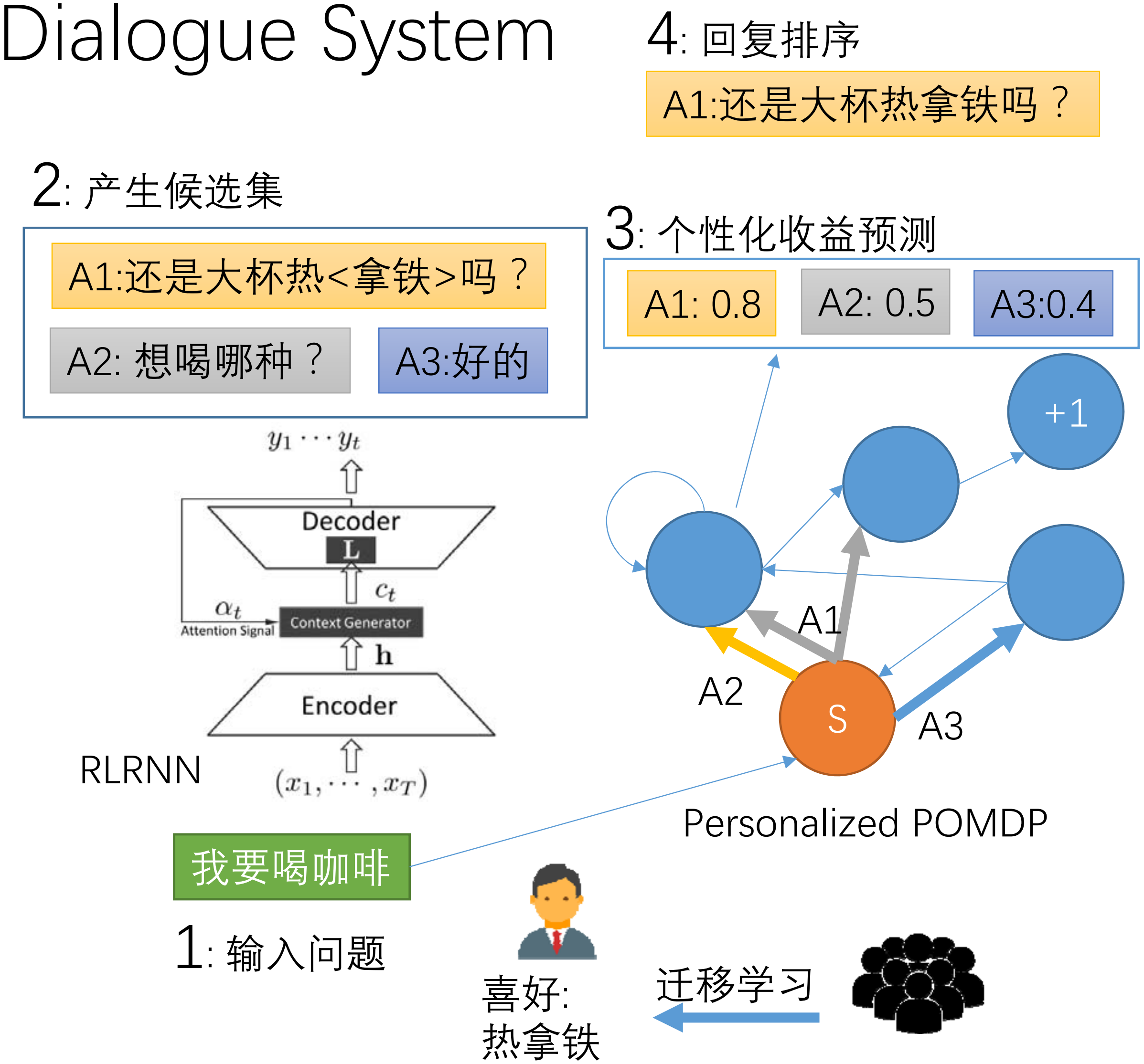
Research at Big Data Institute

Personalized Task-Oriented Dialogue System

Reinforcement Transfer Learning (RTL)



