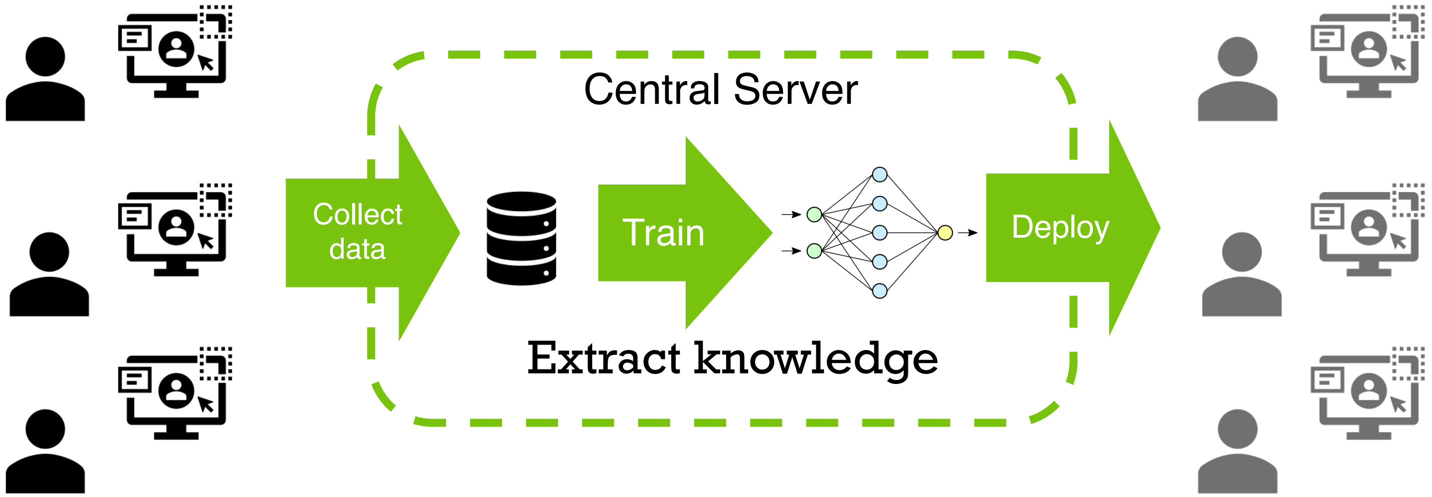
Federated Adversarial Debiasing for Fair and Transferable Representations

Junyuan Hong

CSE Graduate Seminar, MSU

Oct 7, 2021

Centralized Learning



How to transfer knowledge across domains w/o exchanging domain data

What are domains?

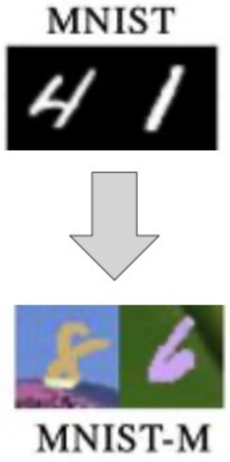
What knowledge to transfer?

Why not exchange data?

