Auxiliary Learning with Joint Task and Data Scheduling

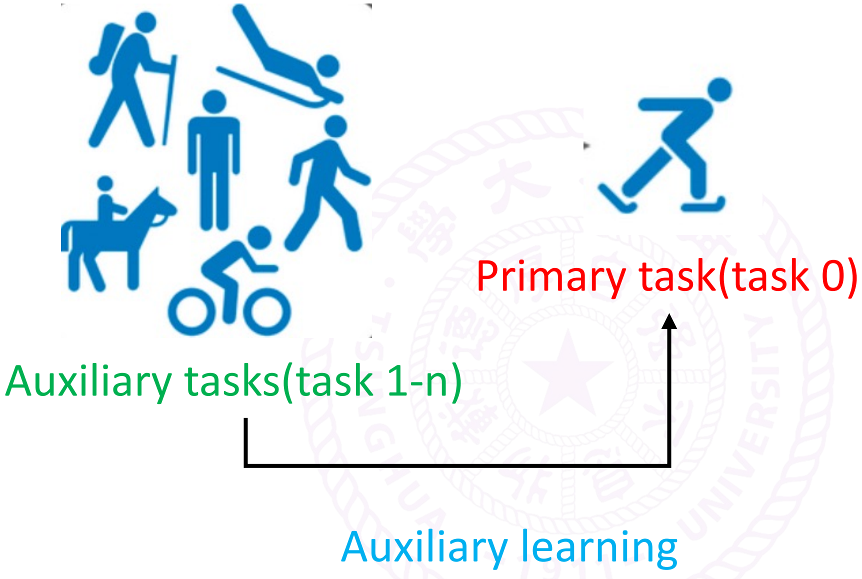
Hong Chen[1] Xin Wang[1,2] Chaoyu Guan[1] Yue Liu[1] Wenwu Zhu[1]
Department of Computer Science and Technology, Tsinghua University[1]
THU-Bosch JCML center, Tsinghua University[2]

Presentation Outline

- Background
- Method
- Results
- Conclusion

Background

- Auxiliary Learning
  - One primary task, several auxiliary tasks to help the primary task
  - Most widely adopted way：
    - Combine different auxiliary losses in a linear way
    - Tune the weights to avoid negative transfer
    - $\sum_{i=0}^{n} w_i L_i$
    - better performance on primary task



Background

- Auxiliary Learning
  - Not only auxiliary task, but also each data sample within each auxiliary task should be considered
  - Beneficial information of samples are different
  - Noisy samples should be excluded

Data sample	Target task Bird classification	Auxiliary task Beak detection
	useful	not useful
	useful	useful

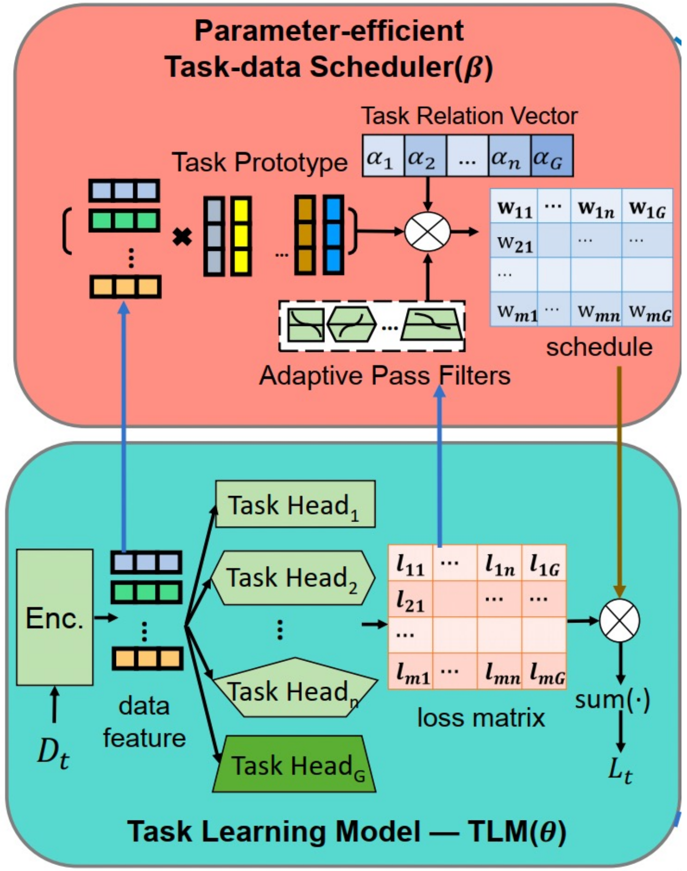
What is needed ?

