YES IJCAI Young Elite Symposium

面向多潜在域的稳健模型复用

邵杰晶

南京大学

Background

Pre-trained models have shown remarkable growth in both **capability** and **quantity**.

Foundation Model

Segment Anything
Research by Meta AI
Meta

CHAT GPT

Model Market

Hugging Face

Reusing pre-trained models is becoming a major paradigm.

What this talk is about

Model Reuse is becoming one of the main development modes of machine learning.

It suffers one serious issue

- In Model Reuse, the usage of pre-trained models may even deteriorate performance, which means, it can be outperformed by its learning-from-scratch counterpart in quite many cases.

Contribution of this work

In this work, we consider the for robust model reuse for multiple latent domains, where robustness means it will not be worse than its learning-from-scratch counterpart. We formulate it as a bi-level optimization with an attention reuse mechanism. Both theoretical and experimental results show quite promising results.

[Towards Robust Model Reuse in the Presence of Latent Domains. IJCAI'21]

Reusing Pre-Trained Models

Parameter-accessible

Fine-Tuning

FitNets [Adriana et al. 2015]

LoRA [Hu et al. 2021]

Parameter-inaccessible

Distillation [Hinton et al. 2015]

Ensemble [Zhou. 2021]

NeC4.5 [Zhou and Jiang. 2004]

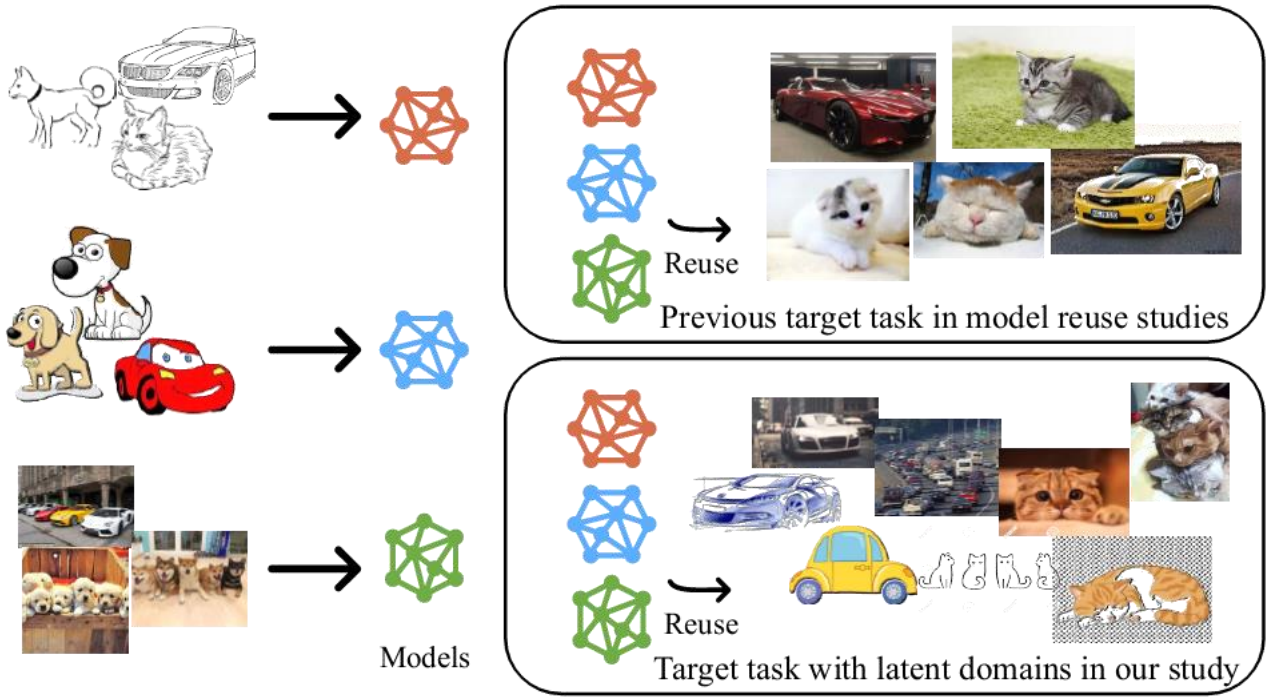
There are many approaches to implement model reuse.

However, these studies typically work on a homogenous target distribution.

The model reuse for heterogenous target has not been studied in literatures.

Towards Target with Multiple Domains

It is desirable to consider the multiple domains.



Previous model reuse studies typically assume target data from **a single distribution**.

However, in practice, the target samples often arise from **multiple latent distributions**.

There are latent domains: photos, line drawings, and cartoons returned by web image search for “car” and “cat”. What is more difficult is one usually only gets the category label (car or cat), rather than the domain labels (photo or cartoon).

Towards Target with Multiple Domains

It is desirable to consider the robust model reuse.

When the target distribution contains multiple latent domains, it is often the case that none of the models in the pre-training model pool can handle all domains, making it challenging to identify appropriate pre-training models for reuse. It results in inferior model reuse compared to learning from scratch.

Performance degeneration!

Motivation 1: Model reuse should explore the diversity of data to exploit the pre-trained models.