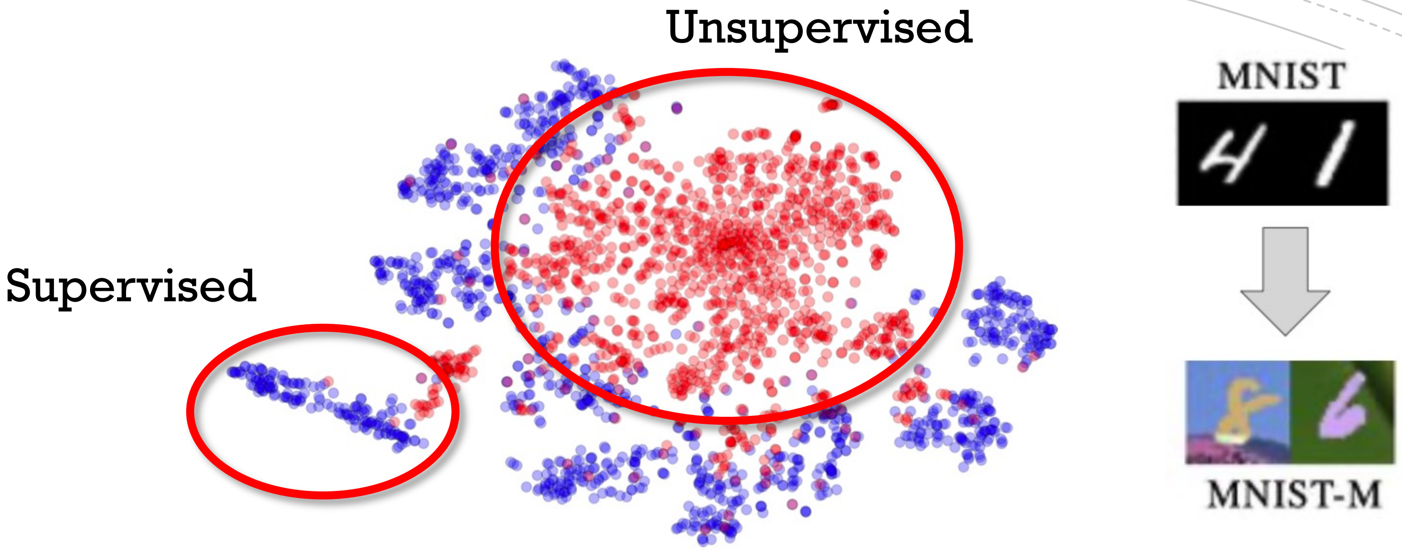
How?

What are domains?

- Key: Distributional shift
- Examples:
  - Data from different social groups
    - Genders, races
  - Data from different sensors
    - Webcam v.s. pro. cam
    - Grey-scale v.s. color images



What knowledge to transfer?



Representation bias: gray-scale v.s. color digit images (MNIST and MNIST-M) extracted by CNN models.
Credit: Ganin, Y., & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. *ICML*

- Supervision
  - Lack of labels -> non-adapted
  - Lack of data -> unfair

Why not exchange users' data?

- Privacy
  - *"Though it (GDPR) was drafted and passed by the European Union (EU), it imposes obligations onto organizations anywhere, so long as they target or collect data related to people in the EU."*
  - General Data Protection Regulation (GDPR) since May 25, 2018
  - https://gdpr.eu/what-is-gdpr/

How to transfer knowledge across domains w/o exchanging domain data?

- Reduce domain/distributional gap
  - Exchange data for gap-aware training
- Transfer the knowledge of domain gap instead of data
  - Without exchange data

How to transfer knowledge across domains w/o exchanging domain data

What are domains? -> Distributional shift

What knowledge? -> Supervision, etc.

Why not exchange data? -> Privacy

How? -> Reduce gap by sharing gap knowledge

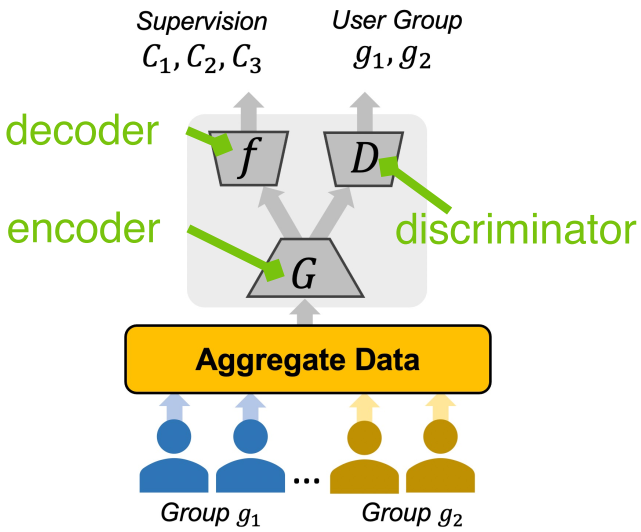
Revisit: Reduce gap by adversarial debiasing

- Extract representations $z = G(x)$ from two groups. Thus, $z \sim p_1$ or $z \sim p_2$
- Measure the group discrepancy (domain gap):

$$\mathbf{D}_{p_1,p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))],$$

- Update encoder to reduce domain gaps

$$G = \arg\min_G \mathrm{D}_{p_1,p_2}$$



Central methods debias aggregated raw data (Ganin, et al. 2015)