问题输入：用户问题
问题输出：系统回复
目标：帮助用户用最少的时间点咖啡



Demo

- 操作提示:
  - “PERSONALIZE”: 进入个性化模型
  - “RESET”: 开始点一杯新的咖啡
  - “RESETUSER”: 删除当前用户的口味偏好

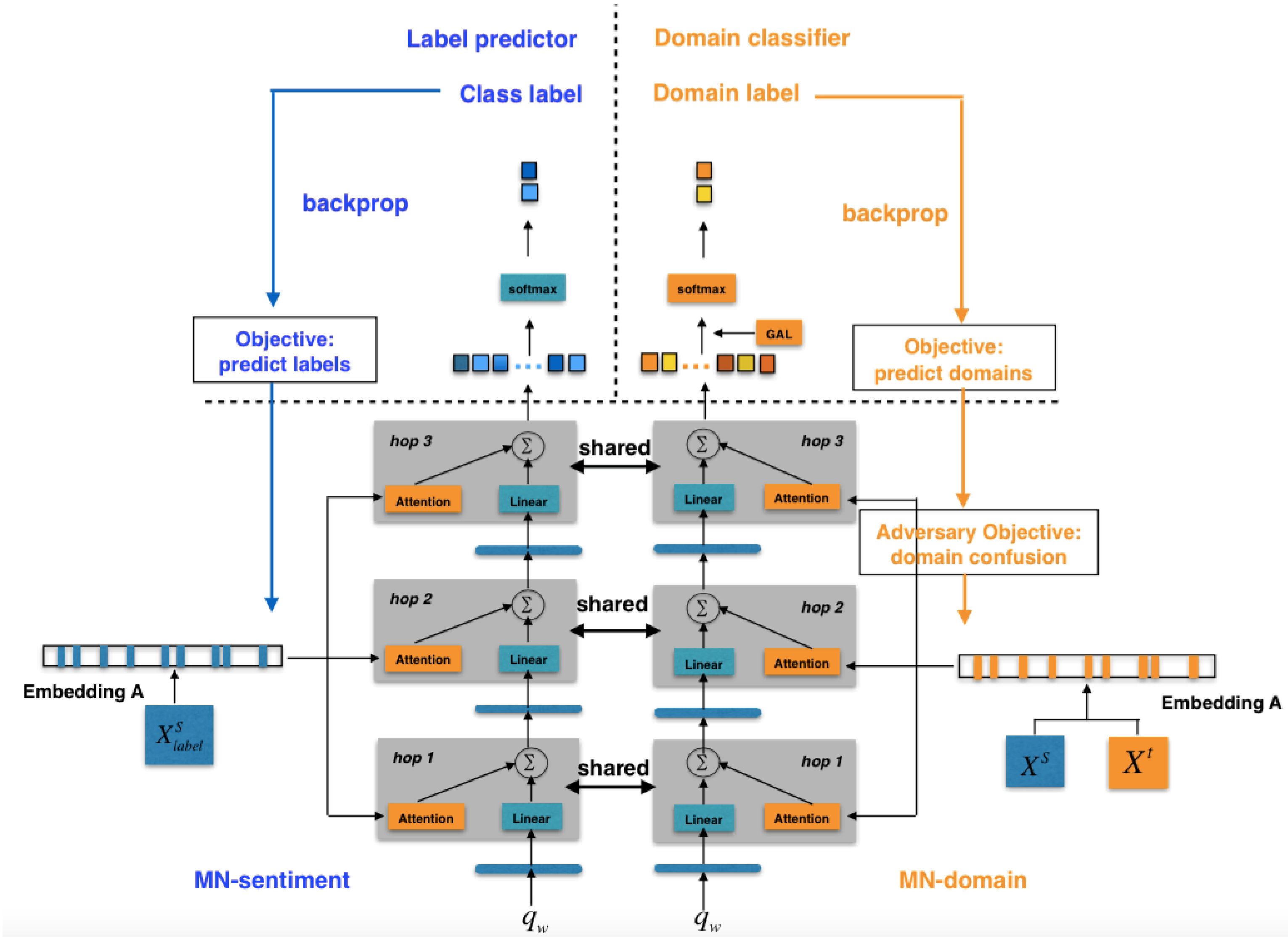
A new user

当系统不知道用户的口味时候，最优动作是**询问**。

An old customer

当系统了解用户口味时候，最优动作是根据用户口味提出**建议**，以简化流程节省时间。

End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification



End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification

Tasks	Positive sentiment words	Negative sentiment words
Electronics -kitchen	good great amazing excellent better best nice cool perfect happy fantastic outstanding cheaper easy beautiful convenient well fine wonderful worthwhile pleased affordable fast cheap flawless unbelievable reliable satisfied impressive pretty compatible nicely comfort powerful brilliant worth unbreakable fancy impressed compact handy elegant quick love durable	bad worst worse uncomfortable useless confused unreliable