Motivation 2: Model reuse should ensure that the learned model is never worse than learning from scratch

The Basic Setup

- Suppose we have a collection of pre-trained model $\{h_1,\ h_2 \ldots, h_K\}$.
- Target distribution $\mathcal{D}_T$ contains multiple latent domains: $\mathcal{D}_T = \sum_\lambda \lambda_i \mathcal{D}_i$.
- We consider a semi-supervised setting where we have a limited labeled set D_l and a large unlabeled set D_u from the target distribution $\mathcal{D}_T$.

The goal: To learn a target model h with the help of pre-trained models $\{h_1,\ h_2 \ldots, h_K\}$, which often outperform, and will not be worse than learning from scratch with D_l.

A Direct Approach

- Suppose we know the reusability or weights w for these pre-trained models, then one can have a very direct approach.

$$\min_{h} \sum_{(x_i,y_i)\in D_l} L(h(x_i), y_i) + \sum_{x_j\in D_u} L(h(x_j), \sum_{k=1}^{K} w_k h_k(x_j))$$

Empirical risk on labeled data

Structural regularization on unlabeled data

Weights for pre-trained models

Learn the target model h via the combined structural risk minimization.

Brief Overview to Weighted Teaching

Weighted teaching is a general scheme for knowledge transfer, has widely applied in…

- **Model Reuse**

  Rapid Performance Gain through Active Model Reuse. IJCAI'19

  Handling Concept Drift via Model Reuse. ML'2020

- **Multi-Task Learning**

  Adaptive smoothed online multi-task learning. NIPS'16.

  Multi-task learning with labeled and unlabeled tasks. ICML'17

- **Domain Adaptation**

  Domain adaptation from multiple sources: A domain-dependent regularization approach. TNNLS'12

Positive Results

- The error of weighted teacher (Reusing Error) could be bounded as:

$$\min_{w} E_{(x,y)\sim\mathcal{D}_T} L\left(\sum_{k=1}^{K} w_k\, h_k(x), y\right) \leq \min_{k}[d_\alpha(\mathcal{D}_k||\mathcal{D}_T)\epsilon_k]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$$

Distribution divergence between source distribution and the target distribution.

Source error

If there exists a source distribution that is close to the target distribution, then the teaching error can be bounded within a small value.

However

- When the target distribution is far from the single source distribution, the teaching error is too risk to provide the guarantee. Specifically, we assume the target distribution $\mathcal{D}_T$ is a linear combination of the source distribution $\mathcal{D}_T = \sum_\lambda \lambda_i \mathcal{D}_i$.
- The reusing error has a upper bound:

$$\min_{w} \max_{\mathcal{D}_T} E_{(x,y)\sim\mathcal{D}_T} L\left(\sum_{k=1}^{K} w_k\, h_k(x), y\right) \leq \sum_{k=1}^{K} w_i [d_\alpha(\mathcal{D}_k||\mathcal{D}_T)\epsilon_k]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$$

Worst case consideration [Li and Zhou, ICML2011; Balsubramani and Freund, COLT2015]

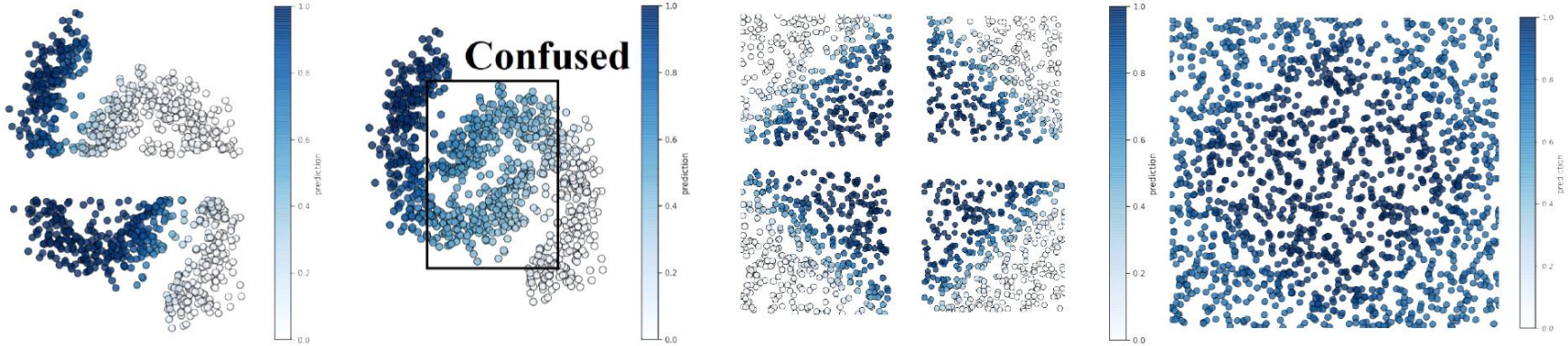
Distribution divergence between source distribution and the target distribution.

Source error

There is no guarantee for weighted reuse in the worst case consideration.

However

Empirical observation:

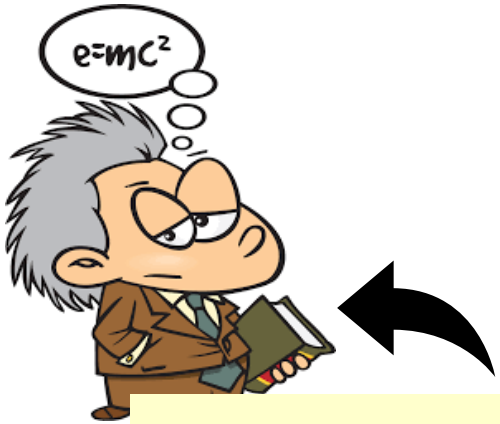


(a) 2 sources (b) $\frac{1}{2}\sum_{k=1}^{2} h_k(\boldsymbol{x})$ (c) 4 sources (d) $\frac{1}{4}\sum_{k=1}^{4} h_k(\boldsymbol{x})$

This finding motivates us to construct a more flexible reuse.

