Federated Learning with Only Positive Labels

Felix X. Yu et al

Presented by Chenpei Huang

Published in ICML ‘20

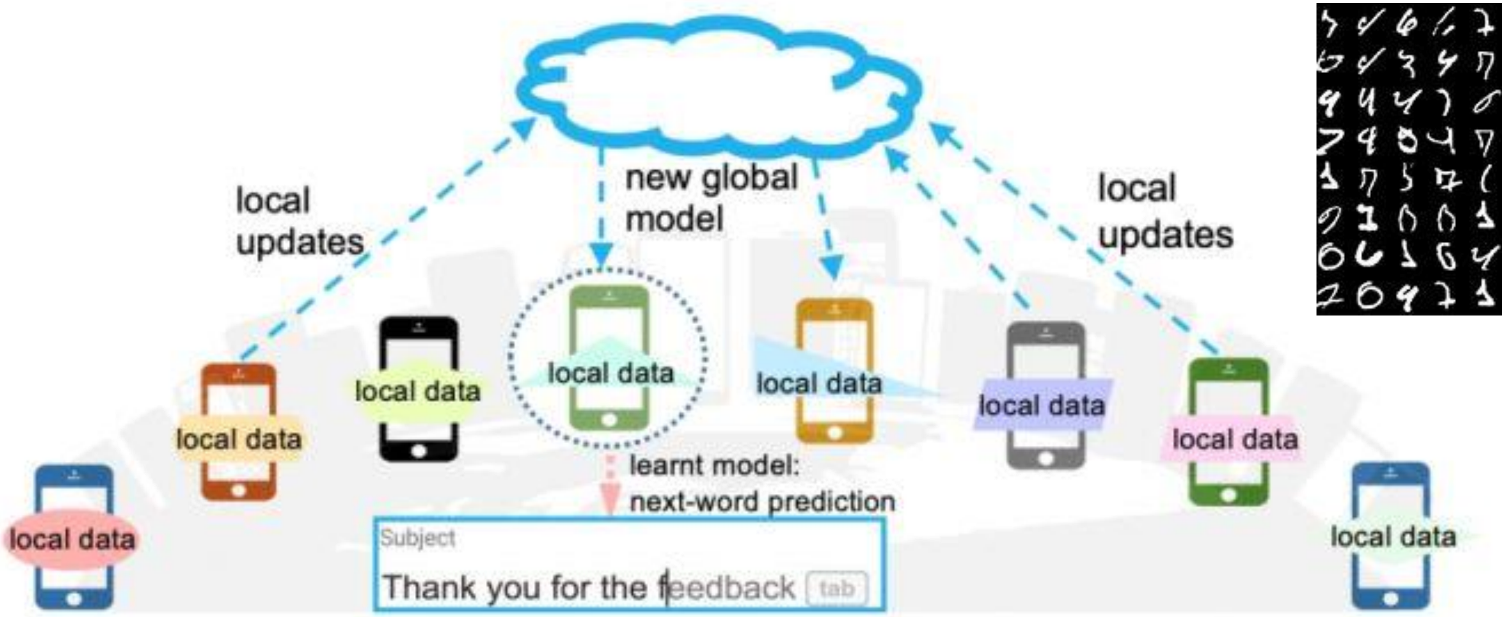
Outline

- Introduction
- Algorithm
- Analysis
- Evaluation
- Conclusion

Introduction

Federated Learning:

- Server sends the current global model to users
- Each user update the model with its local data, and send it to server
- Server average (FedAvg) the deltas and updates global model



UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Introduction

Learning-based user identification

- Examples: face, voiceprint, fingerprint, etc.
- Goal: learn discriminative features
- Challenge: large dataset and privacy concerns

Multiple users?

