Towards Trustworthy AI with Perspectives on Fairness, Interpretability, Privacy, and Security

2023-05-18

Sungwon Han

Microsoft KAIST ibs 데이터사이언스그룹 Data Science Group

Trustworthy AI

Trustworthy **“Deep Learning”**

⇒ Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers.

How do deep neural networks train and make inference?

It varies depending on the task… Let's consider a supervised, image classification problem.
We aim to learn a mapping function f that takes an image x and produces a label y.

$$f_W$$

Parameterized by weights (W)

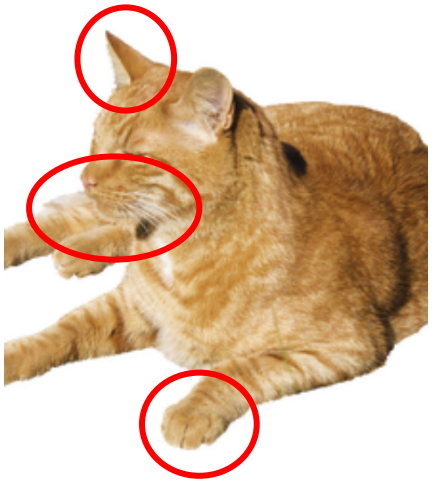
$$y$$

Image
(input)

Category label
(output)

How is the prediction generated?

There are various features that can be used to distinguish the image!
Which features would deep learning model select for the prediction?



Ears? Whiskers? Paw?

Texture?

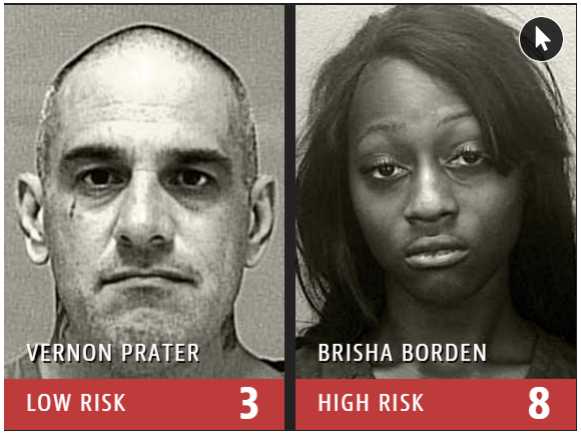
How about this?

Image Credit: ImageNet-trained CNNs are biased towards texture (ICLR 2019)

AI with Deep Learning

Consider another example..

AI model is trying to predict the recidivism risk from given individuals.
Which features would deep learning model select for the prediction?



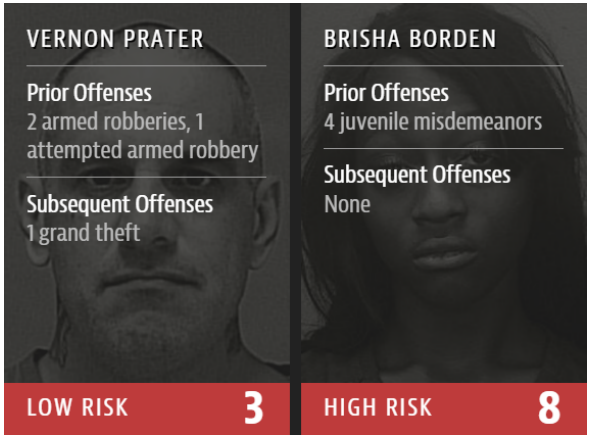


Image Credit: Pro Publica - Machine Bias

“Trustworthy” AI

⇒ Describe AI that is lawful, ethically adherent, and technically robust.

Trustworthy AI

Increasing public interest on trustworthy AI

- Domains in high-stakes decisions are already using AI models, bringing real risks. How do we ensure safety is built into these systems?

- High-risk domains include those related to safety infra and products, education, employment, justice, immigration, and climate protection (as defined in EU AI Act proposal, Article 6-51).

“

On artificial intelligence,
trust is a must, not a nice to have.

”

Requirements of Trustworthy AI

Ethics Guidelines for AI from European Commission

Interrelationship of the seven requirements:

- All are of equal importance and should be implemented and evaluated throughout the AI system's lifecycle
- For some applications, they some requirements may be of lesser or greater relevance.

Index

Ethics Guidelines for AI from European Commission

1. Human agency and oversight
2. Diversity, non-discrimination
3. Transparency
4. Privacy and data governance
5. Technical robustness, security
6. Societal and environmental wellbeing
7. Accountability

Index

In today's presentation…

1. Human agency and oversight
2. **Diversity, non-discrimination**
3. **Transparency**
4. **Privacy and data governance**

➔ Sec. 1) Algorithmic fairness (WWW'23)

➔ Sec. 2) Interpretability (NeurIPS'22)

➔ Sec. 3) Federated learning (ECCV'22)

DualFair: Fair Representation Learning at Both Group and Individual Levels via Contrastive Self-supervision (WWW' 2023)

Introduction

Fairness in machine learning

Why we care about fairness?

*Recidivism rate

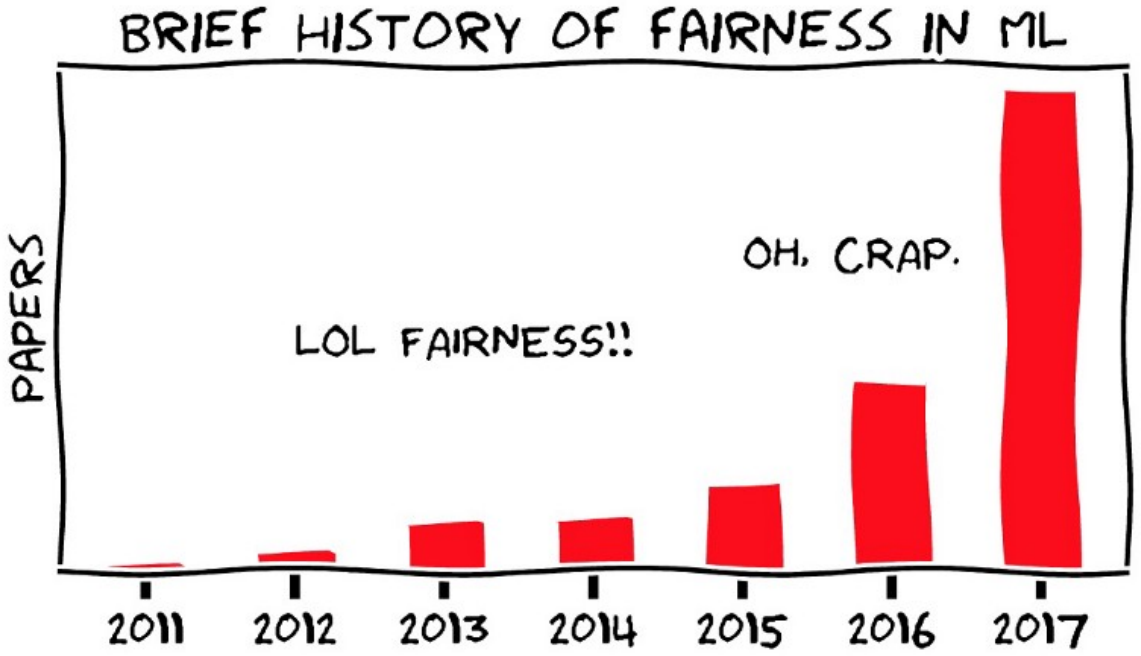
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

(Source: ProPublica analysis of data from Broward County, Fla.)