There is no guarantee for weighted reuse in the worst case consideration.

Human Intelligence



A flexible reuse!

For each instance, we try to find the corresponding models that are more likely to give the correct predictions.

Theoretical questions

Practical questions

Basic Assumption

- **Smooth Assumption**

  Similar instances should have similar concept compositions within their latent domains

It was widely employed in the semi-supervised literatures. [Chapelle 2006; Zhu 2009; Zhou 2013]

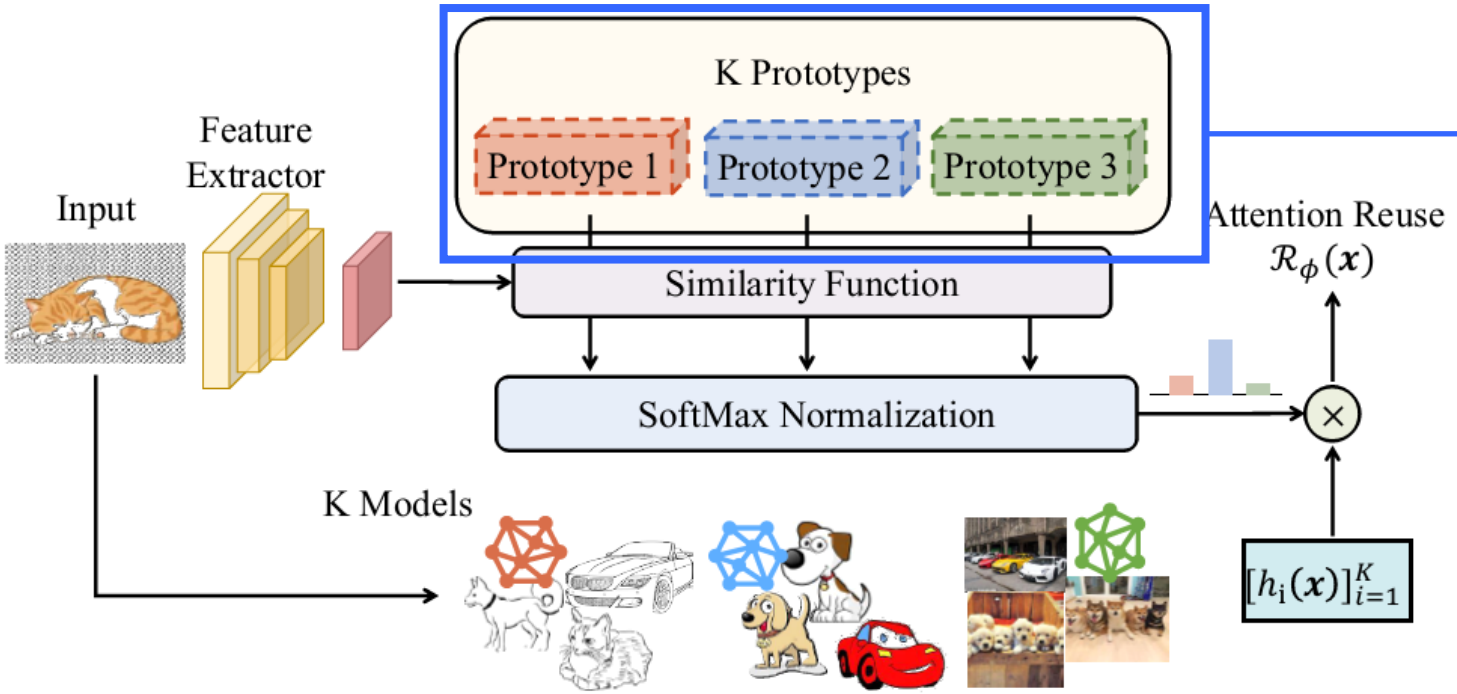
- **Our Assumption**

  If model h could correctly predict x, it is likely to correctly predict instances similar to x.

We construct a description of the marginal distribution for each model, which is named as *Prototype*. The aggregation predictions are based on attention via the *Prototypes*.

Attention Reuse

We construct an attention reuse based on an instance-aware weight mechanism $g:\mathcal{X}\rightarrow\mathcal{W}$