sad unacceptable poor impossible misleading unhappy waste upset disappointing thrilled disappointed disappointment negative terrible messy unsuitable worthless horrible poorly pricy defective dangerous fragile incorrectly stressful confusing expensive frustrating difficult unexpected painful ridiculous

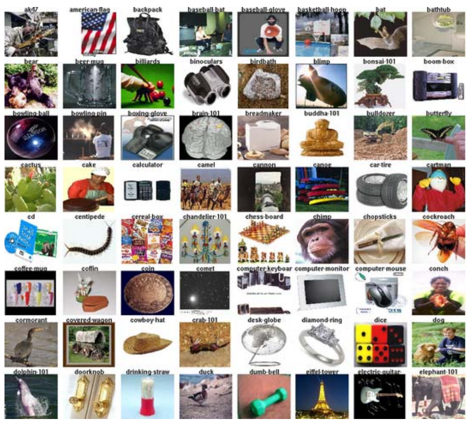
Figure 3: Samples of pivots captured by our method in the E→K task.

Distant Domain Transfer Learning

Source Domain: Face Recognition

Target Domain: Airplane Recognition

Distant Concepts



Source Domain: Object Recognition

Target Domain: Poverty estimation from Satellite images

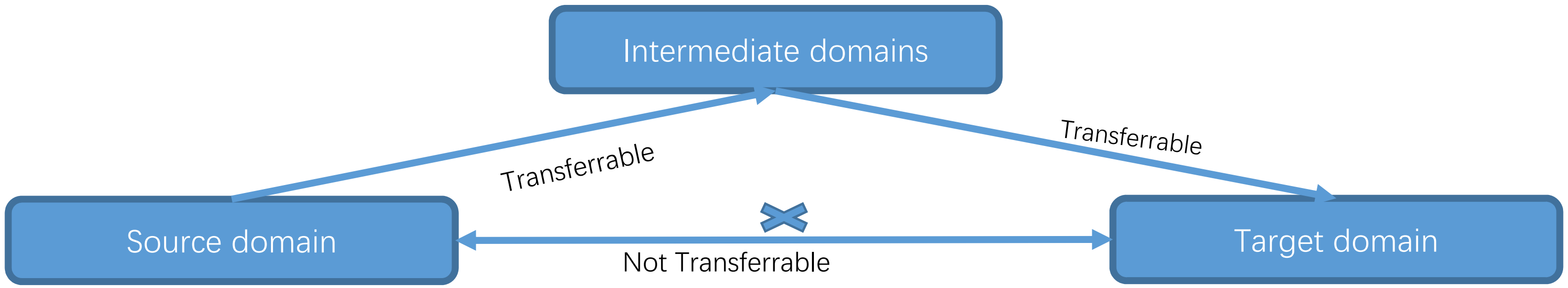
Distant Domain Transfer Learning

- Supervised learning: consider only one problem domain.
- Transfer learning: source and target domains should be similar.
- Distant domain transfer learning: source and target domains can be distant.

