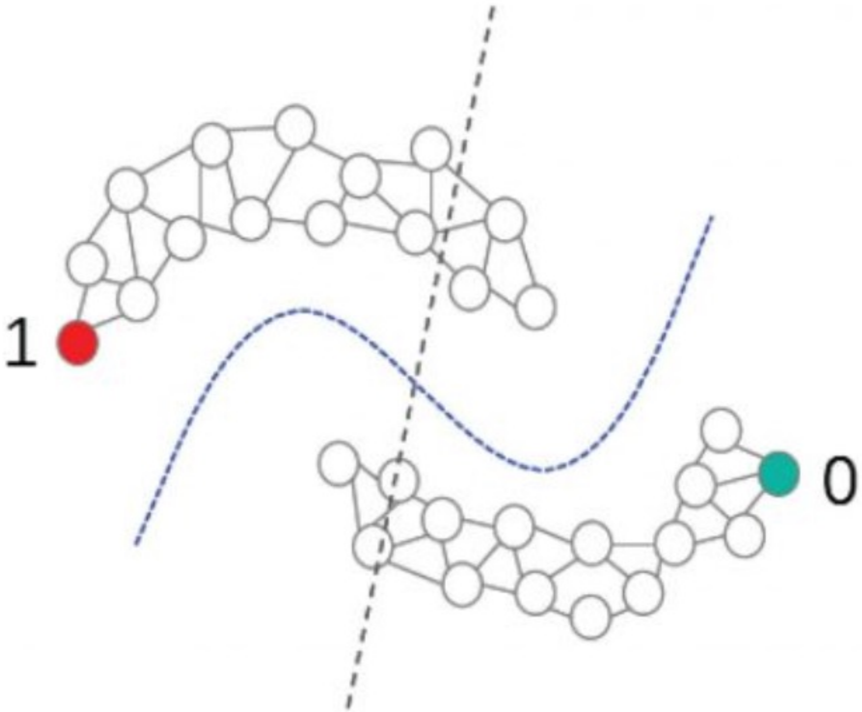
Open Problems in Deep Learning

Neural Networks Design And Application

Open problems in deep learning

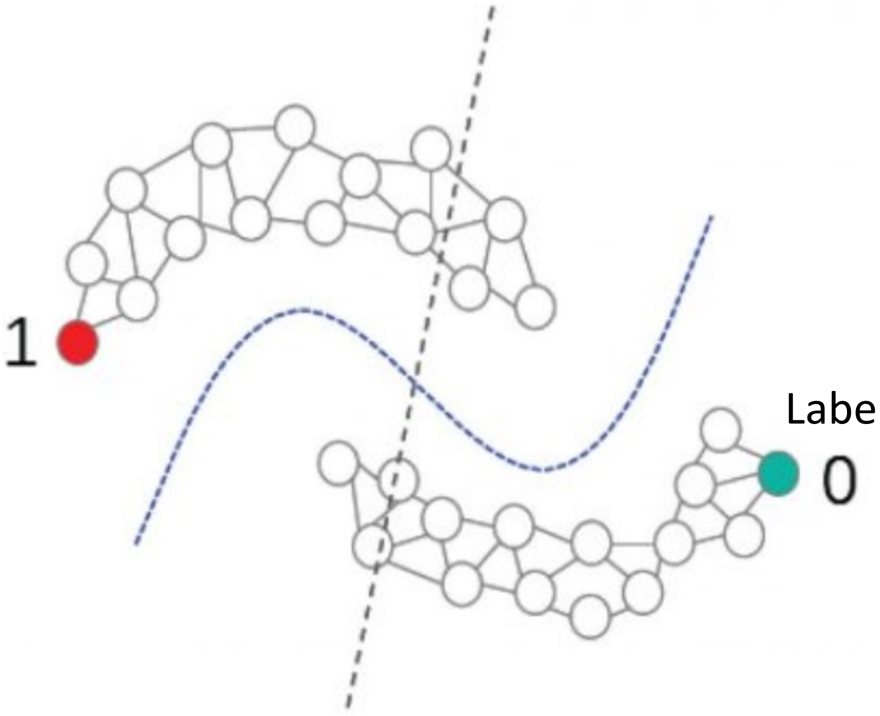
- Weak supervised learning
- Causal modeling
- Curriculum learning
- Over-parameterized modeling
- Self-supervised learning
- Meta-learning
- Federated learning
- ……..