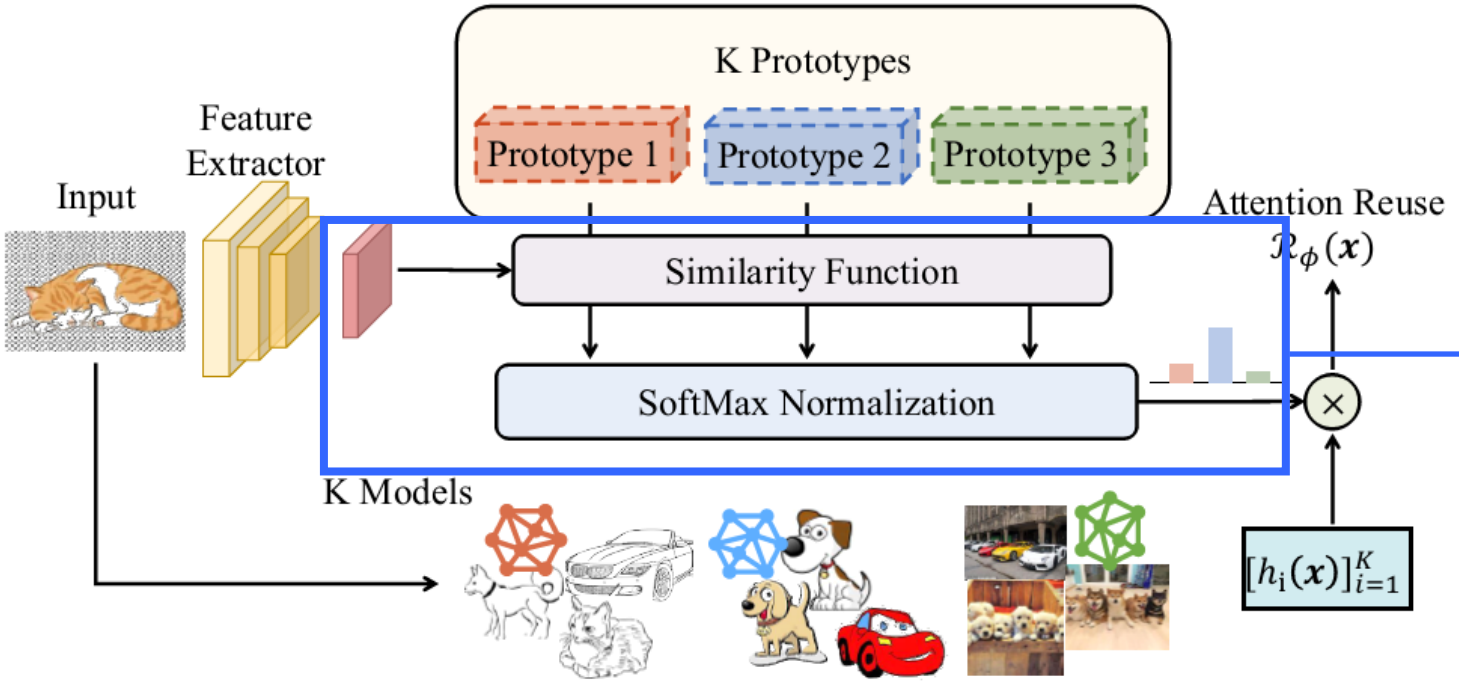


Parameterized prototypes represents the marginal distribution of each source

$$\mathcal{R}_\phi(\mathbf{x}) = \text{Attention}(\mathbf{x}, P, H) = \text{SoftMax}(\psi(\mathbf{x})^T P)H$$

Attention Reuse

We construct an attention reuse based on an instance-aware weight mechanism $g:\mathcal{X}\rightarrow\mathcal{W}$