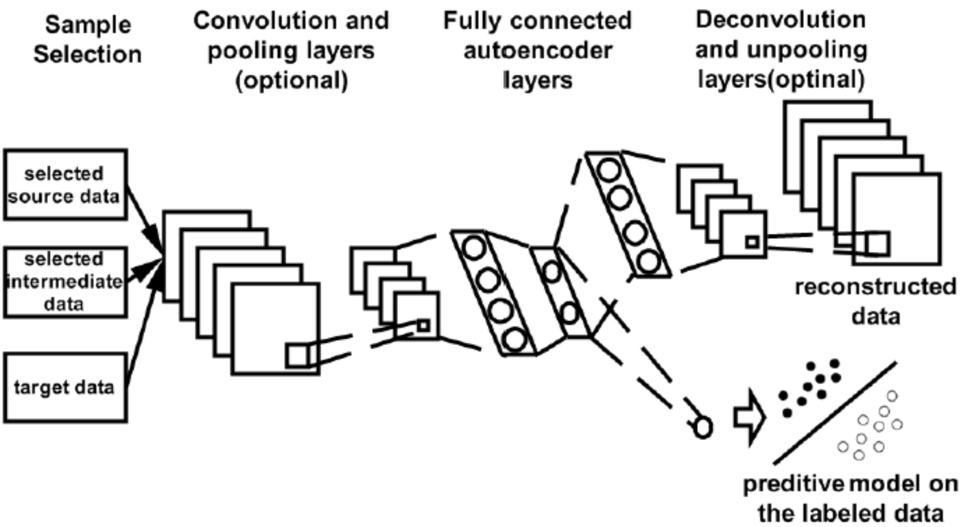
More previous knowledge could be used

Distant Domain Transfer Learning

- Some auxiliary intermediate domain data are selected to bridge the given source and target domains, and performs knowledge transfer along the bridge.



Distant Domain Transfer Learning



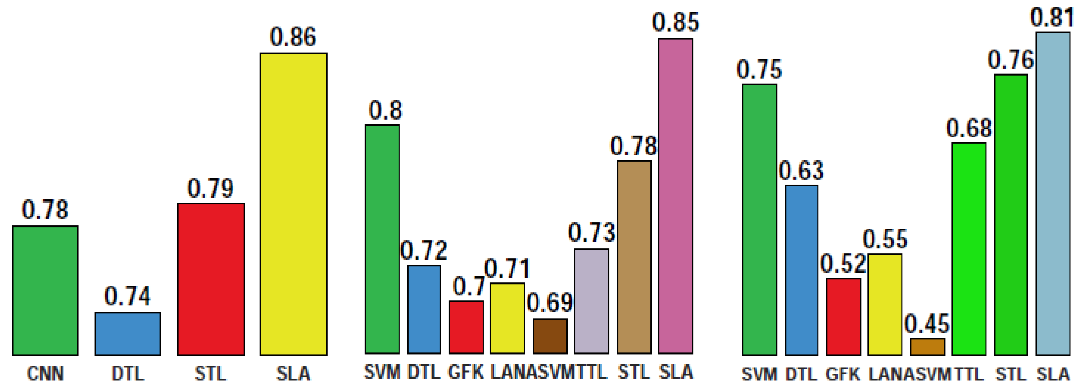
$$\mathcal{J}_1(f_e, f_d, \boldsymbol{v}_S, \boldsymbol{v}_I) = \frac{1}{n_S}\sum_{i=1}^{n_S} v_S^i \|\hat{\boldsymbol{x}}_S^i - \boldsymbol{x}_S^i\|_2^2 + \frac{1}{n_I}\sum_{i=1}^{n_I} v_I^i \|\hat{\boldsymbol{x}}_I^i - \boldsymbol{x}_I^i\|_2^2 + \frac{1}{n_T}\sum_{i=1}^{n_T} \|\hat{\boldsymbol{x}}_T^i - \boldsymbol{x}_T^i\|_2^2 + R(\boldsymbol{v}_S, \boldsymbol{v}_I), \tag{1}$$

$$R(\boldsymbol{v}_S, \boldsymbol{v}_I) = -\frac{\lambda_S}{n_S}\sum_{i=1}^{n_S} v_S^i - \frac{\lambda_I}{n_I}\sum_{i=1}^{n_I} v_I^i.$$

