A Mathematical Framework of Multi-source Transfer Learning

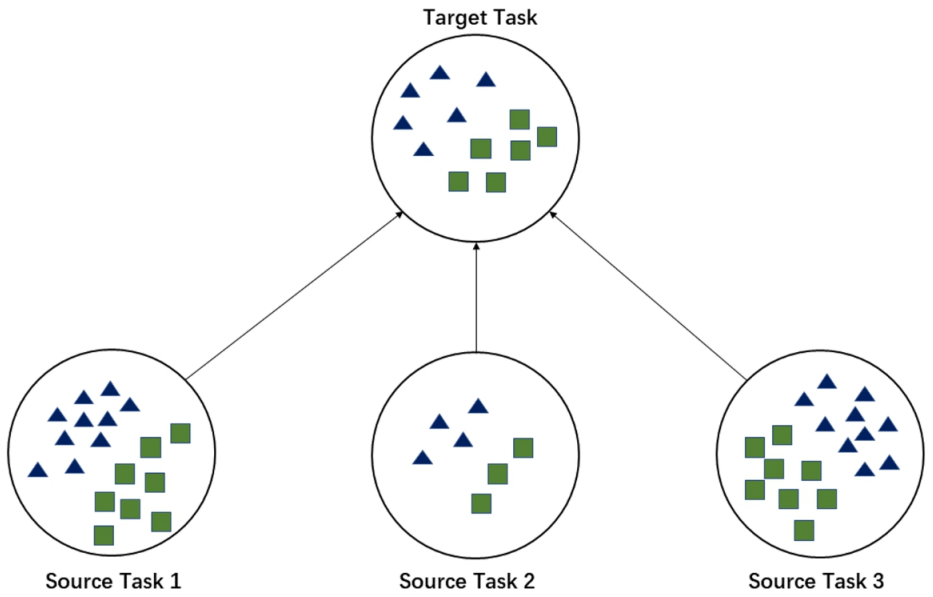
Ziyan Zheng

Related paper:
"**A Mathematical Framework for Quantifying Transferability in Multi-source Transfer Learning,**" Xinyi Tong, Xiangxiang Xu, Shao-Lun Huang and Lizhong Zheng, *Advances in Neural Information Processing Systems* 34 (2021): 26103-26116.
"**Multi-source Transfer Learning for Signal Detection over a Fading Channel with Co-channel Interference,**" Ziyan Zheng, Xinyi Tong, Xinchun Yu, Xiangxiang Xu and Shao-Lun Huang, *ICC 2022-IEEE International Conference on Communications*. IEEE, 2022.

Motivation

◆ Intuition

What kind of source would provide more useful knowledge to the target task?

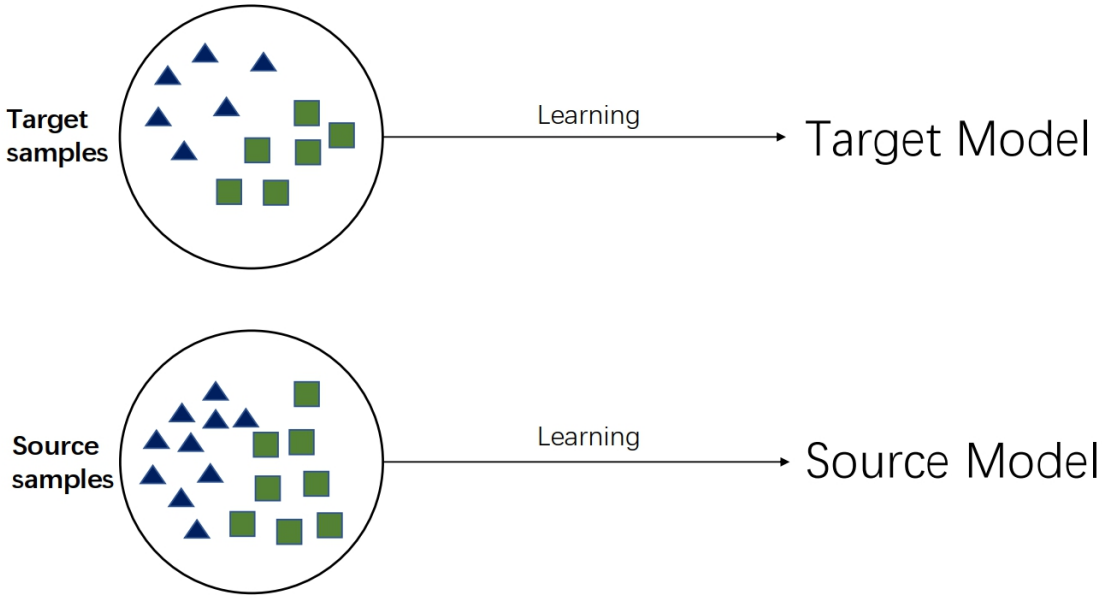


- Source task 1: quite good
- Source task 2: less samples → Sample size is important
- Source task 3: not like target task → Similarity is important

Source Task 1 has higher transferability.