⇒ Schedule for both task and data

$$\sum_{i=0}^{n}\sum_{j=1}^{m} w_{ij} l_{ij}$$

Method

➢Parameter-efficient Task-data Scheduler

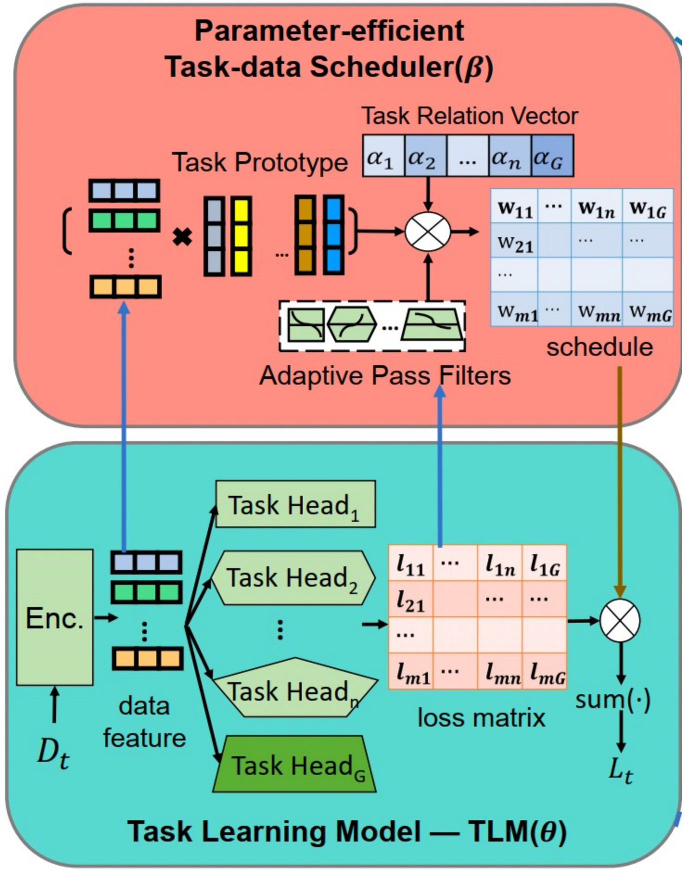


- **Hypothesis 1.** *Data sample x_i^t in auxiliary task T_k is beneficial to the target task T_G, if task T_k is beneficial to the target task T_G and the pair (x_i^t, y_{ik}^t) is beneficial to task T_k.*

- **Hypothesis 2.** (x_i^t, y_{ik}^t) *is beneficial to task T_k, if x_i^t contains useful features for T_k and y_{ik}^t is a correct label.*

Method

➢Parameter-efficient Task-data Scheduler



- $w_{ij} = \sigma(\alpha_j) * \sigma(P_j^T c_{ij}) * \sigma(a_j l_{ij} + b_j)$
- **(task importance + sample feature similarity with prototype + loss judgement)**
- $\boldsymbol{\beta} = \{\boldsymbol{\alpha}, \boldsymbol{P}, \boldsymbol{a}, \boldsymbol{b}\}$ $O(dn)$ learnable parameters, d (feature size)<< m