$$\mathcal{J}_2(f_c, f_e, f_d) = \frac{1}{n_S}\sum_{i=1}^{n_S} v_S^i \ell(y_S^i, f_c(\boldsymbol{h}_S^i)) + \frac{1}{n_T}\sum_{i=1}^{n_T} \ell(y_T^i, f_c(\boldsymbol{h}_T^i)) + \frac{1}{n_I}\sum_{i=1}^{n_I} v_I^i g(f_c(\boldsymbol{h}_I^i)), \tag{2}$$

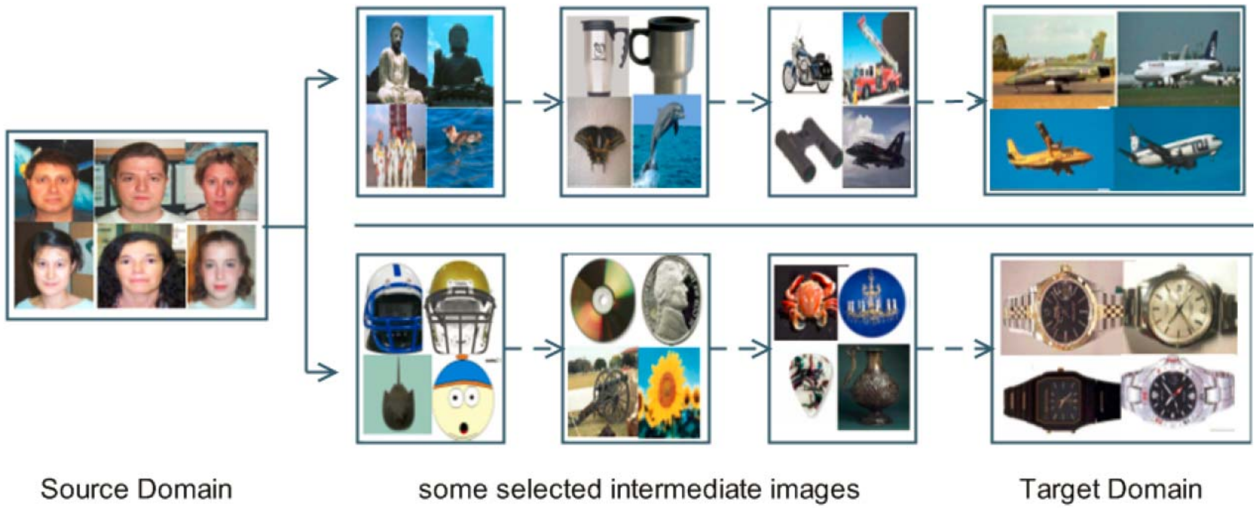
$$\min_{\boldsymbol{\Theta}, \boldsymbol{v}} \mathcal{J} = \mathcal{J}_1 + \mathcal{J}_2, \qquad \text{s.t. } v_S^i, v_I^i \in \{0, 1\}$$

Distant Domain Transfer Learning



(a) Caltech (Tensor) (b) Caltech (SIFT) (c) AwA (SIFT)

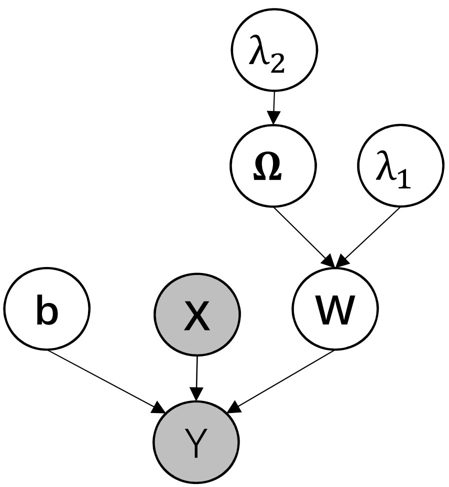
Figure 3: Average accuracies of different learning algorithms on the Caltech-256 and AwA datasets.