Ganin, Y., & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. *ICML*

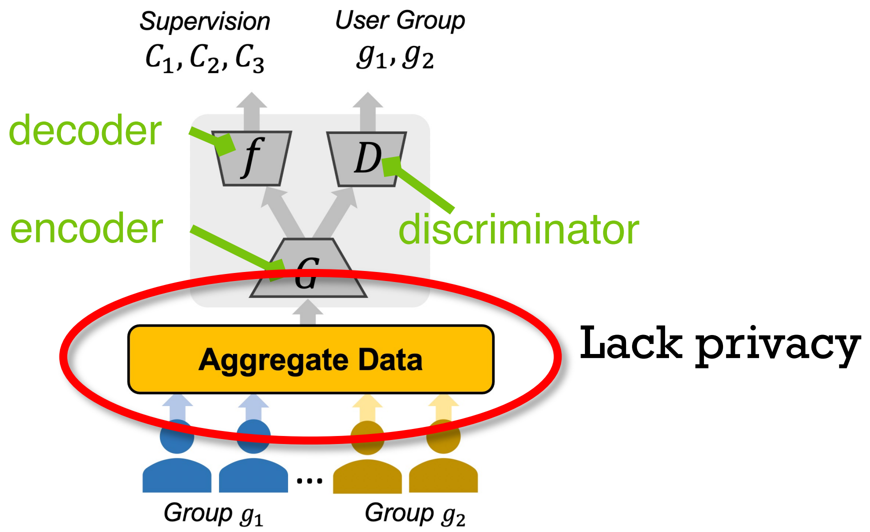
Revisit: Reduce gap by adversarial debiasing

- Extract representations $z = G(x)$ from two groups. Thus, $z \sim p_1$ or $z \sim p_2$
- Measure the group discrepancy:

$$\mathbf{D}_{p_1,p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))],$$

- Update encoder to reduce domain gaps

$$G = \arg\min_G \mathrm{D}_{p_1,p_2}$$



Central methods debias aggregated raw data (Ganin, et al. 2015)

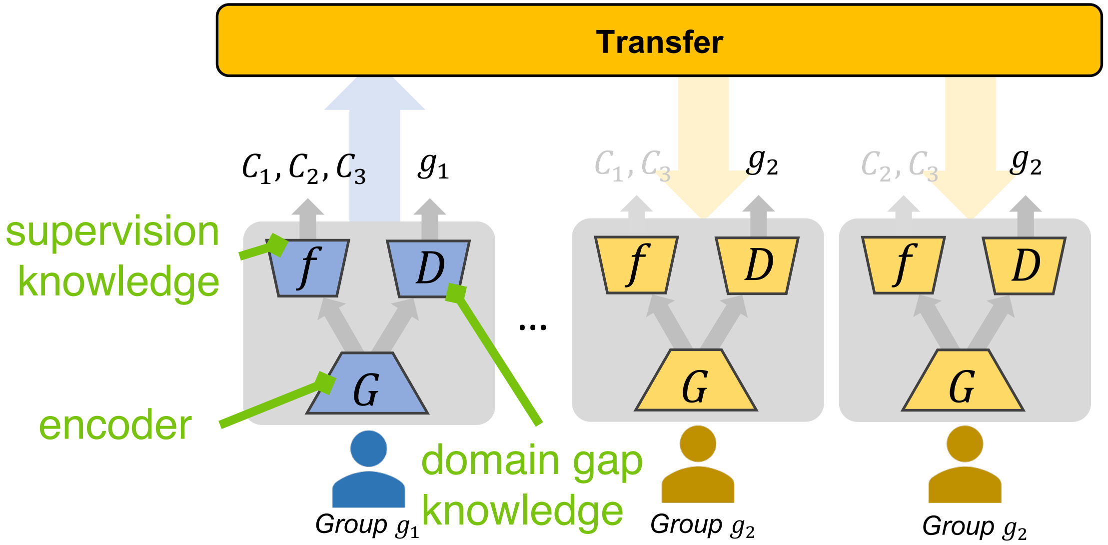
Ganin, Y., & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. *ICML*

Adversarial Debiasing w/o exchanging data

- **Transfer knowledge by models instead of data**

$$G = \arg\min_{G} \mathrm{D}_{p_1, p_2}$$

- **Privacy**: Each user trains discriminators using local data only and encoders are supervised by shared discriminators.



