Relational Surrogate Loss Learning

Tao Huang[1,2] Zekang Li[3] Hua Lu[4] Yong Shan[3] Shusheng Yang[4]
Yang Feng[3] Fei Wang[5] Shan You[2] Chang Xu[1]
[1]School of Computer Science, Faculty of Engineering, The University of Sydney [2]SenseTime Research
[3]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
[4]Huazhong University of Science and Technology [5]University of Science and Technology of China
youshan@sensetime.com

Motivation

Loss Functions:

1. Designed as proxies of evaluation metrics.
2. Requires expertise on specific tasks.
3. Often hard to align well to the metric.

Surrogate Loss Learning:

1. Approximate the evaluation metrics using a deep neural network (DNN).
2. Replace the conventional loss function with the learned DNN.

Limitations:

1. It is difficult for the surrogate loss to fully recover the metric values.
2. Since the surrogate loss is easy to overfit on the training samples, it needs to be trained with model alternatively.
3. The learned surrogate loss cannot generalize to different models and tasks.

Intuition

Evaluation metrics (losses) are used to distinguish whether a model is better or worse than another.

Solution: we only need to keep the **relative rankings** of samples between the loss and metric.

Learning Surrogate Loss with Correlation-based Objective

Spearman's rank correlation:

$$\rho_S(\boldsymbol{a},\boldsymbol{b}) = \frac{\mathrm{Cov}(\mathbf{r}_{\boldsymbol{a}},\mathbf{r}_{\boldsymbol{b}})}{\mathrm{Std}(\mathbf{r}_{\boldsymbol{a}})\mathrm{Std}(\mathbf{r}_{\boldsymbol{b}})} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{r}_{\boldsymbol{a}i} - E(\mathbf{r}_{\boldsymbol{a}}))(\mathbf{r}_{\boldsymbol{b}i} - E(\mathbf{r}_{\boldsymbol{b}}))}{\mathrm{Std}(\mathbf{r}_{\boldsymbol{a}})\mathrm{Std}(\mathbf{r}_{\boldsymbol{b}})} \tag{1}$$

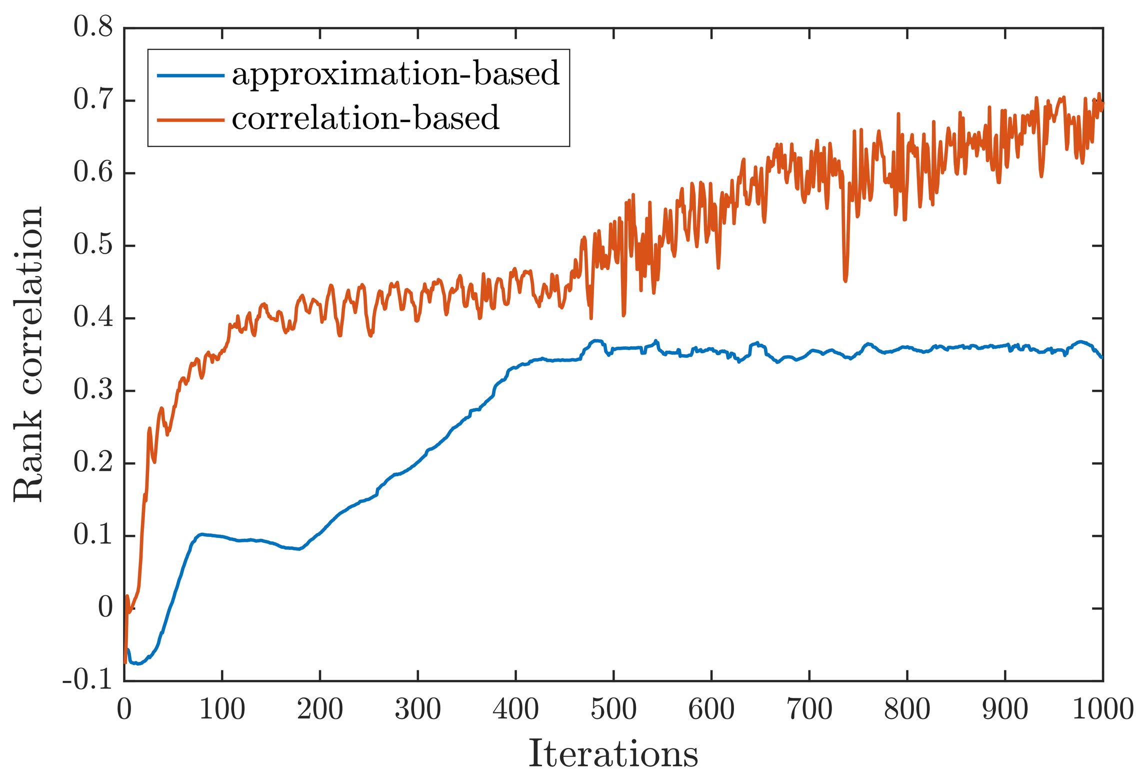
$\mathbf{r}_{\boldsymbol{a}}$: rank vector of $\boldsymbol{a}$; $\mathrm{Cov}(\mathbf{r}_{\boldsymbol{a}},\mathbf{r}_{\boldsymbol{b}})$: covariance of the rank vectors;
$\mathrm{Std}(\mathbf{r}_{\boldsymbol{a}})$: standard derivation of $\mathbf{r}_{\boldsymbol{a}}$.