1. Relations between tasks

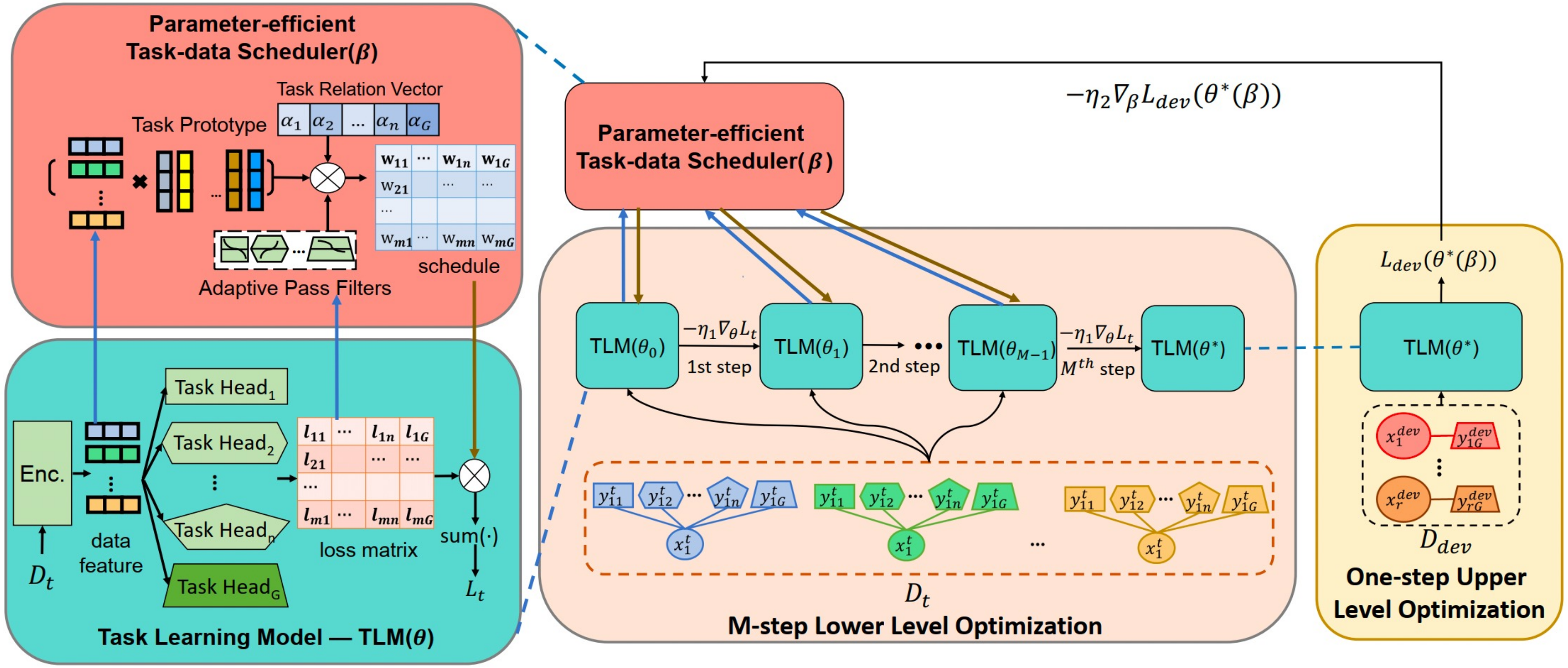
   ⇒ Task relation vector α_i
2. Importance of each data sample to each auxiliary task
   a. Whether data sample contains useful features

      ⇒ Task Prototype P_j
   b. Whether the data label is correct

      ⇒ Adaptive Pass Filters a_j, b_j

Method

➢Joint TML and Scheduler Optimization



Bi-level Optimization:

$$\beta^* = \arg\min_{\beta} L_{dev}(\theta^*(\beta)),$$

$$s.t.\ \theta^* = \arg\min_{\theta} L_t(\theta, \beta).$$

Lower level: Optimize TML

Upper level: Optimize Scheduler

Method

- Joint TML and Scheduler Optimization
  - Lower Optimization

$$\nabla_\theta L_t(\theta,\beta) = \sum_{k\in U}\sum_{i=1}^{m} w_{ik}\nabla_\theta l_k(f_k(x_i^t), y_{ik}^t;\theta).$$

weighted gradient

  - Upper Optimization

$$\begin{aligned}
&\nabla_\beta L_{dev}(\theta^*(\beta)) = \nabla_\theta L_{dev}\cdot\nabla_\beta\theta^* \\
&= -\nabla_\theta L_{dev}\cdot(\nabla_\theta^2 L_t)^{-1}\cdot\nabla_\beta\nabla_\theta L_t|_{(\beta,\theta^*(\beta))}. \\
&\approx\ -\nabla_\theta L_{dev}\cdot\sum_{i=0}^{K}(I-\nabla_\theta^2 L_t)^i\cdot\nabla_\beta\nabla_\theta L_t.
\end{aligned}$$

implicit theorem

Neumann Series

Results

1. CUB: bird classification as primary task, bird attribute classification as auxiliary tasks(1+312 tasks)
2. Pet, CF-10, CF-100: image classification as primary task, rotation prediction as auxiliary task(1+1 tasks)
3. ML-1M: rating prediction as primary task, ctr prediction as auxiliary task(1+1 tasks)

➢Full Supervision

Table 2. Performance of different methods under the fully-supervised setting(CF-10, CF-100, ML-1M respectively represents the CIFAR10, CIFAR100, MovieLens-1M dataset).