Problem Formulation

◆ How to establish the framework for transfer learning?



Models using the source knowledge:

$$(1-\alpha)\cdot \text{Target Model} + \alpha \cdot \text{Source Model}$$

Optimal α ?

Discrete Case

◆ How to establish the framework for transfer learning?

Sample $x \in \mathcal{X}$ & Label $y \in \mathcal{Y}$

We hope to learn the target distribution $P_{XY}^{(0)}$

- Target samples $\{(x_\ell^{(0)}, y_\ell^{(0)})\}_{\ell=1}^{n_0}$  i.i.d from $P_{XY}^{(0)} \to \hat{P}_{XY}^{(0)}$
- Source samples $\{(x_\ell^{(1)}, y_\ell^{(1)})\}_{\ell=1}^{n_1}$  i.i.d from $P_{XY}^{(1)} \to \hat{P}_{XY}^{(1)}$

We use $(1-\alpha)\hat{P}_{XY}^{(0)} + \alpha\hat{P}_{XY}^{(1)}$ to estimate $P_{XY}^{(0)}$.

Note: Training the linear combination of the cross-entropy loss can also lead to the convex combination model above.

Discrete Case - Testing Loss

◆ How to evaluate the model?

$$L_{\text{test}}^{(\alpha)} \triangleq \mathbb{E}\left[\chi^2\left(P_{XY}^{(0)}, (1-\alpha)\hat{P}_{XY}^{(0)} + \alpha\,\hat{P}_{XY}^{(1)}\right)\right]$$

Why not use log-loss?

What is the optimal α?

$$\alpha^* = \arg\min_{\alpha} \; L_{\text{test}}^{(\alpha)}$$

Definition 1. *For random variable X and Y, given a reference distribution R_{XY} for any distribution P_{XY} and Q_{XY}, the referenced χ^2-distance between them is defined as*

$$\chi_R^2\left(P_{XY}, Q_{XY}\right) \triangleq \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{\left(P_{XY}(x,y) - Q_{XY}(x,y)\right)^2}{R_{XY}(x,y)}. \tag{7}$$

In particular, $\chi_R^2\left(P_{XY}, Q_{XY}\right)$ becomes Pearson χ^2-distance denoted by $\chi^2\left(P_{XY}, Q_{XY}\right)$ when $R_{XY} = P_{XY}$.

Discrete Case - Transferability

◆ How to evaluate the model?

$$L_{\text{test}}^{(\alpha)} \triangleq \mathbb{E}\left[\chi^2\left(P_{XY}^{(0)}, (1-\alpha)\hat{P}_{XY}^{(0)} + \alpha\,\hat{P}_{XY}^{(1)}\right)\right]$$

Transferability measure

$$\alpha^* = \arg\min_{\alpha} \; L_{\text{test}}^{(\alpha)}$$

Theorem

The testing loss can be computed as

$$L_{\text{test}}^{(\alpha)} = \alpha^2\chi^2\left(P_{XY}^{(0)}, P_{XY}^{(1)}\right) + \frac{(1-\alpha)^2}{n_0}V^{(0)} + \frac{\alpha^2}{n_1}V^{(1)}, \tag{3}$$

and the optimal α^* is

$$\alpha^* = \frac{\frac{1}{n_0}V^{(0)}}{\chi^2(P_{XY}^{(0)}, P_{XY}^{(1)}) + \frac{1}{n_0}V^{(0)} + \frac{1}{n_1}V^{(1)}}, \tag{4}$$

where $V^{(0)} = |\mathcal{X}||\mathcal{Y}| - 1$ and $V^{(1)} = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{P_{XY}^{(1)}(x,y)\left(1-P_{XY}^{(1)}(x,y)\right)}{P_{XY}^{(0)}(x,y)}$.

