Paper: FEW-SHOT DOMAIN ADAPTATION BY CAUSAL MECHANISM TRANSFER [3]

Jiaxin Liu

Group Reading

April 19, 2021

Overview

Domain Adaptation (DA)[1]

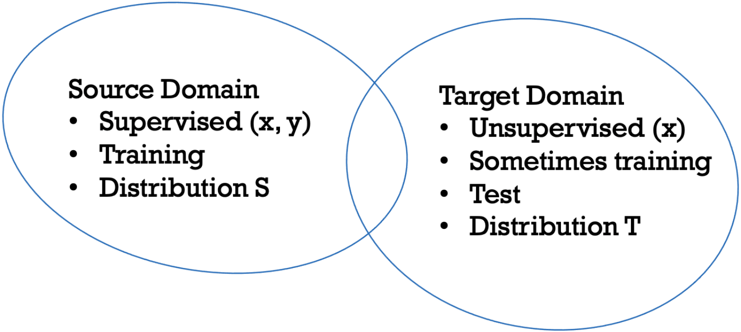


Figure: Domain adaptation.

Transfer Assumption (TA)

1. Example

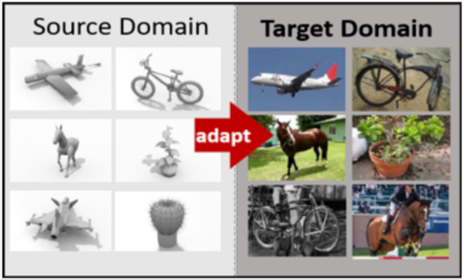


Figure: Example of the DA.

2. TAs: rely on distribution similarities.

Background: (non)linear independent component analysis (ICA)

- Linear ICA ($x = As + n$)

$$x_i(t) = \sum_{j=1}^{n} a_{ij} s_j(t) \quad \text{for all } i, j = 1, \ldots, n$$

  - $x_i(t)$ is i-th observed signal in time t.
  - a_{ij} constant parameters describing "mixing"
  - Assuming independent, non-Gaussian "sources" s_j

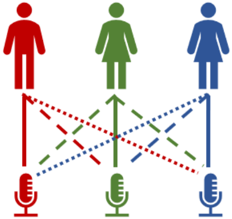


Figure: A common example application (cocktail party problem).

Background: (non)linear independent component analysis (ICA)

- Nonlinear ICA ($x = f(s|\theta) + n$)

$$x_i(t) = f_i(s_1(t), \ldots, s_n(t)) \quad \text{for all } i, j = 1, \ldots n$$

  - Nonlinear ICA is not identifiable

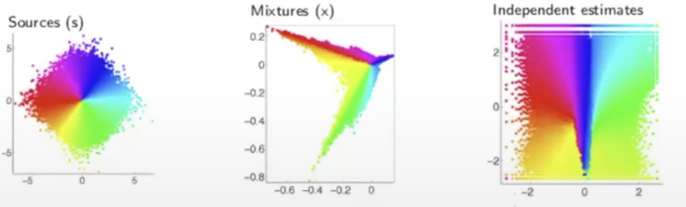


Figure: An example of nonlinear ICA.

Background: (non)linear independent component analysis (ICA)

- Nonlinear ICA ($x = f(s|\theta) + n$)

$$x_i(t) = f_i(s_1(t), \ldots, s_n(t)) \quad \text{for all } i, j = 1, \ldots n$$

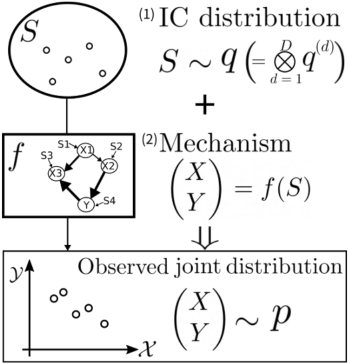


Figure: Nonparametric generative model of nonlinear ICA.

Problem setup

- Multi-source domains $[K]$:
  - input space: $\mathscr{X} \in \mathbb{R}^{D-1}$
  - label space: $\mathscr{Y} \in \mathbb{R}$
  - overall data space: $\mathscr{Z} := \mathscr{X} \times \mathscr{Y} \in \mathbb{R}^{D}$
  - labeled data point: $Z = (X, Y)$.

Problem setup

- Multi-source domains $[K]$:
  - input space: $\mathscr{X} \in \mathbb{R}^{D-1}$
  - label space: $\mathscr{Y} \in \mathbb{R}$
  - overall data space: $\mathscr{Z} := \mathscr{X} \times \mathscr{Y} \in \mathbb{R}^{D}$
  - labeled data point: $Z = (X, Y)$.
- Target domain:
  - limited number (n_{Tar}) of labeled data $\rightarrow$ "few-shot"
  - target distribution: p_{Tar}
  - sampled $D_{Tar} := \{Z_i\}_{i=1}^{n_{Tar}} \overset{i.i.d.}{\sim} p_{Tar}$.