Hong, J., Zhu, Z., Yu, S., Wang, Z., Dodge, H. H., & Zhou, J. (2021). Federated Adversarial Debiasing for Fair and Transferable Representations. *KDD*

Adversarial Debiasing w/o exchanging data

- **Transfer knowledge by models instead of data**

$$G = \arg\min_G \mathrm{D}_{p_1,p_2}$$

Locally learn gap knowledge w/o adversarial data

user 1 (group 1) $\mathbf{D}_{p_1,p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))]$

user 2 (group 2) $\mathbf{D}_{p_1,p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))]$

Missing adversary's information

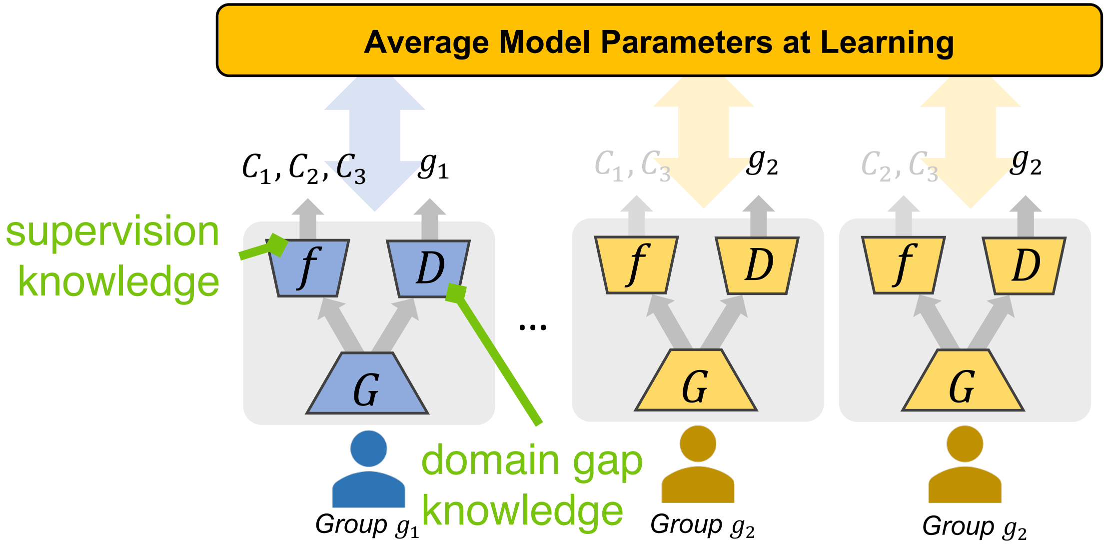
Ganin, Y., & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. *ICML*

Federated Adversarial Debiasing (**FADE**) w/o exchanging data

- **Transfer knowledge by models instead of data**

$$G = \arg\min_G \mathrm{D}_{p_1, p_2}$$

- **Federated learning**: Frequently exchange knowledge during learning



Local discrepancy w/o adversarial data

user 1 (group 1) $\mathbf{D}_{p_1, p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \cancel{\mathbb{E}_{p_2}[\log(1 - D(z))]}$

user 2 (group 2) $\mathbf{D}_{p_1, p_2} = \max_D \cancel{\mathbb{E}_{p_1}[\log D(z)]} + \mathbb{E}_{p_2}[\log(1 - D(z))]$