Weak supervision



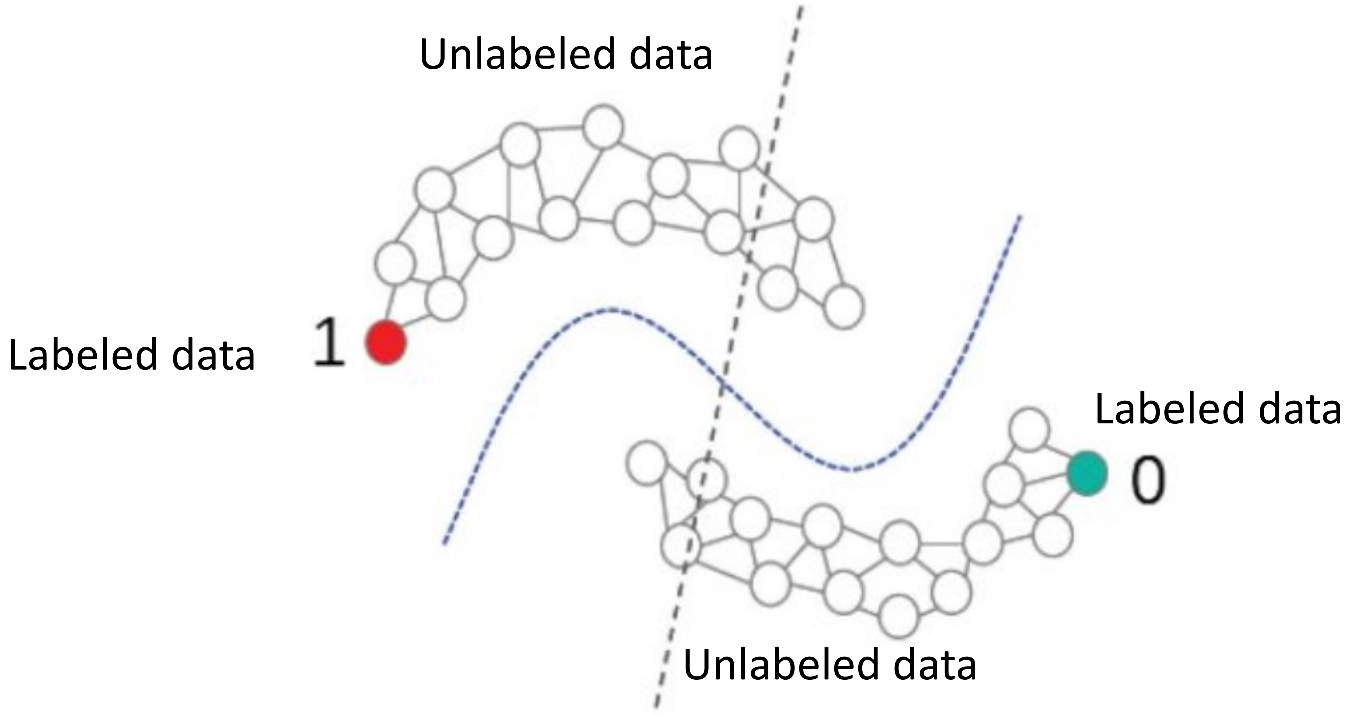
Weak supervision

Labeled data



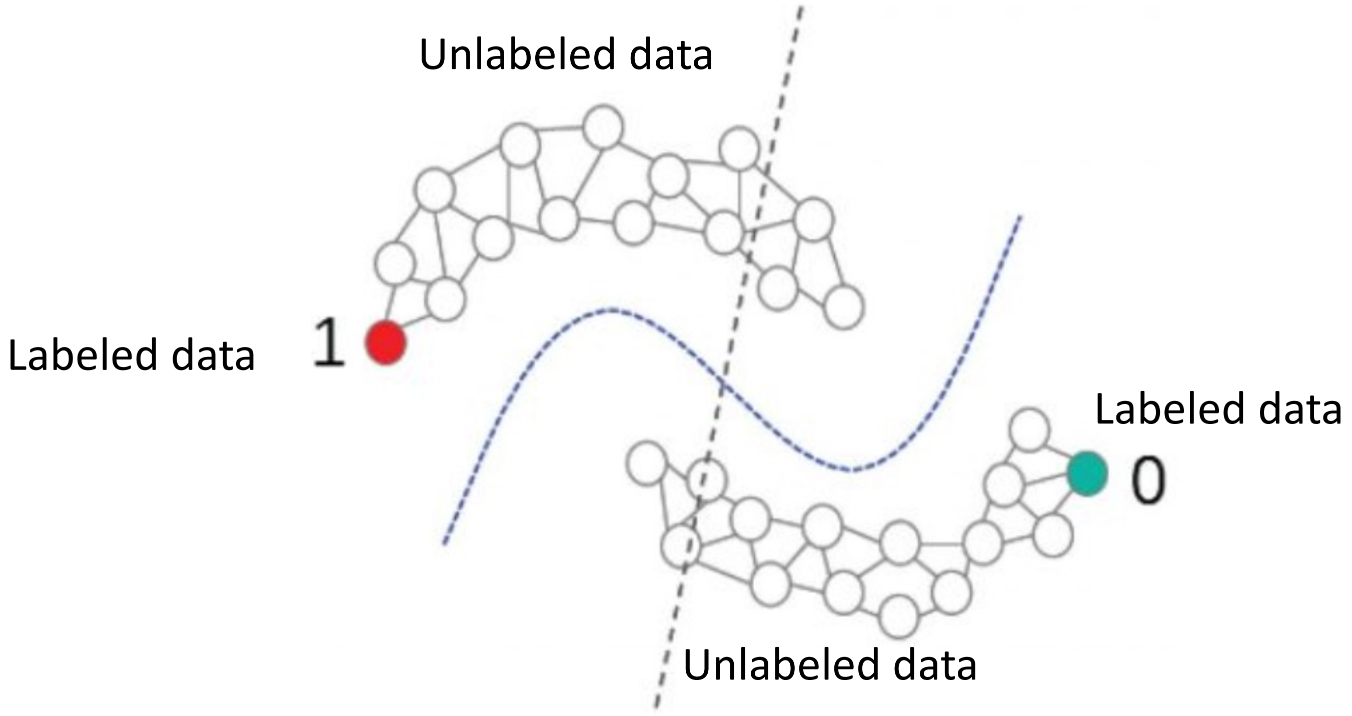
Labeled data

Weak supervision

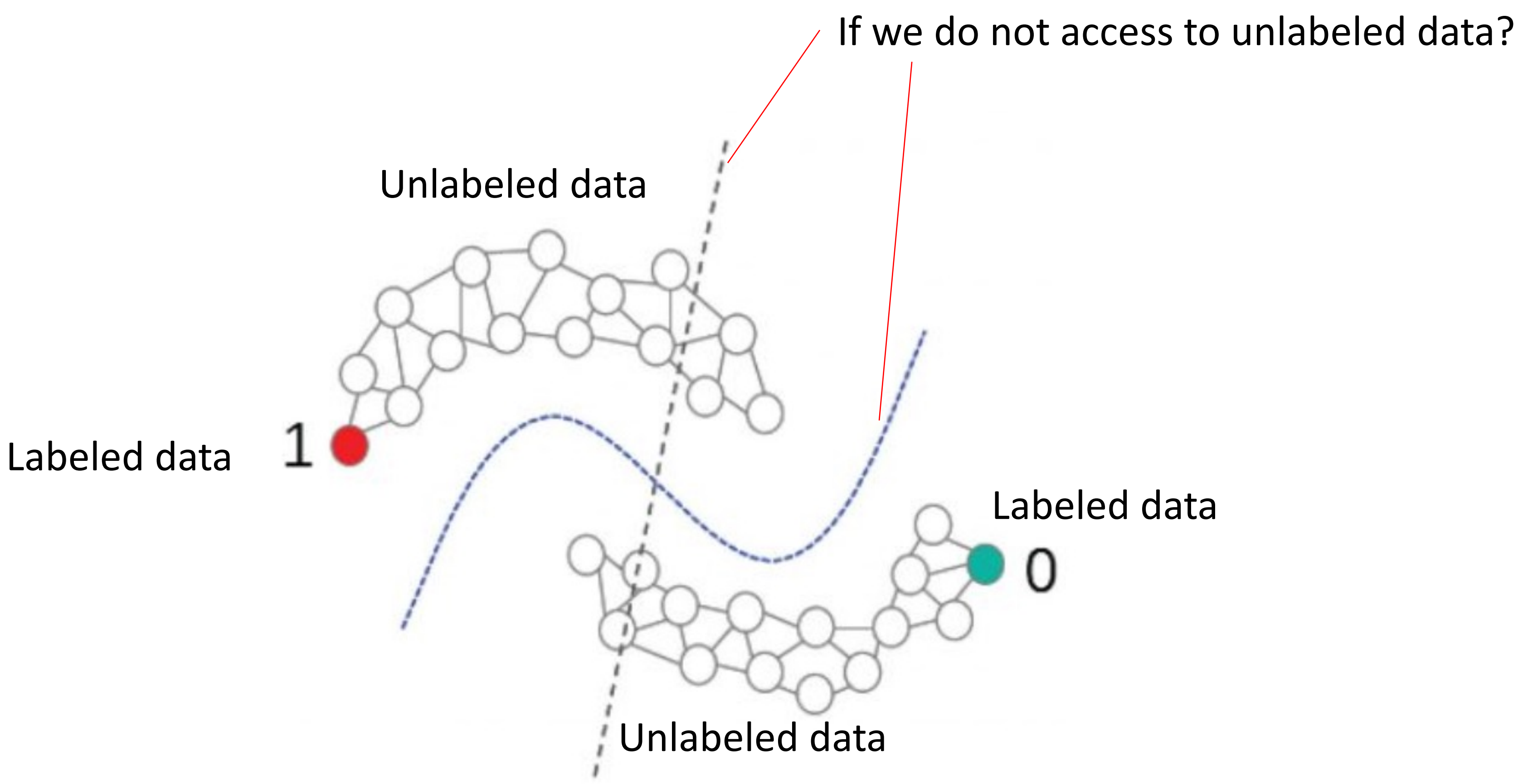


Weak supervision

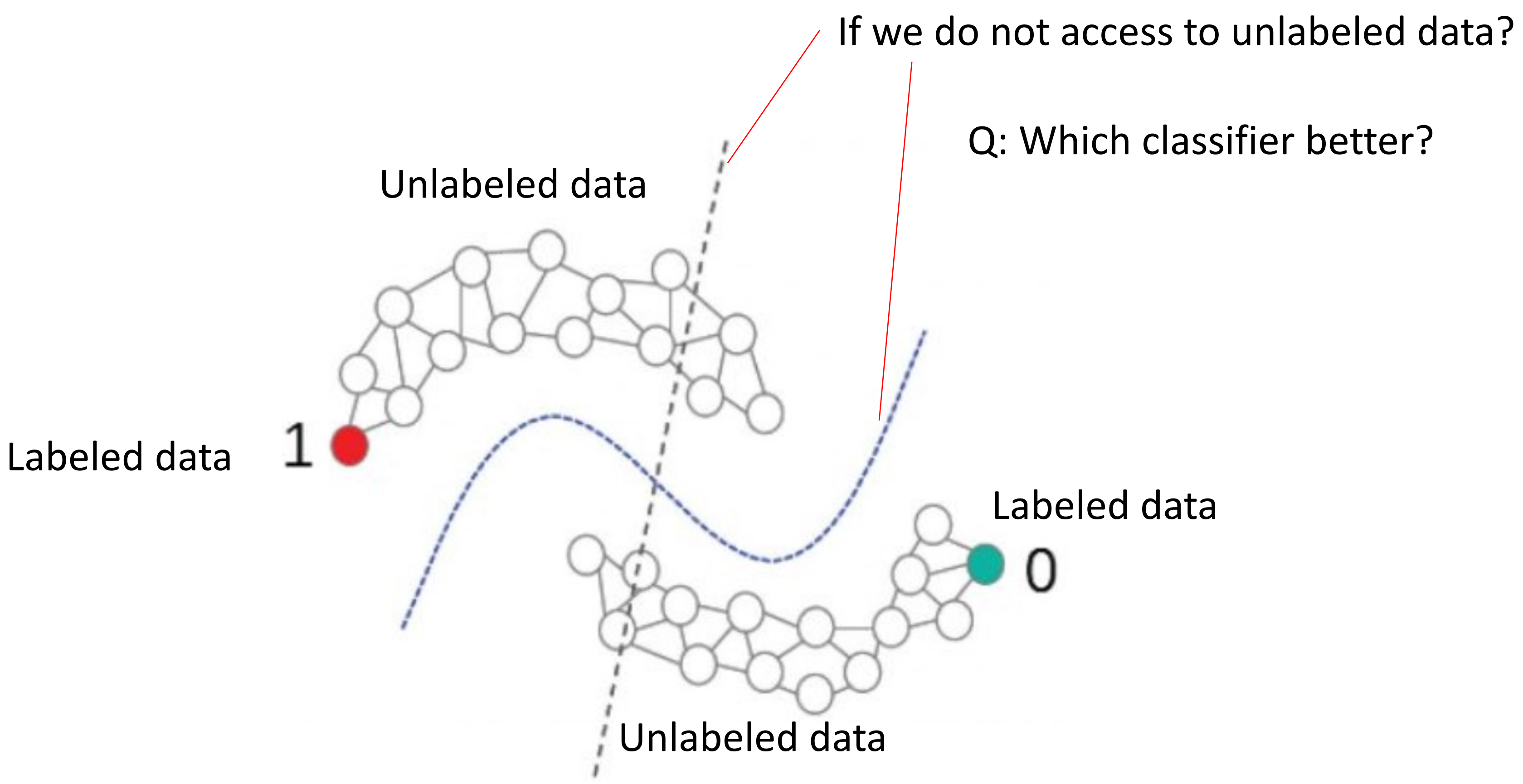
If we do not access to unlabeled data?



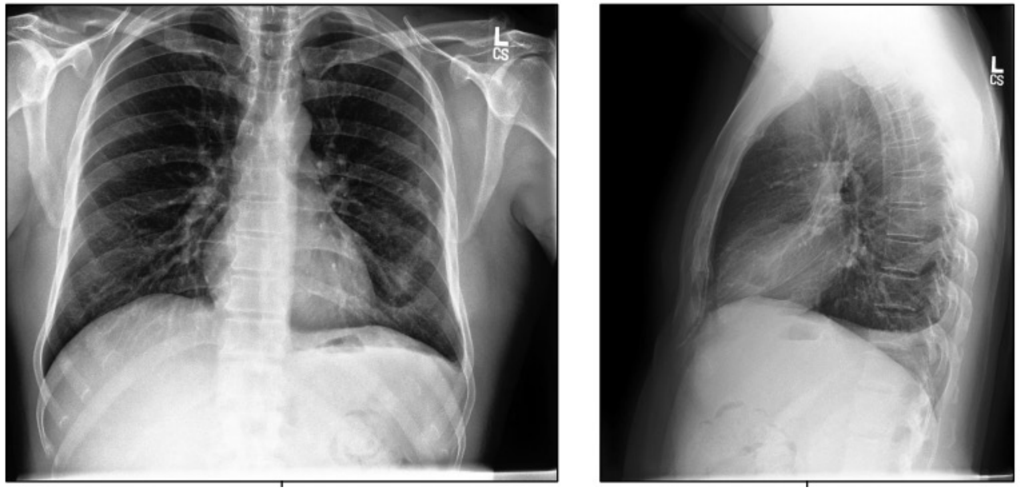
Weak supervision



Weak supervision



Weak supervision

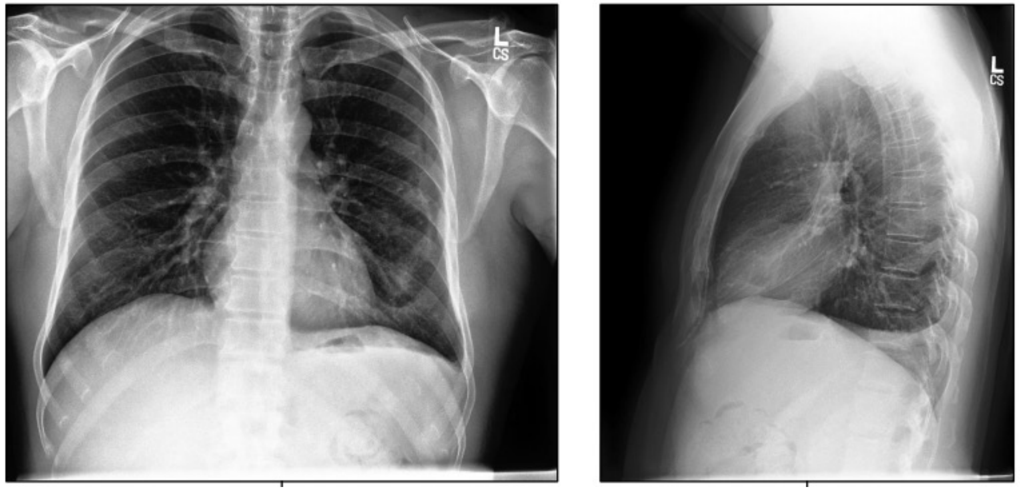


X-ray image analysis

Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590-597. 2019.
https://arxiv.org/pdf/1901.07031.pdf

Weak supervision

Some classification tasks are very difficult to acquire a lot of labeled data

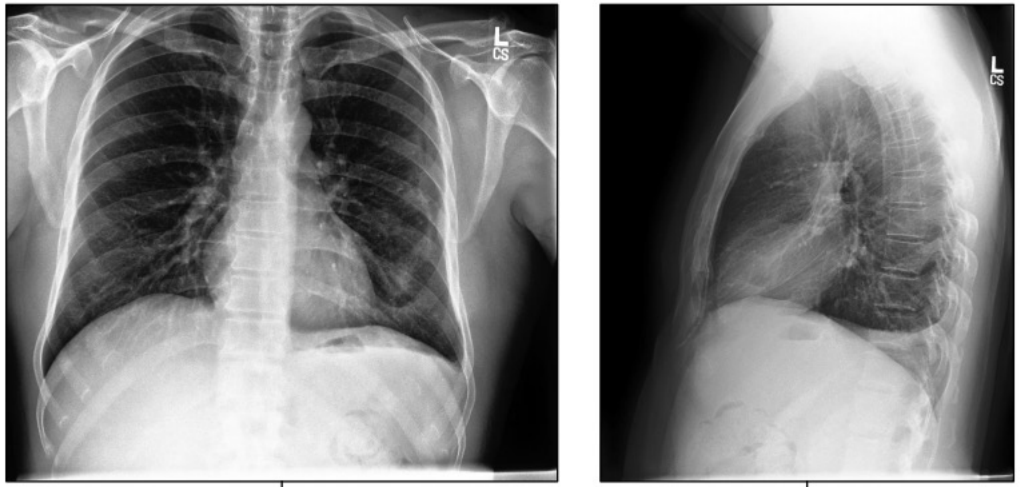


X-ray image analysis

Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590-597. 2019.
https://arxiv.org/pdf/1901.07031.pdf

Weak supervision

Some classification tasks are very difficult to acquire a lot of labeled data, but may provide many unlabeled data

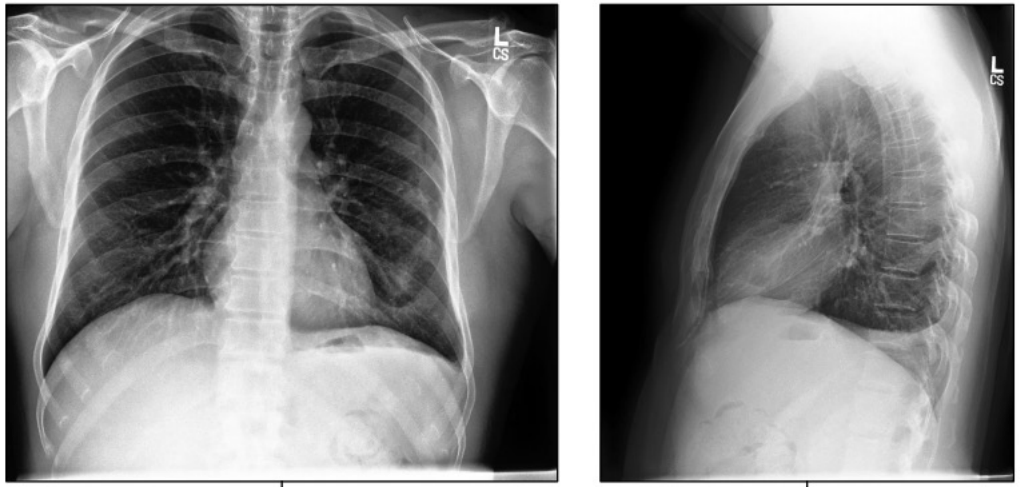


X-ray image analysis

Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590-597. 2019.
https://arxiv.org/pdf/1901.07031.pdf

Weak supervision

Some classification tasks are very difficult to acquire a lot of labeled data, but may provide many unlabeled data

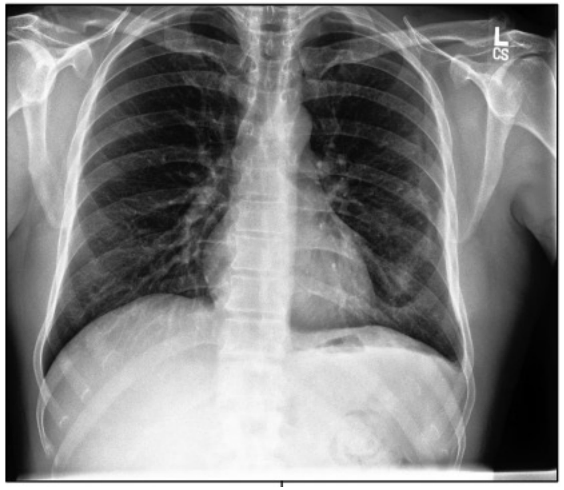
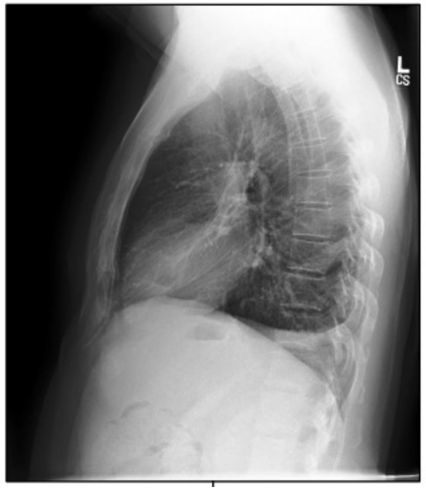


X-ray image analysis

Identify: which x-ray image may imply life threatening diseases?

Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590-597. 2019.
https://arxiv.org/pdf/1901.07031.pdf

Weak supervision

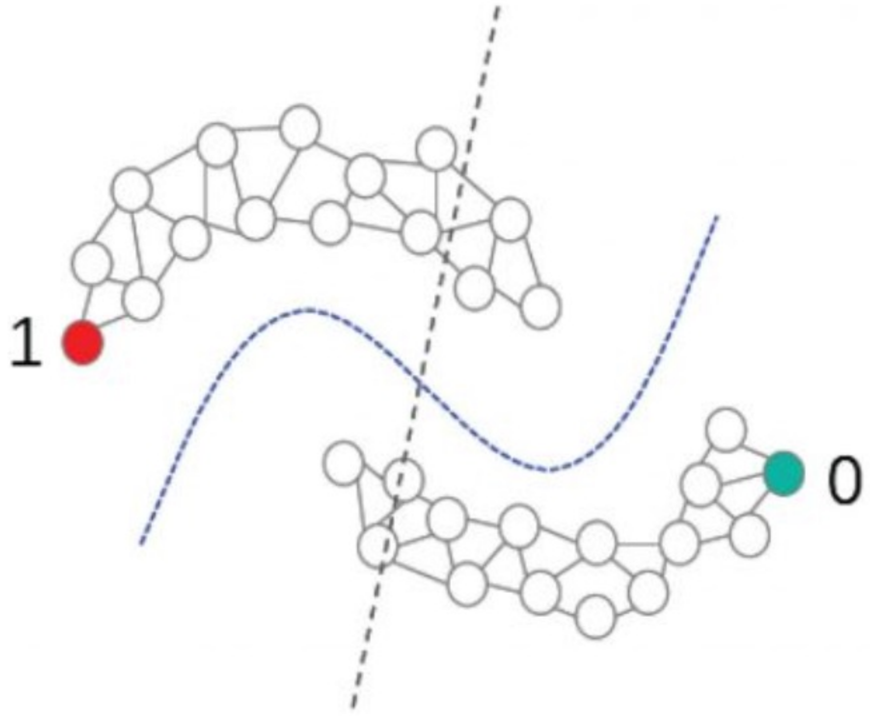
Leaderboard

Will your model perform as well as radiologists in detecting different pathologies in chest X-rays?

Rank	Date	Model	AUC	Num Rads Below Curve
1	Aug 31, 2020	DeepAUC-v1 *ensemble* https://arxiv.org/abs/2012.03173	**0.930**	2.8
2	Sep 01, 2019	Hierarchical-Learning-V1 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.930	2.6
3	Oct 15, 2019	Conditional-Training-LSR *ensemble*	0.929	2.6
4	Dec 04, 2019	Hierarchical-Learning-V4 (ensemble) *Vingroup*	0.929	2.6

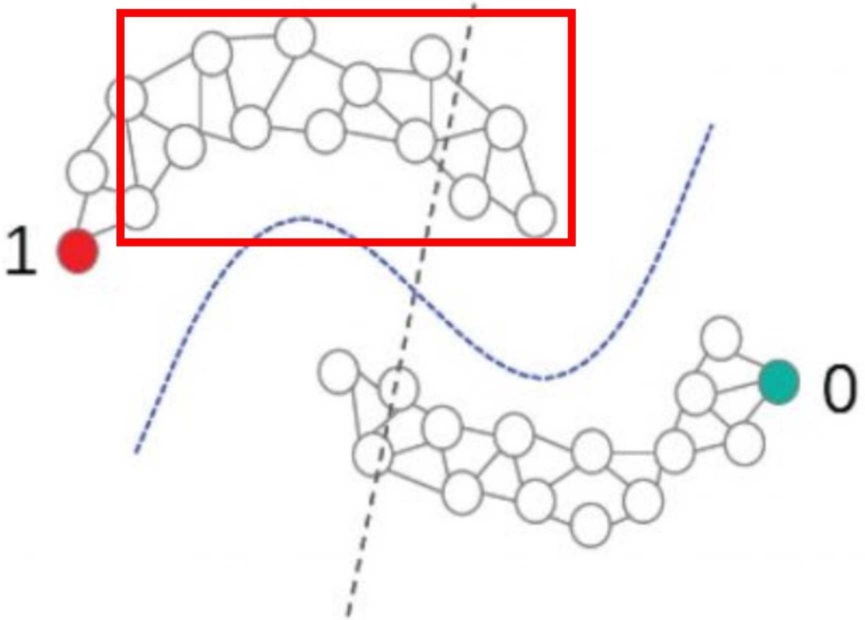
https://stanfordmlgroup.github.io/competitions/chexpert/

Weak supervision



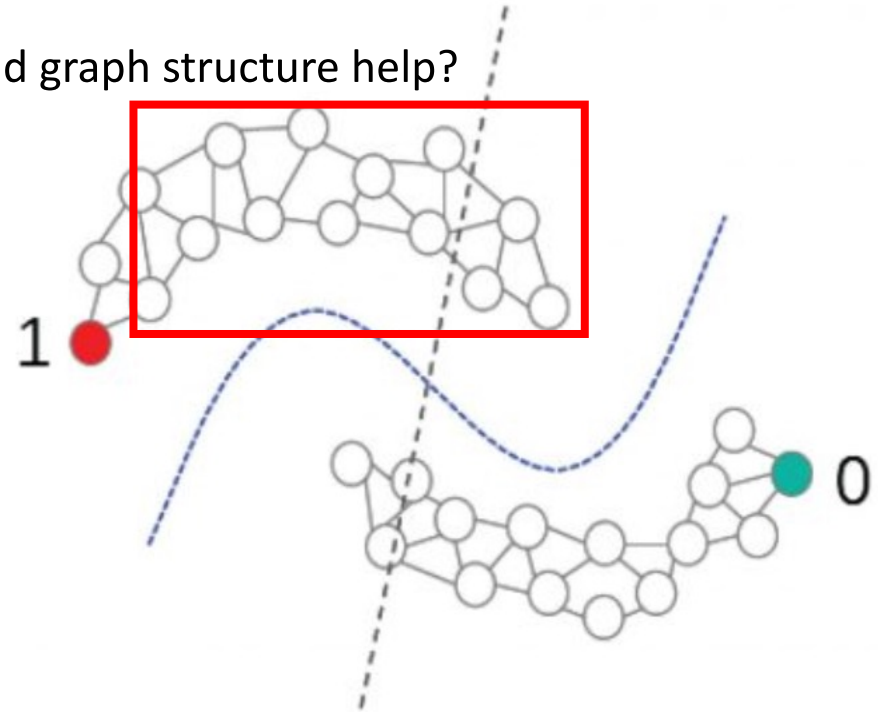
Weak supervision

Q: how to use unlabeled data?



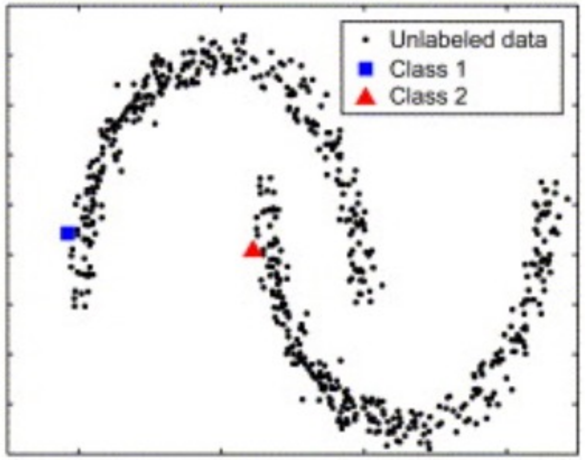
Weak supervision

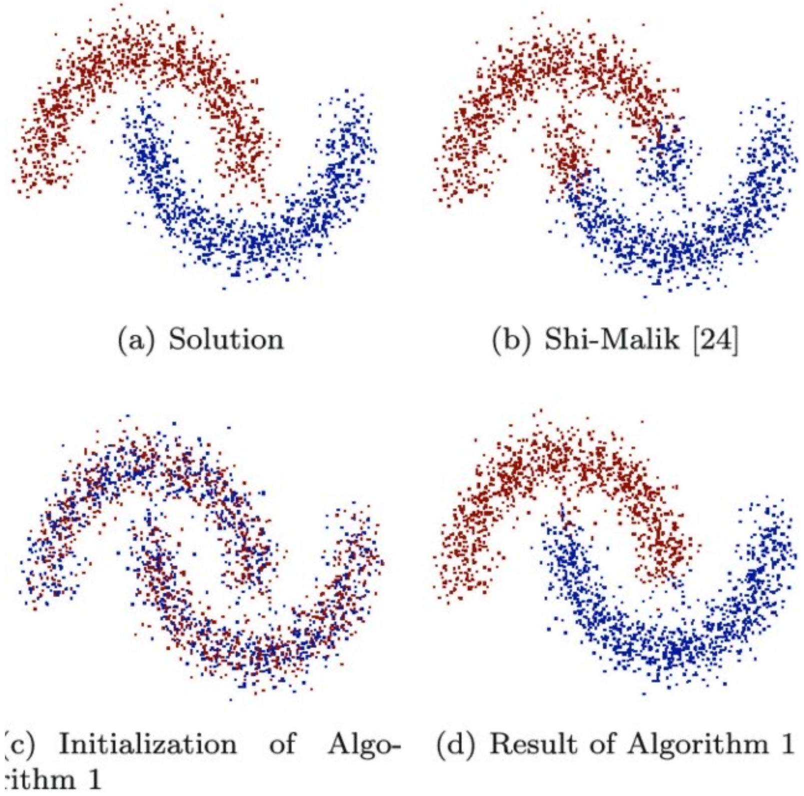
Q: Would graph structure help?



Weak supervision

Two moon dataset

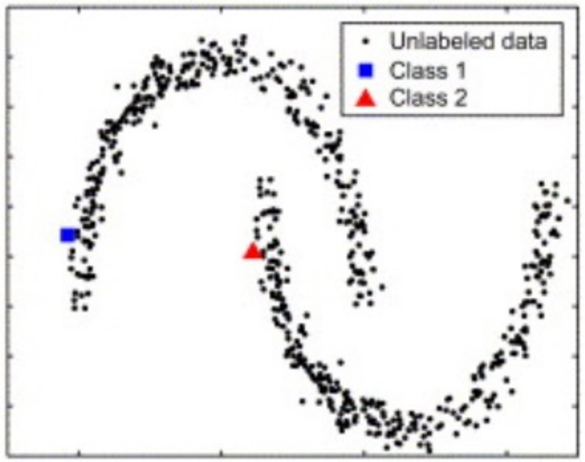


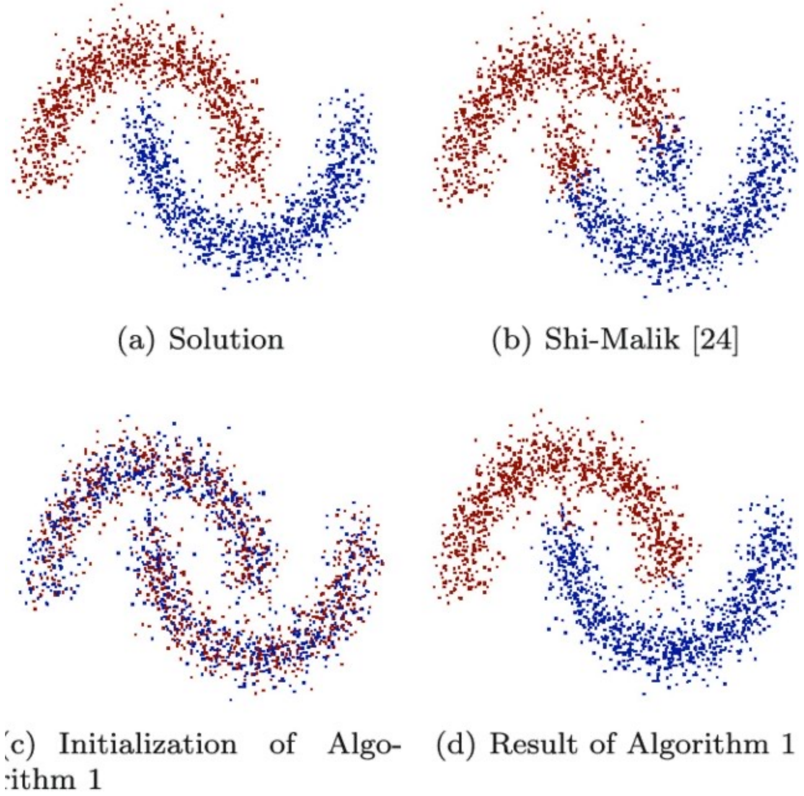


Multi-class Transductive Learning Based on ℓ 1 Relaxations of Cheeger Cut and Mumford-Shah-Potts Model. May 2014
Journal of Mathematical Imaging and Vision 49(1). DOI:10.1007/s10851-013-0452-5