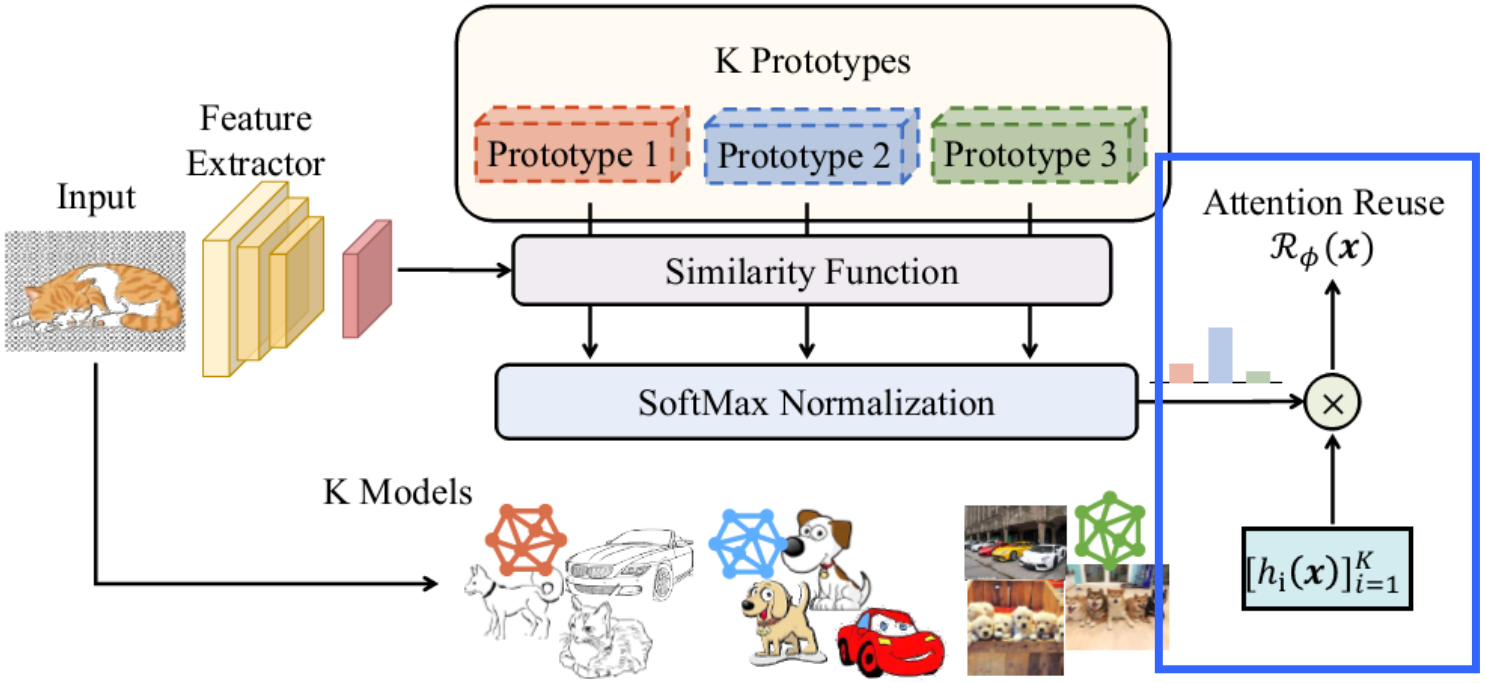


The **similarity matching function** assigns weights to the pre-trained modes based on the embedding of the current instance

$$\mathcal{R}_\phi(\mathbf{x}) = \text{Attention}(\mathbf{x}, P, H) = \text{SoftMax}(\psi(\mathbf{x})^T P)H$$

Attention Reuse

We construct an attention reuse based on an instance-aware weight mechanism $g:\mathcal{X}\rightarrow\mathcal{W}$

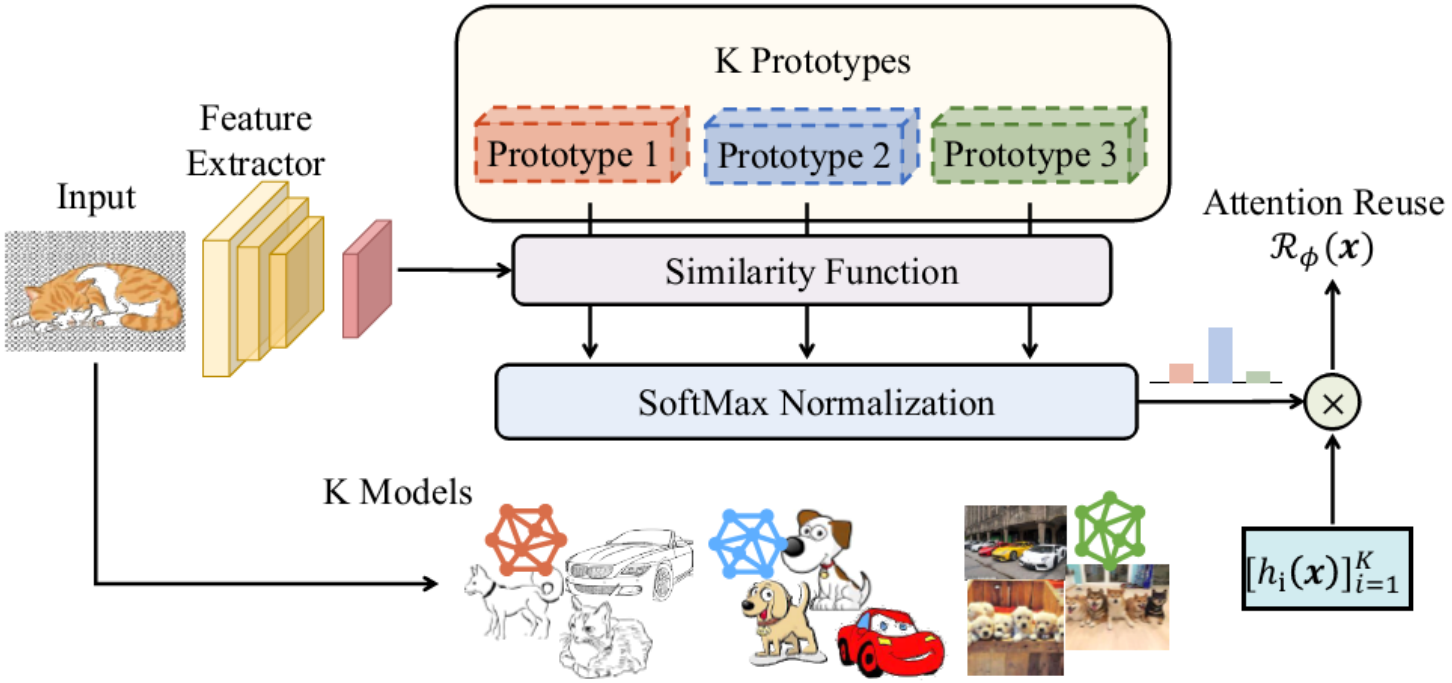


The output aggregation is calculated by the assigned weights and raw predictions.

$$\mathcal{R}_\phi(\mathbf{x}) = \text{Attention}(\mathbf{x}, P, H) = \text{SoftMax}(\psi(\mathbf{x})^T P)H$$

Attention Reuse

We construct an attention reuse based on an instance-aware weight mechanism $g:\mathcal{X}\rightarrow\mathcal{W}$



How to optimize the attention mechanism (*prototypes P*) ?