Metric	Accuracy(%)				RMSE
Method	**CUB**	**Pet**	**CF-10**	**CF-100**	**ML-1M**
STL	73.86	61.45	71.60	74.14	0.9112
NAL	73.42	66.09	70.42	73.38	0.9101
Uncertainty	72.54	67.14	70.93	68.10	0.9103
GCS	73.70	66.30	70.64	74.14	0.9098
AuxL	74.32	66.30	71.23	73.80	0.9097
N-JTDS	71.38	67.28	70.57	75.06	0.9181
JTDS(ours)	**77.04**	**70.01**	**72.59**	**75.68**	**0.9087**

➢Semi-Supervision

Table 3. Image classification accuracy(%) of different methods under the semi-supervised setting.

Method	**CUB**	**Pet**	**CIFAR100**
STL	38.35	30.21	55.16
NAL	48.15	38.00	57.52
Uncertainty	46.66	43.57	57.70
GCS	46.50	36.80	55.66
AuxL	50.19	36.42	55.94
N-JTDS	46.50	45.53	57.54
JTDS	**51.21**	**53.49**	**58.56**

1. Auxiliary tasks are more beneficial when target task lacks in labels
2. Joint task-data scheduling is effective

Results

➢Robustness to label Noise

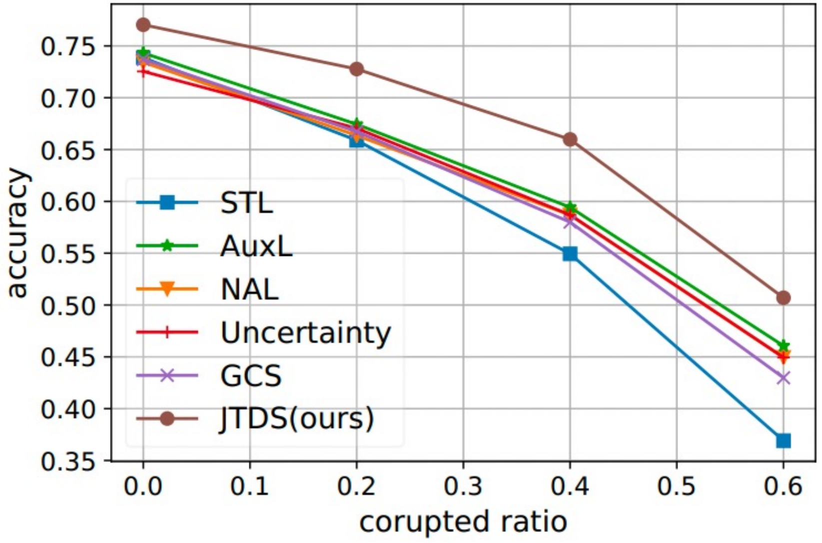
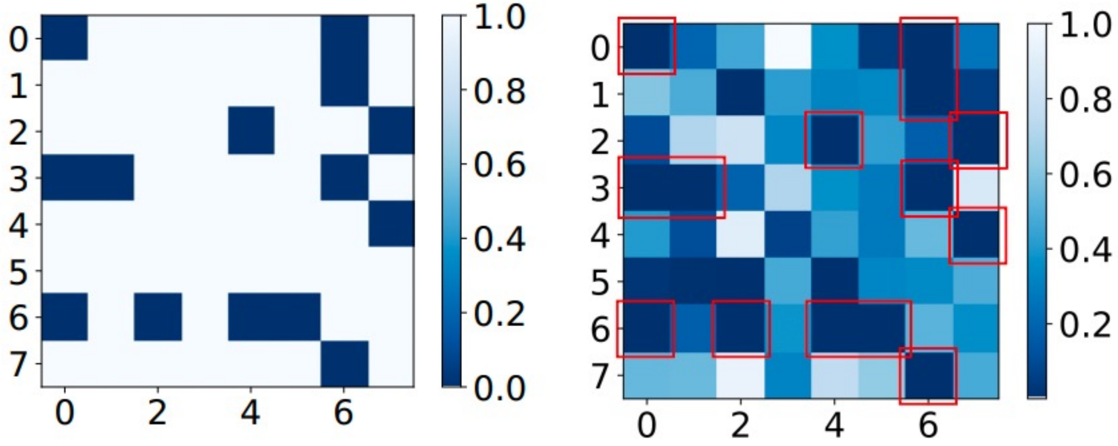


Figure 2. Accuracy of different models under different ratios of corrupted labels