Weak supervision

Two moon dataset



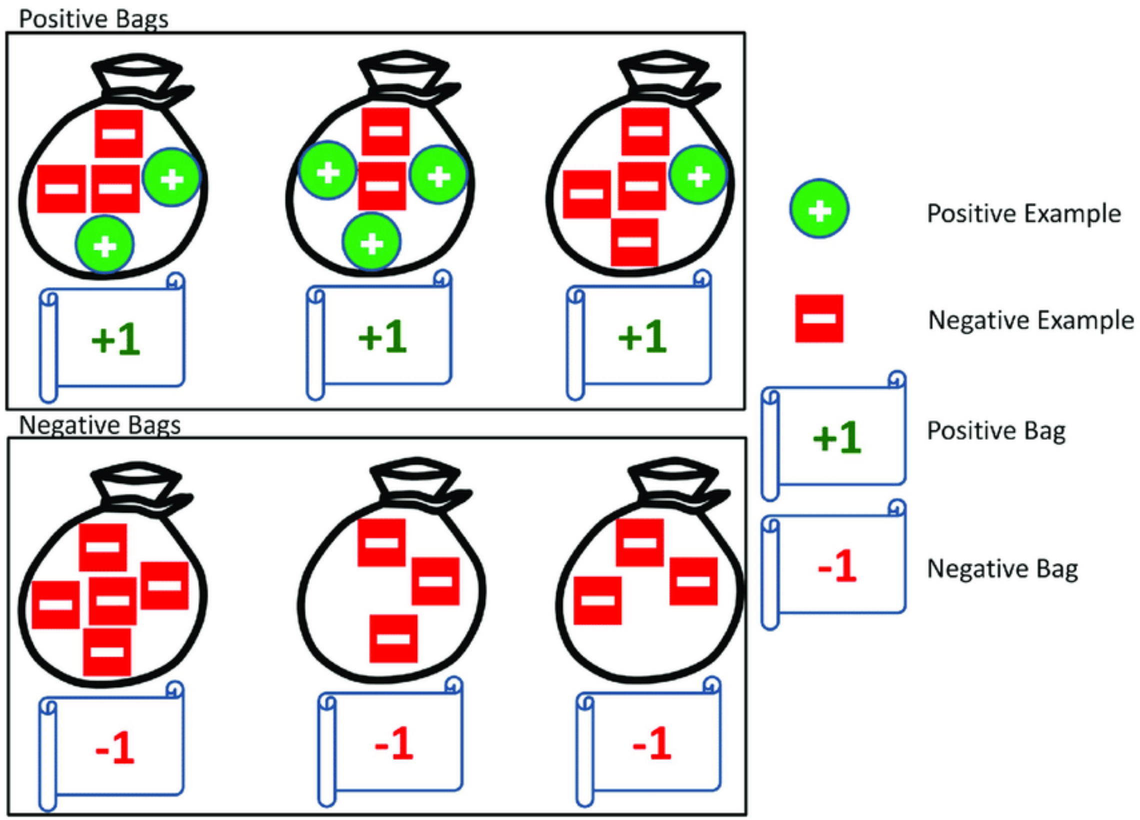


Labeled + unlabeled data ➔ semi-supervised learning

Multi-class Transductive Learning Based on ℓ 1 Relaxations of Cheeger Cut and Mumford-Shah-Potts Model. May 2014 Journal of Mathematical Imaging and Vision 49(1). DOI:[10.1007/s10851-013-0452-5](10.1007/s10851-013-0452-5)

Weak supervision

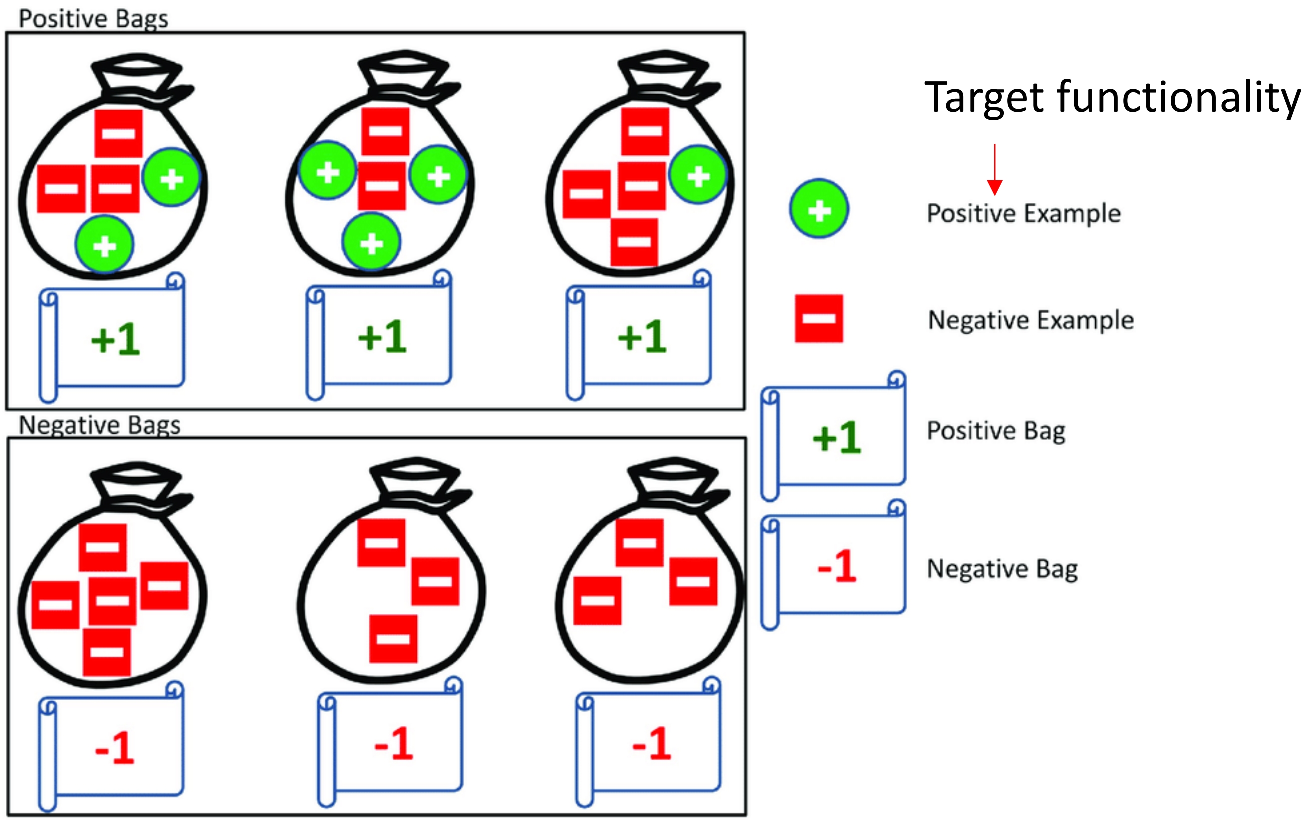
Multi-instance learning (another type of weak supervised learning)



Amino acid composition predicts prion activity. April 2017
PLoS Computational Biology 13(4):e1005465. DOI:10.1371/journal.pcbi.1005465

Weak supervision

Multi-instance learning (another type of weak supervised learning)



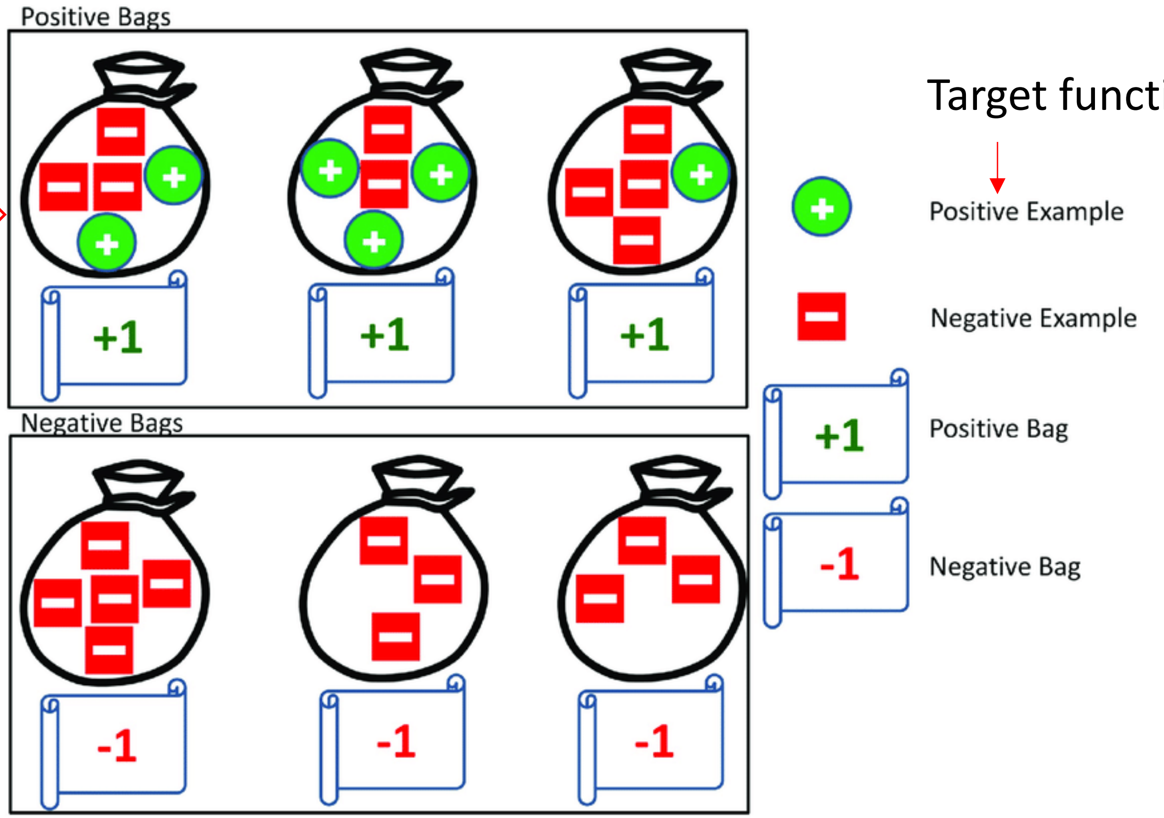
Amino acid composition predicts prion activity. April 2017
PLoS Computational Biology 13(4):e1005465. DOI:10.1371/journal.pcbi.1005465

Weak supervision

Multi-instance learning (another type of weak supervised learning)

Positive bag
(providing functionality)

Target functionality



Amino acid composition predicts prion activity. April 2017
PLoS Computational Biology 13(4):e1005465. DOI:10.1371/journal.pcbi.1005465

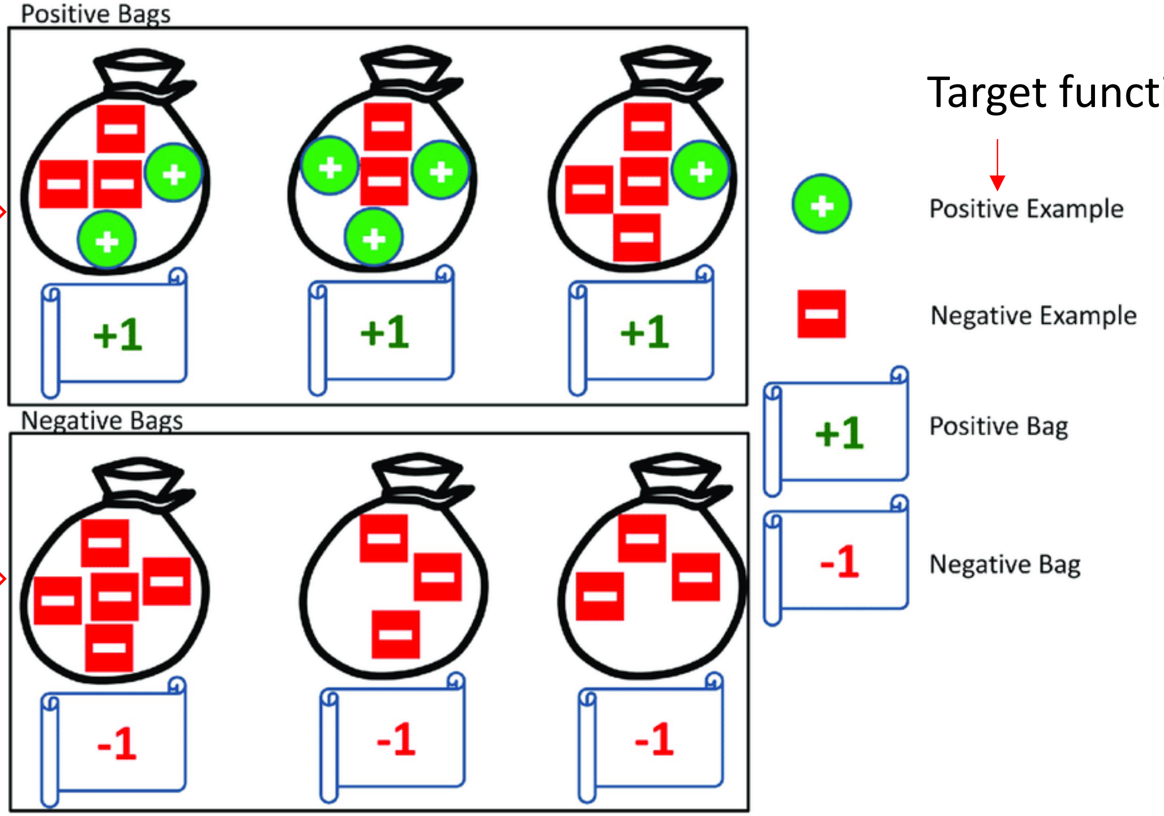
Weak supervision

Multi-instance learning (another type of weak supervised learning)