Problem setup

- Multi-source domains $[K]$:
  - input space: $\mathscr{X} \in \mathbb{R}^{D-1}$
  - label space: $\mathscr{Y} \in \mathbb{R}$
  - overall data space: $\mathscr{Z} := \mathscr{X} \times \mathscr{Y} \in \mathbb{R}^{D}$
  - labeled data point: $Z = (X, Y)$.
- Target domain:
  - limited number (n_{Tar}) of labeled data $\to$ "few-shot"
  - target distribution: p_{Tar}
  - sampled $D_{Tar} := \{Z_i\}_{i=1}^{n_{Tar}} \overset{i.i.d.}{\sim} p_{Tar}$.
- Problem:
  - goal: find $g^* : \mathbb{R}^{D-1} \to \mathbb{R}^{D}$ which performs well for p_{Tar}
  - target risk: $R(g) := \mathbb{E}_{p_{Tar}} l(g, Z) \Rightarrow g^* \in \arg\min_{g \in G} R(g)$
  - empirical risk: $\hat{R} := \frac{1}{n_{Tar}} \sum_{i=1}^{n_{Tar}} l(g, Z_i) \Rightarrow \hat{g} \in \arg\min_{g \in G} \hat{R}(g)$.
  - $\hat{R}(g) \to R(g)$? Can we use the source data?
  - Source distributions: $\{p_k\}_{k=1}^{K}$
  - independent samples: $D_k := \{Z_{k,i}^{Src}\}_{i=1}^{n_k} \overset{i.i.d.}{\sim} p_k (k \in [K], n_k \in N)$.

Key assumption

All domains follow nonlinear ICA models with identical mixing functions.

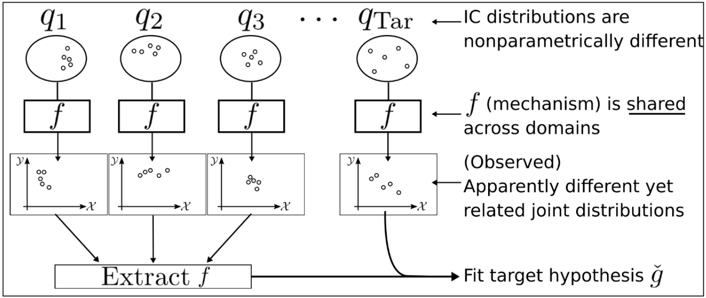


Figure: An example of nonlinear ICA.

- transformation $f : \mathbb{R}^D \rightarrow \mathbb{R}^D$

Key assumption

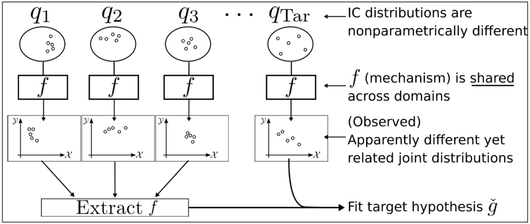


Figure: An example of nonlinear ICA.

- IC distributions: $q_{Tar}, q_k \in Q(k \in [K])$, $S_{k,i}^{Src} \overset{i.i.d.}{\sim} q_k$
- $f : \mathbb{R}^D \to \mathbb{R}^D$
- $Z_{k,i}^{Src} = f(S_{k,i}^{Src})$

Mechanism Transfer

Algorithm 1 Proposed method: mechanism transfer

Input: Source domain data sets $\{\mathcal{D}_k\}_{k\in[K]}$, target domain data set $\mathcal{D}_{\mathrm{Tar}}$, nonlinear ICA algorithm ICA, and a learning algorithm $\mathcal{A}_{\mathcal{G}}$ to fit the hypothesis class $\mathcal{G}$ of predictors.

// Step 1. Estimate the shared transformation.

$\hat{f} \leftarrow \mathrm{ICA}(\mathcal{D}_1, \ldots, \mathcal{D}_K)$

// Step 2. Extract and shuffle target independent components

$\hat{s}_i \leftarrow \hat{f}^{-1}(Z_i), \quad (i = 1, \ldots, n_{\mathrm{Tar}})$

$\{\bar{s}_{\boldsymbol{i}}\}_{\boldsymbol{i}\in[n_{\mathrm{Tar}}]^D} \leftarrow \mathrm{AllCombinations}(\{\hat{s}_i\}_{i=1}^{n_{\mathrm{Tar}}})$

// Step 3. Synthesize target data and fit the predictor.