Introduction

Fairness in machine learning

Why we care about fairness?



cidivism rate

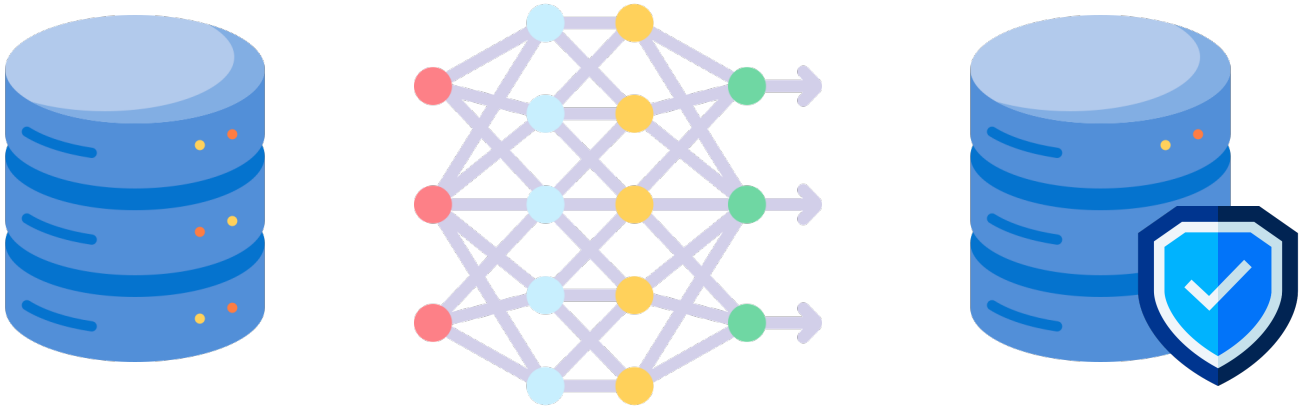
	MERICAN
Labeled Higher	44.9%
Labeled Lower	28.0%

ard County, Fla.)

Background

Fair representation learning

Debias sensitive information and generate low-dimensional representation



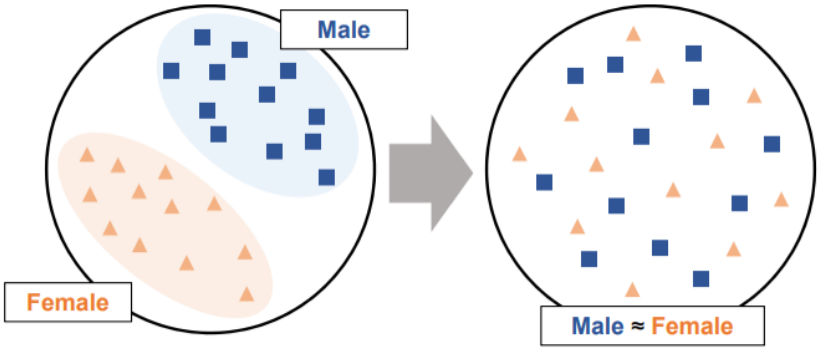
Private dataset
with Sensitive information

Public dataset
without Sensitive information

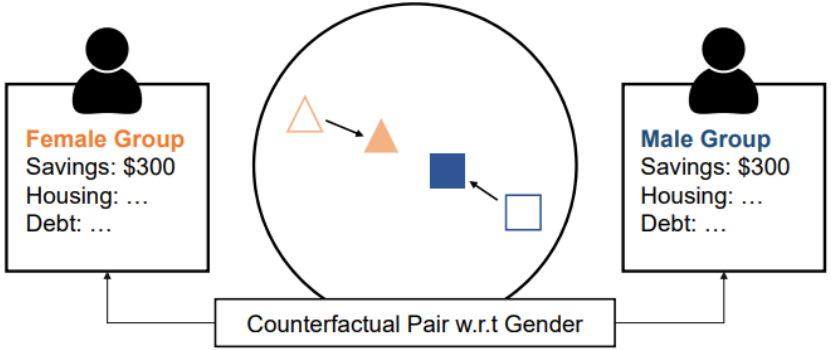
Motivation: Two fairness criteria

Fairness should be achieved at both group and individual-level

Group fairness vs. Counterfactual fairness in representation learning



Group fairness: group-level fairness
Ex) Groups are indistinguishable in embeddings



Counterfactual fairness: individual-level fairness
Ex) Counterfactual pair from groups should be close

Motivation: Contrastive learning

Contrastive learning for representation learning

Generalized-InfoNCE objective can be decomposed into two terms:

$$L_c(\mathbf{x}, \mathcal{X}_+, \mathcal{X}_-) = -\frac{1}{|\mathcal{X}_+|}\Big(\underbrace{\log \sum_{\mathbf{x}' \in \mathcal{X}_+} \exp\left(\mathrm{sim}(f(\mathbf{x}), f(\mathbf{x}'))/\tau\right)}_{\text{Alignment loss}} \underbrace{- \log \sum_{\mathbf{x}' \in (\mathcal{X}_+ \cup \mathcal{X}_-)} \exp(\mathrm{sim}(f(\mathbf{x}), f(\mathbf{x}'))/\tau)}_{\text{Distribution loss}}\Big)$$

- Alignment loss encourages the embeddings of positive pairs to be placed closer.
- Distribution loss matches all instances' embeddings into the prior with high entropy.

$$L_{\text{gen-}c}(\mathbf{x}, \mathbf{x}_+, \mathcal{X}_-) = -L_{\text{align}}(\mathbf{x}, \mathbf{x}_+) + \mathrm{SWD}(\tilde{\mathcal{Z}}, \mathcal{Z}_{prior})$$

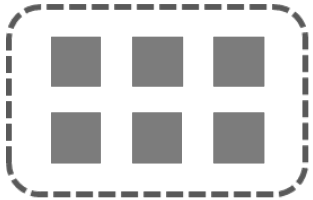
Main idea: Fair contrastive learning

Contrastive objective for both group & counterfactual fairness

Treat counterfactual pair alike and ensure non-distinguishable embeddings among groups

Maximize Agreement

Distribution Matching



$\sim \mathcal{Z}_{prior}$

■ : Input sample (Anchor)

▲ : Counterfactual pair of input (Positive)

■ : Random sample from the same group with Anchor (Negative)

$L = -L_{\text{align}}(\blacksquare, \blacktriangle) + \text{SWD}(\blacksquare, \mathcal{Z}_{prior})$

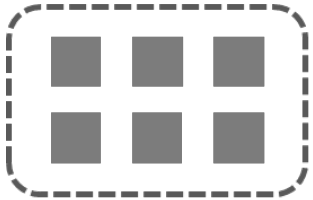
Main idea: Fair contrastive learning

Contrastive objective for both group & counterfactual fairness

Treat counterfactual pair alike and ensure non-distinguishable embeddings among groups

Maximize Agreement

Distribution Matching



$\sim \mathcal{Z}_{prior}$

■ : Input sample (Anchor)

▲ : Counterfactual pair of input (Positive)

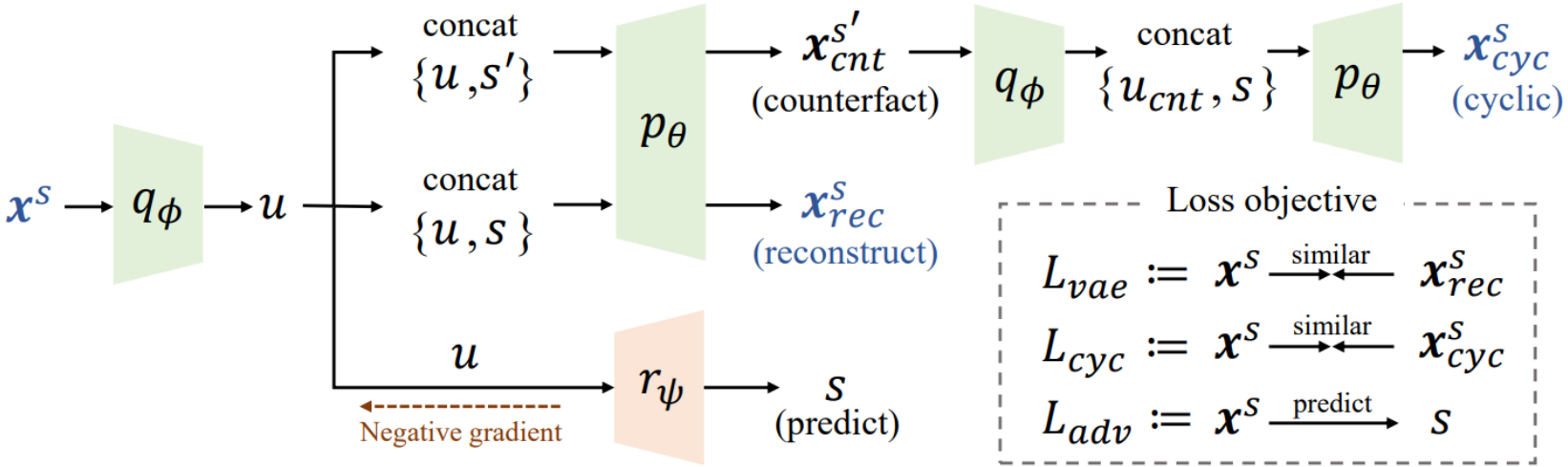
■ : Random sample from the same group with Anchor (Negative)