Positive bag
(providing functionality)

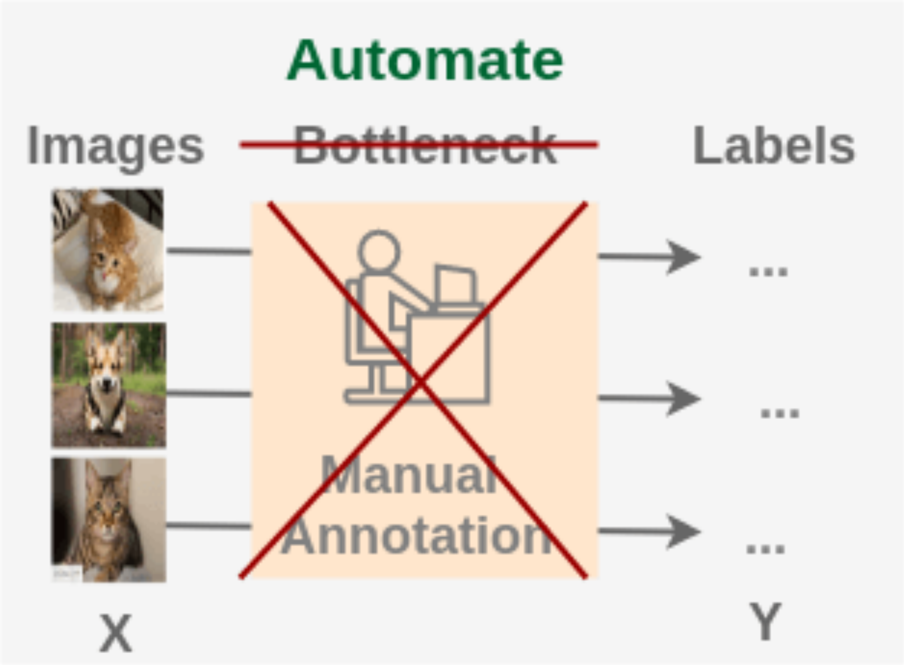
Negative bag
(not providing functionality)

Target functionality



Amino acid composition predicts prion activity. April 2017
PLoS Computational Biology 13(4):e1005465. DOI:10.1371/journal.pcbi.1005465

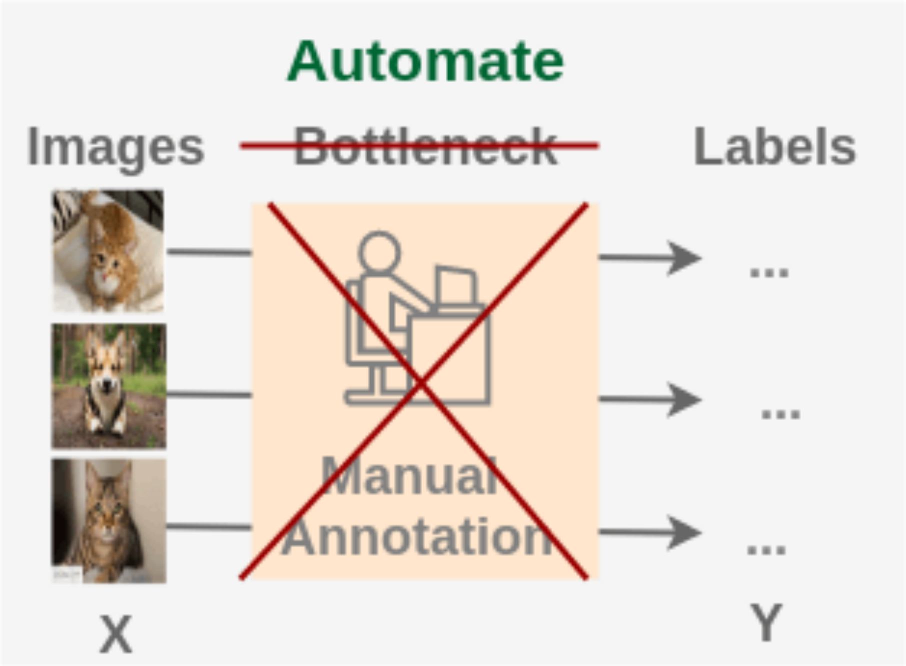
Self-supervision



Without human effort for labeling

https://amitness.com/2020/02/illustrated-self-supervised-learning/

Self-supervision

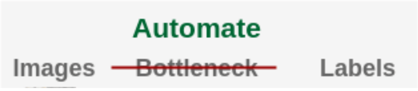


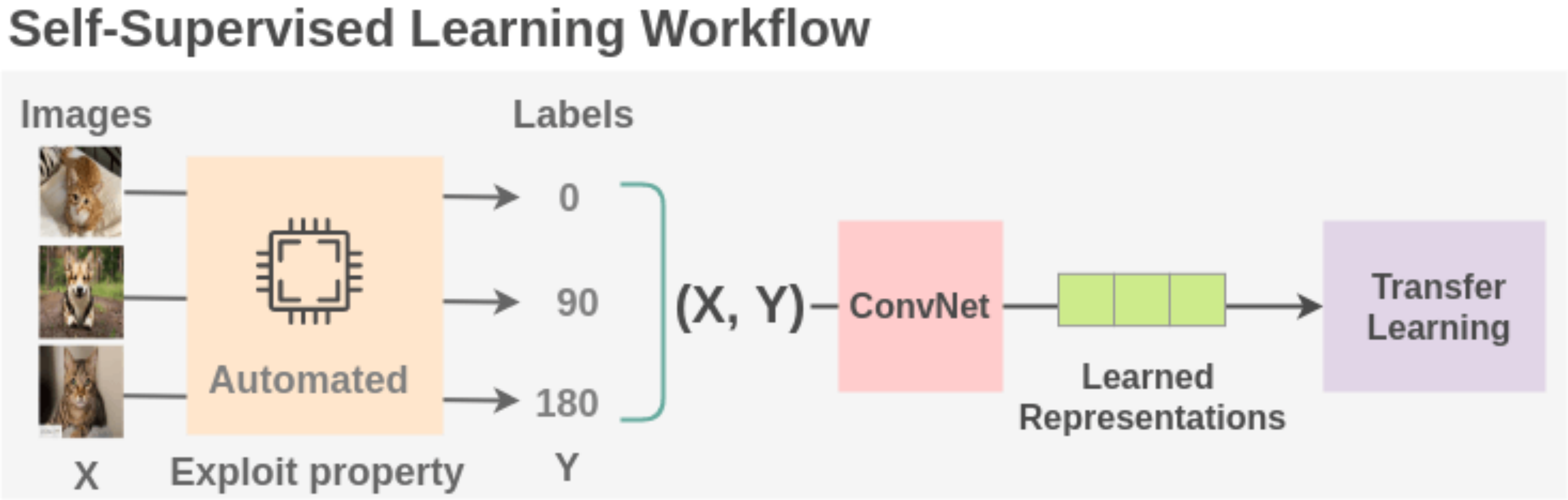
Without human effort for labeling

Q: can we automate labeling to get approximation to label information?

https://amitness.com/2020/02/illustrated-self-supervised-learning/

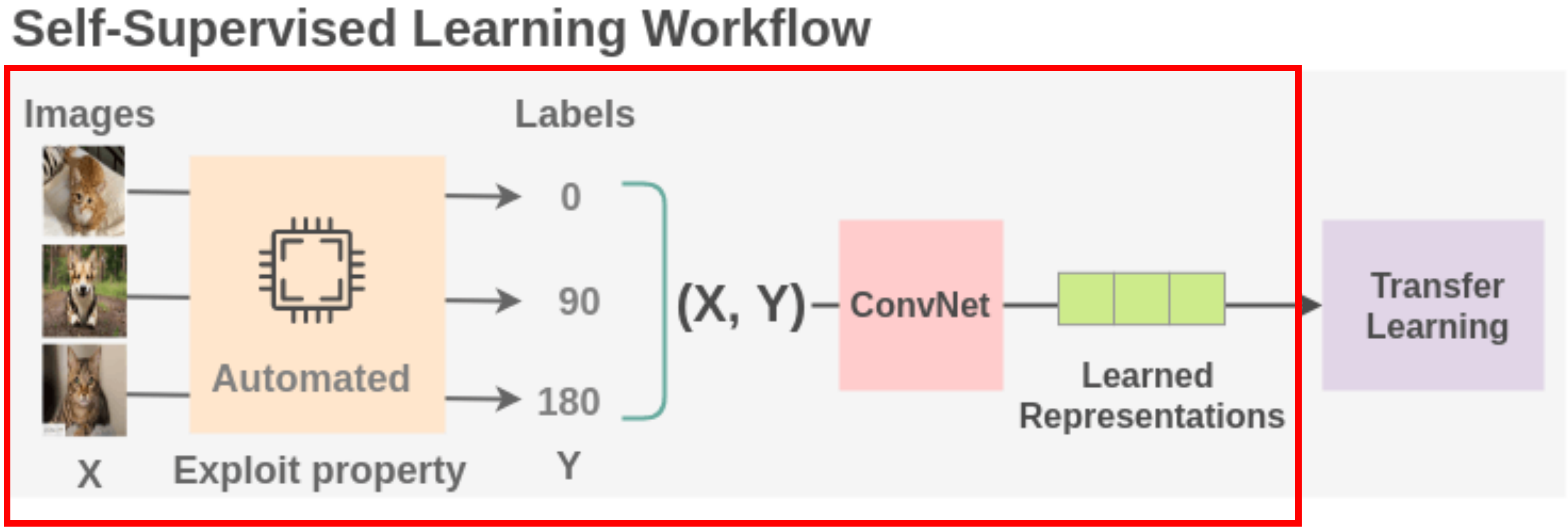
Self-supervision





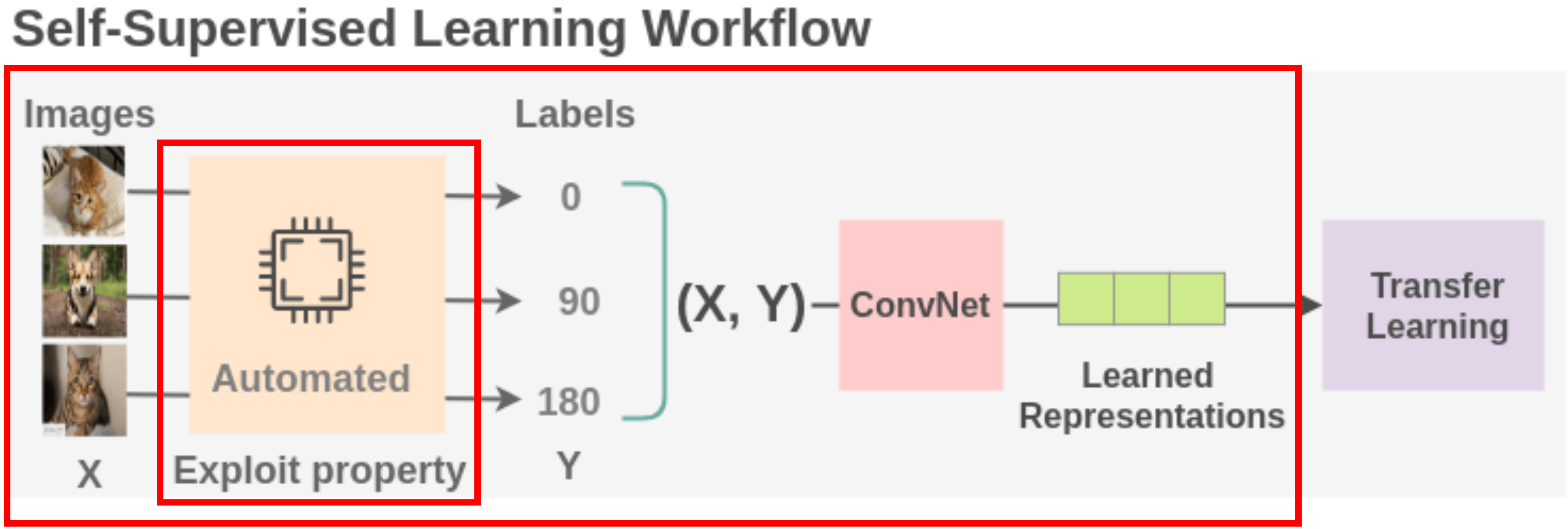
https://amitness.com/2020/02/illustrated-self-supervised-learning/

