From Latent Heterogeneity to Out-of-Distribution Generalization

Heterogeneous Risk Minimization(ICML 2021)
Kernelized Heterogeneous Risk Minimization(NeurIPS 2021)

Jiashuo Liu

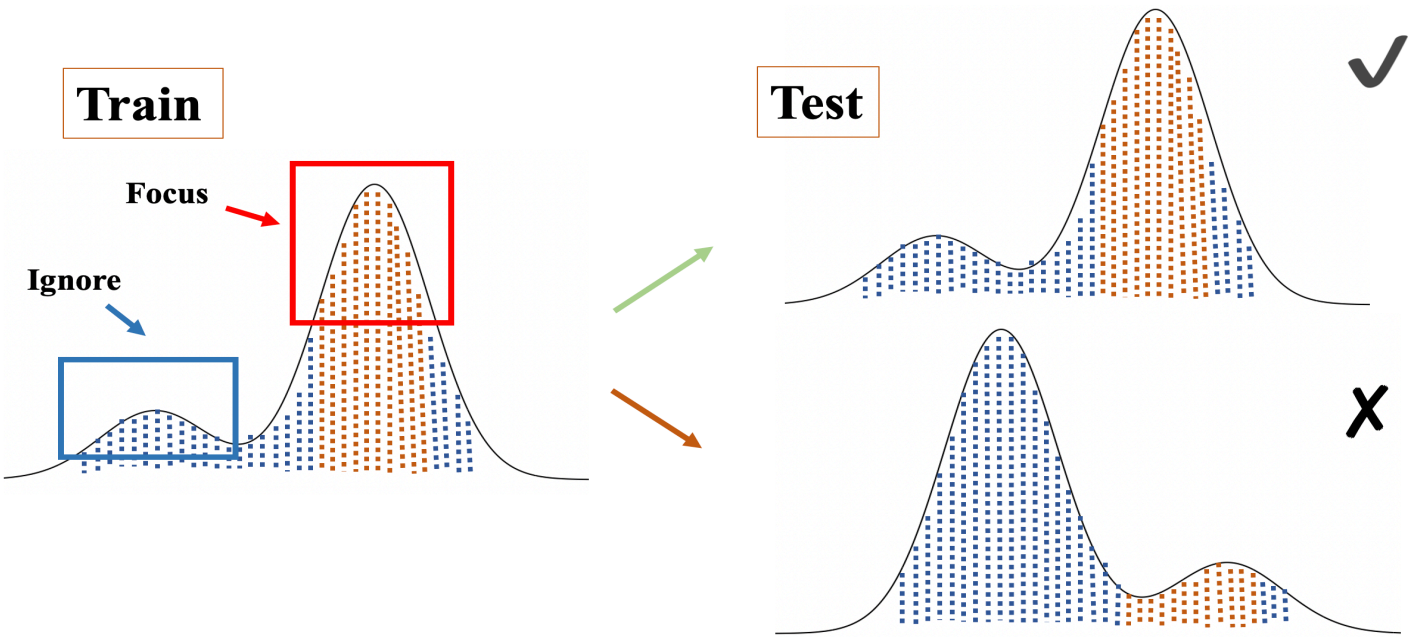
TrustWorthy-AI Group, CST, Tsinghua University

November 5, 2021

Empirical Risk Minimization(ERM)

$$\theta_{ERM} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(\theta; X_i, Y_i) \tag{1}$$