$$L = -L_{\text{align}}(\blacksquare, \blacktriangle) + \text{SWD}(\blacksquare, \mathcal{Z}_{prior})$$

Main idea: C-VAE

Generation of counterfactual samples with C-VAE

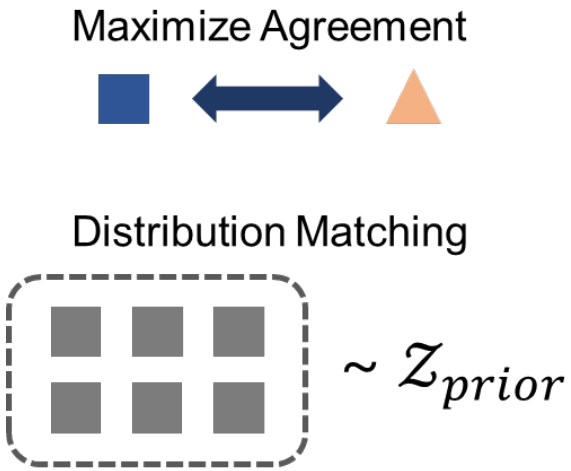
Variational autoencoder with adversarial training for counterfactual sample generator



Main idea: Fair contrastive learning

Contrastive objective for both group & counterfactual fairness

Treat counterfactual pair alike and ensure non-distinguishable embeddings among groups



■ : Input sample (Anchor)

▲ : Counterfactual pair of input (Positive)

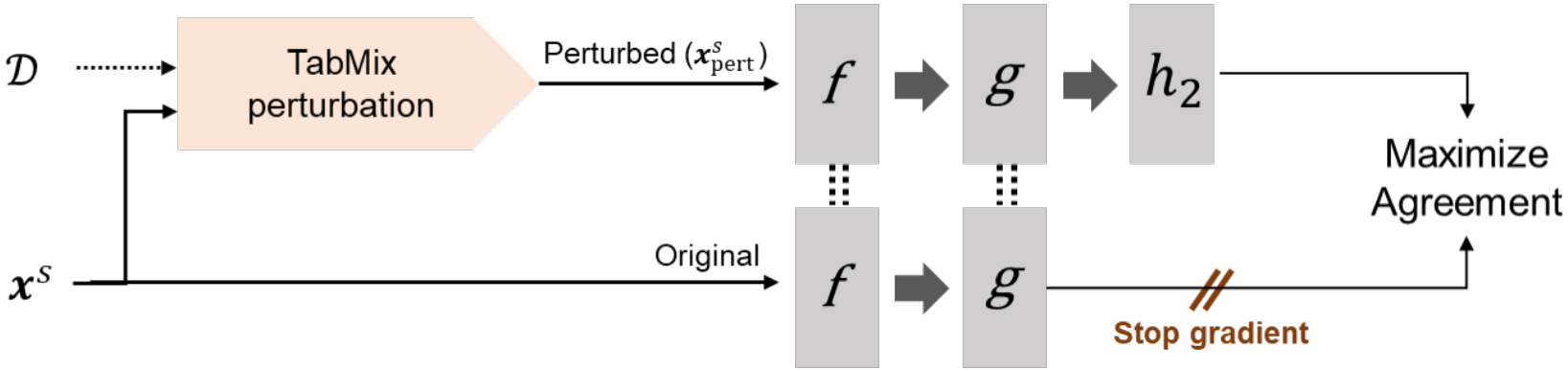
■ : Random sample from the same group with Anchor (Negative)

$$L = -L_{\text{align}}(\blacksquare, \blacktriangle) + \text{SWD}(\blacksquare, \mathcal{Z}_{prior})$$

Main idea: Self-knowledge distillation

Self-knowledge distillation to maintain representation quality

Reduce the discrepancy between original & perturbations to learn data semantics



$$L_{\text{total}} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}^s \in \mathcal{D}} (L_{\text{fair-cl}}(\mathbf{x}^s) + L_{\text{self-kd}}(\mathbf{x}^s))$$

Experiments

Datasets

Six fairness-required datasets with various kinds of downstream tasks for evaluation

Dataset	# samples	# attr.	Sensitive attr.	Split	Task
Adult	48,842	14	gender, race	2:1	classification
Credit	1,000	20	gender	4:1	classification
COMPAS	6,172	7	gender, race	4:1	classification
LSAC	22,407	12	gender, race	4:1	classification
Students	649	33	gender	4:1	regression
Communities	1,994	128	race	4:1	regression

Experiments

Performance evaluation

Performance comparison summaries among fairness-aware baselines and DualFair. Averaged rank for each evaluation metric across six datasets is reported.

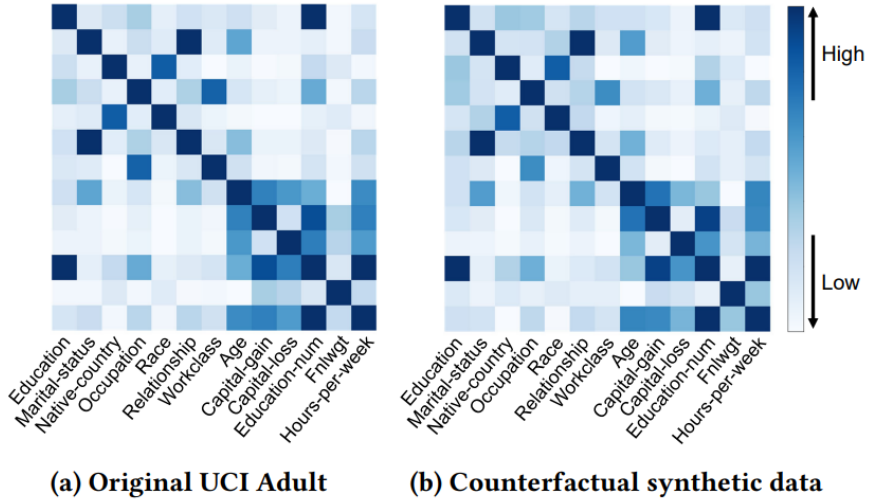
Method	AUC/RMSE	ΔDP	ΔEO	ΔCP	Total
VFAE	3.2	5.8	5.0	4.2	4.6
LAFTR	**2.7**	6.2	5.5	2.8	4.3
MIFR	6.0	3.0	3.8	4.6	4.4
L-MIFR	5.3	3.2	3.3	3.8	3.9
C-InfoNCE	3.2	4.5	3.3	5.2	4.1
WeaC-InfoNCE	4.7	2.8	4.3	4.8	4.2
DualFair	3.0	**2.5**	**2.5**	**2.2**	**2.6**

Experiments

Quality of counterfactual samples

The original relationship between features is well-maintained in the counterfactual samples.

