Divergence-aware Federated Self-Supervised Learning

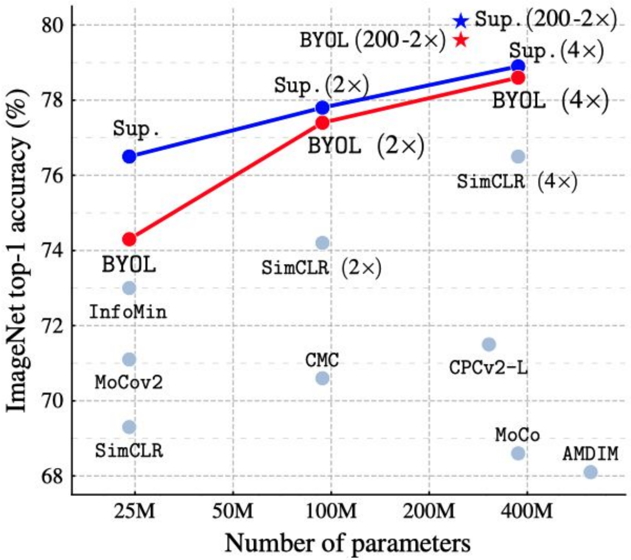
Weiming Zhuang[1,2], Yonggang Wen[1], Shuai Zhang[2]

[1]*Nanyang Technological University*, [2]*SenseTime Research*

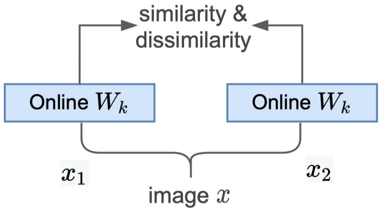
Self-supervised Learning with Siamese Network

Achieved remarked performance

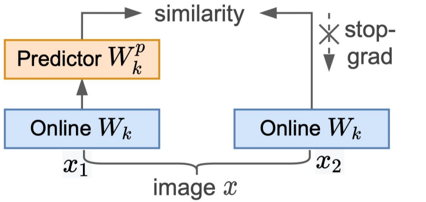
Goal: train useful visual representations from unlabeled data



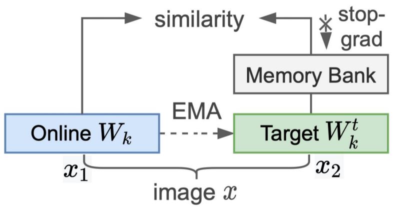
A large number of existing works



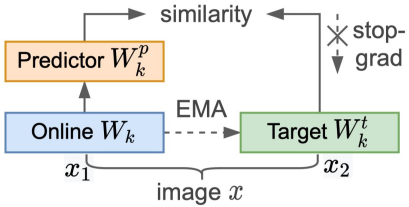
SimCLR
(Chen, et al. ICML'20)



SimSiam
(Chen, et al. CVPR'21)



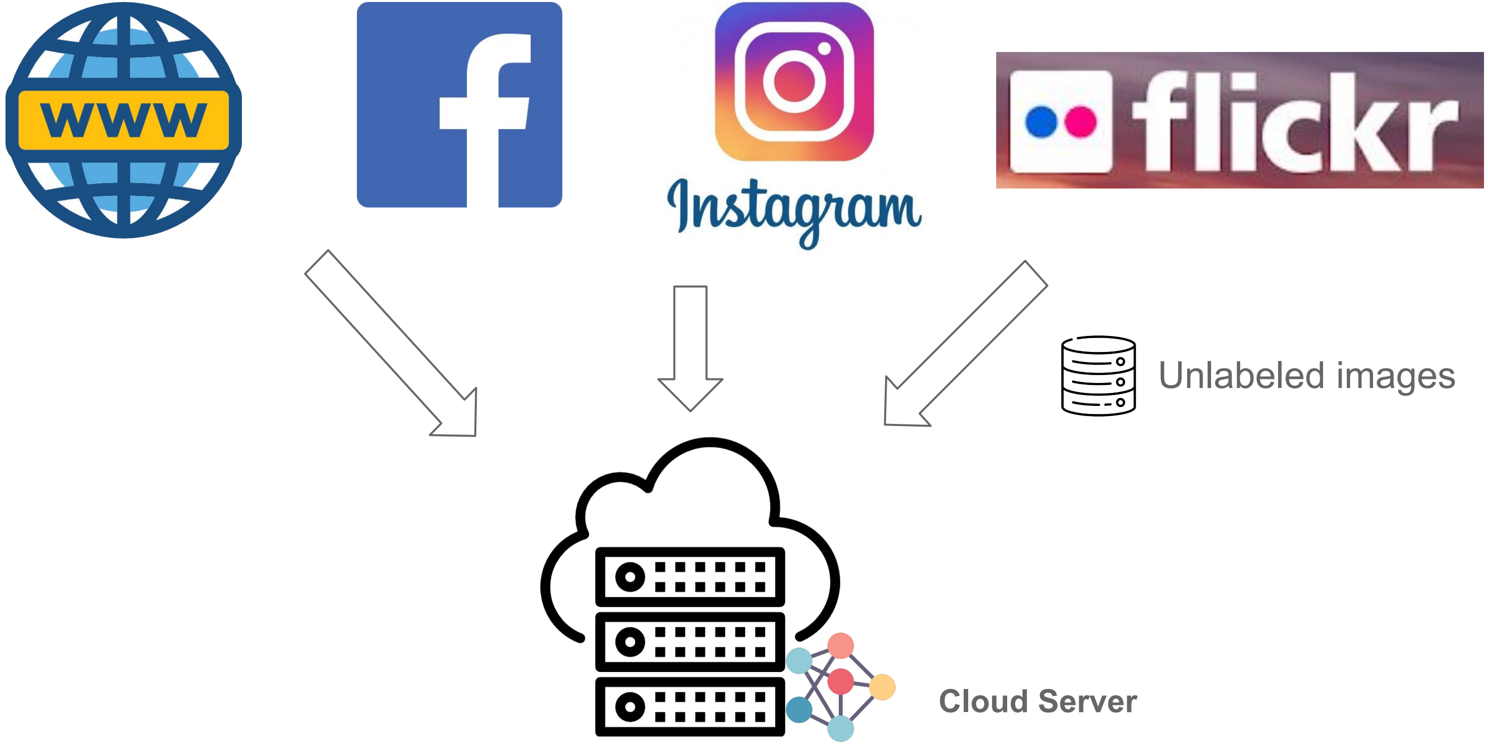
MoCo V1
(He, et al. CVPR'20)