(a) label correctness (b) scheduled weights

Figure 3. Corrupted Sample Detection

Results

➢Learned schedules

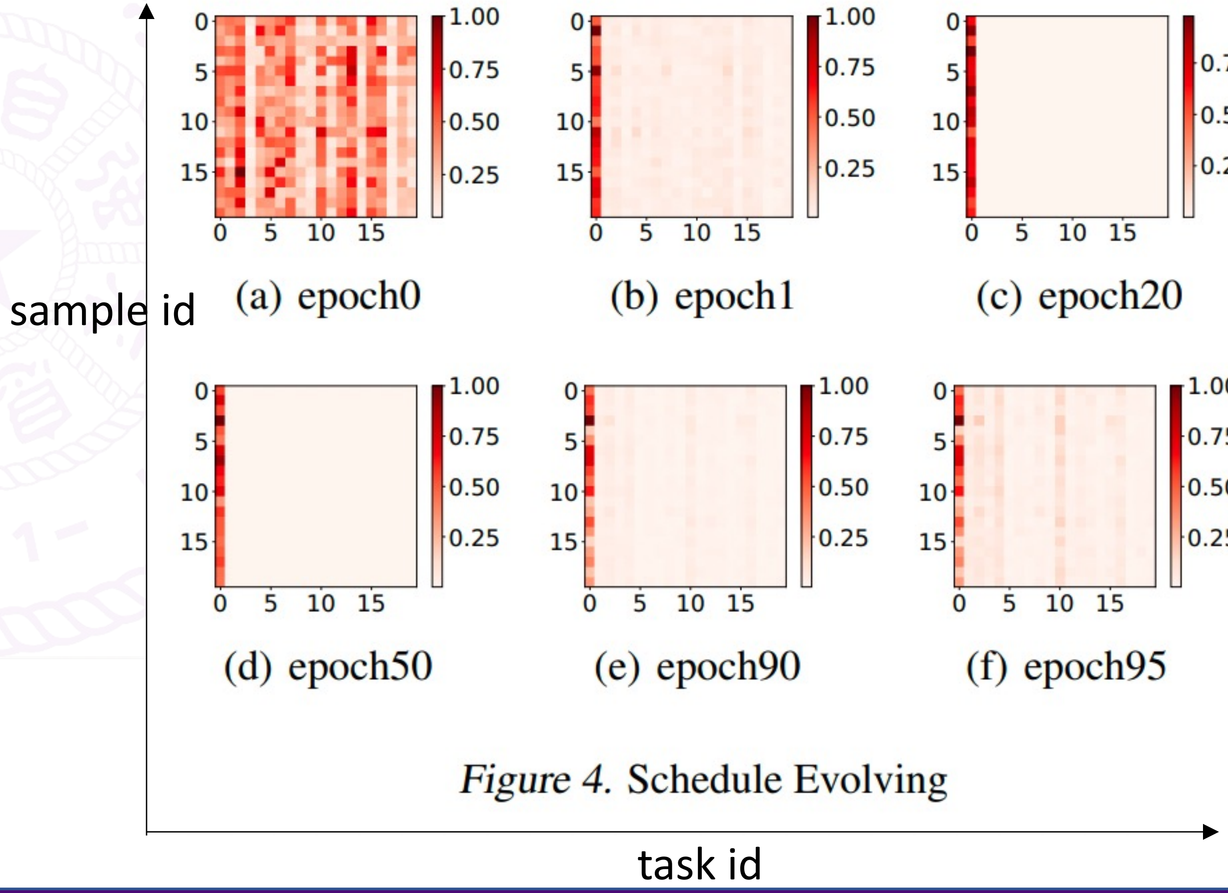


Figure 4. Schedule Evolving

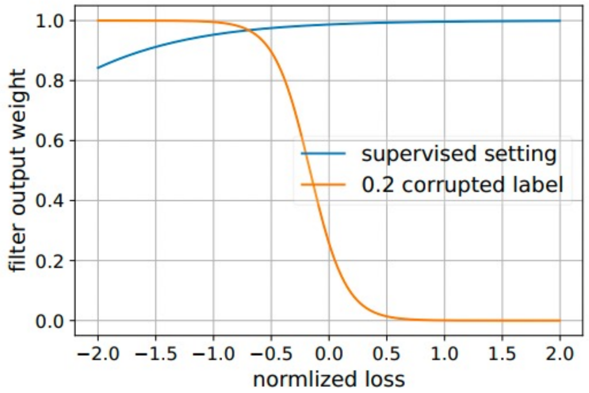


Figure 6. Target task filter function for supervised and corrupted settings.

Conclusion

- Propose task and data scheduling for auxiliary learning
- Propose an parameter-efficient task-data scheduler
- Give a complete solution accommodating various scenarios with efficiently approximating bi-level optimization

Thanks for listening!

My email: h-chen20@mails.tsinghua.edu.cn