Sparse Task Relation Learning in Multi-Task Learning

- Existing works assume that task relations are dense
- When the number of tasks is large, usually a task cannot be helpful to all of other tasks
- Task relations exhibit sparse patterns
- A Framework to Learn Task Relations:

$$\min_{\mathbf{W},\mathbf{b},\mathbf{\Omega}\succeq\mathbf{0}} \sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i} l\left(\mathbf{w}_i^T\phi(\mathbf{x}_j^i)+b_i, y_j^i\right)+\frac{\lambda_1}{2}\mathrm{tr}(\mathbf{W}\mathbf{\Omega}^{-1}\mathbf{W}^T)+\lambda_2 g(\mathbf{\Omega})$$



Concrete Instances:
Regularization with Given Covariance
Trace/Schatten Norm Regularization
Squared Trace/Schatten Norm Regularization
Cluster Norm Regularization
Regularization with Sparse Inverse of $\mathbf{\Omega}$
......

Physical meaning of $\mathbf{\Omega}$
$\mathbf{\Omega}$ can be viewed as a covariance matrix to describe the pairwise task relations

Sparse Task Relation Learning in Multi-Task Learning

- $\mathbf{\Omega}$ corresponds to the task relations
- Expect to learn a sparse $\mathbf{\Omega}$
- l_1 norm leads to sparsity
- The objective function:

$$\min_{\mathbf{W},\mathbf{b},\mathbf{\Omega}\succeq\mathbf{0}} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} l\left(\mathbf{w}_i^T \phi(\mathbf{x}_j^i) + b_i, y_j^i\right) + \frac{\lambda_1}{2} \mathrm{tr}(\mathbf{W}\mathbf{\Omega}^{-1}\mathbf{W}^T) + \lambda_2 \|\mathbf{\Omega}\|_1$$

  - Convex problem

Theorem 1 *Suppose ω_{ij}, the (i,j)th element in $\mathbf{\Omega}$, is equal to 0. Then the optimal $\mathbf{w}_i$ is not spanned by $\phi(\mathbf{X}_j)$ and similarly $\mathbf{w}_j$ is not spanned by $\phi(\mathbf{X}_i)$, where $\phi(\mathbf{X}_j) = (\phi(\mathbf{x}_1^j), \ldots, \phi(\mathbf{x}_{n_j}^j))$ is the data matrix for the jth task.*

Theorem 2 *The optimal $\mathbf{\Omega}$ of problem (6) satisfies $\mathbf{\Omega} \succeq \frac{\mu_m(\mathbf{W})}{\sqrt{2m\tau}}\mathbf{I}$.*

Sparse Task Relation Learning in Multi-Task Learning

- Alternating optimization
- Fix **Ω** Optimize **W** and **b**

$$\min_{\mathbf{W},\mathbf{b}} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} \left(\mathbf{w}_i^T \phi(\mathbf{x}_j^i) + b_i - y_j^i\right)^2 + \frac{\lambda_1}{2} \mathrm{tr}(\mathbf{W}\boldsymbol{\Omega}^{-1}\mathbf{W}^T)$$

  - Analytical solution

- Fix **W** and **b** Optimize **Ω**

$$\min_{\boldsymbol{\Omega} \succeq 0} \frac{\lambda_1}{2} \mathrm{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R}) + \lambda_2 \|\boldsymbol{\Omega}\|_1$$