BYOL
(Grill, et al. NIPS'20)

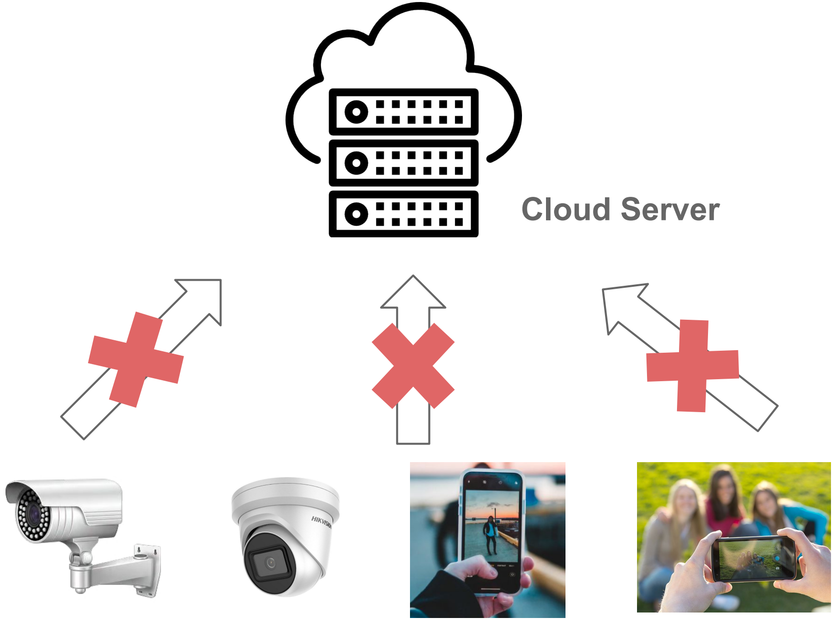
SwAV (Caron, et al. NIPS'20), DirectPred (Tian, et al. ICML'21), and more …

Learn From Centralized Publicly Available Data



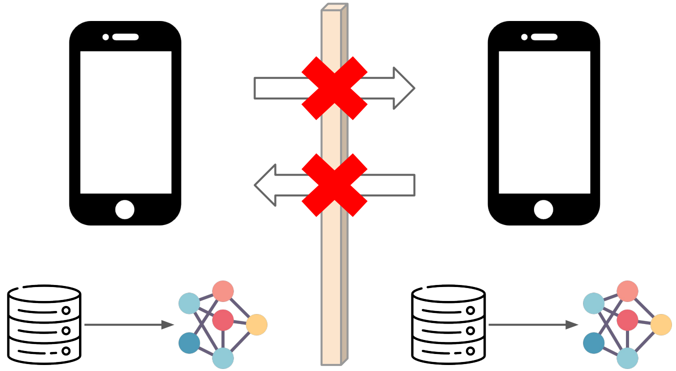
Decentralized Data May Not be Centralized

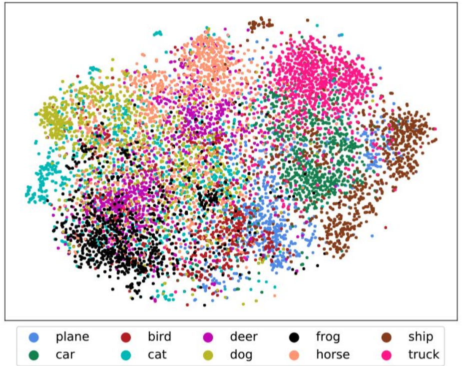
Data Protection Regulations



Existing Solutions

Standalone Training Has Bad Performance





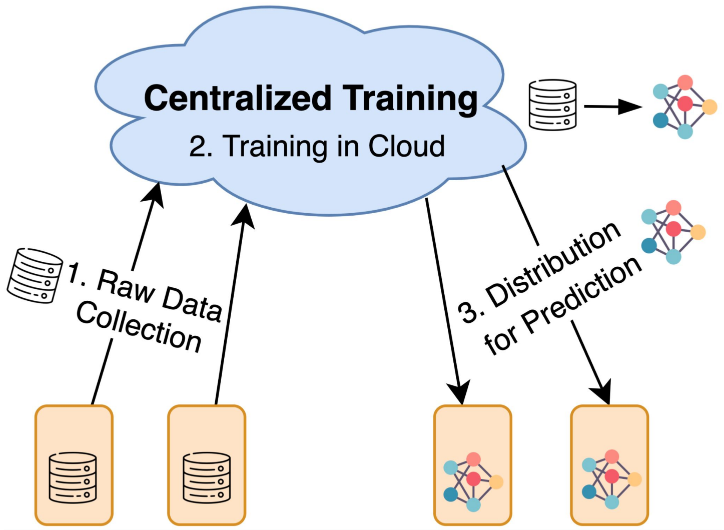
T-SNE of visualization of representations learned from single client with 1 CIFAR-10 class [1]