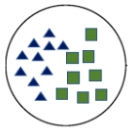
The affecting factors

- $\chi^2\left(P_{XY}^{(0)}, P_{XY}^{(1)}\right) \rightarrow$ distance!
- n_0 & $n_1 \rightarrow$ sample size!
- $|\mathcal{X}||\mathcal{Y}| - 1 \rightarrow$ Model complexity!

Consistent with our intuition

Multi-source Transfer Learning

◆ Extend to the multi-source case

- Target task:  i.i.d from $P_{XY}^{(0)} \to \hat{P}_{XY}^{(0)}$
- Source task 1:  i.i.d from $P_{XY}^{(1)} \to \hat{P}_{XY}^{(1)}$
- $\cdots$
- Source task k:  i.i.d from $P_{XY}^{(k)} \to \hat{P}_{XY}^{(k)}$

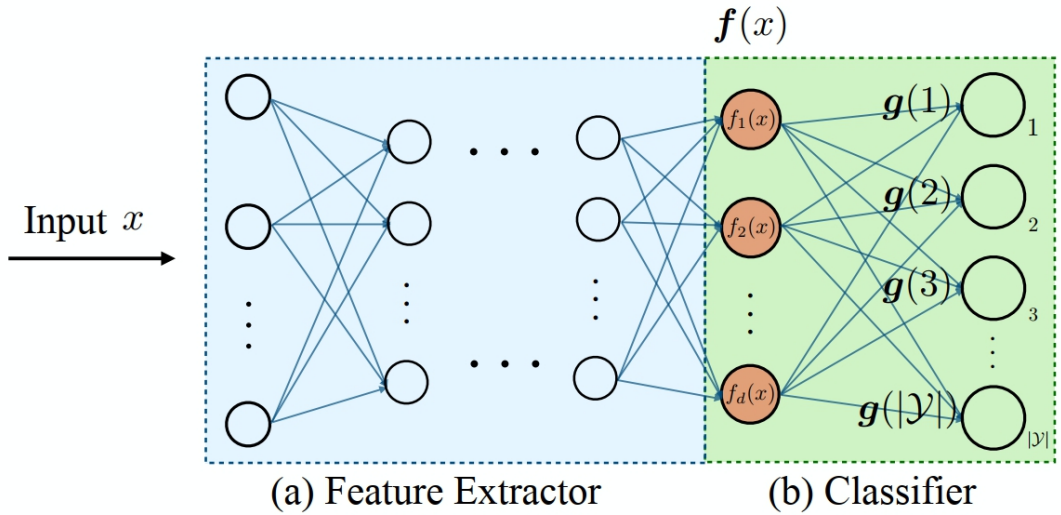
Use $\alpha_0 \hat{P}_{XY}^{(0)} + \alpha_1 \hat{P}_{XY}^{(1)} + \cdots + \alpha_k \hat{P}_{XY}^{(k)}$ to estimate $P_{XY}^{(0)}$

The testing loss becomes $L_{\text{test}} = \chi^2 \left(P_{XY}^{(0)}, \sum_{i=0}^{k} \alpha_i P_{XY}^{(i)} \right) + \sum_{i=0}^{k} \frac{\alpha_i^2}{n_i} V^{(i)}$, where $\sum_{i=0}^{k} \alpha_i = 1$

Each $\alpha_i{}^*$ can be computed by solving a non-negative quadratic programming problem.

Continuous Case

- To make the theory practical, 2 things are needed to solve
  - How the neural network models the distribution
  - How to avoid the high dimensionality $|\mathcal{X}||\mathcal{Y}|$



Discriminative Model

$$\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g})}(y|x) \triangleq P_Y^{(0)}(y)\left(1+\boldsymbol{f}^{\mathrm{T}}(x)\boldsymbol{g}(y)\right),$$

Continuous Case

When $\boldsymbol{f}$ is fixed, the classifier $\boldsymbol{g}$ can be trained with the referenced $\boldsymbol{\chi^2}$-loss

$$\hat{\boldsymbol{g}}_i = \arg\min_{\boldsymbol{g}} \chi^2_R(\hat{P}^{(i)}_{XY}, P^{(0)}_X \tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X}),$$

where $\chi^2_R(P, Q) \triangleq \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{(P(x,y)-Q(x,y))^2}{P^{(0)}_X(x)P^{(0)}_Y(y)}$.