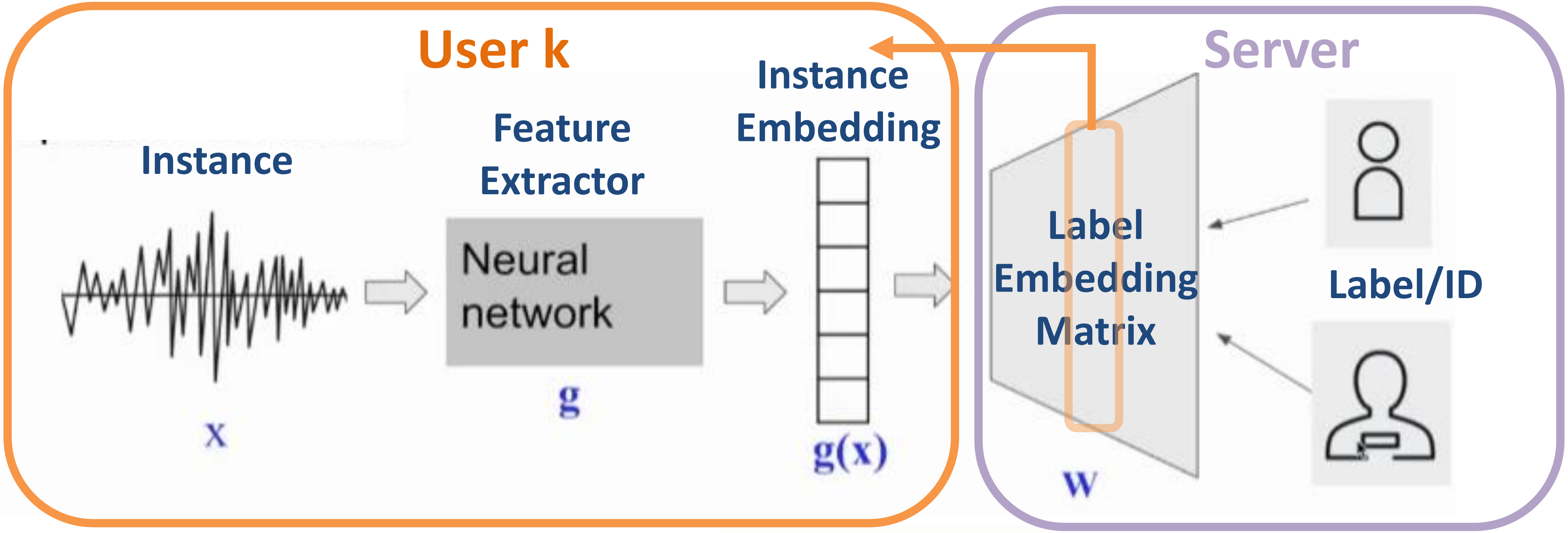
Distributed data?

Sensitive data?

Use federated learning!!!

Introduction

Federated learning for user identification



Score or logit: $f(x) = Wg(x)$

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Introduction

Score or logit: $\mathbf{f(x) = Wg(x)}$

- One wants large scores for positive instance and label pairs
- One want small scores for negative instance and label pairs

$$\ell_{\mathrm{cl}}(f(\boldsymbol{x}),y) = \underbrace{\alpha \cdot \left(\boldsymbol{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{w}_y)\right)^2}_{\ell_{\mathrm{cl}}^{\mathrm{pos}}(f(\boldsymbol{x}),y)} + \underbrace{\beta \cdot \sum_{c\neq y}\left(\max\left\{0,\nu - \boldsymbol{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{w}_c)\right\}\right)^2}_{\ell_{\mathrm{cl}}^{\mathrm{neg}}(f(\boldsymbol{x}),y)},$$

Own $g_\theta(x)$

Own ω_y

Other $g_\theta(x)$

Other ω_c

Introduction

If only trained on positive loss…

$$g(x) = w_1 = \ldots = w_K \text{ for any } x. \Longrightarrow$$

Positive Loss = 0
Negative Loss max
Not work

$$\ell_{\mathrm{cl}}(f(\boldsymbol{x}), y) = \underbrace{\alpha \cdot \left(\boldsymbol{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_y)\right)^2}_{\ell_{\mathrm{cl}}^{\mathrm{pos}}(f(\boldsymbol{x}), y)} + \underbrace{\beta \cdot \sum_{c \neq y} \left(\max\{0, \nu - \boldsymbol{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_c)\}\right)^2}_{\ell_{\mathrm{cl}}^{\mathrm{neg}}(f(\boldsymbol{x}), y)},$$

$\Downarrow$

Minimize positive loss while keeping label embeddings spread-out

Outline

- Introduction
- Algorithm
- Analysis
- Evaluation
- Conclusion

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Algorithm

Federated Averaging with Spreadout (FedAwS):