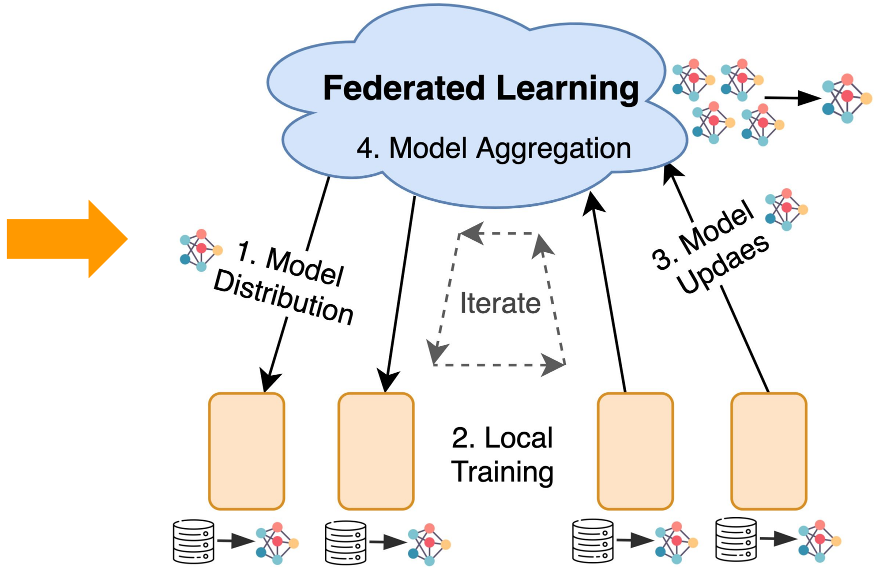
Standalone Training:
Self-supervised learning on a single client without sharing data

Bad performance due to Non-IID Data
Non-IID: Non-Independently & Identically Distributed

[1] Zhuang, Weiming, et al. "Collaborative Unsupervised Visual Representation Learning from Decentralized Data." ICCV. 2021.