average models frequently

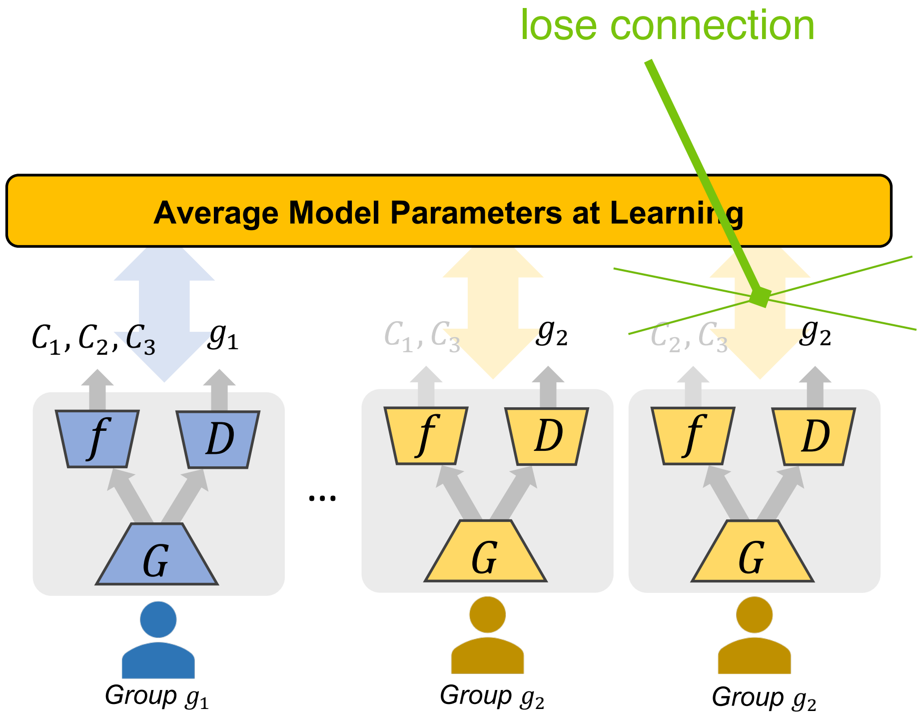
Global discrepancy

$$\mathbf{D}_{p_1, p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))]$$

Hong, J., Zhu, Z., Yu, S., Wang, Z., Dodge, H. H., & Zhou, J. (2021). Federated Adversarial Debiasing for Fair and Transferable Representations. *KDD*

Federated Adversarial Debiasing (**FADE**) w/o exchanging data

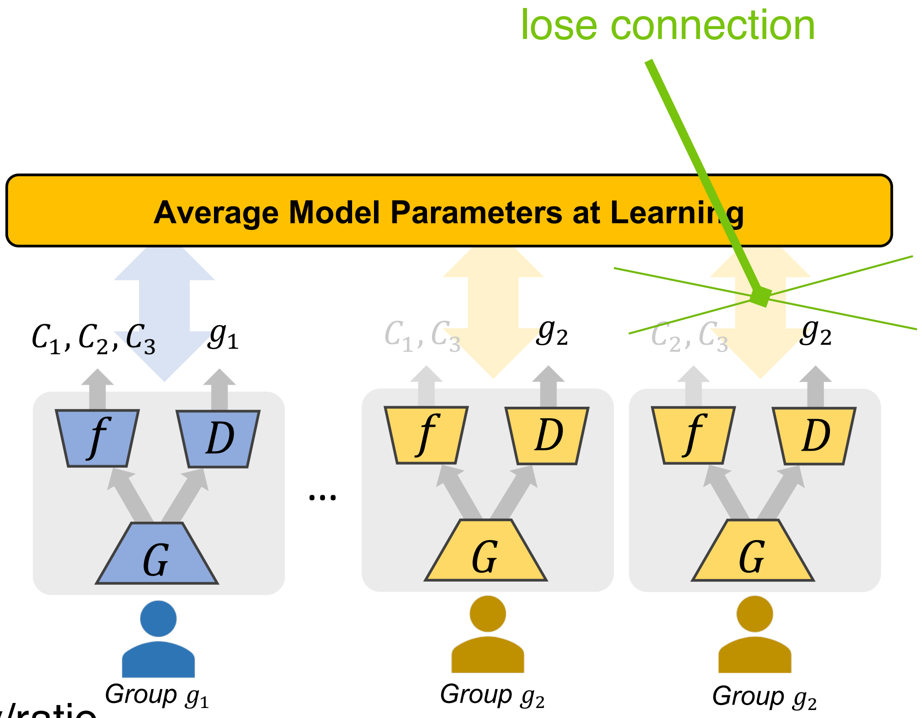
- Autonomous: Users are allowed not to upload their local models per iteration, due to
  - slow network connection
  - temporarily limited computation budgets
- A lot of uncertainty



Hong, J., Zhu, Z., Yu, S., Wang, Z., Dodge, H. H., & Zhou, J. (2021). Federated Adversarial Debiasing for Fair and Transferable Representations. *KDD*