Self-supervision



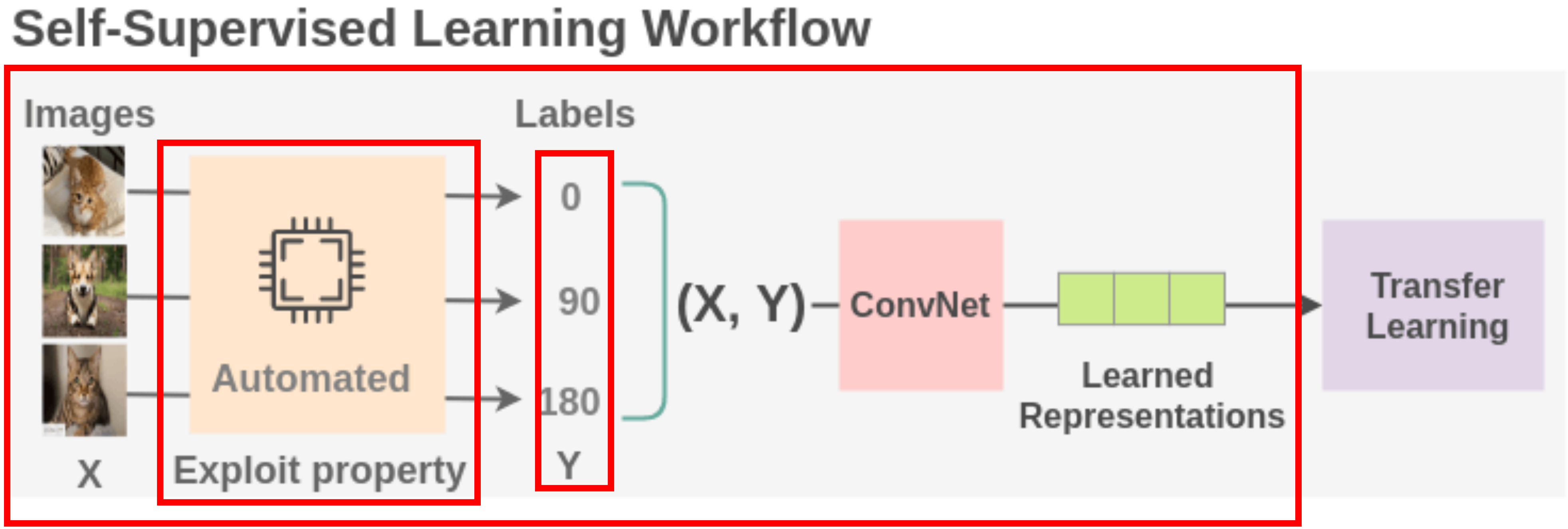
https://amitness.com/2020/02/illustrated-self-supervised-learning/

Self-supervision



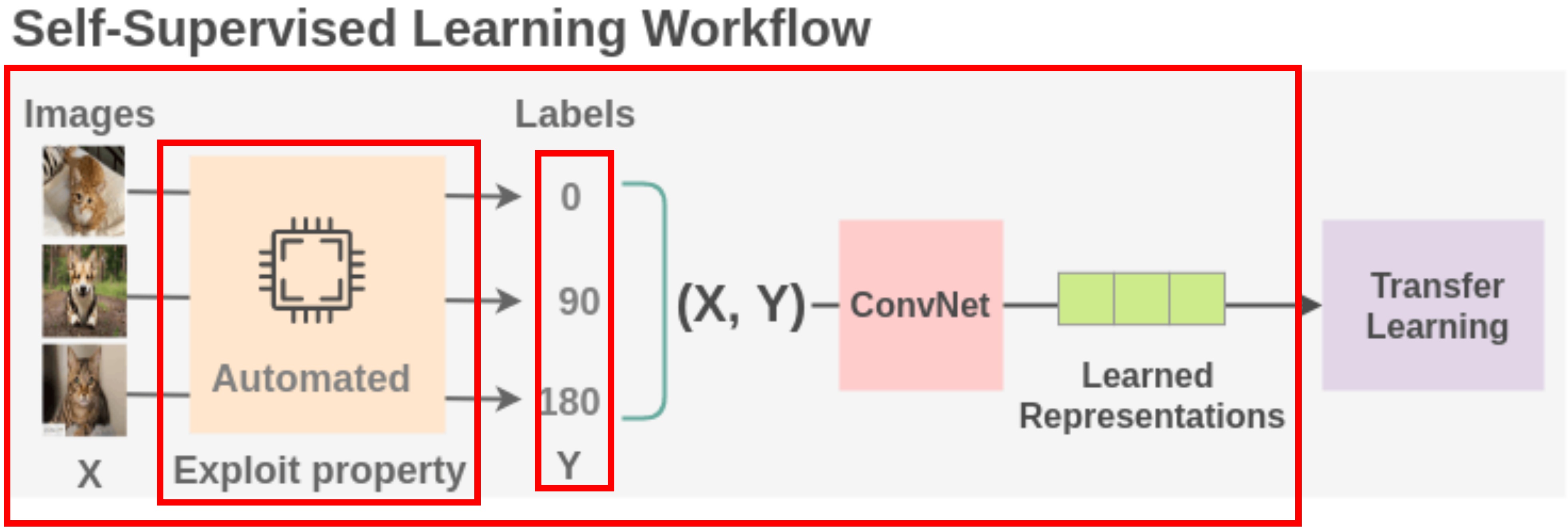
https://amitness.com/2020/02/illustrated-self-supervised-learning/

Self-supervision



https://amitness.com/2020/02/illustrated-self-supervised-learning/

Self-supervision



Hypothesis: if the approximated labels behave exactly the same with the underlying labels, then we recover a fully supervised learning

https://amitness.com/2020/02/illustrated-self-supervised-learning/

Self-supervision: contrastive learning

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f\left(x^+\right)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}}\right)\right]$$

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

An anchor sample

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right]$$

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

An anchor sample

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

An anchor sample

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

An anchor sample

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample

Data augmentation: small variation → not change semantic → not change label

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

An anchor sample

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample → random sample data and directly treat them as negative samples

Data augmentation: small variation → not change semantic → not change label

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

A representation function

An anchor sample

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{\boxed{f}(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

A representation function What is it?

An anchor sample

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{\boxed{f}(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

A representation function

What is it?

An anchor sample

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{\boxed{f}(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample

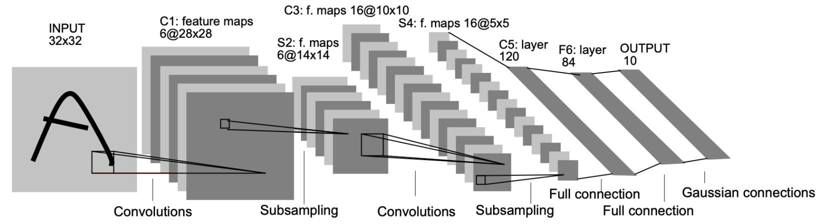


Fig. 1. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

A representation function What is it?

An anchor sample

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{\boxed{f}(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample

Each layer's feature map can be used as representation of an input data sample

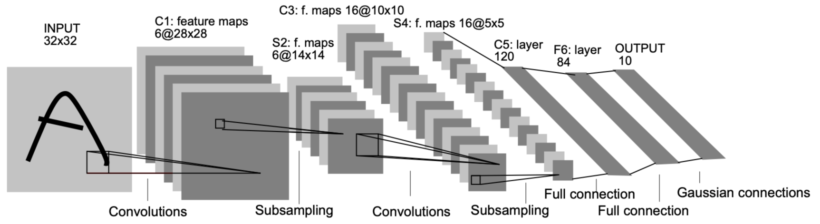


Fig. 1. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

A representation function What is it?

An anchor sample

A positive sample

A negative sample

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{\boxed{f}(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}}\right)\right]$$

Each layer's feature map can be used as representation of an input data sample

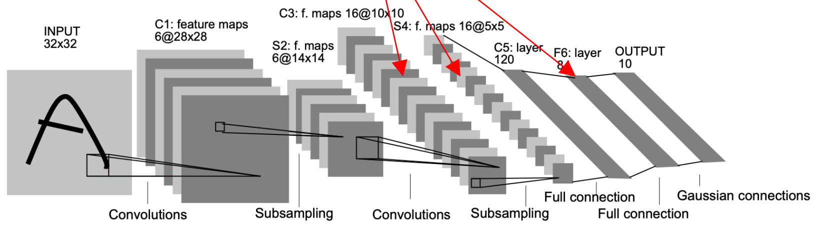


Fig. 1. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

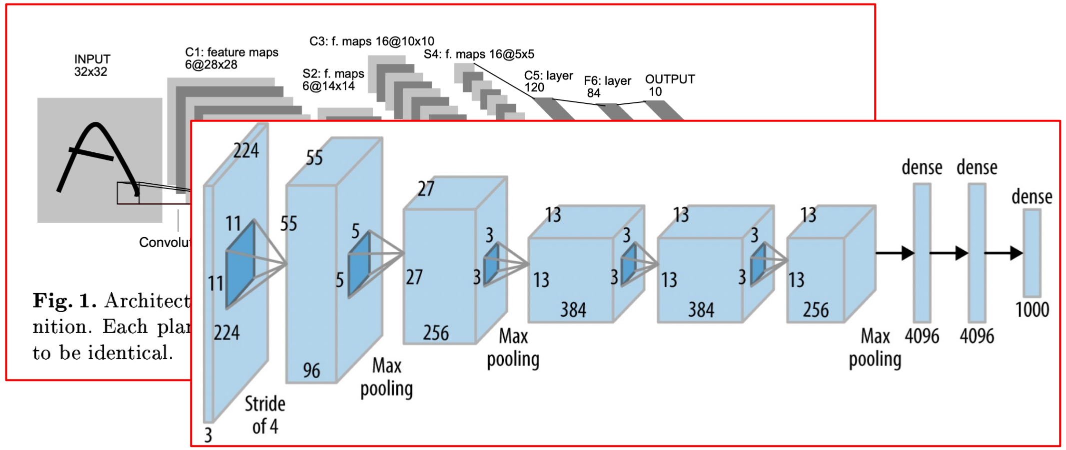
A representation function What is it?

An anchor sample

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{\boxed{f}(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample



Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

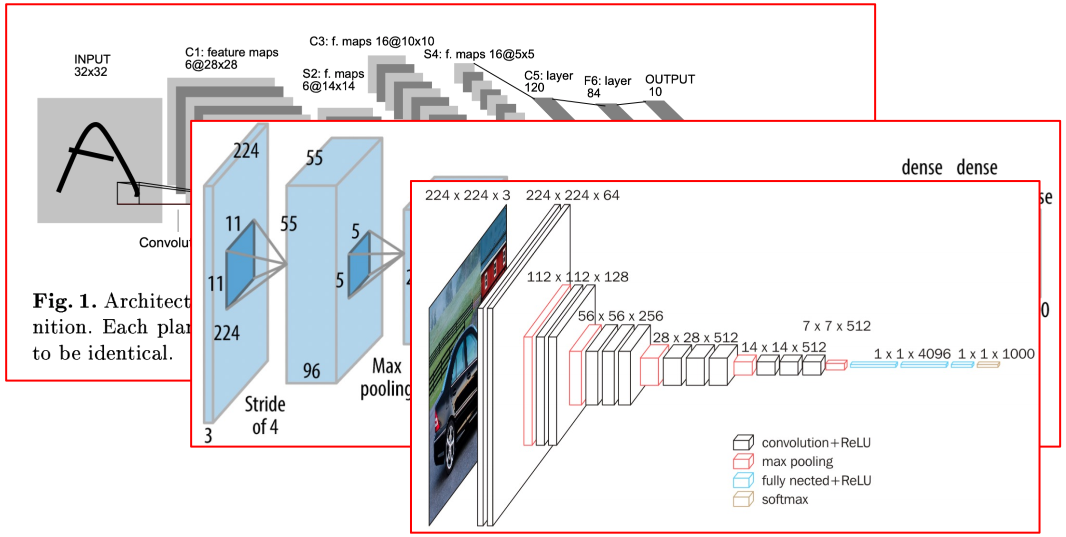
A representation function What is it?