	Adult		Credit		Compas		LSAC	
Training set	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Original	0.91	0.78	0.76	0.68	0.74	0.68	0.85	0.65
Counterfactual	0.89	0.76	0.75	0.62	0.73	0.67	0.84	0.63



(a) Original UCI Adult (b) Counterfactual synthetic data

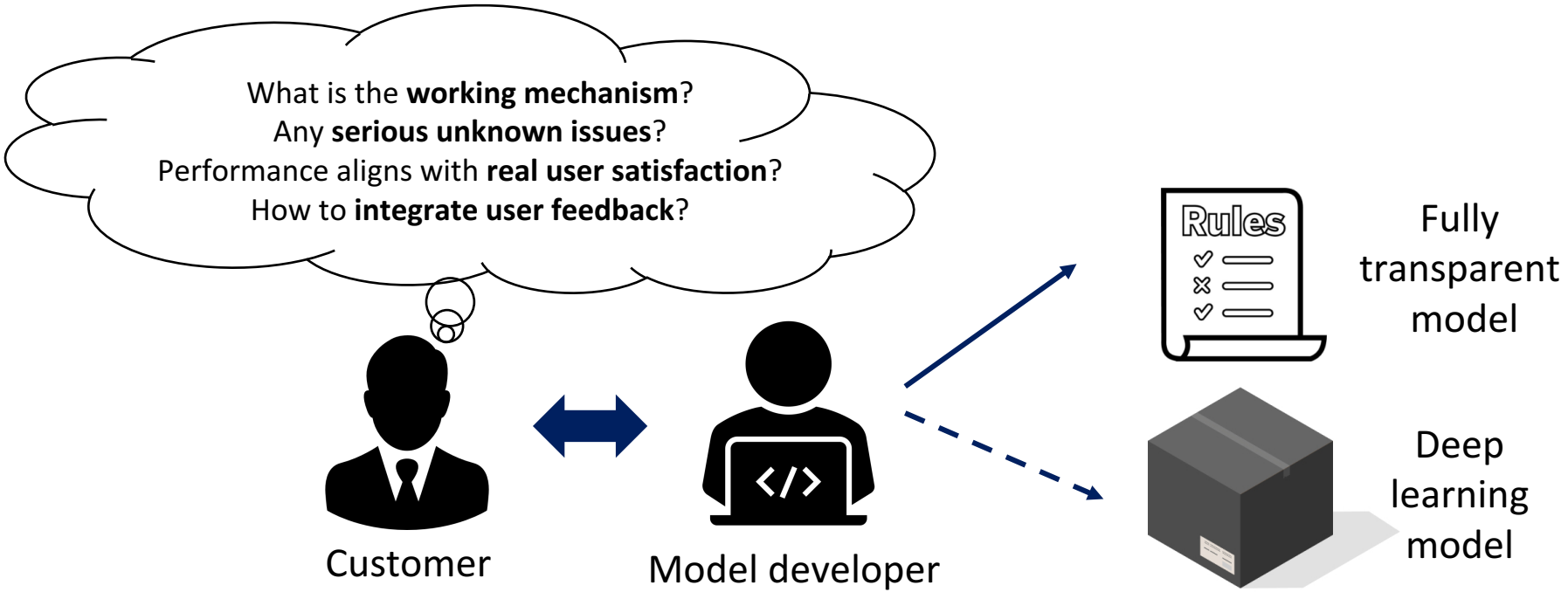
Conclusion

- We propose a self-supervised learning framework, **DualFair**, that simultaneously **debiases sensitive attributes at both group and individual levels**.
- We introduce the C-VAE model to generate counterfactual samples and propose **fairness-aware contrastive loss** to meet the two fairness criteria jointly.
- We design the **self-knowledge distillation loss** to maintain representation quality by minimizing the embedding discrepancy between original and perturbed instances.
- Experiments confirm that DualFair generates a fair embedding with high representation quality. We further show **a synergistic effect of the two fairness criteria**.

Self-explaining deep models with logic rule reasoning
(NeurIPS' 2022)

Background

Trust Issues with Deep learning models



Motivation

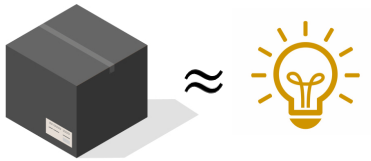
Limitation of Post-Hoc Explanations

Post-hoc explanations

Trust?

Feedback?

Can we trust the explanations?



- Always an approximation [1]
- "General uneasiness" of practitioners [2]

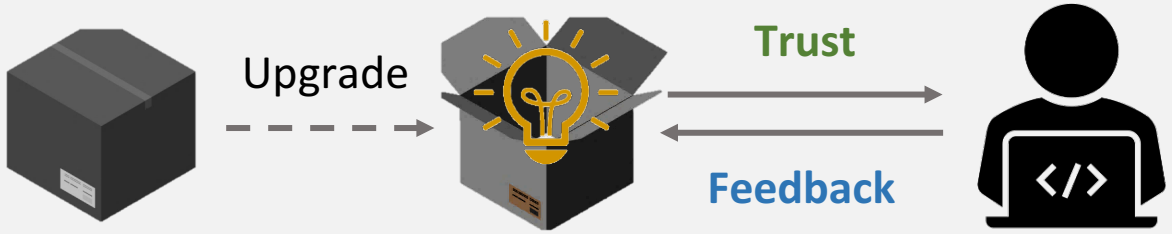
How to integrate user feedback?

- No systematic method for direct control
- Requires model retraining
- No guarantee for satisfying user demands

[1] "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead", *Nature Machine Intelligence*, 2019
[2] "Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs", *ACM HCI* 2020

Main Approach

SELOR: Self-**E**xplaining with **LO**gic rule **R**easoning

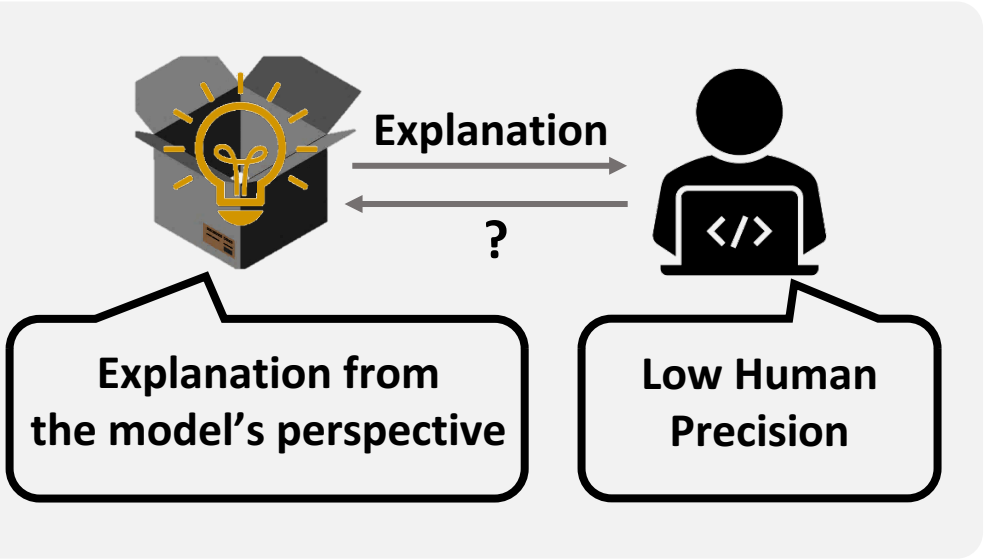


Lays the **foundation** for close collaboration

- **Trust**: explanations **faithful** to the model
- **Feedback**: explanations as **handle** for control

Main Approach

SELOR: Self-**E**xplaining with **LO**gic rule **R**easoning



Human Precision:
Whether the explanation naturally leads to the prediction according to human perception

Low Human Precision:

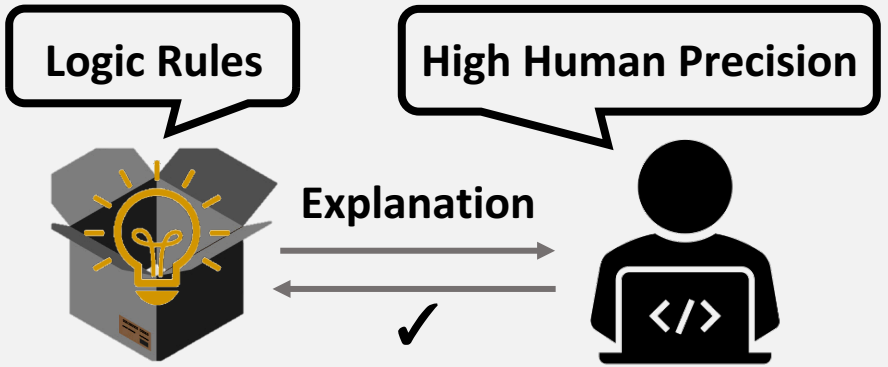
is, an => positive sentiment

High Human Precision:

Awesome => positive sentiment

Main Approach

SELOR: Self-**E**xplaining with **LO**gic rule **R**easoning



- Close to human decision logic
- Widely applied for making predictions
- Require minimum human effort

awesome AND tasty

Antecedent

(condition to apply)

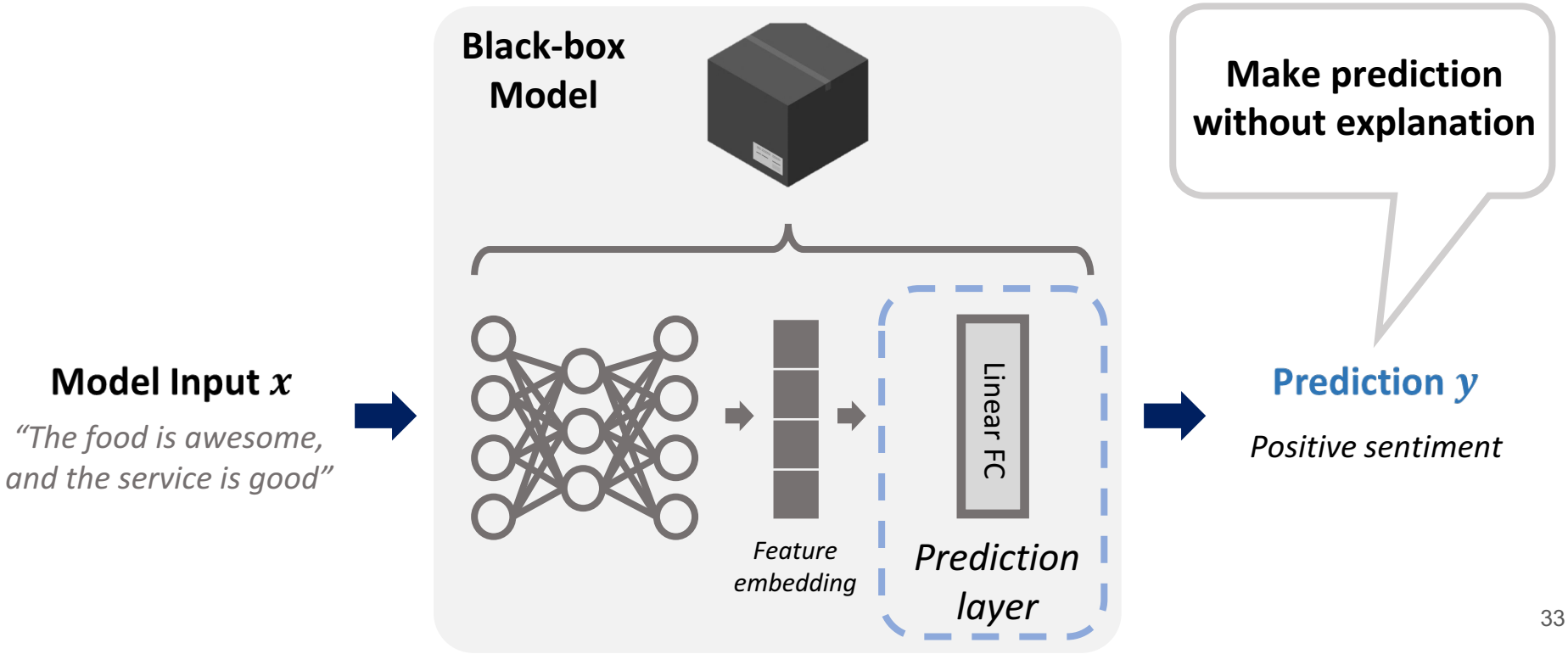
positive sentiment

Consequent

(prediction result)

Main Model Framework

SELOR Framework: Black-box Model

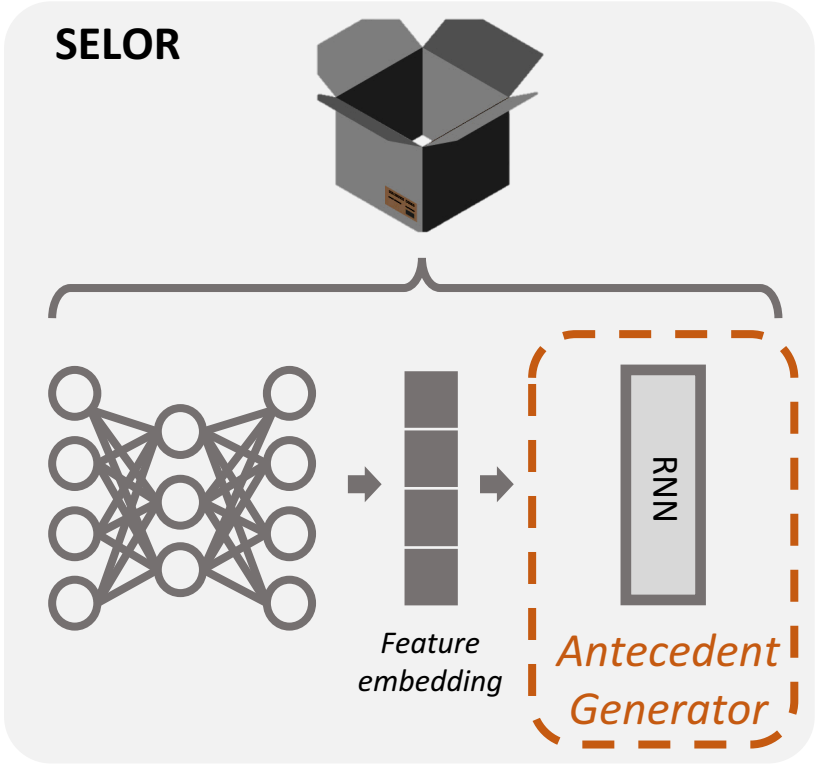


Main Model Framework

SELOR Framework: Antecedent Generator

Model Input x

"The food is awesome, and the service is good"



