Learning to Rank for Active Learning via Multi-Task Bilevel Optimization

Zixin Ding[1], Si Chen[2], Ruoxi Jia[2], Yuxin Chen[1]

[1]University of Chicago, [2]Virginia Tech

Problem Setting

$$\mathcal{S}_1^* \in \underset{\mathcal{S}_0 \subseteq \mathcal{S}_1 \subseteq \mathcal{X}, |\mathcal{S}_1 \setminus \mathcal{S}_0| = B}{\arg\max} u(\mathcal{S}_1)$$

- A *labeling function* f to project input $\mathcal{X}$ to a groundtruth label set $\mathcal{Y}$
- Given labeled Set $\mathcal{S}_0$ with $\mathcal{S}_0 \subseteq \mathcal{X}$ and $|\mathcal{S}_0| = k$, find the optimal set $\mathcal{S}_1$ with $k + B$ size that would achieve maximal validation set accuracy in *one round*.
- A *groundtruth utility function mapping*

$$u : 2^{\mathcal{X}} \to \mathbb{R}$$

where $u(\xi)$ quantifies the utility of a subset $\xi \subseteq \mathcal{X}$
- Learn the utility function $\hat{u}$ by ranking and optimal transport distance between labeled set and validation set.
- First Stage: Learning the utility function.
- Second Stage: Greedily follow the learned utility function with predicted maximal utility.

Collect utility samples to train **ranking function** with **Optimal Transport distance** as a regularizer with **bilevel training.**

How to learn a deep surrogate model for one-round data acquisition in active learning?

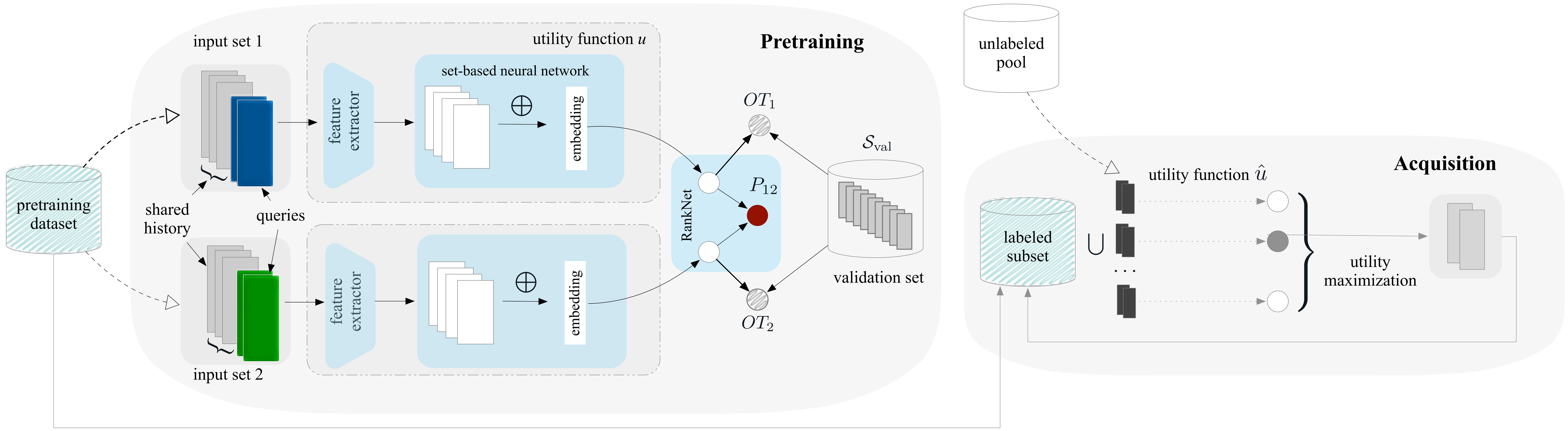

Figure 1: Overview of the **RAMBO** Algorithm (**R**anking-based **A**ctive learning via **M**ultitask **B**ilevel **O**ptimization)

Utility Model Learning

Definition 1. *[Surrogate Utility Model] Let $\mathcal{X}$ be the instance domain, and ξ be any sampled subset drawn from distribution $\mathcal{D}$ over $\mathcal{X}$. A surrogate utility model $\hat{u}(\xi)$ is a set function: $2^{\mathcal{X}} \to \mathbb{R}$, optimized to predict the true utility $u(\xi)$ on a training set $\xi \sim \mathcal{D}$:*

$$\hat{u} = \underset{\tilde{u}_w}{\arg\min}\, \hat{\mathbb{E}}_{\xi\sim\mathcal{D}}[\mathcal{L}(\tilde{u}_w(\xi), u(\xi)]$$

where $\mathcal{L}(\cdot,\cdot)$ denotes the loss function, and $\tilde{u}_w$ is a parametric set function to approximate u.