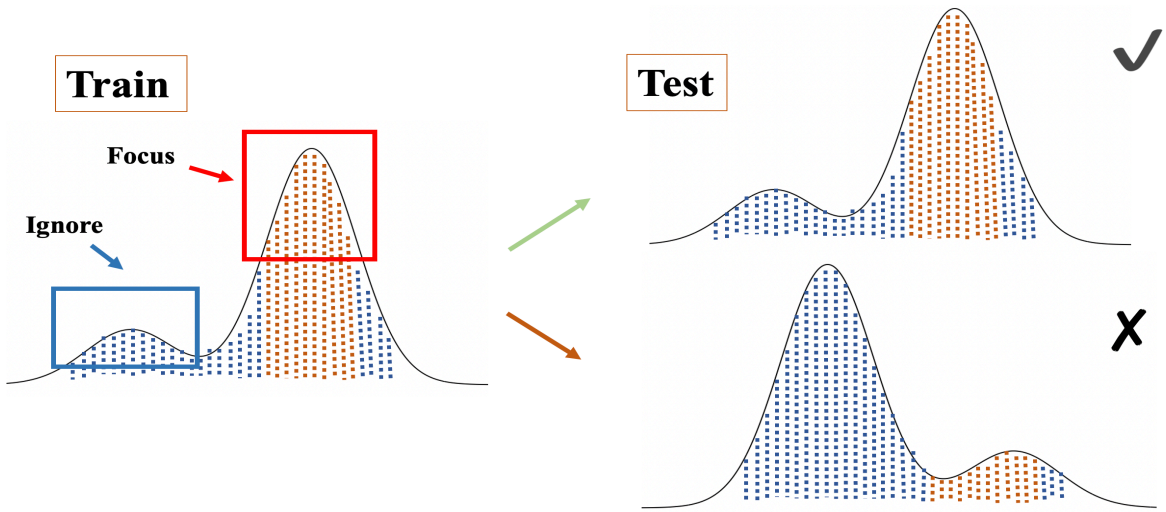
- Optimize the **average error** of data points.
- Focus on the **major group** of data.
- Ignore the **minor group** of data $\rightarrow$ Break down under distributional shifts



Latent Heterogeneity in Data

Data are collected from multiple sources, which induces latent heterogeneity.

- ERM excessively focuses on the majority and ignores the minor components in data.
- Overall Good = Majority Perfect + Minority Bad
- Majority and Minority can change across different data sources/environments.
- Latent Heterogeneity renders ERM break down under distributional shifts.



Insights: We should leverage the latent heterogeneity in data and develop more rational risk minimization approach to achieve Majority Good and Minority Good.

Out-of-Distribution Generalization Problem(OOD Problem)

Out-of-Distribution Generalization Problem(OOD Problem) is proposed in order to guarantee the generalization ability under distributional shifts, which can be formalized as:

$$\theta_{OOD} = \arg\min_{\theta} \max_{e \in \mathrm{supp}(\mathcal{E})} \mathcal{L}^e(\theta; X, Y) \tag{2}$$

where

- $\mathcal{E}$ is the random variable on indices of all possible environments, and for each environment $e \in \mathrm{supp}(\mathcal{E})$, the data distribution is denoted as $P^e(X, Y)$.
- The data distribution $P^e(X, Y)$ can be quite different among environments in $\mathrm{supp}(\mathcal{E})$.
- $\mathcal{L}^e(\theta; X, Y)$ denotes the risk of predictor θ on environment e, whose formulation is given by:

$$\mathcal{L}^e(\theta; X, Y) = \mathbb{E}_{X,Y \sim P^e}[\ell(\theta; X, Y)] \tag{3}$$

- OOD problem hopes to optimize the worst-case risk of all possible environments or distributions in $\mathrm{supp}(\mathcal{E})$

Invariance Assumption

To deal with the potential distributional shifts, one common assumption made in invariant learning is the **Invariance Assumption**.

Assumption (Invariance Assumption)

There exists random variable $\Phi^*(X)$ such that the following properties hold:

1. Invariance property: *for all $e_1, e_2 \in \text{supp}(\mathcal{E})$, we have*

$$P^{e_1}(Y|\Phi^*(X)) = P^{e_2}(Y|\Phi^*(X)) \tag{4}$$

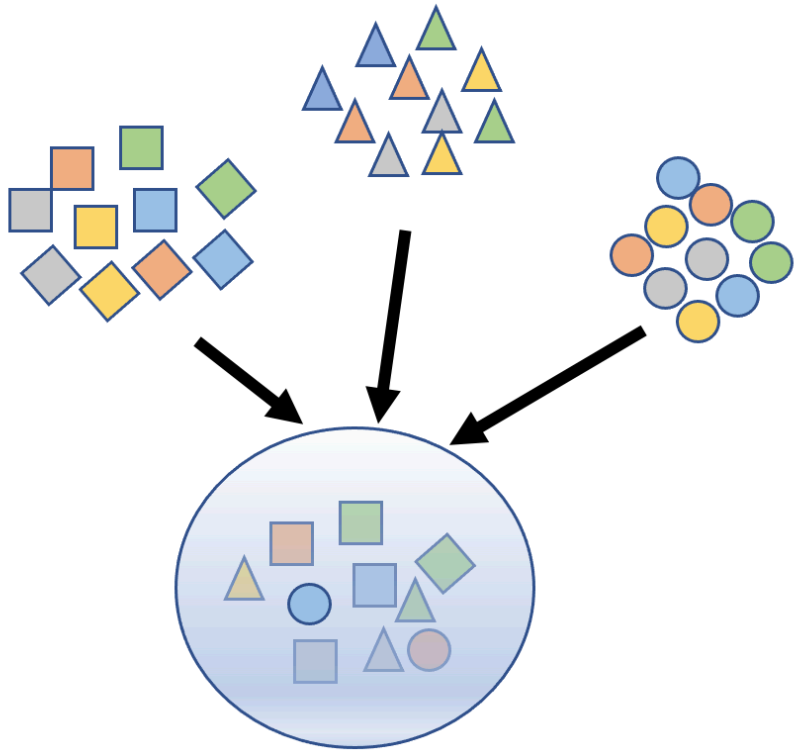
2. Sufficiency property: $Y = f(\Phi^*) + \epsilon,\ \epsilon \perp X$.

Here we make some demonstrations on the Invariance Assumption:

- The first property assumes that the relationship between $\Phi^*(X)$ and Y remains invariant across environments, which is also referred to as causal relationship.
- The second property assumes that $\Phi^*(X)$ can provide all information of the target label Y.
- $\Phi^*(X)$ is referred to as **(Causally) Invariant Predictors**.

Limitations: No Training Environments

Modern datasets are frequently assembled by merging data from multiple sources without explicit source labels, which means there are not multiple environments but only one pooled dataset.



Heterogeneous Risk Minimization[2](HRM)

Assumption (Heterogeneity Assumption)

For random variable pair (X, Φ^*) and Φ^* satisfying the Invariance Assumption, using functional representation lemma[1], there exists random variable Ψ^* such that $X = X(\Phi^*, \Psi^*)$, then we assume $P^e(Y|\Psi^*)$ can arbitrary change across environments $e \in \mathrm{supp}(\mathcal{E})$.

Problem (Heterogeneous Risk Minimization Problem)

Given heterogeneous dataset $D = \{D^e\}_{e \in \mathrm{supp}(\mathcal{E}_{latent})}$ without environment labels, the task is to generate environments $\mathcal{E}_{learn}$ with minimal $|\mathcal{I}_{\mathcal{E}_{learn}}|$ and learn invariant model under learned $\mathcal{E}_{learn}$ with good OOD performance.

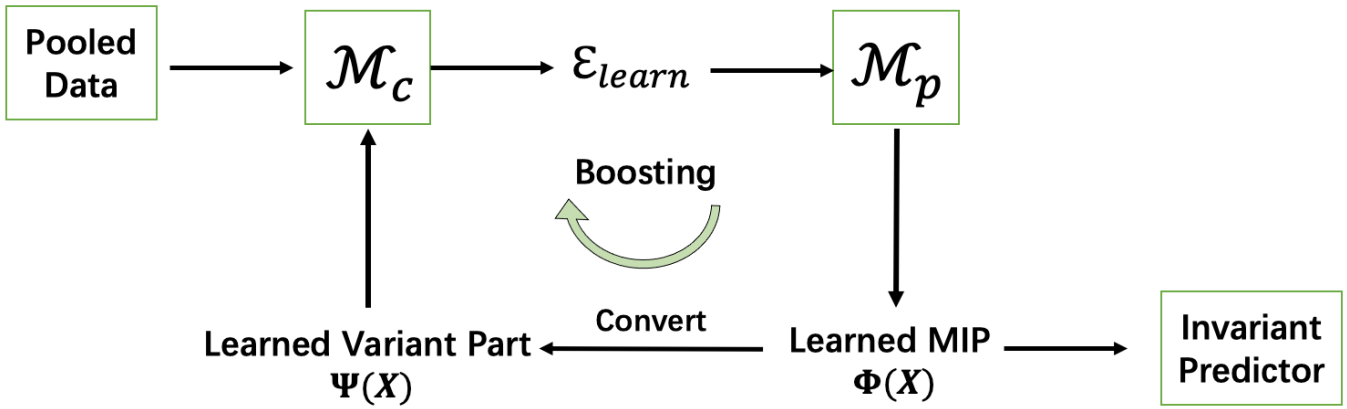
- Heterogeneous Risk Minimization temporarily focuses on a simple setting, where $X = [\Phi^*, \Psi^*]^T$ in raw feature level and Φ^*, Ψ^* satisfy the Invariance Assumption.