- Target task:  $\rightarrow \hat{\boldsymbol{g}}_0 \quad (\mathbb{E}[\hat{\boldsymbol{g}}_0] = \boldsymbol{g}_0)$
- Source task 1:  $\rightarrow \hat{\boldsymbol{g}}_1 \quad (\mathbb{E}[\hat{\boldsymbol{g}}_1] = \boldsymbol{g}_1)$
- $\cdots$
- Source task k:  $\rightarrow \hat{\boldsymbol{g}}_k \quad (\mathbb{E}[\hat{\boldsymbol{g}}_k] = \boldsymbol{g}_k)$

Continuous Case

Use the combination

$$\alpha_0 \tilde{P}_{Y|X}^{(\boldsymbol{f},\hat{\boldsymbol{g}}_0)} + \alpha_1 \tilde{P}_{Y|X}^{(\boldsymbol{f},\hat{\boldsymbol{g}}_1)} + \cdots + \alpha_k \tilde{P}_{Y|X}^{(\boldsymbol{f},\hat{\boldsymbol{g}}_k)}$$

as the estimation, then the testing loss becomes

$$\begin{aligned} L_{\text{test}} = \chi_R^2 \left(P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_0)}, \sum_{i=0}^{k} \alpha_i P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_i)} \right) + \sum_{i=0}^{k} \frac{\alpha_i^2}{n_i} \tilde{V}^{(i)} \\ + \chi_R^2 \left(P_{XY}^{(0)}, P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_0)} \right) \end{aligned}$$

which is consistent with the theory in the discrete case.

Transferability

$$(\alpha_0^*, \alpha_1^*, \cdots, \alpha_k^*) = \underset{(\alpha_0^*,\alpha_1^*,\cdots,\alpha_k^*):\sum_{i=0}^{k} \alpha_i = 1}{\arg\min} L_{\text{test}}$$

Algorithm

An iterative algorithm

- $(\boldsymbol{f}, \boldsymbol{g}) \longleftarrow$ Training Loss with given $\alpha_0, \alpha_1, \cdots, \alpha_k$
- $(\alpha_0, \alpha_1, \cdots, \alpha_k) \longleftarrow$ Testing Loss with given $\boldsymbol{f}, \boldsymbol{g}$
- Until Converge

Algorithm 1 Multi-Source Knowledge Transfer Algorithm

1: **Input:** target and source data samples $\{(x_l^{(i)}, y_l^{(i)})\}_{l=1}^{n_i}$ $(i = 0, \cdots, k)$
2: Randomly initialize $\boldsymbol{\alpha}^*$
3: **repeat**
4: $(\boldsymbol{f}^*, \boldsymbol{g}^*) \leftarrow \arg\min_{\boldsymbol{f},\boldsymbol{g}} L^{(\boldsymbol{\alpha}^*, \boldsymbol{f}, \boldsymbol{g})}$
5: $\boldsymbol{\alpha}^* \leftarrow \arg\min_{\boldsymbol{\alpha} \in \mathcal{A}_k} L_{\text{test}}^{(\boldsymbol{\alpha})}$
6: **until** $\boldsymbol{\alpha}^*$ converges
7: $(\boldsymbol{f}^*, \boldsymbol{g}^*) \leftarrow \arg\min_{\boldsymbol{f},\boldsymbol{g}} L^{(\boldsymbol{\alpha}^*, \boldsymbol{f}, \boldsymbol{g})}$
8: **return** $\boldsymbol{f}^*, \boldsymbol{g}^*$

Train neural networks

$$L^{(\boldsymbol{\alpha},\boldsymbol{f},\boldsymbol{g})} \triangleq \sum_{i=0}^{k} \alpha_i \chi^2_{R_{XY}} \left(\hat{P}_{XY}^{(i)}, P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g})} \right)$$

Solve a non-negative quadratic programming problem

$$L_{\text{test}}^{(\boldsymbol{\alpha})} = \chi^2_{R_{XY}} \left(P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_0)}, \sum_{i=0}^{k} \alpha_i P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_i)} \right) + \sum_{i=0}^{k} \frac{\alpha_i^2}{n_i} \tilde{V}^{(i)} + \chi^2_{R_{XY}} \left(P_{XY}^{(0)}, P_X^{(0)} \tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_0)} \right)$$

Determine by the MAP decision rule

$$\hat{y}(x) = \arg\max_{y \in \mathcal{Y}} \tilde{P}_{Y|X}^{(\boldsymbol{f}^*, \boldsymbol{g}^*)}(y|x) = \arg\max_{y \in \mathcal{Y}} P_Y^{(0)}(y) \left(1 + \boldsymbol{f}^{*\mathrm{T}}(x) \boldsymbol{g}^*(y)\right)$$