Federated Learning (FL) Primer

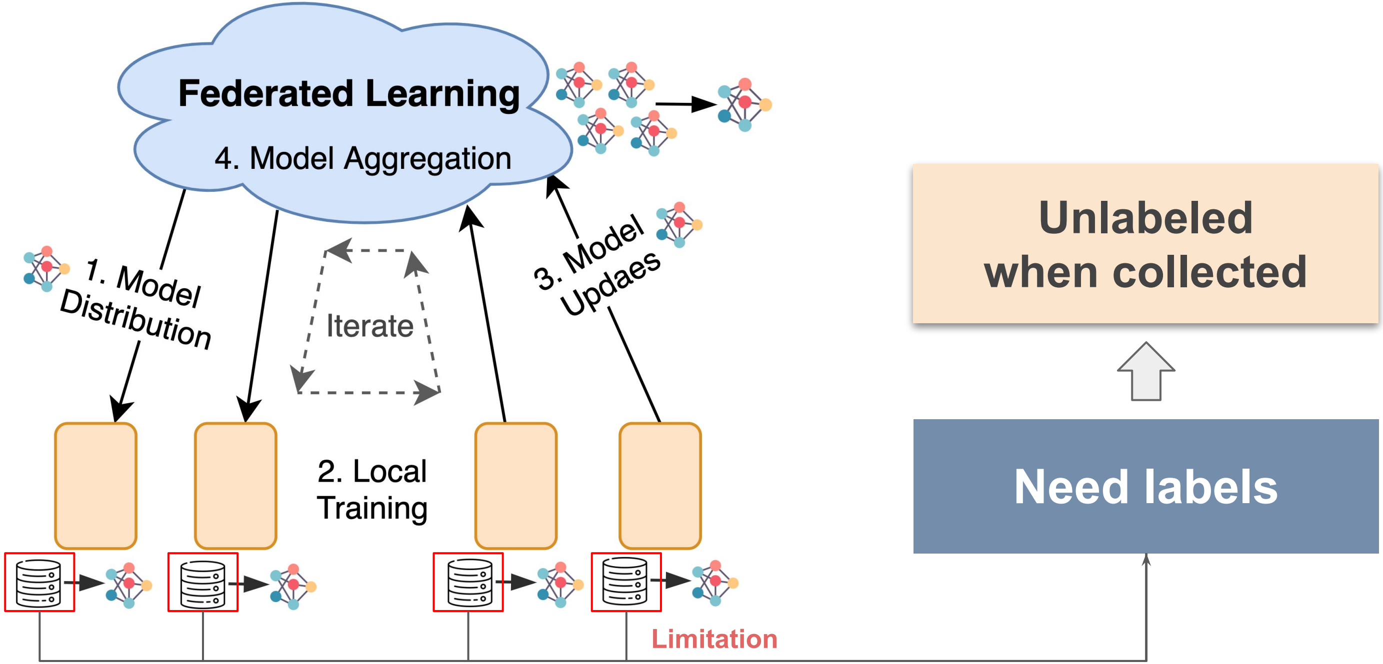




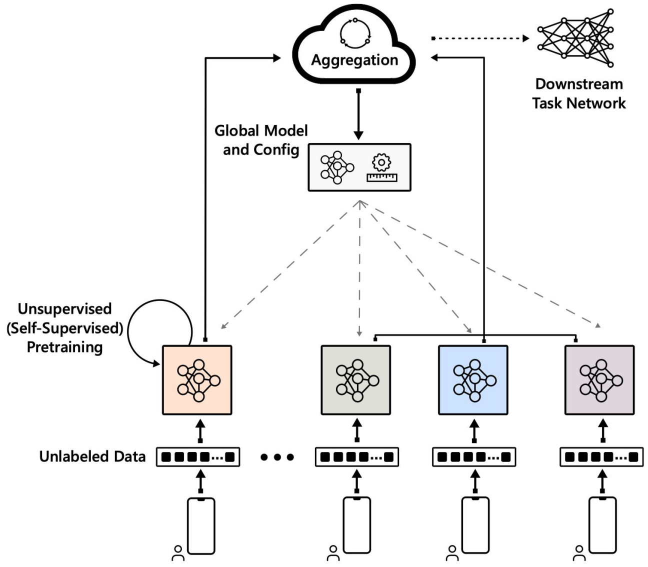
Share **data**

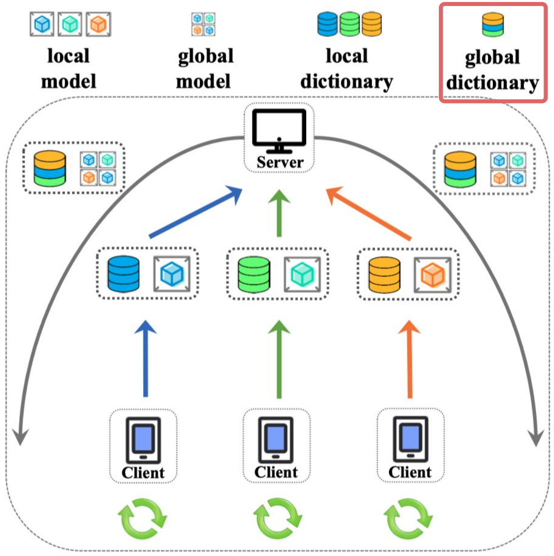
Share **model updates,** not **data**

Federated Learning Relies on Data Labels



Federated Unsupervised Learning



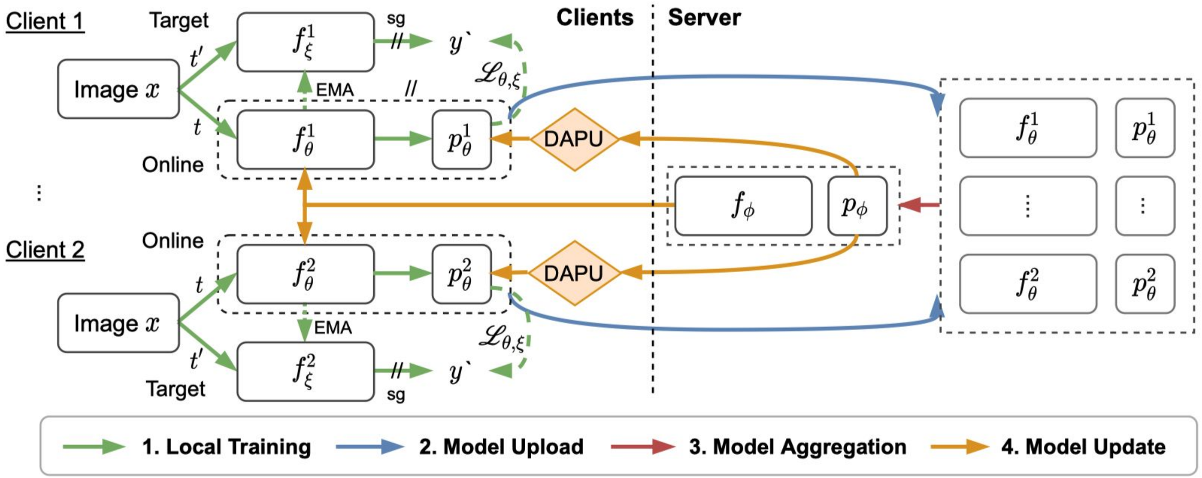


Bypass Non-IID data problem [1]

Potential privacy risks from sharing features to form a global dictionary [2]

[1] van Berlo et. al., "Towards federated unsupervised representation learning." *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*. 2020.
[2] Zhang et al. "Federated unsupervised representation learning." *arXiv preprint arXiv:2010.08982* (2020).

Federated Self-Supervised Learning (FedSSL)



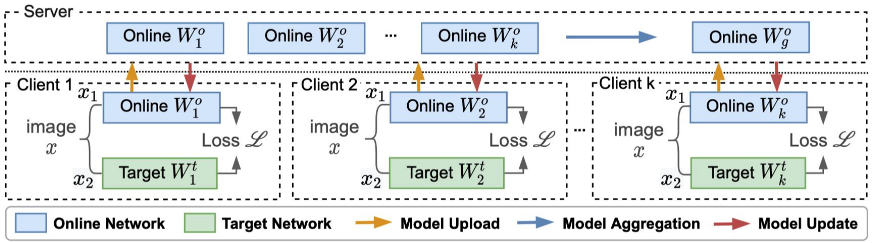
Based on BYOL [1],

but not yet revealed **deep insights into the fundamental building blocks**

[1] Zhuang, Weiming, et al. "Collaborative Unsupervised Visual Representation Learning from Decentralized Data." ICCV. 2021.