[1] El Gamal, A. and Kim, Y.-H. Network information theory. Network Information Theory, 12 2011.

[2] Jiashuo Liu, Zheyuan Hu, Peng Cui *et al.* Heterogeneous Risk Minimization. *In ICML 2021*.

Heterogeneous Risk Minimization(HRM)

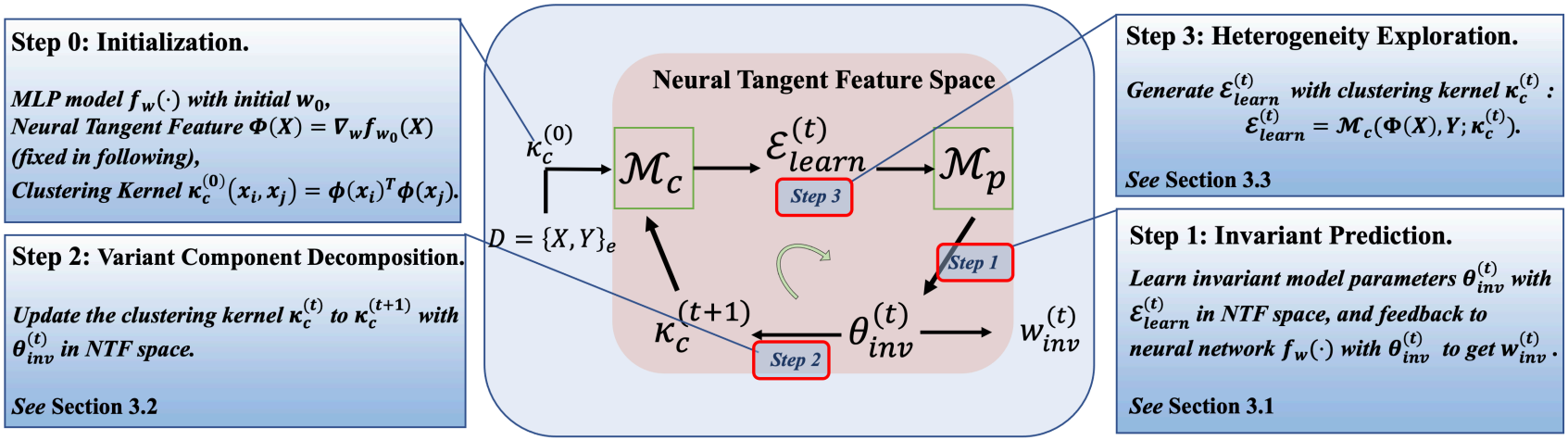
The Heterogeneous Risk Minimization framework contains two modules, named **Heterogeneity Identification** module $\mathcal{M}_c$ and **Invariant Prediction** module $\mathcal{M}_p$.



- Focus on the raw feature setting, and simply adopt feature selection techniques as $\Phi(X) = M \odot X$ and $\Psi(X) = (1 - M) \odot X$.
- Cannot deal with complicated data, where Φ^* and Ψ^* are hidden components.

The Goal of KerHRM: Extend HRM to Complicated Data

Kernelized Heterogeneous Risk Minimization(KerHRM[3])



- Step 0:

$$f_w(X) \approx f_{w_0}(X) + \nabla_w f_{w_0}(X)^T (w - w_0) \tag{5}$$

$$= f_{w_0}(X) + \Phi(X)^T (w - w_0) \tag{6}$$

$$\approx f_{w_0}(X) + USV^T (w - w_0) \tag{7}$$

$$= f_{w_0}(X) + \Psi(X)\left(V^T (w - w_0)\right) = f_{w_0}(X) + \Psi(X)\theta \tag{8}$$

where $\Psi(X) \in \mathbb{R}^k$ is called the reduced Neural Tangent Features(Reduced NTFs), which convert the complicated data, non-linear setting into raw feature data, linear setting.

[3] Jiashuo Liu, Zheyuan Hu, Peng Cui *et al.* Kernelized Heterogeneous Risk Minimization. *In NeurIPS 2021*.