Multi-task Learning

Loss function over a pair of utility samples ξ_1, ξ_2:

$$\mathcal{L}_{Total} = L_{Rank} + \lambda_{OT} L_{OT}$$

Ranking loss:

$$\mathcal{L}_{\text{Rank}}(\xi_1, \xi_2) = -\bar{P}_{12}\log P_{12} - (1-\bar{P}_{12})\log(1-P_{12}).$$

Optimal Transport (OT) distance regulatory loss:

$$\mathcal{L}_{\text{OT}}(\xi_1, \xi_2) = \lambda_1(\hat{OT}_1 - OT_1)^2 + \lambda_2(\hat{OT}_2 - OT_2)^2 - \lambda_3(\min(\hat{OT}_1, 0) + \min(\hat{OT}_2, 0))$$

Bilevel Training

Longer length Utility Samples as training samples for outer optimization tasks

$$\min_{\lambda}\; E(w(\lambda), \lambda) \quad \text{s.t.} \quad w(\lambda) = \underset{\hat{w}\in\mathbb{R}^d}{\arg\min}\, \mathcal{L}(\hat{w})$$

Shorter length Utility Samples as training samples for inner optimization tasks

$$\mathcal{L}(\hat{w}) = \sum_{\{(S_1^i, u(S_1^i)), (S_2^i, u(S_2^i))\}\in D_{tr}} \mathcal{L}_{\text{Total}}(\hat{w}) + \Omega_\lambda(\hat{w})$$

Ablation Study

Bilevel	Optimal Transport	RankNet	Accuracy
✓	✓	✓	**83.1 ± 0.1**
✓	✓	×	81.9 ± 0.2
✓	×	✓	81.2 ± 0.4
✓	×	×	81.8 ± 0.2
×	✓	✓	81.0 ± 0.3
×	✓	×	81.7 ± 0.2
×	×	✓	80.9 ± 0.3
×	×	×	81.6 ± 0.1
-	-	-	81.2 ± 0.2

Bilevel	Optimal Transport	RankNet	Accuracy
✓	✓	✓	**77.3 ± 0.2**
✓	✓	×	76.1 ± 0.3
✓	×	✓	76.2 ± 0.4
✓	×	×	70.5 ± 0.3
×	✓	✓	75.5 ± 0.3
×	✓	×	75.5 ± 0.3
×	×	✓	76.0 ± 0.8
×	×	×	74.6 ± 0.7
-	-	-	74.7 ± 0.3

Ablation study on three submodules with acquisition budget $B = 5000$, and pretraining set $k = 200$ for FashionMNIST (left) and $k = 3500$ for CIFAR10 (right). The last row corresponds to the random baseline.

Experimental Results

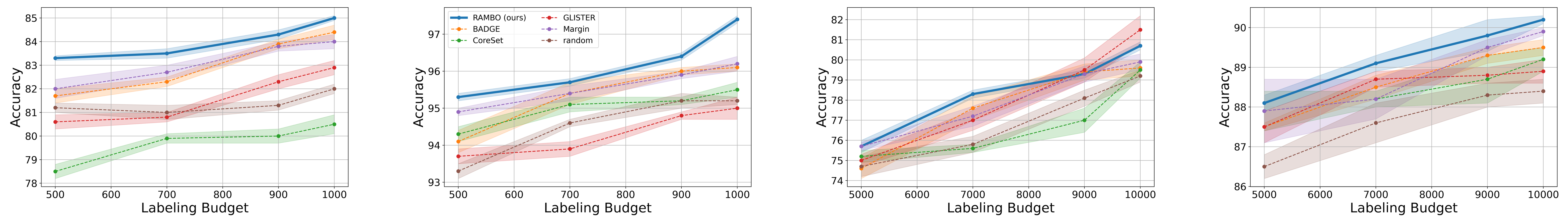

Figure 2: Accuracy vs. Labeling budget. Results are given in % for (from left to right) FashionMNIST, MNIST, CIFAR10, and SVHN.

Sensitivity Analysis

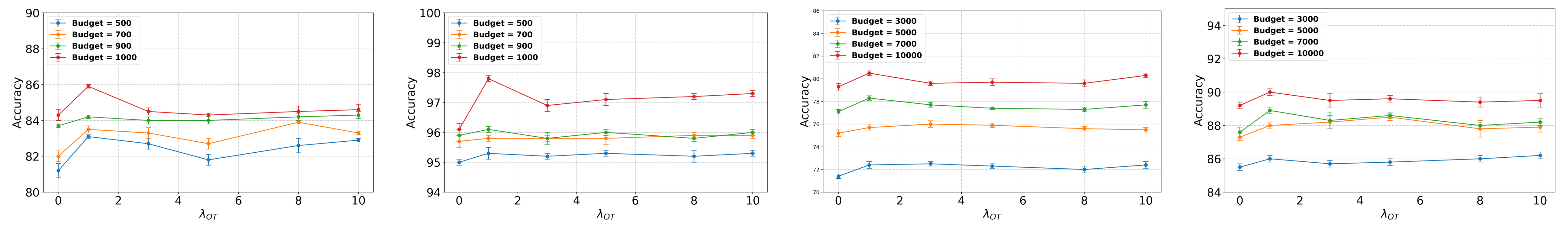

Figure 3: Different choices of λ_{OT} for pretraining set size $k = 200$ for FashionMNIST, MNIST, CIFAR10 and SVHN by different acquisition budget.