An anchor sample

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{\boxed{f}(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample



Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

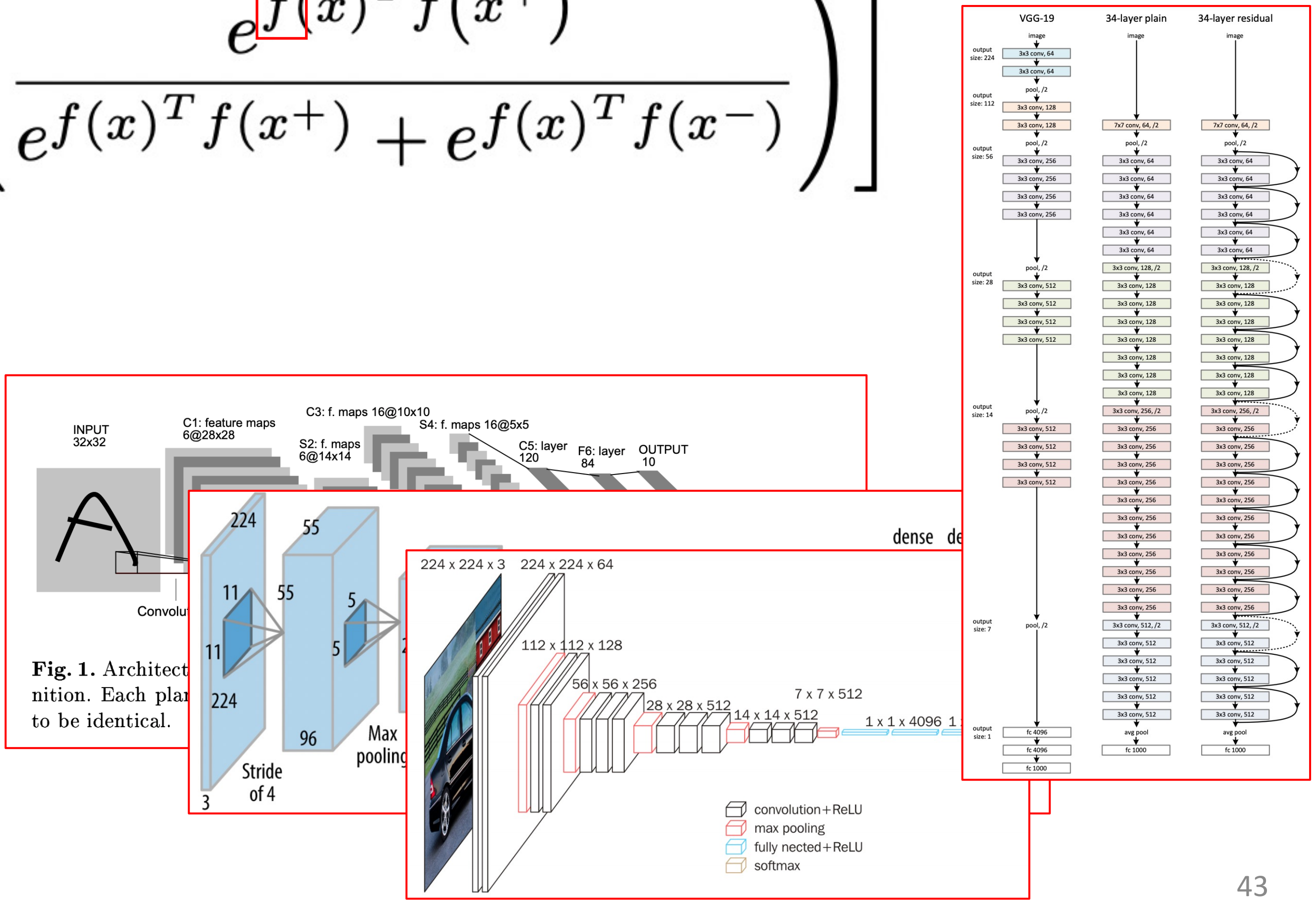
A representation function What is it?

An anchor sample

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{\boxed{f}(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample



Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

Go over each data

An anchor sample

A representation function

What is it?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{\boxed{f}(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}}\right)\right]$$

A positive sample

A negative sample

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

Go over each data

An anchor sample

A representation function

What is it?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{\boxed{f}(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right]$$

Minimize over **f**?

A positive sample

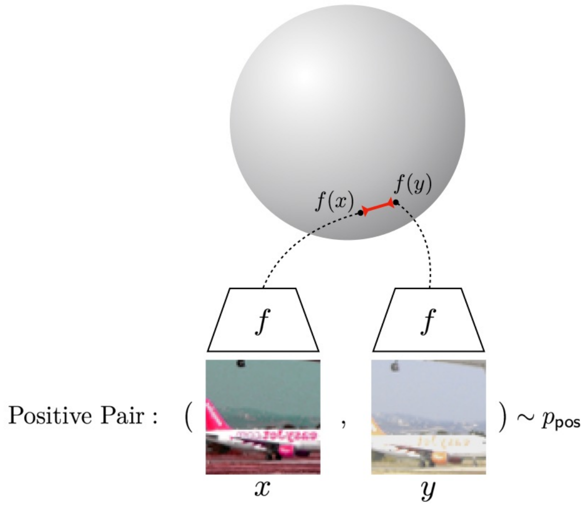
A negative sample

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

Uniformity: Preserve maximal information.

Figure 1: Illustration of alignment and uniformity of feature distributions on the output unit hypersphere. STL-10 (Coates et al., 2011) images are used for demonstration.



Alignment: Similar samples have similar features. (Figure inspired by Tian et al. (2019).)

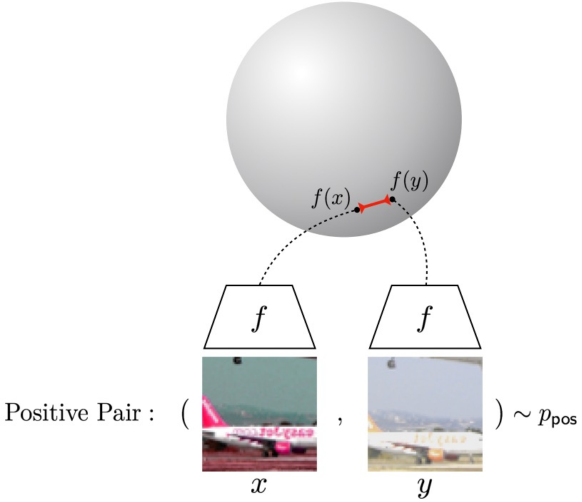
Wang, Tongzhou, and Phillip Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." In *International Conference on Machine Learning*, pp. 9929-9939. PMLR, 2020.

Self-supervision: contrastive learning

Data samples from different classes

Uniformity: Preserve maximal information.

Figure 1: Illustration of alignment and uniformity of feature distributions on the output unit hypersphere. STL-10 (Coates et al., 2011) images are used for demonstration.

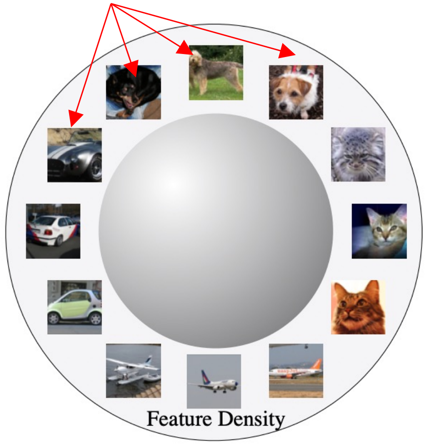


Alignment: Similar samples have similar features.
(Figure inspired by Tian et al. (2019).)

Wang, Tongzhou, and Phillip Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." In *International Conference on Machine Learning*, pp. 9929-9939. PMLR, 2020.

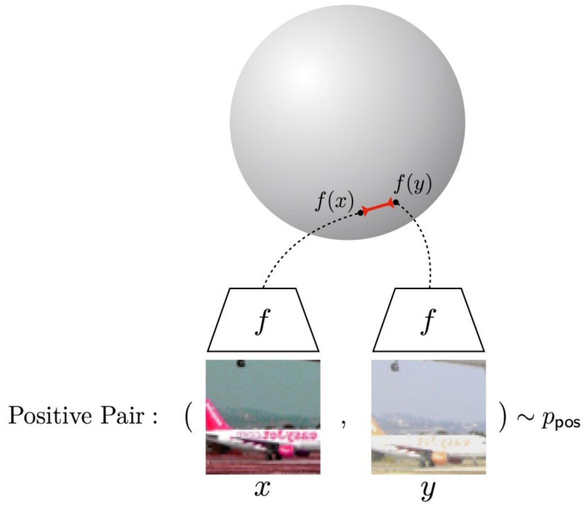
Self-supervision: contrastive learning

Data samples from different classes



Uniformity: Preserve maximal information.

Figure 1: Illustration of alignment and uniformity of feature distributions on the output unit hypersphere. STL-10 (Coates et al., 2011) images are used for demonstration.

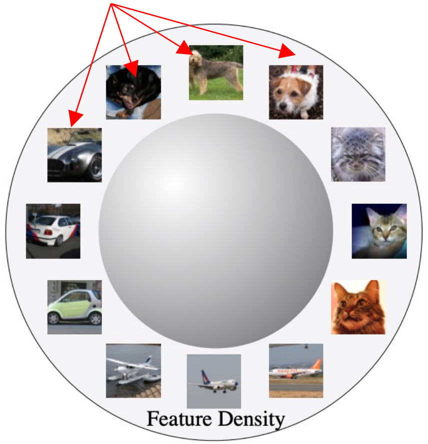


Alignment: Similar samples have similar features. (Figure inspired by Tian et al. (2019).)

Wang, Tongzhou, and Phillip Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." In *International Conference on Machine Learning*, pp. 9929-9939. PMLR, 2020.

Self-supervision: contrastive learning

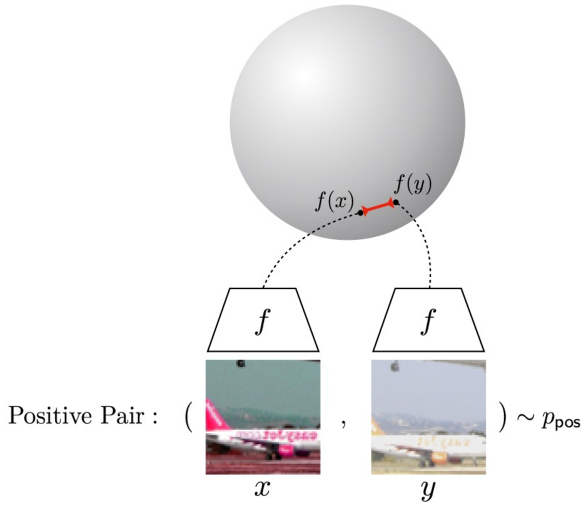
Data samples from different classes



Uniformity: Preserve maximal information.

Figure 1: Illustration of alignment and uniformity of feature distributions on the output unit hypersphere. STL-10 (Coates et al., 2011) images are used for demonstration.

Data samples from the same class:

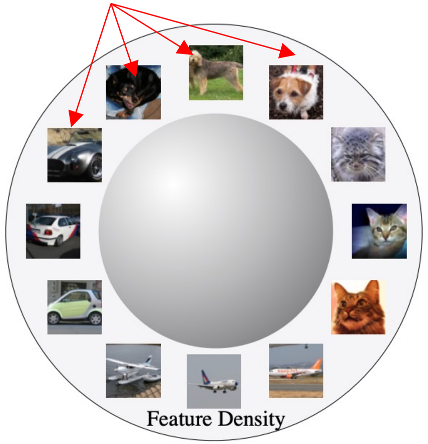


Alignment: Similar samples have similar features.
(Figure inspired by Tian et al. (2019).)

Wang, Tongzhou, and Phillip Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." In *International Conference on Machine Learning*, pp. 9929-9939. PMLR, 2020.

Self-supervision: contrastive learning

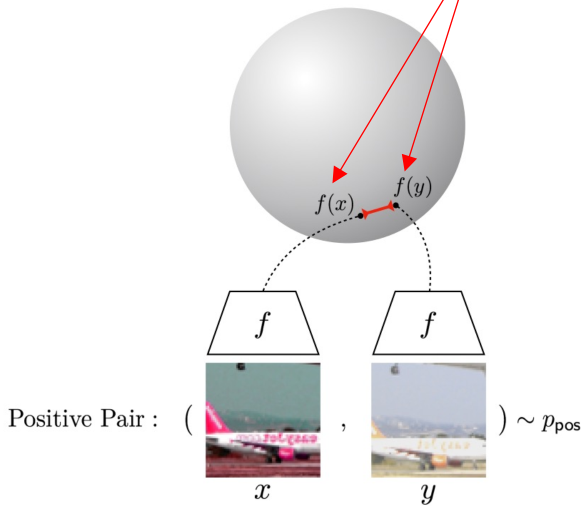
Data samples from different classes



Uniformity: Preserve maximal information.

Figure 1: Illustration of alignment and uniformity of feature distributions on the output unit hypersphere. STL-10 (Coates et al., 2011) images are used for demonstration.

Data samples from the same class: stay close to each other



Alignment: Similar samples have similar features. (Figure inspired by Tian et al. (2019).)

Wang, Tongzhou, and Phillip Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." In *International Conference on Machine Learning*, pp. 9929-9939. PMLR, 2020.