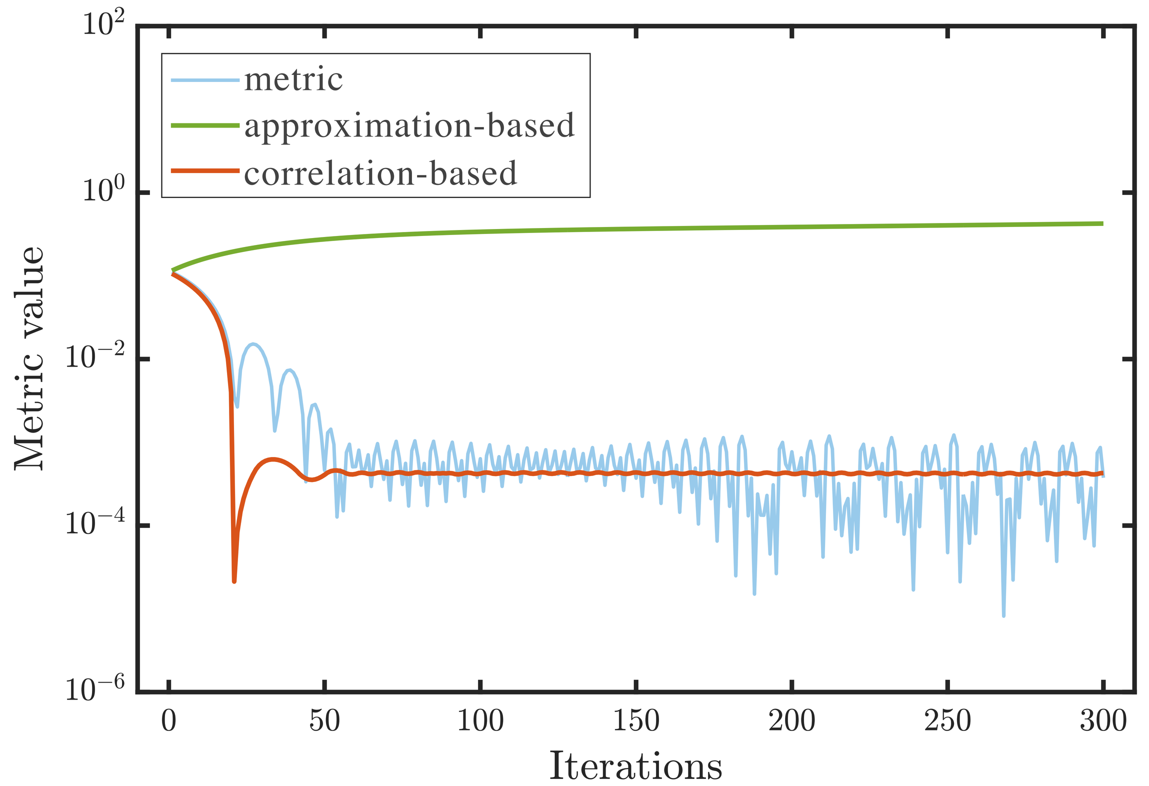
Correlation-based objective:

$$\mathcal{O}_S(\mathcal{L}(\boldsymbol{y},\hat{\boldsymbol{y}};\boldsymbol{\theta}_l),\mathcal{M}(\boldsymbol{y},\hat{\boldsymbol{y}})) = \rho_S(\mathcal{L}(\boldsymbol{y},\hat{\boldsymbol{y}};\boldsymbol{\theta}_l),\mathcal{M}(\boldsymbol{y},\hat{\boldsymbol{y}})) \tag{2}$$

$\boldsymbol{y}$: model predictions; $\hat{\boldsymbol{y}}$: ground-truth labels;
$\mathcal{L}$: surrogate loss with learnable weights $\boldsymbol{\theta}_l$; $\mathcal{M}$: evaluation metric.

Compare with approximation-based objective:



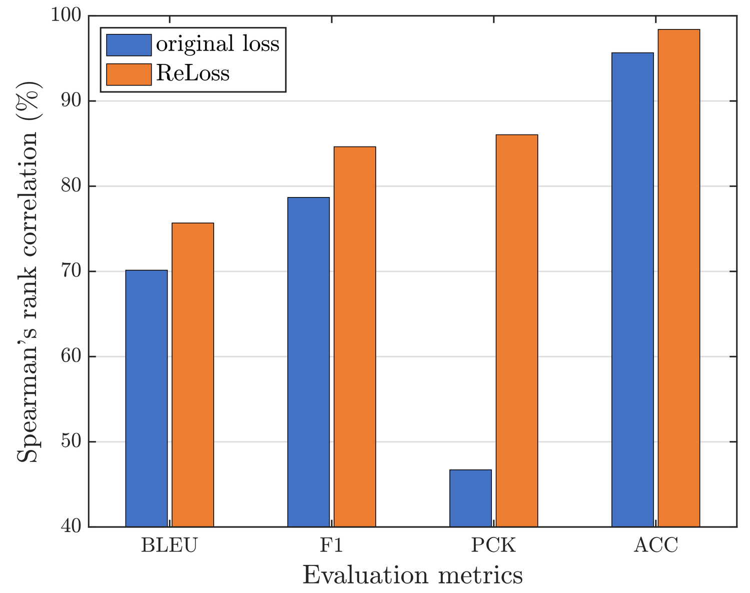


Surrogate loss learned with correlation-based objective

1. has higher rank correlations to the evaluation metric;
2. achieves better performance compared to approximation-based objective;
3. smoother convergent curve compared to original loss (metric).

Experiments

ReLoss vs. conventional losses:



ReLoss achieves better correlations compared to the original loss functions.

Image classification:

Dataset	Model	CE	ReLoss
CIFAR-10	ResNet-56	94.32	**94.57**
CIFAR-100	ResNet-56	73.61	**74.15**
ImageNet	ResNet-50	76.5	**76.8**
ImageNet	MobileNet V2	71.8	**72.2**

Human pose estimation:

Method	Backbone	MSE	ReLoss
SimpleBaseline	ResNet-50	70.4	**71.9**
HRNet	ResNet-50	74.4	**74.8**

Neural machine translation:

Model	Dataset	ori. loss	ReLoss
NAT-Base	EN-RO	29.24	**30.07**
NAT-Base	RO-EN	28.97	**29.68**

Machine reading comprehension:

Method	ROUGE-L	BLEU-4	F1
MacBERT-base	51.4	50.3	53.9
NAT-Base	**51.8**	**50.6**	**54.2**