$\bar{z}_{\boldsymbol{i}} \leftarrow \hat{f}(\bar{s}_{\boldsymbol{i}})$

$\check{g} \leftarrow \mathcal{A}_{\mathcal{G}}(\{\bar{z}_{\boldsymbol{i}}\}_{\boldsymbol{i}})$

Output: $\check{g}$: the predictor in the target domain.

Figure: Algorithm for the proposed method: mechanism transfer.

Step1: Estimate f using the source domain data

- Get f via generalized contrastive learning (GCL [2])
  - Binary classification function:

$$r_{\hat{f},\psi}(z,u) := \sum_{d=1}^{D} \psi_d(\hat{f}^{-1}(z)_d, u)$$

  - $\hat{f} : \mathbb{R}^D \to \mathbb{R}^D$ estimator of f
  - u: auxiliary information
  - classification task: $(Z_k^{Src}, k) \to +$, $(Z_k^{Src}, k')(k' \neq k) \to -$.

Step1: Estimate f using the source domain data

- Get f via generalized contrastive learning (GCL [2])
  - Binary classification function:

$$r_{\hat{f},\psi}(z,u) := \sum_{d=1}^{D} \psi_d(\hat{f}^{-1}(z)_d, u)$$

  - $\hat{f}: \mathbb{R}^D \to \mathbb{R}^D$ estimator of f
  - u: auxiliary information
  - classification task: $(Z_k^{Src}, k) \to +$, $(Z_k^{Src}, k')(k' \neq k) \to -$.
- Domain-contrastive learning criterion to estimate f:

$$\underset{\hat{f}\in\mathscr{F},\{\psi_d\}_{d=1}^{D}\subset\Psi}{\arg\min} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\phi\left(r_{\hat{f},\psi}(Z_{k,i}^{Src}, k)\right) + \mathbb{E}_{k'\neq k}\phi\left(-r_{\hat{f},\psi}(Z_{k,i}^{Src}, k')\right)\right)$$

Step1: Estimate f using the source domain data

- Domain-contrastive learning criterion to estimate f:

$$\underset{\hat{f}\in\mathscr{F},\{\psi_d\}_{d=1}^{D}\subset\Psi}{\arg\min} \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\phi\left(r_{\hat{f},\psi}(Z_{k,i}^{Src},k)\right)+\mathbb{E}_{k'\neq k}\phi\left(-r_{\hat{f},\psi}(Z_{k,i}^{Src},k')\right)\right)$$

  - logistic loss $\phi(m) := \log(1+\exp(-m))$

- Domain-contrastive learning criterion to estimate f:

$$\underset{\hat{f}\in\mathscr{F},\{\psi_d\}_{d=1}^{D}\subset\Psi}{\arg\min} \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\phi\left(r_{\hat{f},\psi}(Z_{k,i}^{Src},k)\right)+\mathbb{E}_{k'\neq k}\phi\left(-r_{\hat{f},\psi}(Z_{k,i}^{Src},k')\right)\right)$$

  - logistic loss $\phi(m) := \log(1+\exp(-m))$

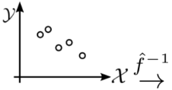
Algorithm 1 Proposed method: mechanism transfer

Input: Source domain data sets $\{\mathcal{D}_k\}_{k\in[K]}$, target domain data set $\mathcal{D}_{\mathrm{Tar}}$, nonlinear ICA algorithm ICA, and a learning algorithm $\mathcal{A}_{\mathcal{G}}$ to fit the hypothesis class $\mathcal{G}$ of predictors.

// Step 1. Estimate the shared transformation.

$\hat{f} \leftarrow \mathrm{ICA}(\mathcal{D}_1,\ldots,\mathcal{D}_K)$

Figure: Algorithm for step 1.



(a) Labeled target data

Figure: (a) The algorithm is given labeled target domain data.

Step2: Extract and inflate the target ICs using $\hat{f}$

- Extract ICs:

$$\hat{s}_i = \hat{f}^{-1}(Z_i)$$

- Inflate the set of IC:

$$\bar{s}_i = \left(\hat{s}_{i_1}^{(1)}, \ldots, \hat{s}_{i_D}^{(D)}\right), \quad i = (i_1, \ldots, i_D) \in [n_{Tar}]^D$$

Step2: Extract and inflate the target ICs using $\hat{f}$

- Extract ICs:

$$\hat{s}_i = \hat{f}^{-1}(Z_i)$$

- Inflate the set of IC:

$$\bar{s}_i = \left(\hat{s}_{i_1}^{(1)}, \ldots, \hat{s}_{i_D}^{(D)}\right), \quad i = (i_1, \ldots, i_D) \in [n_{Tar}]^D$$

- // Step 2. Extract and shuffle target independent components

  $\hat{s}_i \leftarrow \hat{f}^{-1}(Z_i), \quad (i = 1, \ldots, n_{\mathrm{Tar}})$

  $\{\bar{s}_i\}_{i \in [n_{\mathrm{Tar}}]^D} \leftarrow \mathrm{AllCombinations}(\{\hat{s}_i\}_{i=1}^{n_{\mathrm{Tar}}})$

Figure: Algorithm for step 2.