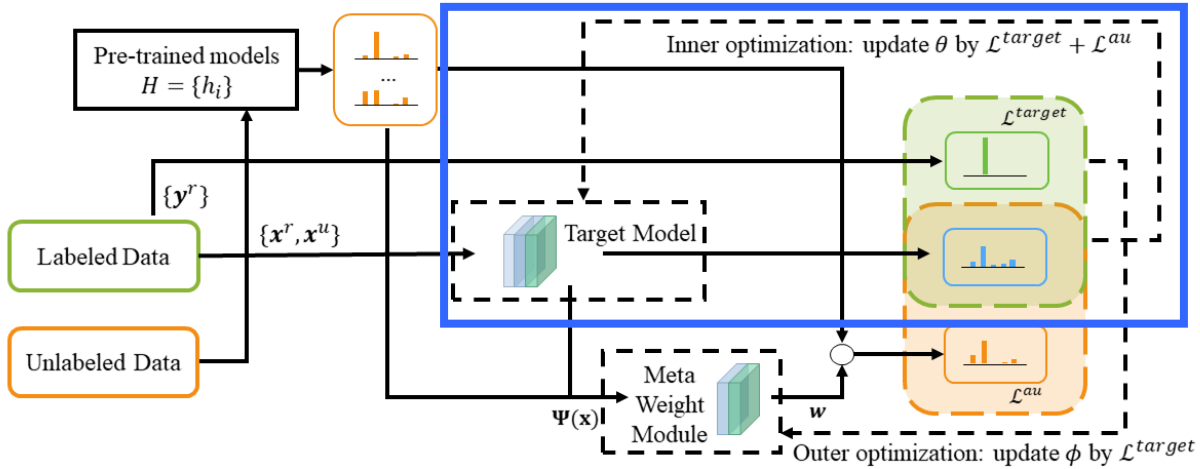
$$\mathcal{R}_\phi(\mathbf{x}) = \text{Attention}(\mathbf{x}, P, H) = \text{SoftMax}(\psi(\mathbf{x})^T P)H$$

Optimization

- We utilize pre-trained models via the exploration of target samples.
- We use a two-level objective to build a decoupled student model via model reuse.

$$\min_{\phi} \sum_{i=1}^{n} L(h(\mathbf{x}_i; \hat{\theta}), y_i) \tag{5}$$

$$\text{s.t. } \hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} L\left(h(\mathbf{x}_i; \theta), y_i\right) + \sum_{i=n+1}^{n+m} \Omega_{\phi}\left(\mathbf{x}_i; \theta, H\right)$$



- Firstly, we derive target model $\theta(\phi)$ via the current attention module ϕ.

Inner objective: a semi-supervised learning under the guide of attention reuse.

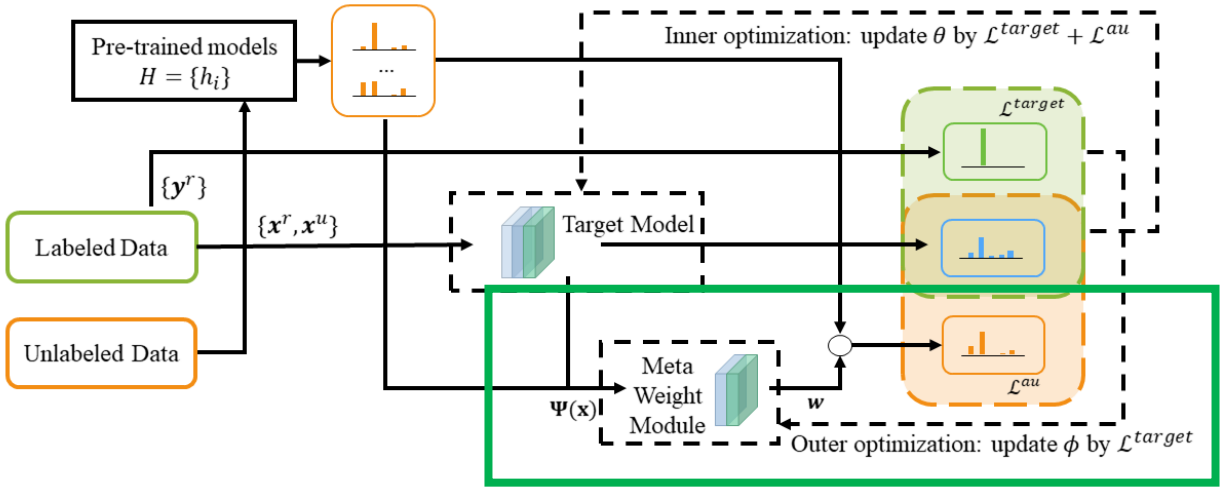
- Then, we evaluate the target model on labeled data and update ϕ via second-order gradient.

Optimization

- We utilize pre-trained models via the exploration of target samples.
- We use a two-level objective to build a decoupled student model via model reuse.

$$\min_{\phi} \sum_{i=1}^{n} L(h(\mathbf{x}_i; \hat{\theta}), y_i) \tag{5}$$

$$\text{s.t. } \hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} L\left(h(\mathbf{x}_i; \theta), y_i\right) + \sum_{i=n+1}^{n+m} \Omega_{\phi}\left(\mathbf{x}_i; \theta, H\right)$$



- Firstly, we derive target model $\theta(\phi)$ via the current attention module ϕ. Inner objective: a

  Inner objective: a semi-supervised learning under the guide of attention reuse.