Experiment: CIFAR-10

Binary classification task

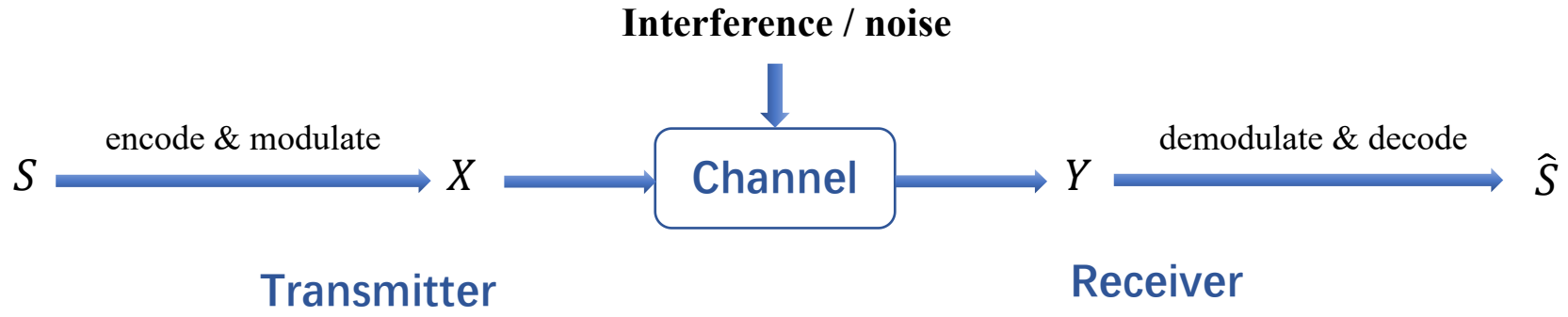
10 classes → 5 disjoint subdatasets, each contains 2 classes

choose one as the target task (task 0), and the other four as source tasks for transferring knowledge, referred to as task 1, 2, 3, 4.

Table 2: Test accuracies (%) on the target task, compared with the combining coefficients α determined by 20 rounds of random searches (RS).

Target Sample Size	6	20	100
Acc. with only target samples	70.9	74.4	81.5
Average acc. by 20 RS	67.8	73.9	75.4
Highest acc. by 20 RS	74.4	78.0	80.8
Acc. by Algorithm 1	**78.9**	**81.2**	**83.7**

Application: Signal detection in time-varying channels



◆ Signal detection task -- detect X from Y

Standard model-based approaches may be inapplicable if:

- Precise channel model is unavailable
- Optimal receiver is complicated or unknown

e.g. → Existence of co-channel intereference

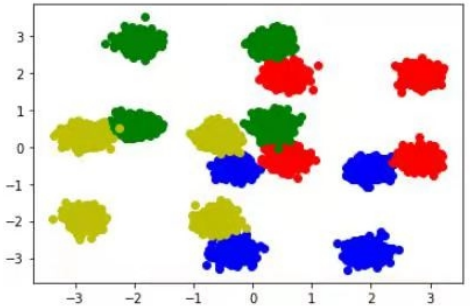
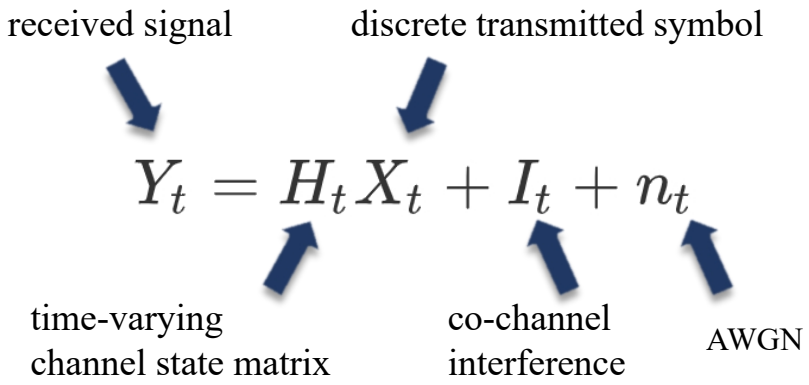


Fig 1. A constellation of transmitted QPSK-modulated signals with QPSK-modulated interference