Select Antecedent

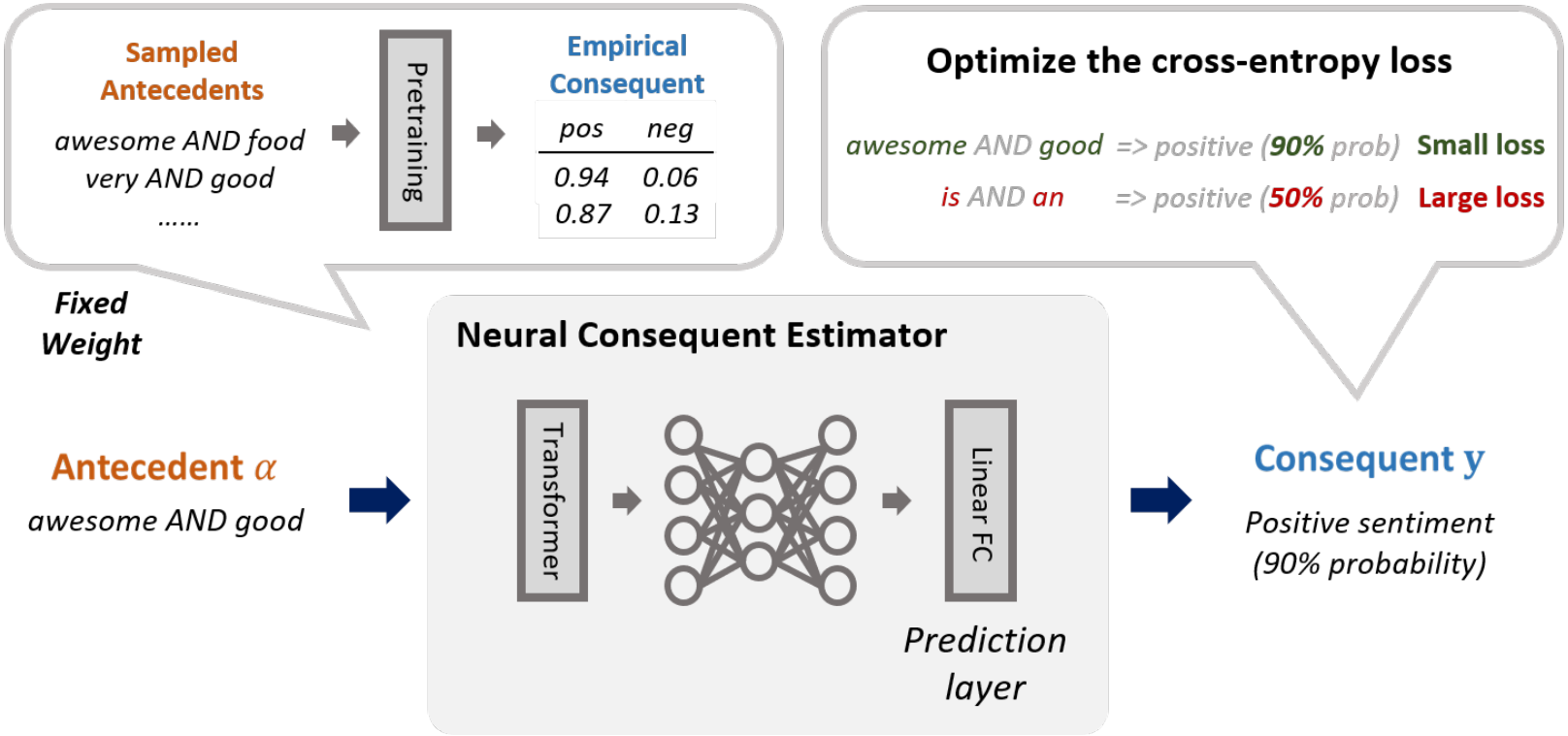
Antecedent α

The
service AND food
the AND food
……

……
the AND is AND good
and AND the AND is

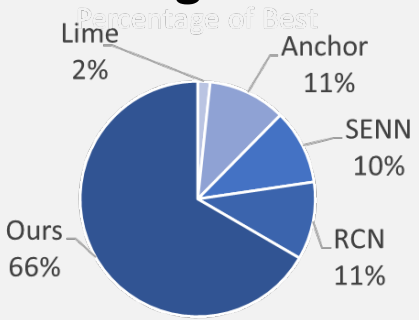
Main Model Framework

SELOR Framework: Consequent Estimator



Results

High Human Precision



User study
Percentage of best

+500%

(Adult dataset)

Good Prediction Performance

SELOR ≈ Black-box

Training Cost

- **Efficient, differentiable** training
- **Slightly slower** than black-box model

Conclusion

- We propose **SELOR**, which incorporates self-explanatory capabilities into a deep model to **provide high human precision by explaining logic rules.**

- SELOR does not require predefined rule sets and can be learned in a differentiable way.

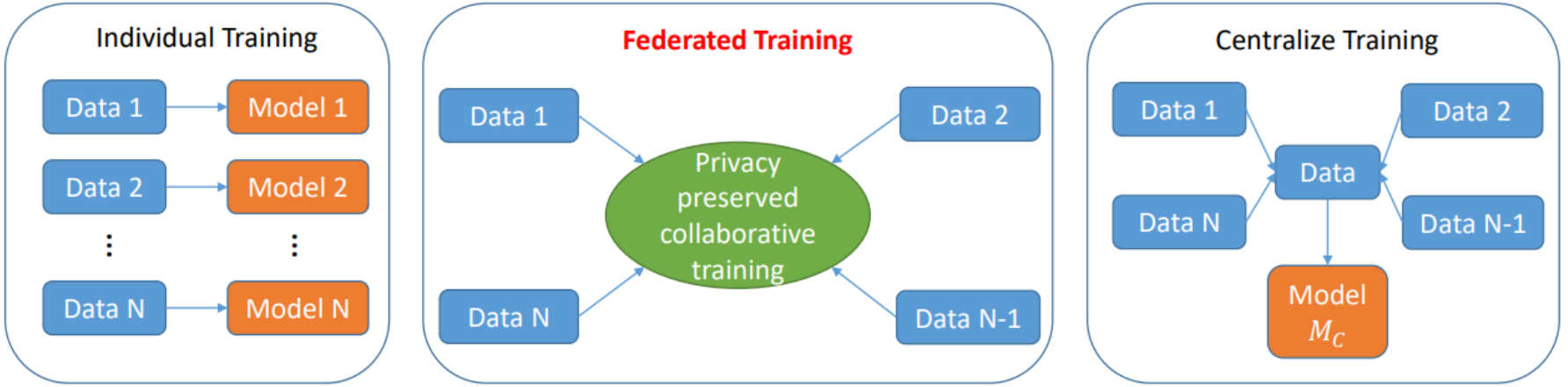
- Extensive tests involving human evaluation show that our method achieves **high prediction performance and human precision in explanation.**

FedX: Unsupervised Federated Learning with Cross Knowledge Distillation (ECCV' 2022)

Background

Federated learning

- Concerns on user data privacy and confidentiality.
- Inability to build an ML model due to inadequate data or training cost on ML implementation of the computational cost involved for training an ML model.