- The trained label embeddings should be geometric saperated.
- Add regulation term to spread them out by a margin v
- User sends the updated feature extractor and own label embedding to server. Server averages the feature extractor and compute the regulation.

$$\operatorname{reg}_{\mathrm{sp}}(W) = \sum_{c \in [C]} \sum_{c' \neq c} \left(\max \left\{ 0, \nu - \boldsymbol{d}(\boldsymbol{w}_c, \boldsymbol{w}_{c'}) \right\} \right)^2 .$$

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Algorithm

FedAwS: two challenges

- ➢ hyperparameter v is hard to determine
- ➢ the number of user (C) is huge => expensive computation

Solution: stochastic "hard" negative mining

Choose Top-k closest label embeddings in one subset

Choose v to be Top-(k+1) closest distance

$$\mathrm{reg}_{\mathrm{sp}}(W) = \sum_{c \in [C]} \sum_{c' \neq c} \left(\max\left\{0, \nu - \boldsymbol{d}(\boldsymbol{w}_c, \boldsymbol{w}_{c'})\right\}\right)^2. \quad \Rightarrow \quad \mathrm{reg}_{\mathrm{sp}}^{\mathrm{top}}(W) = \sum_{c \in \mathcal{C}^t} \sum_{\substack{y \in \mathcal{C}', \\ y \neq c}} -\boldsymbol{d}^2(\boldsymbol{w}_c, \boldsymbol{w}_y) \cdot [\![y \in \mathcal{N}_k(c)]\!],$$

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Outline

- Introduction
- Algorithm
- Analysis
- Evaluation
- Conclusion

Analysis

I. Reason why a simple spreadout will work

II. Similar in shape, the cosine contrastive loss

III. Relate FedAwS with cosine contrastive loss

Error probability ← **Distance** → **Contrastive loss**

FedAwS

Analysis

I. Reason why a simple spreadout will work

Proposition 1. *Let the minimum distance between the class embeddings be* $\rho := \inf_{i \neq j} \mathbf{d}(\boldsymbol{w}_i, \boldsymbol{w}_j)$*, and the expected distance between the embeddings of an instance* $\boldsymbol{x}$ *and its true class* y *be* $\epsilon = \mathbb{E}_{(\boldsymbol{x},y)\sim \mathrm{P_{XY}}} \mathbf{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_y)$*. Then, the probability of misclassification satisfies*

$$\mathrm{P}\left(\exists z \neq y \text{ s.t. } \mathbf{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_y) \geq \mathbf{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_z)\right) \leq 2\epsilon/\rho.$$

Proof:

$$\mathrm{P}\left(\exists z \neq y \text{ s.t. } \mathbf{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_y) \geq \mathbf{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_z)\right)$$

$$\leq \mathrm{P}\left(\mathbf{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_y) \geq \frac{\rho}{2}\right)$$

Markov Inequality

$$\leq \frac{2\mathbb{E}_{(\boldsymbol{x},y)\sim \mathrm{P_{XY}}} \mathbf{d}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_y)}{\rho} = \frac{2\epsilon}{\rho}.$$

The error probability is bounded by the intra-class distance divided by the minimum inter-class distance!

Analysis

II. Cosine contrastive loss

Definition 1 (Cosine contrastive loss)**.** *Given an instance and label pair* $(\boldsymbol{x}, y)$ *and the scorer* $f(\boldsymbol{x})$ *in* (1)*, the cosine contrastive loss takes the following form.*

$$\ell_{\mathrm{ccl}}(f(\boldsymbol{x}), y) = \left(\boldsymbol{d}_{\cos}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_y)\right)^2 + \sum_{c \neq y} \left(\max\left\{0, \nu - \boldsymbol{d}_{\cos}(g_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{w}_c)\right\}\right)^2. \quad (11)$$

$$\ell_{\mathrm{ccl}}(f(\boldsymbol{x}), y) = (1 - s_y)^2 + \sum_{c \neq y} \left(\max\left\{0, \nu - 1 + s_c\right\}\right)^2$$

$$\boldsymbol{d}_{\cos}(\boldsymbol{u}, \boldsymbol{u}') = 1 - \boldsymbol{u}^\top \boldsymbol{u}' \quad \forall\, \boldsymbol{u},\ \boldsymbol{u}' \in \mathbb{R}^d.$$

