Federated Adversarial Debiasing for Fair and Transferable Representations

Junyuan Hong
CSE Graduate Seminar, MSU
Oct 7, 2021
Centralized Learning

- Collect data
- Train
- Extract knowledge
- Deploy

Central Server
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?

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

*Representation bias*: gray-scale vs. color digit images (MNIST and MNIST-M) extracted by CNN models.

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/](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):
  $D_{p_1,p_2} = \max_D \mathbb{E}_{p_1} [\log D(z)] + \mathbb{E}_{p_2} [\log (1 - D(z))]$
- Update encoder to reduce domain gaps
  $G = \arg \min_G D_{p_1,p_2}$

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

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:
  \[ D_{p_1,p_2} = \max_D \mathbb{E}_{p_1}[\log D(z)] + \mathbb{E}_{p_2}[\log(1 - D(z))], \]
- Update encoder to reduce domain gaps
  \[ G = \arg \min_G D_{p_1,p_2} \]

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

Adversarial Debiasing w/o exchanging data

- **Transfer knowledge by models instead of data**
  \[ G = \arg \min_G D_{p_1, p_2} \]
- **Privacy**: Each user trains discriminators using local data only and encoders are supervised by shared discriminators.

Adversarial Debiasing w/o exchanging data

- Transfer knowledge by models instead of data

\[ G = \arg \min_G D_{p_1, p_2} \]

Locally learn gap knowledge w/o adversarial data

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

Missing adversary’s information

Federated Adversarial Debiasing (FADE) w/o exchanging data

- Transfer knowledge by models instead of data
  \[ G = \arg \min_G D_{p_1,p_2} \]
- Federated learning: Frequently exchange knowledge during learning

Local discrepancy w/o adversarial data

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

Global discrepancy

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

---

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

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

Global discrepancy w/ unstable connections

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

Minimize domain gap to transfer supervision knowledge:

\[ G = \arg \min_G D_{p_1, p_2} \]

Estimated domain gap (discrepancy)

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

General case

**Theorem 4.1.** The condition \( p_1(z) = p_2(z) \) is a sufficient condition for minimizing \( \tilde{D}_{p_1, p_2} \) 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)
  \[ \tilde{D}_{p_1,p_2} = \max_D \alpha_1 \mathbb{E}_{p_1}[\log D(z)] + \alpha_2 \mathbb{E}_{p_2}[\log(1 - D(z))] \]

- General case

  **Theorem 4.1.** The condition \( p_1(z) = p_2(z) \) is a sufficient condition for minimizing \( \tilde{D}_{p_1,p_2} \) 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 \( \epsilon \) 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
  \[ \tilde{D}_{p_1,p_2} = O(\alpha_1 \epsilon / (\alpha_1 + \alpha_2)) \text{ when } \alpha_1 \ll \alpha_2. \]

- More unbalanced users are, more biased gap knowledge is transferred.
- Mitigate imbalance by scaling up large loss
  \[ L_{\text{adv}}^{\text{adv}}(D,G) = -\frac{1}{2} \left( \frac{1}{N} \mathbb{E}_{x \sim D} \left[ \text{adv}(G(z \cdot x)) \right] \right)^2, \]
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

- 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

<table>
<thead>
<tr>
<th>Property</th>
<th>Central</th>
<th>FADE (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data privacy</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>(raw data)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autonomous users</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Satisfiable optimization</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>(Theoretic &amp; empirical)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What knowledge to transfer?

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

- **Robustness**
  - Lack of computation resource -> inability of adv. augmentation

- **Class features**
  - Non-iid class distribution in users -> missing class features
Thank You!

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

**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.

https://github.com/illidanlab/FADE