Federated Adversarial Debiasing (**FADE**) w/o exchanging data

- Autonomous: Users are allowed not to upload their local models per iteration.
- Model the uncertainty



Discrepancy w/o losing connections

$$\mathbf{D}_{p_1,p_2} = \max_D \boxed{\mathbb{E}_{p_1}[\log D(z)]} + \boxed{\mathbb{E}_{p_2}[\log(1 - D(z))]},$$

α_1 α_2

Uploading probability/ratio from group 2

Global discrepancy w/ unstable connections

$$\tilde{\mathbf{D}}_{p_1,p_2} = \max_D \alpha_1 \mathbb{E}_{p_1}[\log D(z)] + \alpha_2 \mathbb{E}_{p_2}[\log(1 - D(z))]$$

Hong, J., Zhu, Z., Yu, S., Wang, Z., Dodge, H. H., & Zhou, J. (2021). Federated Adversarial Debiasing for Fair and Transferable Representations. *KDD*

How well FADE transfers?

- Minimize domain gap to transfer supervision knowledge:

$$G = \arg\min_{G} \mathrm{D}_{p_1, p_2}$$

- Estimated domain gap (discrepancy)

$$\tilde{\mathbf{D}}_{p_1, p_2} = \max_{D} \alpha_1 \mathbb{E}_{p_1}[\log D(z)] + \alpha_2 \mathbb{E}_{p_2}[\log(1 - D(z))]$$

- General case

Theorem 4.1. *The condition* $p_1(z) = p_2(z)$ *is a sufficient condition for minimizing* $\tilde{\mathbf{D}}_{p_1,p_2}$ *and the minimal value is* $\alpha_1 \log \alpha_1 + \alpha_2 \log \alpha_2 + (\alpha_1 + \alpha_2)\log(\alpha_1 + \alpha_2)$.

~~How well FADE transfers?~~
How well domain gap knowledge is transferred?

- Estimated domain gap (discrepancy)

$$\tilde{\mathbf{D}}_{p_1,p_2} = \max_D \alpha_1 \mathbb{E}_{p_1}[\log D(z)] + \alpha_2 \mathbb{E}_{p_2}[\log(1 - D(z))]$$

- General case

Theorem 4.1. *The condition $p_1(z) = p_2(z)$ is a sufficient condition for minimizing $\tilde{\mathbf{D}}_{p_1,p_2}$ and the minimal value is $\alpha_1 \log \alpha_1 + \alpha_2 \log \alpha_2 + (\alpha_1 + \alpha_2) \log(\alpha_1 + \alpha_2)$.*