General ways to study co-channel interference / non-Gaussian noise

- Ignore
- Treat as Gaussian noise

Application: Signal detection in time-varying channels

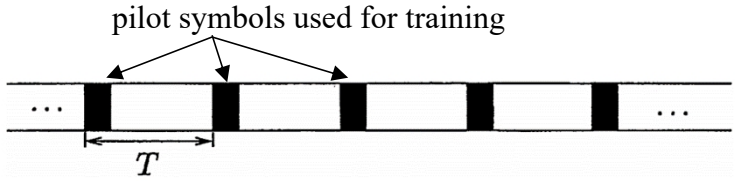


◆ Notations and Assumptions

- $t \in \mathbb{Z}$: time period / packet index
- Each period t contains many symbols to transmit, the first several ones are pilots
- $H_t,\ X_t,\ I_t,\ n_t$ are pairwise independent

◆ Goal

- Detect X_T from Y_T at the latest period $T \in \mathbb{Z}$



◆ Available training samples for the detector

- Source samples: $\left\{\left(x_t^{(i)}, y_t^{(i)}\right)\right\}_{i=1}^{N_t}$ for $t = 1,\ 2,\ \ldots,\ T-1$
- Target samples: $\left\{\left(x_T^{(i)}, y_T^{(i)}\right)\right\}_{i=1}^{N_T}$, note that $N_T \ll N_t\ (\forall t < T)$

Application: Signal detection in time-varying channels

$$Y_t = H_t X_t + I_t + n_t$$

◆ Time-varying flat Rayleigh fading channel with co-channel interference

- H_t -- Gauss-Markov process $H_t = aH_{t-1} + u_t, \quad u_t \sim \mathcal{CN}\left(0, \left(1-a^2\right)\sigma_H^2\right)$
- X_t -- QPSK-modulated signals $X_t = K_t + jQ_t, (K_t, Q_t) \in \{(1,1),(-1,1),(-1,-1),(1,-1)\}$
- I_t -- QPSK-modulated interference $I_t \in \{q+jq, -q+jq, -q-jq, q-jq\}$ for some constant q
- n_t -- AWGN with variance σ_n^2

◆ Update process of the MSTL detector

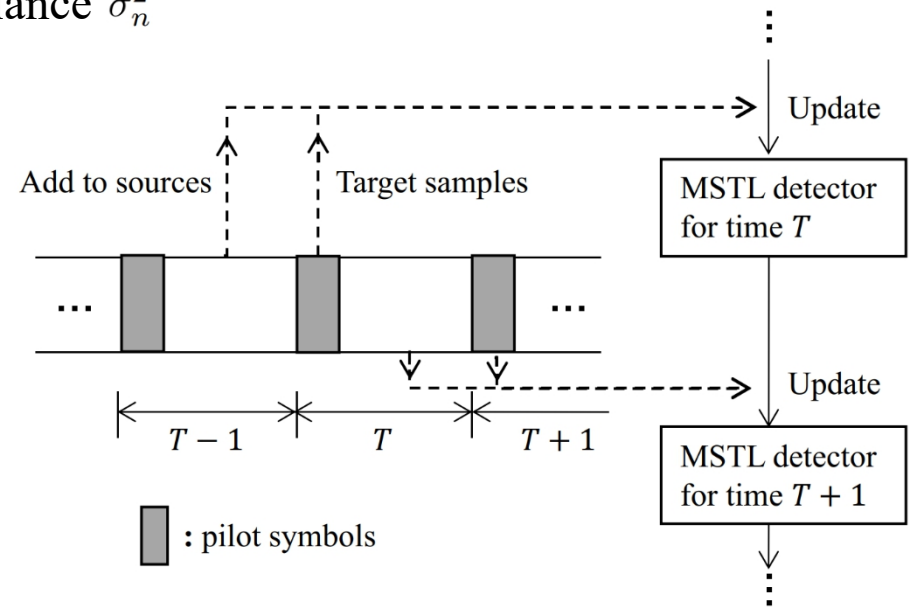


Fig 6. The update process of the detector

- Ziyan Zheng, et al. "Multi-source Transfer Learning for Signal Detection over a Fading Channel with Co-channel Interference." *IEEE International Conference on Communications* (2022).