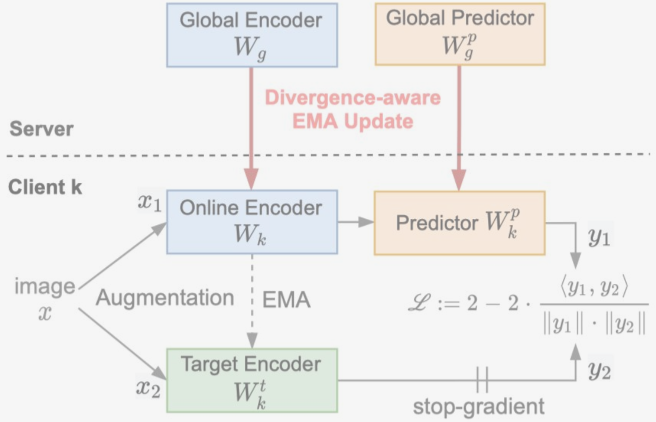
Part 1

FedSSL framework for in-depth Studies



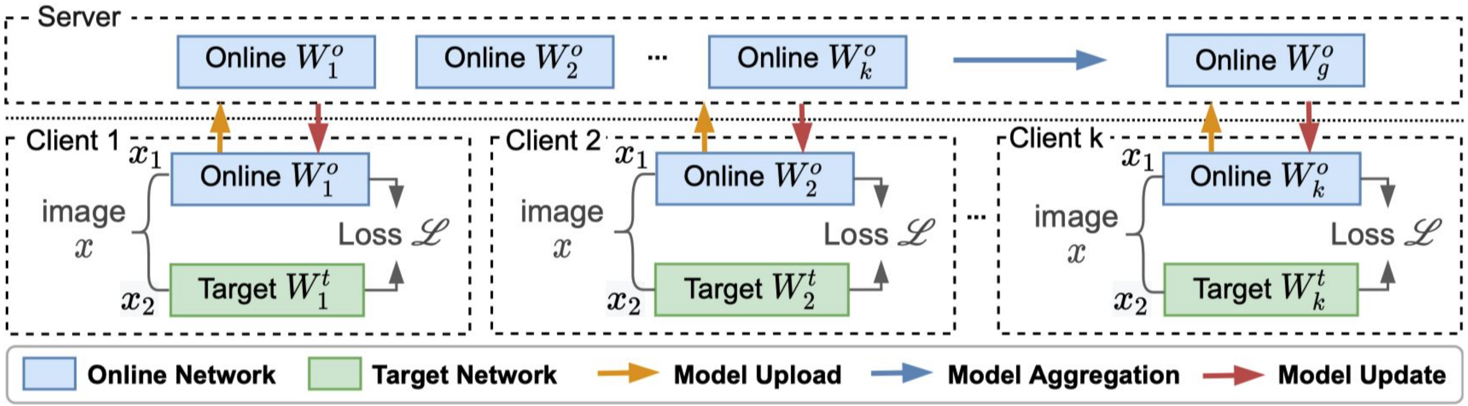
Part 2

Divergence-aware EMA Update



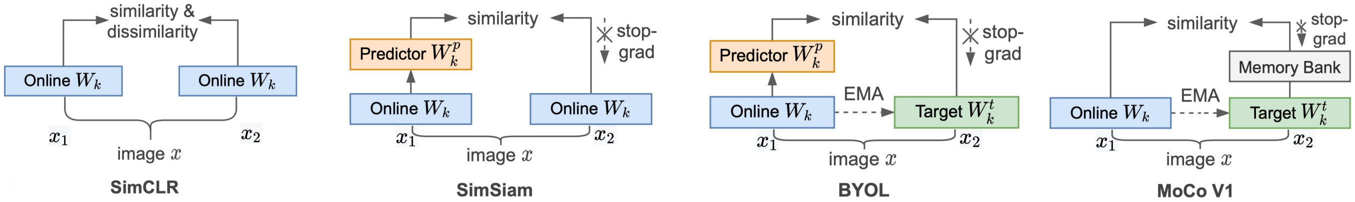
FedSSL Framework

Federated Self-supervised Learning (FedSSL) Framework



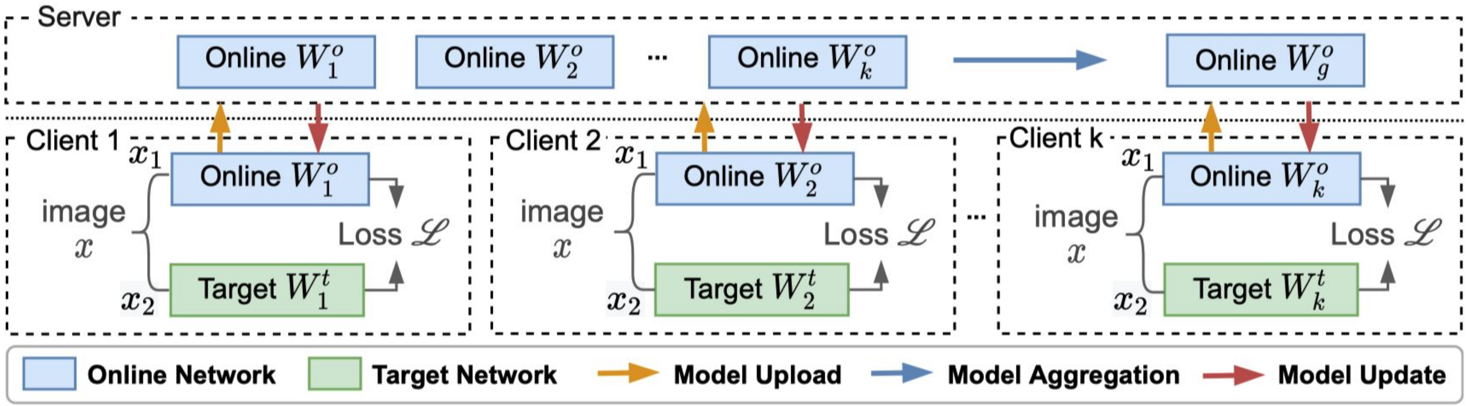
Iterate four key activities until stopping conditions

1. **Local training**: clients conduct self-supervised learning with **Siamese networks**