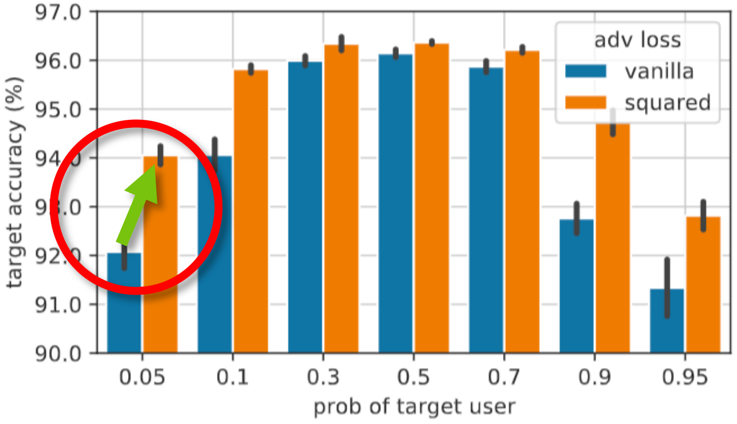
- Unbalanced case

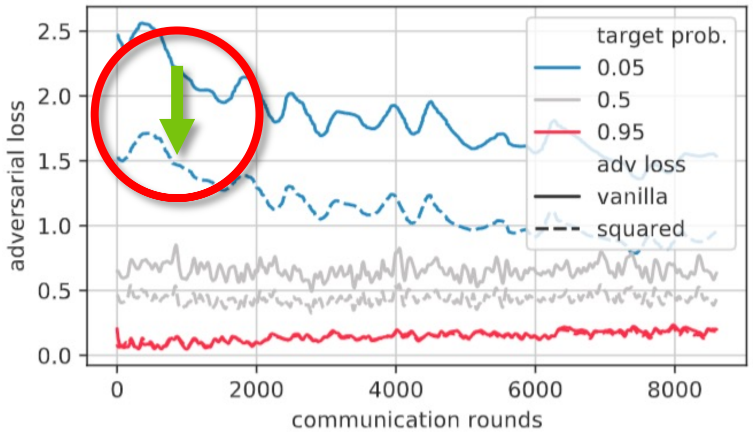
Theorem 4.2. *Let ϵ be a positive constant. Suppose $|\log p_1(x) - \log p_2(x)| \leq \epsilon$ for any x in the support of p_1 and p_2. Then we have $\tilde{\mathbf{D}}_{p_1,p_2} = O(\alpha_1 \epsilon / (\alpha_1 + \alpha_2))$ when $\alpha_1 \ll \alpha_2$.*

  - More unbalanced users are, more biased gap knowledge is
  - Mitigate imbalance by scaling up large loss

$$L_{i,g,2}^{\text{adv}}(D, G) = -\frac{1}{2}\left(L_{i,g}^{\text{adv}}(G, D)\right)^2,$$

Transfer supervision knowledge w/ imbalanced groups





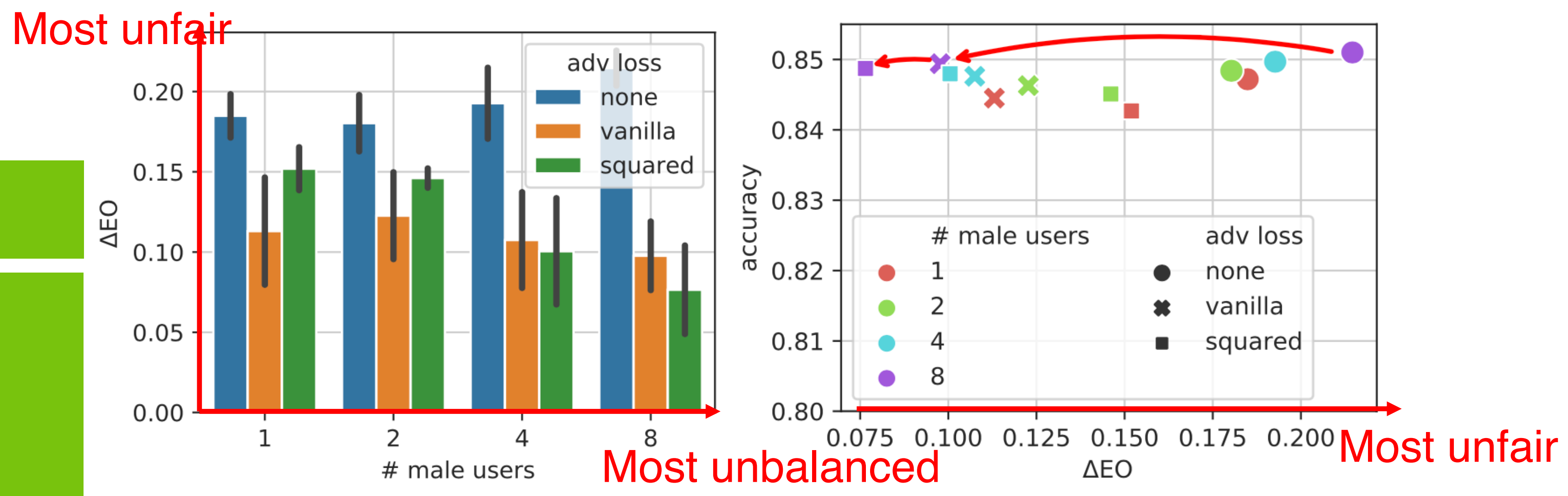
- From supervised USPS domain to MNIST domain
- Squared loss improves the adversarial loss (gap knowledge)