Application: Signal detection in time-varying channels

◆ Comparisons

- FCDNN-1 Conventional fully-connected DNN based on L_2 loss, using N_T pilots for training
- FCDNN-2 Using N_T pilots and $\sum_{t=T-9}^{T-1} N_t$ reserved samples for training
- LMMSE Estimate CSI by N_T pilots

◆ Implementation

- Measure of interference level

$$\text{SINR} \triangleq 10 \lg \frac{\sigma_H^2 |X_t|^2}{\sigma_n^2}$$

  Set $q \propto \sigma_n$ for the consistency of interference and noise

- For each task time T, $N_t = 200$ for $t = T-9,\ T-8,\ \ldots,\ T-1$, discard samples before $t = T-10$

- We vary: (i) Fading coefficient a
  (ii) Number of pilots N_T
  (iii) SINR

● Ziyan Zheng, et al. "Multi-source Transfer Learning for Signal Detection over a Fading Channel with Co-channel Interference." *IEEE International Conference on Communications* (2022).

Application: Signal detection in time-varying channels

◆ Results

- Largest gain occurs in the range $a \in [0.9875,\ 0.9975]$

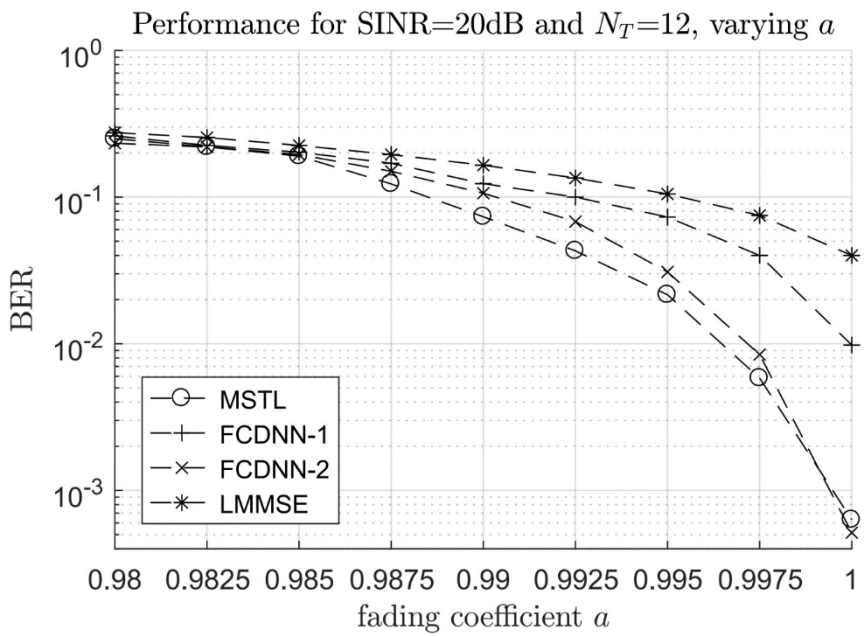


Fig 7

- Low BER for high SINR

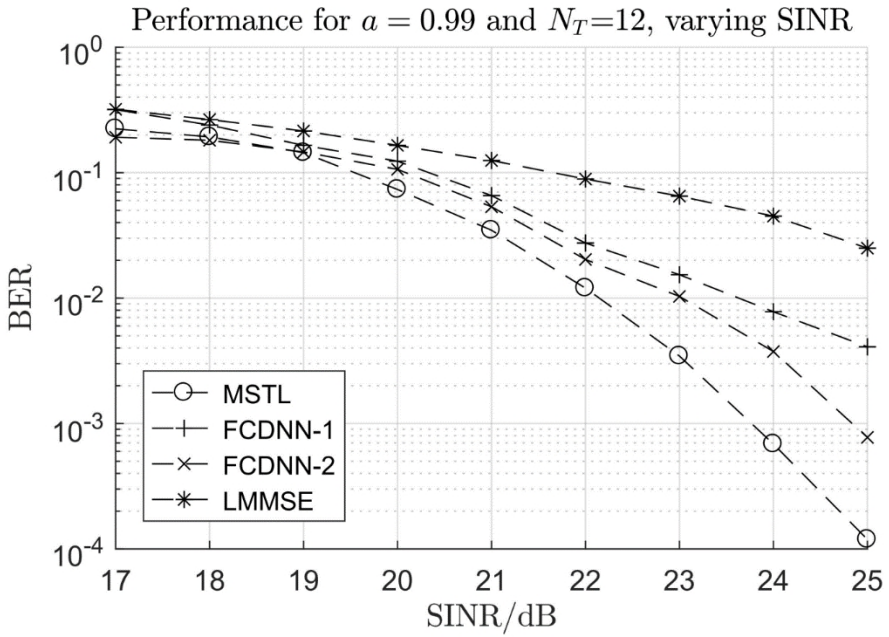


Fig 8

- Ziyan Zheng, et al. "Multi-source Transfer Learning for Signal Detection over a Fading Channel with Co-channel Interference." *IEEE International Conference on Communications* (2022).