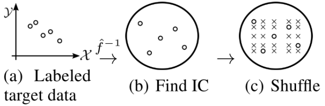


Figure: (b) From labeled target domain data, extract the ICs. (c) By shuffling the values, synthesize likely values of IC.

Step3: Synthesize target data from the inflated ICs

- Target risk:

$$\check{R}(g) := \frac{1}{n_{Tar}^{D}} \sum_{i \in [n_{Tar}]^{D}} l(g, \hat{f}(\bar{s}_i))$$

- Empirical risk minimization:

$$\check{g} \in \arg\min_{g \in G} \{\check{R}(g) + \Omega(g)\}$$

  - $\Omega(g) = \lambda \|g\|^2$, where $\lambda > 0$.

Step3: Synthesize target data from the inflated ICs

- Target risk:

$$\check{R}(g) := \frac{1}{n_{Tar}^{D}} \sum_{i \in [n_{Tar}]^D} l(g, \hat{f}(\bar{s}_i))$$

- Empirical risk minimization:

$$\check{g} \in \arg\min_{g \in G} \{\check{R}(g) + \Omega(g)\}$$

  - $\Omega(g) = \lambda \|g\|^2$, where $\lambda > 0$.

- // Step 3. Synthesize target data and fit the predictor.
  $\bar{z}_i \leftarrow \hat{f}(\bar{s}_i)$
  $\check{g} \leftarrow \mathcal{A}_{\mathcal{G}}(\{\bar{z}_i\}_i)$
  Output: $\check{g}$: the predictor in the target domain.

Figure: Algorithm for step 3.

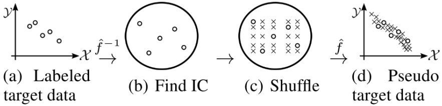


Figure: (d) From the synthesized IC, generate pseudo target data. The generated data is used to fit a predictor for the target domain.

Mechanism Transfer

Algorithm 1 Proposed method: mechanism transfer

Input: Source domain data sets $\{\mathcal{D}_k\}_{k\in[K]}$, target domain data set $\mathcal{D}_{\mathrm{Tar}}$, nonlinear ICA algorithm ICA, and a learning algorithm $\mathcal{A}_{\mathcal{G}}$ to fit the hypothesis class $\mathcal{G}$ of predictors.

// Step 1. Estimate the shared transformation.

$\hat{f} \leftarrow \mathrm{ICA}(\mathcal{D}_1, \ldots, \mathcal{D}_K)$

// Step 2. Extract and shuffle target independent components

$\hat{s}_i \leftarrow \hat{f}^{-1}(Z_i), \quad (i = 1, \ldots, n_{\mathrm{Tar}})$

$\{\bar{s}_{\boldsymbol{i}}\}_{\boldsymbol{i}\in[n_{\mathrm{Tar}}]^D} \leftarrow \mathrm{AllCombinations}(\{\hat{s}_i\}_{i=1}^{n_{\mathrm{Tar}}})$

// Step 3. Synthesize target data and fit the predictor.

$\bar{z}_{\boldsymbol{i}} \leftarrow \hat{f}(\bar{s}_{\boldsymbol{i}})$

$\check{g} \leftarrow \mathcal{A}_{\mathcal{G}}(\{\bar{z}_{\boldsymbol{i}}\}_{\boldsymbol{i}})$

Output: $\check{g}$: the predictor in the target domain.

Figure: Algorithm for the proposed method: mechanism transfer.

Data set

- Gasoline consumption in 18 of OECD countries over 19 years.
  - y: motor gasoline consumption/car
  - $x = [x_1, x_2, x_3]$: per-capita income, motor gasoline price, and the stock of cars per capita.
  - domain: each country is considered as as a domain.

Results

See table 2 in this paper.

- Shai Ben-David et al. “A theory of learning from different domains”. In: *Machine learning* 79.1 (2010), pp. 151–175.
- Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. “Nonlinear ICA using auxiliary variables and generalized contrastive learning”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 859–868.
- Takeshi Teshima, Issei Sato, and Masashi Sugiyama. “Few-shot domain adaptation by causal mechanism transfer”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9458–9469.