FedSSL Framework

Federated Self-supervised Learning (FedSSL) Framework

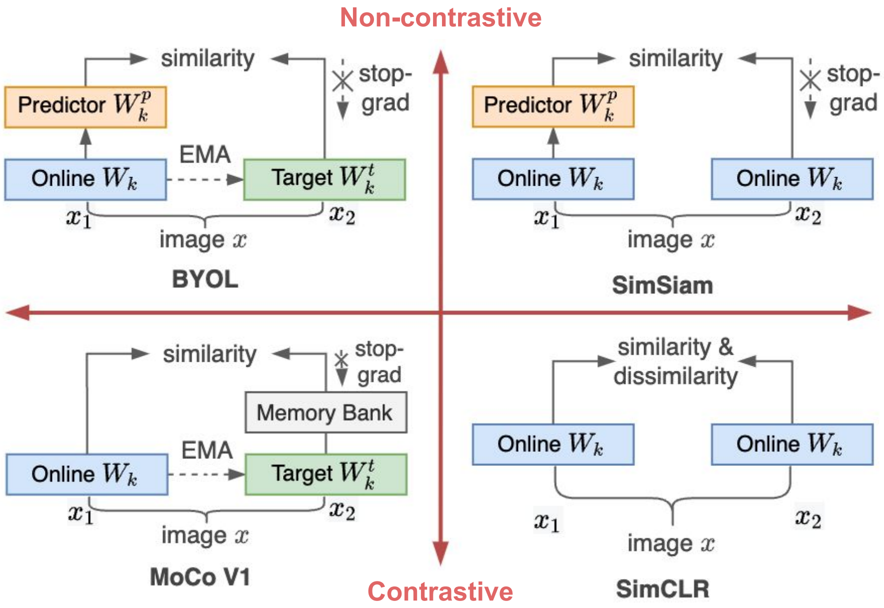


Iterate four key activities until stopping conditions

1. **Local training**: clients conduct self-supervised learning with **Siamese networks**
2. **Model upload**: clients upload trained **online networks** to the server
3. **Model aggregation**: server aggregates them to obtain a new global model
4. **Model update**: server updates clients' **online networks** with the global model

Experiment Setting

Self-supervised Learning Methods

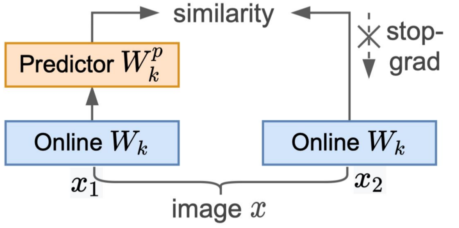


Experimental Setup

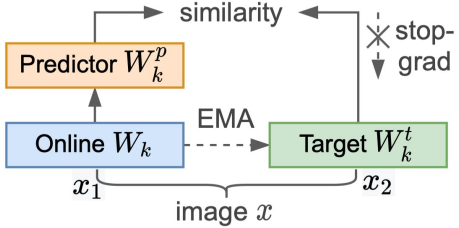
- Dataset: CIFAR10, CIFAR100
- Training Setting:
  - 5 clients, 100 rounds
  - Non-IID:
    - 2 classes/client for CIFAR10
    - 20 classes/client for CIFAR100
- Evaluation
  - Linear evaluation: fine-tune a classifier with frozen model
  - kNN monitoring

Summary of Insights

We focus on **non-contrastive** FedSSL methods



SimSiam

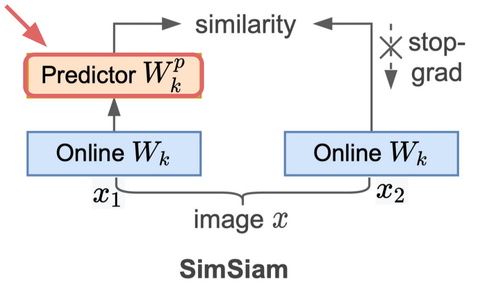


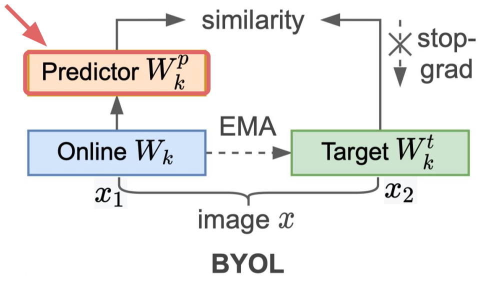
BYOL

Summary of Insights

We focus on **non-contrastive** FedSSL methods:

1. **Predictor** is essential



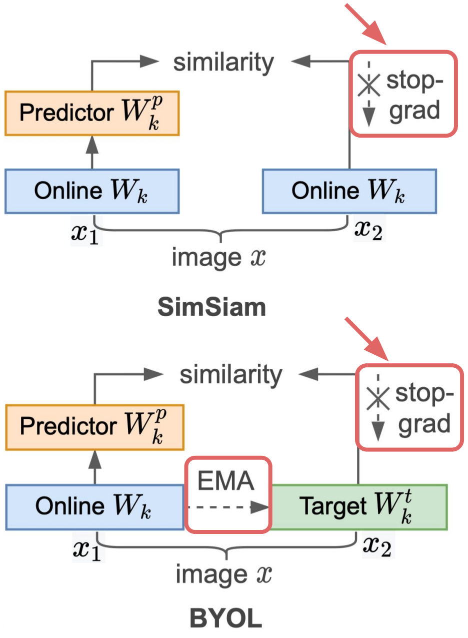


Summary of Insights

We focus on **non-contrastive** FedSSL methods:

1. **Predictor** is essential

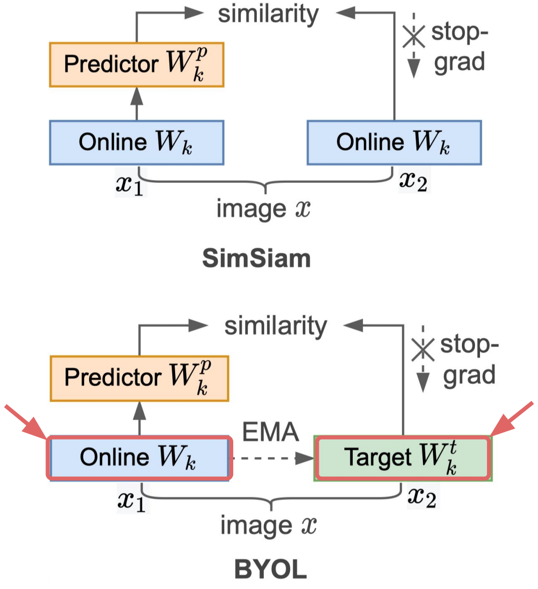
2. **Stop-gradient** and **EMA** help improve performance



Summary of Insights

We focus on **non-contrastive** FedSSL methods:

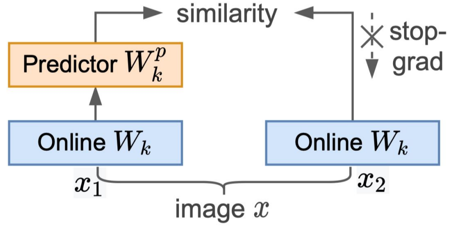
1. **Predictor** is essential
2. **Stop-gradient** and **EMA** help improve performance
3. **Local encoders** should retain local knowledge of non-IID data



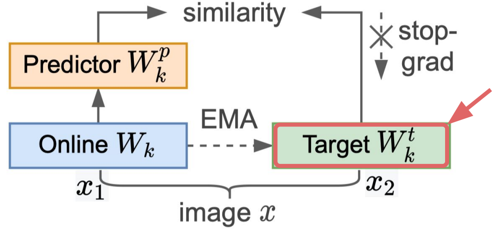
Summary of Insights

We focus on **non-contrastive** FedSSL methods:

1. **Predictor** is essential
2. **Stop-gradient** and **EMA** help improve performance
3. **Local encoders** should retain local knowledge of non-IID data
4. **Target encoder** should gain knowledge from online encoder



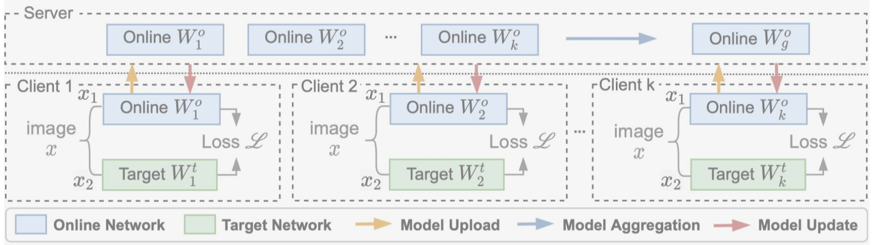
SimSiam



BYOL

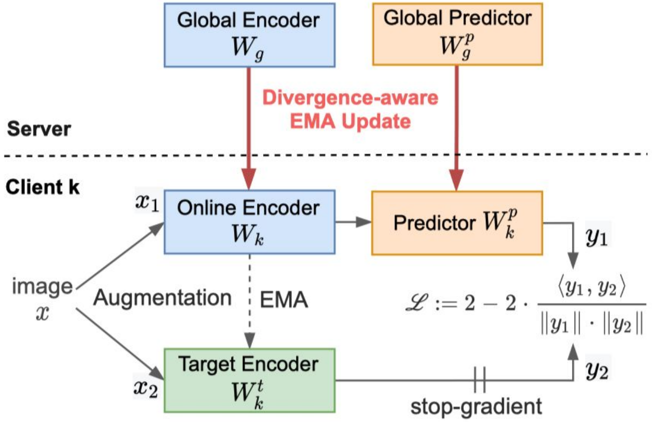
Part 1

FedSSL framework for in-depth Studies



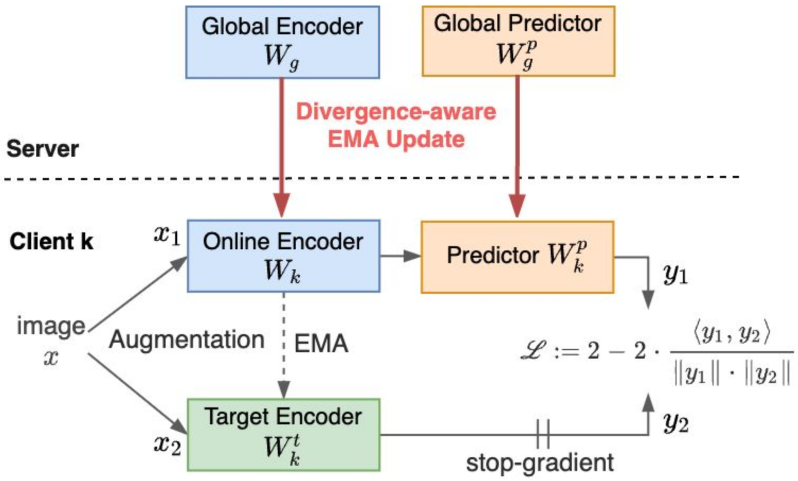
Part 2

Divergence-aware EMA Update



Divergence-aware Exponential Moving Average

Federated Divergence-aware Exponential Moving Average (FedEMA)



Encoder Update: $W_k^r = \mu W_k^{r-1} + (1-\mu) W_g^r,$

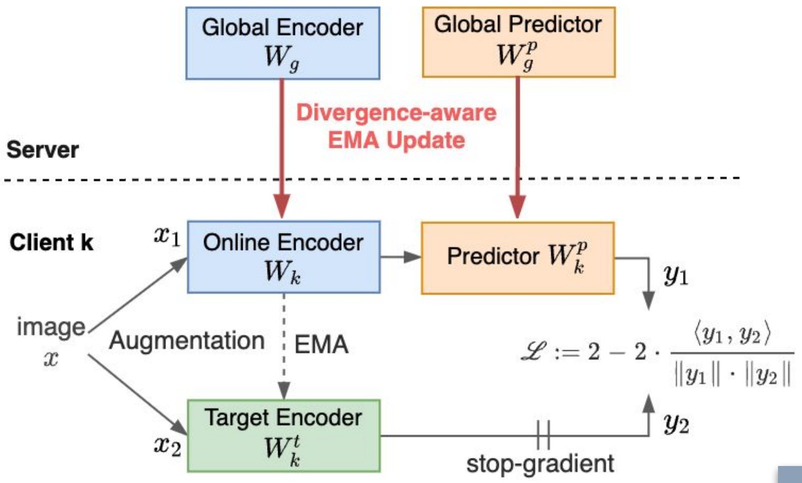
Predictor Update: $W_k^{p,r} = \mu W_k^{p,r-1} + (1-\mu) W_g^{p,r},$