Application: Signal detection in time-varying channels

◆ Results

- Outperforms other algorithms in a large range of pilots length

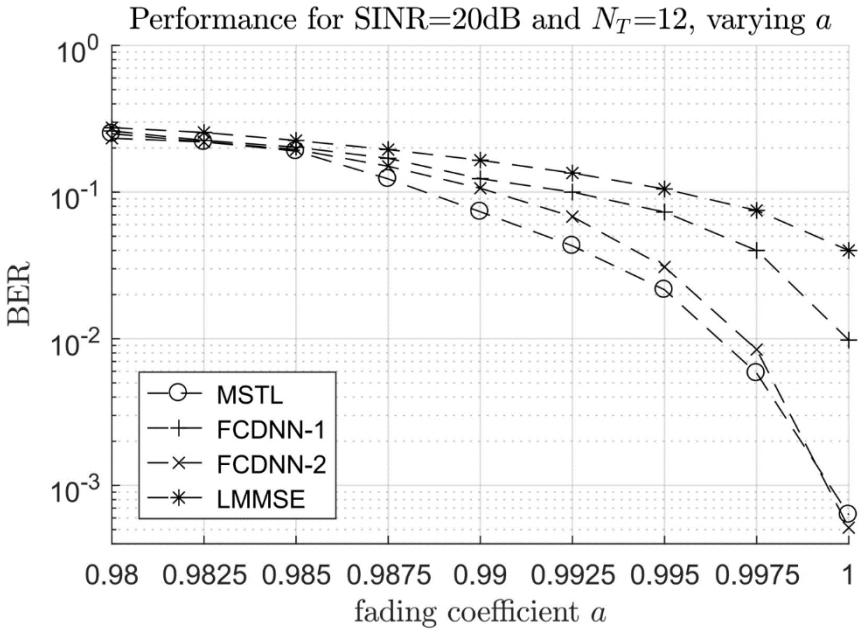


Fig 9

- Monotonous similarity of the sources

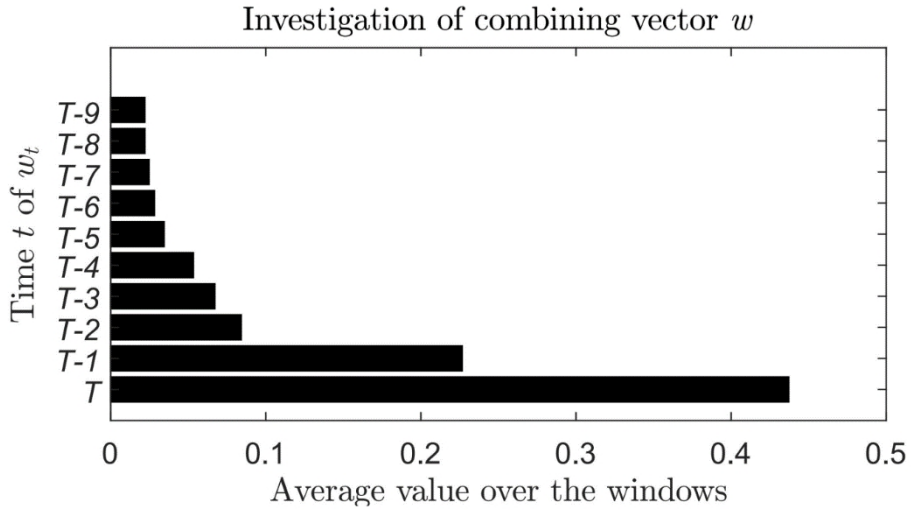


Fig 10

- Ziyan Zheng, et al. "Multi-source Transfer Learning for Signal Detection over a Fading Channel with Co-channel Interference." *IEEE International Conference on Communications* (2022).

Conclusion

- Construct a theoretical framework for multi-source transferability covering distance, sample sizes, and task complexity at the same time
- Extend to continuous data via training neural networks
- A consistent algorithm that works in applications

● Ziyan Zheng, et al. "Multi-source Transfer Learning for Signal Detection over a Fading Channel with Co-channel Interference." *IEEE International Conference on Communications* (2022).

Thanks !