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Analysis

III. Relate FedAwS with cosine contrastive loss

FedAwS objective:

$$\ell_{\mathrm{sp}}(f(\boldsymbol{x}), y) = (1 - s_y)^2 + \sum_{c \neq y} \left(\max\left\{0, \nu - 1 + \boxed{\boldsymbol{w}_y^\top \boldsymbol{w}_c}\right\}\right)^2,$$

Cosine contrastive loss:

$$\ell_{\mathrm{ccl}}(f(\boldsymbol{x}), y) = (1 - s_y)^2 + \sum_{c \neq y} \left(\max\left\{0, \nu - 1 + \boxed{s_c}\right\}\right)^2$$

$$|\Delta_c| \le 2(1 + 2\nu) \cdot \left|\boldsymbol{w}_c^\top \boldsymbol{r}_{\boldsymbol{x},y}\right|.$$

where $\boldsymbol{r}_{\boldsymbol{x},y} = \boldsymbol{w}_y - g_{\boldsymbol{\theta}}(\boldsymbol{x})$

approach 0 during local training

Outline

- Introduction
- Algorithm
- Analysis
- Evaluation
- Conclusion

Evaluation

Evaluation Method: classification with one class settings

Dataset: [CIFAR-10, CIFAR-100], [AmazonCat, WIKIAHRC, Amazon670K]

Baseline:

1. Training with only positive loss
2. Training with positive loss wit fixed label embeddings (avoid collapsing)
3. Softmax (oracle)

Model architecture: resnet, embedding dimension [64/512]

Training setup: 4K labels and users are selected in one round

Evaluation

Result: small dataset

Dataset	Model	Baseline-1	Baseline-2	FedAwS	Softmax (Oracle)
CIFAR-10	RESNET-8	10.7	83.3	86.3	88.4
CIFAR-10	RESNET-32	9.8	92.1	92.4	92.4
CIFAR-100	RESNET-32	1.0	65.1	67.9	68.0
CIFAR-100	RESNET-56	1.1	67.5	69.6	70.0

Observation:

- Training on positive labels gives very poor result due to collapse
- But once label embeddings are fixed, even randomly chosen, results are surprisingly good
- The proposed FedAwS outperforms two baselines and approaches softmax

Evaluation

Result: multi-lable dataset

Dataset	#Features	#Labels	#TrainPoints	#TestPoints	Avg. #I/L	Avg. #L/I
AMAZONCAT	203,882	13,330	1,186,239	306,782	448.57	5.04
WIKILSHTC	1,617,899	325,056	1,778,351	587,084	17.46	3.19
AMAZON670K	135,909	670,091	490,449	153,025	3.99	5.45

K=10, λ=10

		Federated Learning with Only Positives			Oracle	
		Baseline-1	Baseline-2	FedAwS	Softmax	SLEEC
AMAZONCAT	P@1	3.4	64.1	92.1	92.1	90.5
	P@3	3.2	46.8	70.8	77.9	76.3
	P@5	3.1	32.6	58.7	62.3	61.5
AMAZON670K	P@1	0.0	4.3	33.1	35.2	35.1
	P@3	0.0	2.8	29.6	31.6	31.3
	P@5	0.0	2.2	27.4	29.5	28.6
WIKILSHTC	P@1	7.6	7.9	37.2	54.1	54.8
	P@3	4.5	3.4	22.6	38.8	33.4
	P@5	2.8	2.6	16.2	29.9	23.9

Baseline 2 fails because the class# is too big to be separated

Outline

- Introduction
- Algorithm
- Analysis
- Evaluation
- Conclusion

Conclusion

Centralized	FedAwS	FedUV
Share instance and label embedding	Share label emb to server	No emb shared to other users or server
Learn from positive and negative labels	Learn from positive, and spreadout labels	Fit instance emb to ECC code (ensured to separate)

Conclusion

- This work studied a novel learning setting, federated learning with only positive labels, and propose FedAwS that learn without negative instance and label pairs.
- It proves that strong geometric regulation can replace the negative sampling.
- The method achieves near oracle performance.


THANK YOU
UNIVERSITY of HOUSTON | ENGINEERING