Algorithms

- Step 1: $\mathcal{M}_p$ Invariant Learning with Reduced NTFs $\Psi(X)$[4]:

$$\theta_{inv} = \arg\min_{\theta} \sum_{e \in \mathcal{E}_{learn}} \mathcal{L}^e(\theta; \Psi, Y) + \alpha \mathrm{Var}_{\mathcal{E}_{learn}}(\nabla_\theta \mathcal{L}^e) \tag{9}$$

The obtained θ_{inv} captures the invariant component in data, which can be used to wipe out the invariant part inside data.

- Step 2: Variant Component Decomposition with θ_{inv}.
  - The initial similarity of two data points x_i and x_j:

$$\kappa_c^{(0)}(x_i, x_j) = \phi(x_i)^T \phi(x_j) = \langle U_i S, U_j S \rangle \tag{10}$$

  - Wipe out the invariant component with θ_{inv}:

$$\Psi_V^{(t+1)}(x_i) \leftarrow U_i S - \left\langle U_i S, \theta_{inv}^{(t)} \right\rangle \theta_{inv}^{(t)} / \|\theta_{inv}^{(t)}\|^2 \tag{11}$$

  - Obtain a new kernel for clustering:

$$\kappa_c^{(t+1)}(x_i, x_j) = \Psi_V^{(t+1)}(x_i)^T \Psi_V^{(t+1)}(x_j) \tag{12}$$

[4] Here we adopt the regularizer proposed in 'Masanori Koyama, Shoichiro Yamaguchi. When is invariance useful in an Out-of-Distribution Generalization problem ?'

Algorithms

- Step 3: $\mathcal{M}_c$ Heterogeneity Exploration with κ_c

  - Capture the different relationship between Ψ_V^* and Y.
  - Use $P(Y|\Psi_V)$ as the cluster centre: assume the j-th cluster centre $P_{\Theta_j}(Y|\Psi_V(X))$ to be a Gaussian around $f(\Theta_j; \Psi_V()X)$ as:

$$h_j(\Psi_V(X), Y) = P_{Theta_j}(Y|\Psi_V(X)) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(Y - f(\Theta_j; \Psi_V(X)))^2/2]\sigma^2) \tag{13}$$

  - Propose on convex clustering algorithm, which finds a mixture distribution in distribution set $\mathcal{Q}$ defined as:

$$\mathcal{Q} = \{Q : Q = \sum_{k\in[K]} q_j h_j\} \tag{14}$$

and gives the objective function:

$$\min_{Q\in\mathcal{Q}} D_{KL}(\hat{P}_N \| Q) \Leftrightarrow \min_{\Theta,\mathbf{q}} \left\{ \mathcal{L}_c = -\frac{1}{N} \sum_{i\in[N]} \log \left[\sum_{j\in[K]} q_j h_j(\psi_V(x_i), y_i) \right] \right\} \tag{15}$$

Simulation: Colored MNIST

Data Generation[5]:

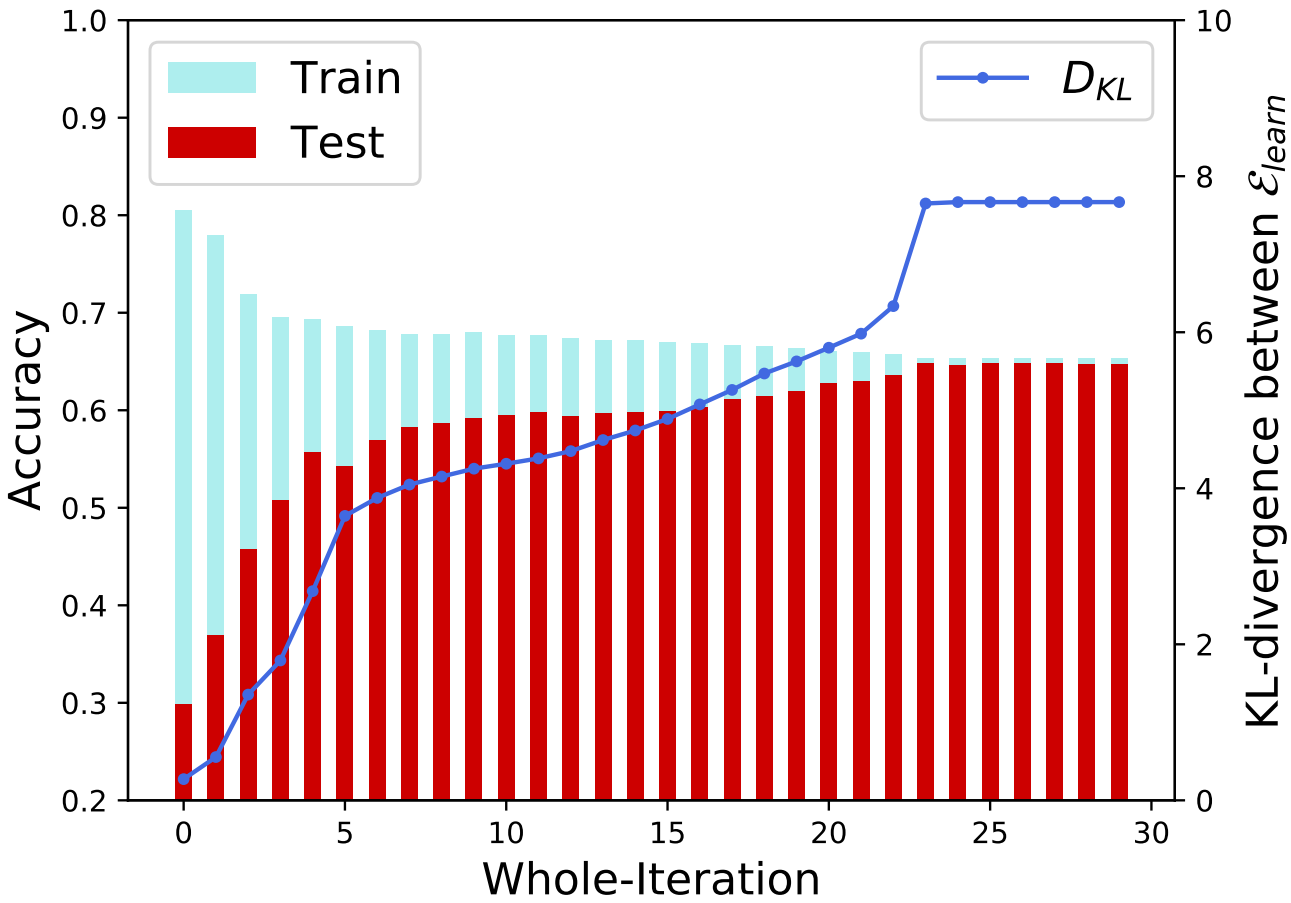
- Firstly, a binary label Y is assigned to each images according to its digits: $Y = 0$ for digits 0∼4 and $Y = 1$ for digits 5∼9.
- Secondly, we sample the color id C by flipping Y with probability e and therefore forms environments, where $e = 0.1$ for the first training environment, $e = 0.2$ for the second training environments and $e = 0.9$ for the testing environment.
- Thirdly, we induce noisy labels by randomly flipping the label Y with probability 0.2.

Settings:

- For training, $e_1 = 0.1, e_2 = 0.2$.
- For testing, $e_3 = 0.9$.
- The correlation between Color and Label is inverse between training and testing.

[5] Martin Arjovsky *et al.* Invariant Risk Minimization

Surprising Results



- D_{KL} denotes $KL(P_1(Y|C)\|P_2(Y|C))$

Conclusion

As for KerHRM, there exist some drawbacks:

- The convergence guarantee of the whole framework, especially the frontend, still remains ambiguous.
- Extend to more complicated data?

As for the OOD Generalization problem, there exist some open problems

- How to formulate the OOD Generalization problem? How to justify its learnability?
- Environment complexity of methods for OOD Generalization.
- Real datasets to evaluate the effect of methods for OOD Generalization?
- Incorporate pre-trained models to deal with complicated data?

Some other materials

- Anual Progress Report on Out-of-Distribution Generalization[6]
- Stable Learning and its Causal Implication[7]

[6] http://pengcui.thumedialab.com/papers/OOD_APR_valse2021.pdf

[7] http://pengcui.thumedialab.com/papers/Stable%20Learning-tutorial-valse2021.pdf

Contact

Jiashuo Liu

Department of Computer Science and Technology
Tsinghua University

(+86) 13015155336
@liujiashuo77
liujiashuo77@gmail.com
ljsthu.github.io
https://github.com/LJSthu