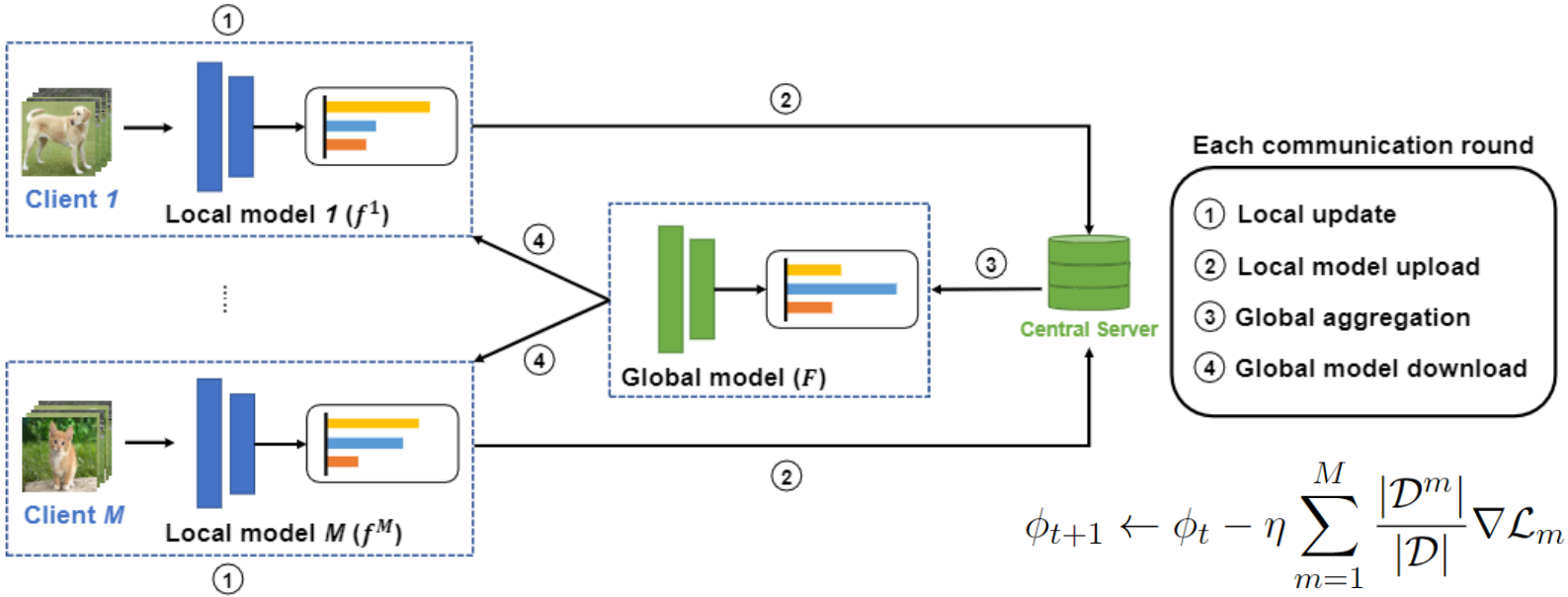
Federated learning: problem statement

- Assume that data are distributed over every different party.
- Parties aim to train a single model F that can apply for various downstream tasks with the help of central server.
- Then, the global objective function to solve is as follows:

$$\arg\min_{\phi} \mathcal{L}(\phi) = \sum_{m=1}^{M} \frac{|\mathcal{D}^m|}{|\mathcal{D}|} \mathcal{L}_m(\phi),$$
$$\text{where } \mathcal{L}_m(\phi) = \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}^m}[l_m(\mathbf{x}, y; \phi)].$$

Basic architecture for federated learning: FedAvg

- Four processes run in each communication round



$$\phi_{t+1} \leftarrow \phi_t - \eta \sum_{m=1}^{M} \frac{|\mathcal{D}^m|}{|\mathcal{D}|} \nabla \mathcal{L}_m(\phi_t)$$

Background

Extension to unsupervised federated learning

- We can run unsupervised representation learning on this federated framework.
- Then, the global objective function to solve is as follows:

$$\arg\min_{\phi} \mathcal{L}(\phi) = \sum_{m=1}^{M} \frac{|\mathcal{D}^m|}{|\mathcal{D}|} \mathcal{L}_m(\phi),$$
$$\text{where } \mathcal{L}_m(\phi) = \mathbb{E}_{\mathbf{x} \in \mathcal{D}^m}[l_m(\mathbf{x}; \phi)].$$

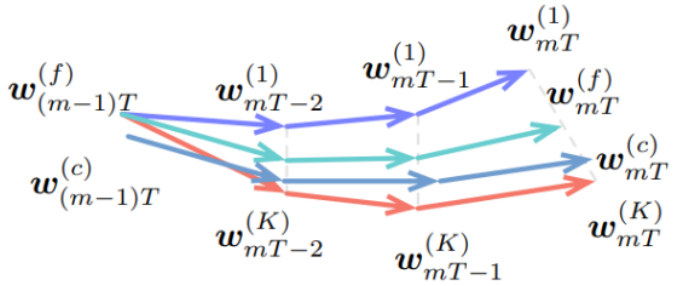
- Loss function could be excerpted from
  - InfoNCE-based model (e.g., SimCLR, MoCo, NNCLR)
  - Asymmetric siamese-based model (e.g., BYOL, SimSiam)
  - …

Motivation

Challenges in federated learning

- Non-IID distribution of local-data leads to the biased results

IID Settings:



Non-IID Settings:

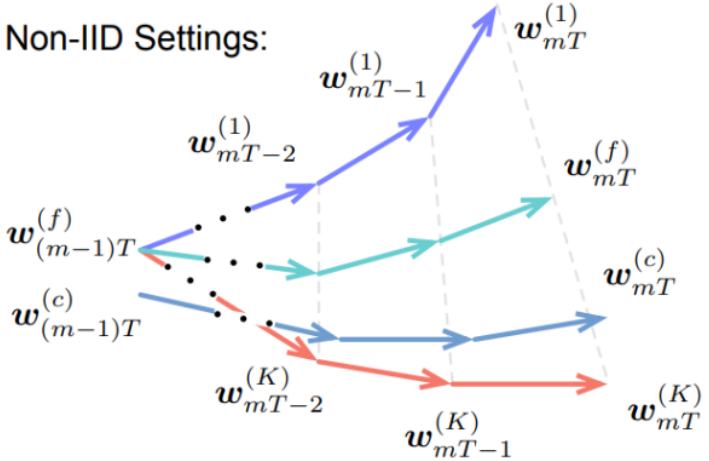
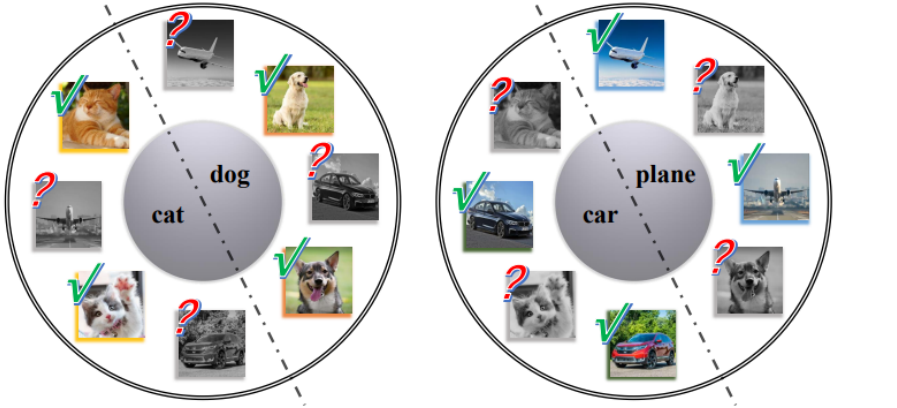


Image credit: Federated Learning with Non-IID Data

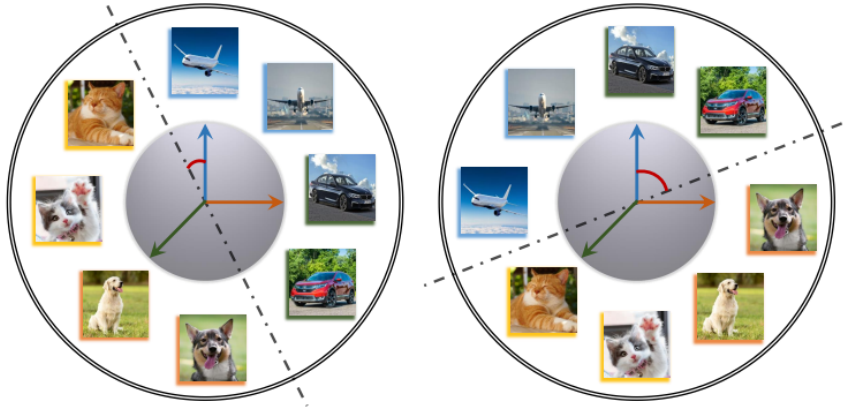
Motivation

Extra challenges in “unsupervised” federated learning

- Non-IID distribution without ground-truth labels amplifies the embedding divergence



(a) Inconsistency of representation spaces.



(b) Misalignment of representations.