Iterate four key activities until stopping conditions

1. **Local training**: clients conduct training using **BYOL**
2. **Model upload**: clients upload trained online network
3. **Model aggregation**: server aggregates them to obtain a new **global encoder** and **predictor**
4. **Model update**: server updates clients' **online encoder and predictor with exponential moving average** of the global model

Divergence-aware Exponential Moving Average

Federated Divergence-aware Exponential Moving Average (FedEMA)



Encoder Update: $W_k^r = \mu W_k^{r-1} + (1-\mu) W_g^r,$

Predictor Update: $W_k^{p,r} = \mu W_k^{p,r-1} + (1-\mu) W_g^{p,r},$

Decay rate: $\mu = \min(\lambda \left\| W_g^r - W_k^{r-1} \right\|, 1),$

Autoscaler to calculate personalized λ_k for client k

$$\lambda_k = \frac{\tau}{||W_g^{r+1} - W_k^r||}, \text{ where } \tau \in [0, 1]$$

Intuition

Large divergence: retain more local knowledge

Small divergence: incorporate more global knowledge

Evaluation

Linear Evaluation

Method	CIFAR-10 (%)		CIFAR-100 (%)	
	IID	Non-IID	IID	Non-IID
Standalone training	82.42 ± 0.32	74.95 ± 0.66	53.88 ± 2.24	52.37 ± 0.93
FedBYOL	84.29 ± 0.18	79.44 ± 0.99	54.24 ± 0.24	57.51 ± 0.09
FedU (Zhuang et al., 2021a)	83.96 ± 0.18	80.52 ± 0.21	54.82 ± 0.67	57.21 ± 0.31
FedEMA ($\lambda = 0.8$)	$\mathbf{85.59 \pm 0.25}$	$\mathbf{82.77 \pm 0.08}$	$\mathbf{57.86 \pm 0.15}$	$\mathbf{61.21 \pm 0.54}$
FedEMA (autoscaler, $\tau = 0.7$)	$\mathbf{86.26 \pm 0.26}$	$\mathbf{83.34 \pm 0.39}$	$\mathbf{58.55 \pm 0.34}$	$\mathbf{61.78 \pm 0.14}$

- Our proposed FedEMA outperforms all other methods
- Autoscaler is more effective than manually tuned λ

Semi-supervised Evaluation

Method	CIFAR-10 (%)		CIFAR-100 (%)	
	1%	10%	1%	10%
Standalone training	61.37 ± 0.13	69.06 ± 0.24	21.37 ± 0.73	39.99 ± 0.87
FedBYOL	70.48 ± 0.30	76.95 ± 0.46	30.21 ± 0.40	47.07 ± 0.14
FedU (Zhuang et al., 2021a)	69.52 ± 0.73	77.06 ± 0.55	29.00 ± 0.27	46.67 ± 0.06
FedEMA ($\lambda = 1$)	$\mathbf{72.78 \pm 0.66}$	$\mathbf{79.01 \pm 0.30}$	$\mathbf{32.49 \pm 0.22}$	$\mathbf{49.82 \pm 0.36}$
FedEMA (autoscaler, $\tau = 0.7$)	$\mathbf{73.44 \pm 0.22}$	$\mathbf{79.49 \pm 0.34}$	$\mathbf{33.04 \pm 0.23}$	$\mathbf{50.48 \pm 0.11}$

- Our proposed FedEMA outperforms all other methods
- Autoscaler is more effective than manually tuned λ

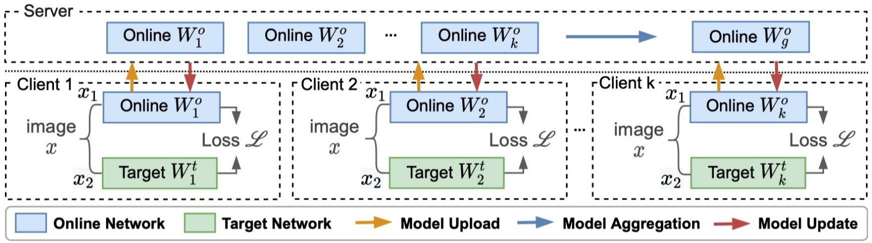
Random Select K From N clients

Method	5/20 clients (%)				8/80 clients (%)			
	CIFAR-10		CIFAR-100		CIFAR-10		CIFAR-100	
	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
FedBYOL	83.25	74.92	49.49	47.09	73.58	63.28	41.19	41.58
FedEMA (ours)	**84.98**	**75.77**	**55.41**	**52.78**	**73.96**	**64.19**	**41.97**	**43.05**

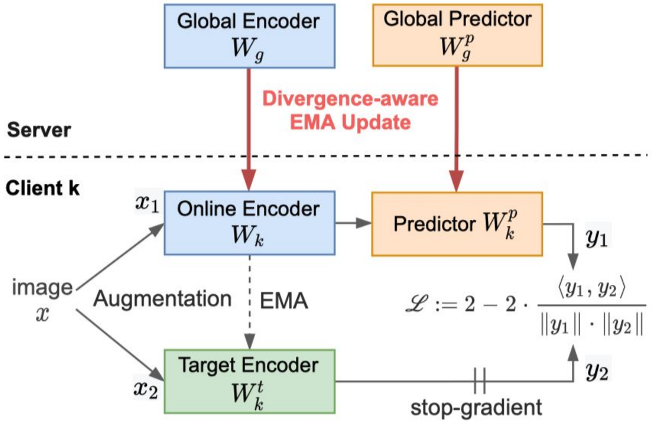
- Our method also works when randomly selecting K from N clients (K/N)

Conclusion

FedSSL Framework for In-depth Studies



FedEMA Algorithm



https://weiming.me/publication/fedssl/

Thank You