- Then, we evaluate the target model on labeled data and update ϕ via second-order gradient.

  Outer objective: update the attention module based on the labeled data.

Theoretical Results: Robustness

Theorem 1. *Note that* $\mathcal{D}_T = \sum_{i=1}^{K} \lambda_i \mathcal{D}_i$, *the upper bound of consistent reuse* $\mathcal{R}_w$ *satisfies, for* $\alpha > 1$,

$$\min_{\mathbf{w}} \ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_T} L(\mathcal{R}_w(\mathbf{x}), y) \leq \sum_{i=1}^{K} w_i [d_\alpha(\mathcal{D}_T \,||\, \mathcal{D}_i)\epsilon]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$$

where $d_\alpha(\mathcal{D}\|\mathcal{D}') = e^{D_\alpha(\mathcal{D}\|\mathcal{D}')}$ *denotes the exponential of the Rényi Divergence of two distributions* $\mathcal{D}$ *and* $\mathcal{D}'$. *In the worst case* $\forall i \in [K], d_\alpha(\mathcal{D}_i \,||\, \mathcal{D}_T) \to \infty$, $\alpha \to 1$, *the upper bound could be tailored as:*

$$\min_{\mathbf{w}} \max_{\mathcal{D}_T} \ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_T} L(\mathcal{R}_w(\mathbf{x}), y) \leq M. \qquad (3)$$

Theorem 2 (Robustness). *Note that* $\mathcal{D}_T = \sum_{i=1}^{K} \lambda_i \mathcal{D}_i$, *the upper bound of attention reuse in the worst case satisfies:*

$$\min_{\phi} \max_{\mathcal{D}_T} \ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_T} L\left(\mathcal{R}_\phi(\mathbf{x}), y\right) \leq \epsilon.$$

- Theorem 1 indicates that **consistent reuse is not effective** to distribution, which is composed with latent domains, particularly **there is no guarantee for consistent reuse in the worst case.**
- Theorem 2 indicates that **attention reuse is robust** to target distribution, which is composed with latent domains. Particularly, the ideal generalization error converges to a small constant, even in the worst case.

Theoretical Results: Generalization

Theorem 3 (Generalization). *Assume L is ζ-Lipschitz continuous w.r.t. ϕ. Let $\phi \in \mathbb{R}^{d'}(d' = K * p)$ be the parameters in a unit ball, and n be the labeled data size. Let $\phi^* = \arg\max_{\phi \in \mathbb{R}^{d'}} R_T(\hat{\theta}(\phi))$ be the optimal parameter in the unit ball, and $\hat{\phi}$ be the empirical optima among a candidate set $\mathcal{A}$. With probability at least $1 - \delta$ we have,*

$$R_T(\hat{\theta}(\phi^*)) \leq R_T(\hat{\theta}(\hat{\phi})) + \frac{3\zeta + \sqrt{4d' \ln(n) + 8\ln(2/\delta)}}{\sqrt{n}}.$$

- Supervised Learning which optimizes high-dimensional (d) parameters , achieves the optimal weight in the order $O(\sqrt{d \ln(d) \ln(n) / n})$
- We learn a lowdimensional ($d' \ll d$) attention module ϕ via bi-level optimization, sharing a order $O(\sqrt{d' \ln(n)/n})$, **favoring a better order than learning from scratch.**

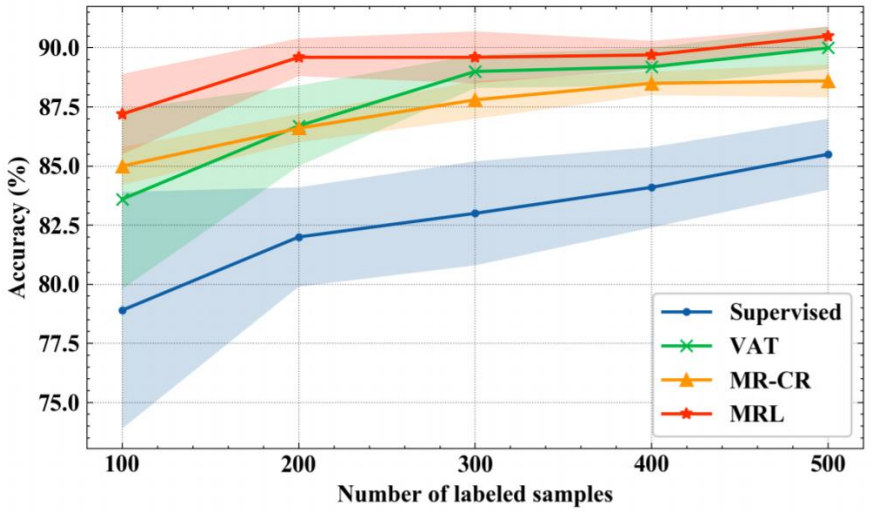
Experimental Results

- Existing Model Reuse methods are no longer tackle the latent domains, even worse than the SSL baselines.
- Our MRL still achieve the consistent performance gain.