Image credit: Federated Unsupervised Representation Learning

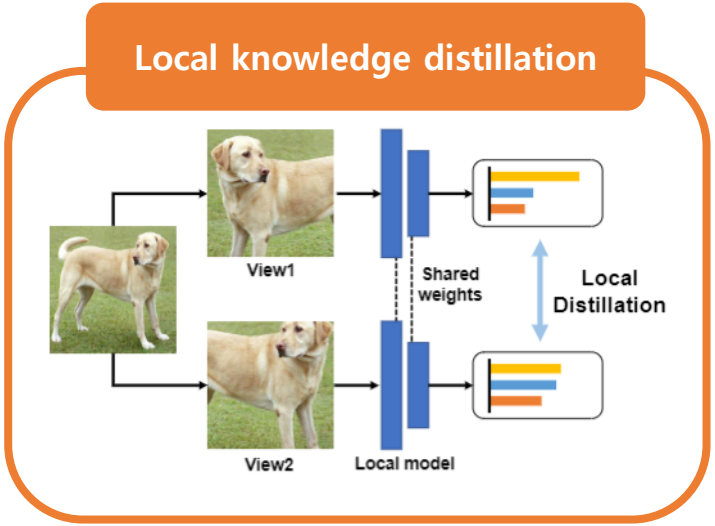
Main Idea: knowledge distillation

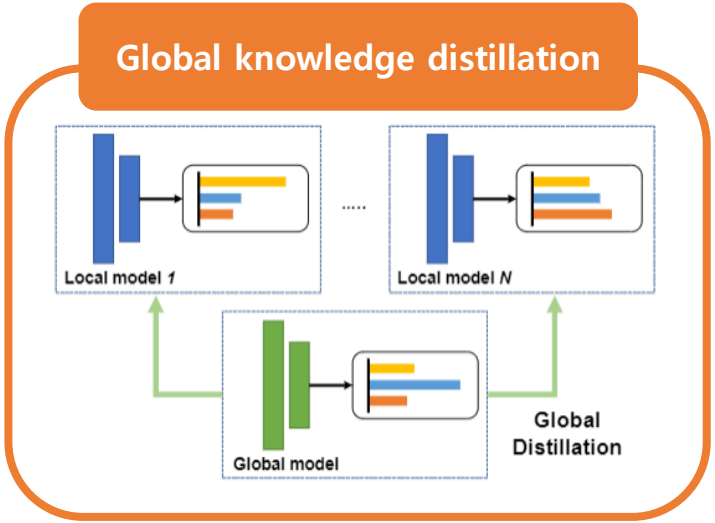
Knowledge distillation to convey global knowledge to local client!

- Knowledge distillation refers to the idea of model compression by teaching a smaller network, step by step, exactly what to do using a bigger already trained network.
- **Global model** has knowledge from entire data distribution (usually has higher performance), which can be regarded as a **teacher model**.
- **Local model** has knowledge only from locally biased data and needs entire data's information, which can be regarded as a **student model**.
- Distill knowledge from global model to regularize the local model's training in an unsupervised fashion!

Main Idea: Two-sided knowledge distillation

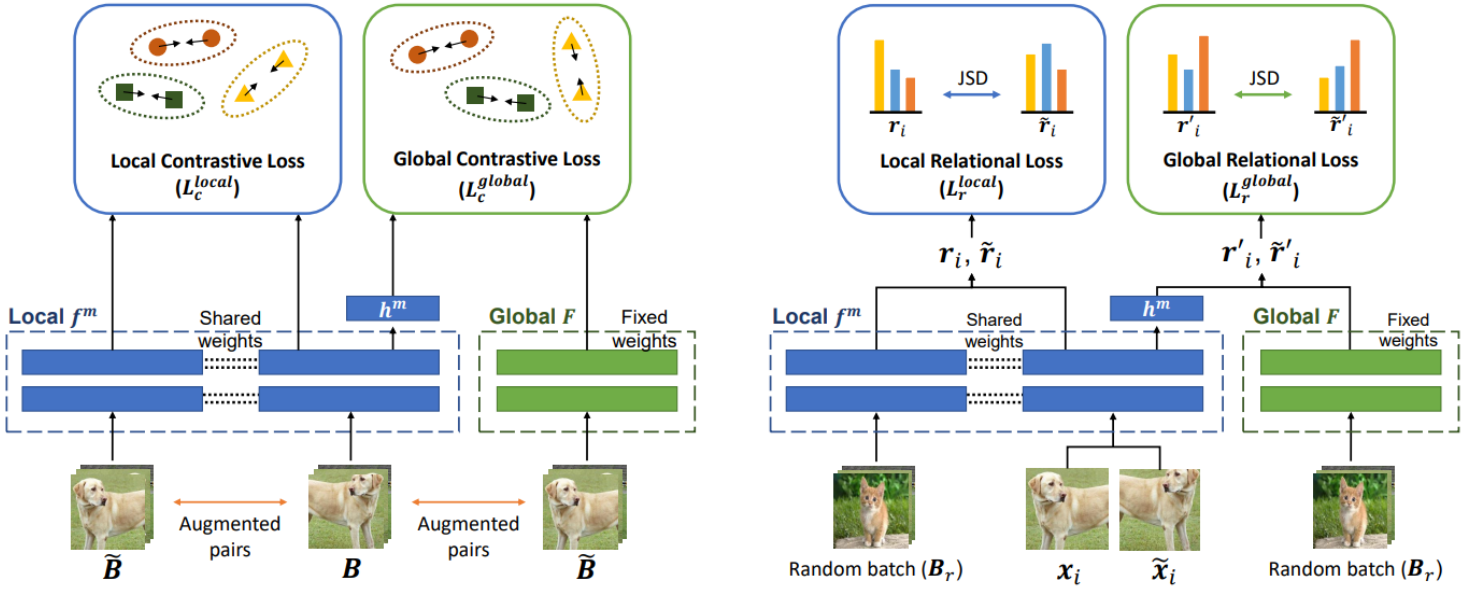
Learns representation from local data and refines the central server's knowledge via **two-sided knowledge distillation**





Model overview

Two-sided contrastive loss and relational loss



(a) Two-sided contrastive loss

(b) Two-sided relational loss

Performance Evaluation

Table 1: Performance improvement with FedX on classification accuracy over three datasets. Both the final round accuracy and the best accuracy show that our model brings substantial improvement for all baseline algorithms.

Method	CIFAR-10		SVHN		F-MNIST	
	Last	Best	Last	Best	Last	Best
FedSimCLR	51.31	52.88	75.19	76.50	77.66	79.44
+ FedX	**56.88**	**57.95**	**77.19**	**77.70**	**81.98**	**82.47**
FedMoCo	56.74	57.82	70.69	70.99	82.31	83.58
+ FedX	**58.23**	**59.43**	**73.57**	**73.92**	**83.62**	**84.65**
FedBYOL	52.24	53.14	65.95	67.32	81.45	82.37
+ FedX	**56.49**	**57.79**	**68.94**	**69.05**	**83.18**	**84.30**
FedProtoCL	51.33	52.12	49.85	50.19	81.76	**83.57**
+ FedX	**55.36**	**56.76**	**69.31**	**69.75**	**82.74**	83.34
FedU	50.79	50.79	66.02	66.22	80.59	82.03
+ FedX	**56.15**	**57.26**	**68.13**	**68.39**	**83.73**	**84.12**

Conclusion

- We propose FedX, a new advance in unsupervised federated learning that learns data representations via a **unique two-sided knowledge distillation** at local and global levels.

- FedX can be applied to extant algorithms to **enhance performance by 1.58–5.52pp** in top-1 accuracy and further **enhance training speed**.

- **FedX preserves privacy** between clients and does not share data directly. It is also **does not require complex communication** for sending data features.

Takeaways

Conclusion

- This presentation introduces four research branches for trustworthy AI:
  (1) Fairness
  (2) Interpretability
  (3) Data privacy
  (4) Security

- There still has a large room of improvements in jointly achieving multiple human-centered properties for trustworthy AI:
  (1) Fairness vs. Performance tradeoff
  (2) Interpretability vs. Performance tradeoff
  (3) Transparency vs. Security tradeoff
  (4) Privacy vs. Security tradeoff
  ….

Microsoft KAIST ibs 데이터사이언스그룹 Data Science Group

Thank you

Sungwon Han (lion4152@gmail.com)