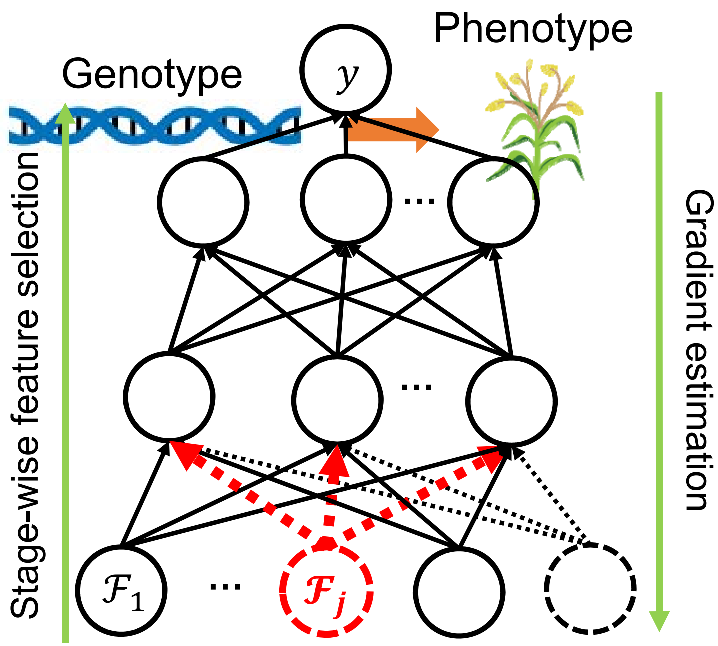
  - Convex problem

Method	Sentiment↑	Landmine↑	MHC-I↑	Parkinson↓	School↓
Number of Tasks	4	29	35	42	149
STL	0.8303±0.0098	0.6854±0.0261	0.6677±0.0234	1.1327±0.0866	1.1453±0.0599
MTL_{Ω}	0.8237±0.0209	0.7015±0.0244	0.6879±0.0227	1.0828±0.0460	1.1004±0.0631
MTL-TNR	**0.8764±0.0078**	0.7236±0.0249	0.7076±0.0203	1.0744±0.0415	1.0247±0.0879
MTL-STNR	**0.8808±0.0085**	**0.7496±0.0287**	0.6943±0.0230	1.0207±0.0162	**0.9048±0.0981**
MTL-CNR	–	0.7228±0.0204	0.7070±0.0084	1.0203±0.0119	1.1040±0.0474
MTL-gLasso	0.8267±0.0279	0.7286±0.0202	**0.7232±0.0297**	1.1182±0.0617	1.1559±0.0538
SPATS	0.8576±0.0103	**0.7420±0.0121**	**0.7299±0.0240**	**0.9894±0.0229**	**0.9110±0.0291**

School	Parkinson	Landmine	MHC-I
73.79%	77.62%	79.19%	81.14%

Deep Neural Networks for High Dimension, Low Sample Size Data

- **Motivating Application:**
  - Phenotype prediction problem using genetic data
  - $N \approx 5000, Dim > 150\ million$
- **Learning from high dimension, low sample size data is challenging.**
  - **High dimension** leads to overfitting.
  - **Low sample size** leads to high-variance gradients.
- **Method (Deep Neural Pursuit)**
  - **Overfitting:** stagewise **feature selection** in the context of neural network.
  - **High-variance gradients:** gradients estimation by dropout multiple times.
- **Results:**
  - DNP performs superior w.r.t. both performances of classification and true feature identification.



Classification AUC

	True Dim	2	5	10	25
LogR-l_1	AUC	0.868	0.826	0.755	**0.661**
	Std	0.003	0.001	0.014	**0.017**
GBFS	AUC	0.748	0.721	0.757	0.565
	Std	0.184	0.130	0.011	0.029
HSIC-Lasso	AUC	**0.948**	**0.881**	0.747	**0.642**
	Std	**0.003**	**0.001**	0.003	**0.007**
DNP	AUC	**0.926**	**0.887**	**0.813**	**0.650**
	Std	**0.005**	**0.023**	**0.025**	**0.016**

LogR-l_1	F1 score	0.141	0.313	0.364	0.266
	Std	0.025	0.045	0.046	0.035
GBFS	F1 score	0.182	0.050	0.429	0.105
	Std	0.183	0.056	0.048	0.052
HSIC-Lasso	F1 score	**1.000**	**0.889**	0.667	0.253
	Std	**0.000**	**0.000**	0.000	0.033
DNP	F1 score	**1.000**	**0.862**	**0.857**	**0.378**
	Std	**0.000**	**0.075**	**0.085**	**0.054**

Feature Identification F1-score