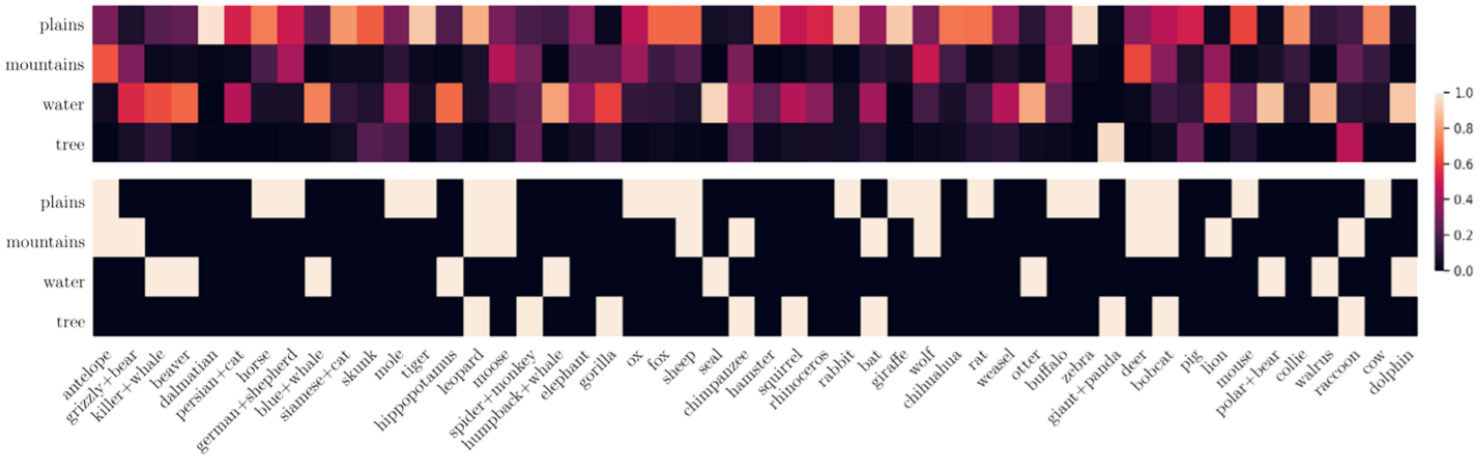
	Digital Recognition							
	MNIST	SVHN	USPS	MS	MU	SU	MSU	Ave. Gain
Supervised	.837 ± .018	.660 ± .013	.794 ± .023	.726 ± .021	.857 ± .014	.691 ± .009	.745 ± .006	-
PL	.870 ± .018	.706 ± .018	.818 ± .024	.762 ± .010	.890 ± .008	.728 ± .014	.774 ± .011	.034
TE	.853 ± .017	.710 ± .017	.822 ± .032	.774 ± .006	.879 ± .015	.736 ± .010	.752 ± .011	.031
VAT	.834 ± .020	.668 ± .023	.827 ± .024	.716 ± .024	.871 ± .026	.706 ± .015	.760 ± .018	.010
MR-BS	.899 ± .008	.664 ± .017	.851 ± .017	.786 ± .003	.936 ± .001	.820 ± .003	.756 ± .007	.057
MR-CR	.877 ± .006	.681 ± .013	.865 ± .008	.746 ± .010	.877 ± .014	.719 ± .012	.758 ± .011	.030
MRL	**.971 ± .002**	**.764 ± .010**	**.898 ± .005**	**.867 ± .007**	**.956 ± .006**	**.837 ± .003**	**.867 ± .004**	**.131**
	Attribute Classification							
	Smart	Slow	Bulbous	Solitary	Nestspot	Lean	Spots	Ave. Gain
Supervised	.820 ± .010	.853 ± .021	.846 ± .011	.812 ± .012	.862 ± .009	.853 ± .009	.820 ± .021	-
PL	.855 ± .008	.886 ± .012	.865 ± .008	.858 ± .007	.879 ± .009	.859 ± .013	.847 ± .011	.026
TE	.855 ± .014	.878 ± .008	.864 ± .006	.849 ± .009	.873 ± .007	.855 ± .006	.862 ± .008	.024
VAT	.855 ± .011	.884 ± .015	.876 ± .007	.857 ± .021	.887 ± .008	.861 ± .006	.871 ± .010	.032
MR-BS	.811 ± .028	.870 ± .011	.844 ± .009	.818 ± .017	.852 ± .014	.851 ± .010	.862 ± .007	.006
MR-CR	.827 ± .013	.874 ± .011	.852 ± .017	.826 ± .016	.852 ± .013	.847 ± .007	.866 ± .006	.011
MRL	**.876 ± .016**	**.901 ± .006**	**.890 ± .011**	**.874 ± .011**	**.896 ± .009**	**.889 ± .005**	**.896 ± .008**	**.051**

Experimental Results

- MRL consistently outperforms the compared methods, especially when labeled samples are few.
- We could find it has reasonably mined the relationship between different instances and latent domains.



Results on varying budgets.



Comparison between the learned attention and the ground truth

Take-Home Message

- ✓ We propose a new problem: Model Reuse for Latent Domains, where the target data are composed with latent domains.
- ✓ We propose a novel method MRL. Both domain characteristics and pre-trained models are considered for the exploration of instances in the target task.
- ✓ Theoretical analysis and empirical studies verify our robustness and effectiveness.

Thank you!

[Towards Robust Model Reuse in the Presence of Latent Domains. IJCAI'21]

Jie-Jing Shao
(shaojj@lamda.nju.edu.cn)