Transfer supervision knowledge across domains

Method	A→D	A→W	D→A	D→W	W→A	W→D	Re→Ar	Re→Cl	Re→Pr	Avg.
Federated methods										
Source only	79.5	73.4	59.6	91.6	58.2	95.8	67.0	46.5	78.2	72.2
non-iid target users w/ 20 (Office) or 45 (OfficeHome) classes per user										
FADE-DANN	85.4 (1.9)	81.8 (1.8)	43.1 (33)	97.7 (0.5)	64.8 (0.5)	99.7 (0.2)	46.4 (37)	34.9 (27)	78.8 (0.1)	70.3
FADE-CDAN	**92.3 (1.2)**	**91.6 (0.5)**	**65.9 (9.3)**	**98.9 (0.2)**	**70.2 (0.8)**	**99.9 (0.1)**	70.3 (1.6)	54.9 (4.6)	**82.2 (0.1)**	80.7
FedAvg-SHOT	83.6 (0.5)	83.1 (0.5)	64.7 (1.4)	91.7 (0.2)	64.7 (2.2)	97.4 (0.5)	**70.7 (0.5)**	**55.4 (0.5)**	80.1 (0.3)	76.8
iid target users										
FADE-DANN	84.2 (1.5)	81.3 (0.4)	66.3 (0.3)	97.5 (1.2)	59.4 (10.6)	99.9 (0.2)	67.3 (0.9)	51.3 (0.4)	79.0 (0.6)	76.2
FADE-CDAN	**93.6 (0.8)**	**92.2 (1.3)**	**71.2 (1.0)**	**98.7 (0.4)**	**71.3 (0.7)**	**100 (0.0)**	70.6 (1.3)	55.1 (1.0)	**82.3 (0.2)**	81.7
FedAvg-SHOT	**96.3 (0.5)**	**94.3 (1.1)**	70.9 (2.0)	98.4 (0.4)	**72.7 (0.9)**	99.8 (0.0)	**74.8 (0.3)**	**60.0 (0.1)**	**84.9 (0.2)**	83.6
Central methods										
ResNet [15]	68.9	68.4	62.5	96.7	60.7	99.3	53.9	41.2	59.9	67.9
Source only [23]	80.8	76.9	60.3	95.3	63.6	98.7	65.3	45.4	78.0	73.8
DANN [11]	79.7	82.0	68.2	96.9	67.4	99.1	63.2	51.8	76.8	76.1
CDAN [28]	92.9	**94.1**	71.0	**98.6**	69.3	**100**	70.9	56.7	81.6	81.7
SHOT [23]	**94.0**	90.1	**74.7**	98.4	**74.3**	99.9	**73.3**	**58.8**	**84.3**	**83.1**

- Variants of FADE outperforms the state-of-the-art source-data-free transfer learning (SHOT) on non-iid target users.

Fair learning with imbalanced female/male users



Adult dataset with fairness on male/female groups

Qualitative comparison

Property	Central	FADE (ours)
Data privacy	❌ (raw data)	✓
Autonomous users	❌	✓
Satisfiable optimization	✓	✓ (Theoretic & empirical)

What knowledge to transfer?

- Supervision
  - Lack of labels -> non-adapted
  - Lack of data -> unfair
  - Hong, J., Zhu, Z., Yu, S., Wang, Z., Dodge, H. H., & Zhou, J. (2021). Federated Adversarial Debiasing for Fair and Transferable Representations. *KDD*
- Robustness
  - Lack of computation resource -> inability of adv. augmentation
  - Hong, J., Wang, H., Wang, Z., & Zhou, J. (2021). Federated Robustness Propagation: Sharing Adversarial Robustness in Federated Learning. *arXiv:2106.10196*.
- Class features
  - Non-iid class distribution in users -> missing class features
  - Zhu, Z., Hong, J., & Zhou, J. (2021). Data-Free Knowledge Distillation for Heterogeneous Federated Learning. ICML

Thank You!

w/ Zhuangdi Zhu, Shuyang Yu, Zhangyang Wang, Hiroko H. Dodge, & Jiayu Zhou
(2021). Federated Adversarial Debiasing for Fair and Transferable Representations. *KDD*

Code

https://github.com/illidanlab/FADE

Visit at poster session

Acknowledgement

This material is based in part upon work supported by the National Science Foundation under Grant IIS-1749940, EPCN-2053272, Office of Naval Research N00014-20-1-2382, and National Institute on Aging (NIA) R01AG051628, R01AG056102, P30AG066518, P30AG024978, RF1AG072449.