Self-supervision: contrastive learning

Random init

Supervised learning

Contrastive learning
(unsupervised, self-supervised)

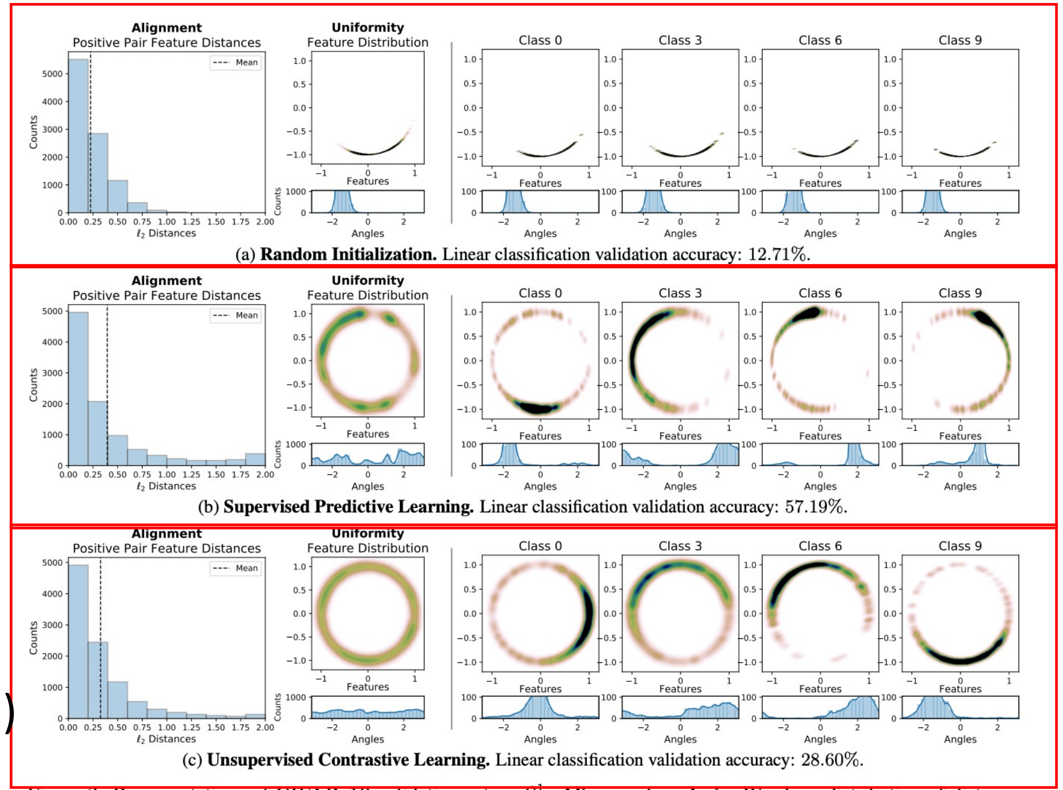


Figure 3: Representations of CIFAR-10 validation set on $\mathcal{S}^1$. **Alignment analysis:** We show distribution of distance between features of positive pairs (two random augmentations). **Uniformity analysis:** We plot feature distributions with Gaussian kernel density estimation (KDE) in $\mathbb{R}^2$ and von Mises-Fisher (vMF) KDE on angles (i.e., $\mathrm{arctan2}(y, x)$ for each point $(x, y) \in \mathcal{S}^1$). **Four rightmost plots** visualize feature distributions of selected specific classes. Representation from contrastive learning is both *aligned* (having low positive pair feature distances) and *uniform* (evenly distributed on $\mathcal{S}^1$).

Wang, Tongzhou, and Phillip Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." In *International Conference on Machine Learning*, pp. 9929-9939. PMLR, 2020.

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\underset{x,x^+,x^-}{\mathbb{E}}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right]$$

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\boxed{\mathbb{E}_{x,x^{+},x^{-}}\left[-\log\left(\frac{e^{f(x)^{T}f(x^{+})}}{e^{f(x)^{T}f(x^{+})}+e^{f(x)^{T}f(x^{-})}}\right)\right]}$$

Review: A negative sample → random sample data and directly treat them as negative samples

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\mathop{\mathbb{E}}_{x,x^{+},x^{-}}\left[-\log\left(\frac{e^{f(x)^{T}f(x^{+})}}{e^{f(x)^{T}f(x^{+})}+e^{f(x)^{T}f(\boxed{x^{-}})}}\right)\right]$$

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

What if “negative” samples are actually from the same class?

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

What if “negative” samples are actually from the same class?

False negative

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau)L_{un}^{\neq}(f) \qquad (8)$$

What if "negative" samples are actually from the same class?

False negative

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau)L_{un}^{\neq}(f) \qquad (8)$$

What if “negative” samples are actually from the same class?

False negative

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

The prob: we have true negative

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau)L_{un}^{\neq}(f) \qquad (8)$$

What if "negative" samples are actually from the same class?

False negative

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

The prob: we have true negative

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau)L_{un}^{\neq}(f) \qquad (8)$$

What if "negative" samples are actually from the same class?

False negative

$$L_{un}^{=}(f) = \mathop{\mathbb{E}}_{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}}\left[\ell(f(x)^T(f(x^+) - f(x^-)))\right]$$
$$\geq \mathop{\mathbb{E}}_{c\sim\nu,x\sim\mathcal{D}_c}\left[\ell(f(x)^T(\mu_c - \mu_c))\right] = 1$$

Arora, Sanjeev, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

The prob: we have
true negative

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau)L_{un}^{\neq}(f) \qquad (8)$$

What if "negative" samples are actually from the same class?

False negative

$$L_{un}^{=}(f) = \mathop{\mathbb{E}}_{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))\right]$$
$$\geq \mathop{\mathbb{E}}_{c\sim\nu, x\sim\mathcal{D}_c}\left[\ell(f(x)^T(\mu_c-\mu_c))\right] = 1$$

$$L_{un}^{\neq}(f)$$
$$= \mathop{\mathbb{E}}_{\substack{c^+,c^-\sim\rho^2 \\ x,x^+\sim\mathcal{D}_{c^+}^2 \\ x^-\sim\mathcal{D}_{c^-}}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))|c^+\neq c^-\right]$$

and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

The prob: we have true negative

What if "negative" samples are actually from the same class?

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau) L_{un}^{\neq}(f) \qquad (8)$$

False negative

$$L_{un}^{=}(f) = \mathop{\mathbb{E}}_{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))\right]$$

$$\geq \mathop{\mathbb{E}}_{c\sim\nu, x\sim\mathcal{D}_c}\left[\ell(f(x)^T(\mu_c-\mu_c))\right] = 1$$

$$L_{un}^{\neq}(f) = \mathop{\mathbb{E}}_{\substack{c^+,c^-\sim\rho^2 \\ x,x^+\sim\mathcal{D}_{c^+}^2 \\ x^-\sim\mathcal{D}_{c^-}}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))\middle|\boxed{c^+\neq c^-}\right]$$

and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

The prob: we have
true negative

$$\mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau) L_{un}^{\neq}(f) \qquad (8)$$

$$\begin{aligned} L_{un}^{=}(f) &= \mathop{\mathbb{E}}_{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}} \left[\ell(f(x)^T(f(x^+)-f(x^-)))\right] \\ &\geq \mathop{\mathbb{E}}_{c\sim\nu, x\sim\mathcal{D}_c} \left[\ell(f(x)^T(\mu_c-\mu_c))\right] = 1 \end{aligned}$$

$$\begin{aligned} &L_{un}^{\neq}(f) \\ &= \mathop{\mathbb{E}}_{\substack{c^+,c^-\sim\rho^2 \\ x,x^+\sim\mathcal{D}_{c^+}^2 \\ x^-\sim\mathcal{D}_{c^-}}} \left[\ell(f(x)^T(f(x^+)-f(x^-)))|c^+\neq c^-\right] \end{aligned}$$

$$\widehat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{L}_{un}(f)$$

Theorem 4.5. *With probability at least $1-\delta$, $\forall f \in \mathcal{F}$*

$$L_{sup}(\widehat{f}) \leq L_{sup}^{\mu}(\widehat{f}) \leq L_{un}^{\neq}(f) + \beta\, s(f) + \eta\, Gen_M$$

where $\beta = c'\frac{\tau}{1-\tau}$, $\eta = \frac{1}{1-\tau}$ and c' is a constant.

...and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

The prob: we have
true negative

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau)L_{un}^{\neq}(f) \qquad (8)$$

$$L_{un}^{=}(f) = \mathop{\mathbb{E}}_{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))\right]$$
$$\geq \mathop{\mathbb{E}}_{c\sim\nu,x\sim\mathcal{D}_c}\left[\ell(f(x)^T(\mu_c-\mu_c))\right] = 1$$

$$L_{un}^{\neq}(f)$$
$$= \mathop{\mathbb{E}}_{\substack{c^+,c^-\sim\rho^2 \\ x,x^+\sim\mathcal{D}_{c^+}^2 \\ x^-\sim\mathcal{D}_{c^-}}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))|c^+\neq c^-\right]$$

$$\hat{f} \in \arg\min_{f\in\mathcal{F}} \hat{L}_{un}(f)$$

Theorem 4.5. *With probability at least* $1-\delta$, $\forall f \in \mathcal{F}$

$$\boxed{L_{sup}(\hat{f})} \leq L_{sup}^{\mu}(\hat{f}) \leq \boxed{L_{un}^{\neq}(f)} + \boxed{\beta\, s(f)} + \boxed{\eta\, Gen_M}$$

where $\beta = c'\frac{\tau}{1-\tau}$, $\eta = \frac{1}{1-\tau}$ *and* c' *is a constant.*

and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

The prob: we have true negative

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau) L_{un}^{\neq}(f) \qquad (8)$$

$$L_{un}^{=}(f) = \mathop{\mathbb{E}}_{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))\right]$$

$$\geq \mathop{\mathbb{E}}_{c\sim\nu, x\sim\mathcal{D}_c}\left[\ell(f(x)^T(\mu_c-\mu_c))\right] = 1$$

$$L_{un}^{\neq}(f)$$

$$= \mathop{\mathbb{E}}_{\substack{c^+,c^-\sim\rho^2 \\ x,x^+\sim\mathcal{D}_{c^+}^2 \\ x^-\sim\mathcal{D}_{c^-}}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))|c^+\neq c^-\right]$$

$$\widehat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{L}_{un}(f)$$

Theorem 4.5. *With probability at least* $1-\delta$, $\forall f \in \mathcal{F}$

$$L_{sup}(\widehat{f}) \leq L_{sup}^{\mu}(\widehat{f}) \leq L_{un}^{\neq}(f) + \beta\, s(f) + \boxed{\eta}\, Gen_M$$

where $\beta = c'\frac{\tau}{1-\tau}$, $\boxed{\eta = \frac{1}{1-\tau}}$ *and* c' *is a constant.*

constant

...and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

The prob: we have
true negative

$$\boxed{\widehat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{L}_{un}(f)}$$

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau)L_{un}^{\neq}(f) \qquad (8)$$

$$L_{un}^{=}(f) = \mathop{\mathbb{E}}_{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}}\left[\ell(f(x)^T(f(x^+) - f(x^-)))\right]$$
$$\geq \mathop{\mathbb{E}}_{c\sim\nu, x\sim\mathcal{D}_c}\left[\ell(f(x)^T(\mu_c - \mu_c))\right] = 1$$

$$L_{un}^{\neq}(f) = \mathop{\mathbb{E}}_{\substack{c^+,c^-\sim\rho^2 \\ x,x^+\sim\mathcal{D}_{c^+}^2 \\ x^-\sim\mathcal{D}_{c^-}}}\left[\ell(f(x)^T(f(x^+) - f(x^-)))|c^+\neq c^-\right]$$

Theorem 4.5. *With probability at least $1-\delta$, $\forall f\in\mathcal{F}$*

$$L_{sup}(\widehat{f}) \leq L_{sup}^{\mu}(\widehat{f}) \leq L_{un}^{\neq}(f) + \beta\, s(f) + \boxed{\eta}\, \boxed{Gen_M}$$

where $\beta = c'\frac{\tau}{1-\tau}$, $\boxed{\eta = \frac{1}{1-\tau}}$ *and c' is a constant.*

$O\left(\sqrt{1/n}\right)$

constant

and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

The prob: we have true negative

$$\underset{x,x^+,x^-}{\mathbb{E}}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau)L_{un}^{\neq}(f) \qquad (8)$$

$$L_{un}^{=}(f) = \underset{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}}{\mathbb{E}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))\right]$$

$$\geq \underset{c\sim\nu, x\sim\mathcal{D}_c}{\mathbb{E}}\left[\ell(f(x)^T(\mu_c-\mu_c))\right] = 1$$

$$L_{un}^{\neq}(f) = \underset{\substack{c^+,c^-\sim\rho^2 \\ x,x^+\sim\mathcal{D}_{c^+}^2 \\ x^-\sim\mathcal{D}_{c^-}}}{\mathbb{E}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))|c^+\neq c^-\right]$$

$$\widehat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{L}_{un}(f)$$

Theorem 4.5. *With probability at least $1-\delta$, $\forall f\in\mathcal{F}$*

$$L_{sup}(\widehat{f}) \leq L_{sup}^{\mu}(\widehat{f}) \leq L_{un}^{\neq}(f) + \boxed{\beta}\, s(f) + \eta\, Gen_M$$

where $\beta = c'\frac{\tau}{1-\tau}$, $\eta = \frac{1}{1-\tau}$ and c' is a constant.

and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

The prob: we have
true negative

$$\underset{x,x^+,x^-}{\mathbb{E}}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau) L_{un}^{\neq}(f) \qquad (8)$$

$$L_{un}^{=}(f) = \underset{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}}{\mathbb{E}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))\right]$$
$$\geq \underset{c\sim\nu, x\sim\mathcal{D}_c}{\mathbb{E}}\left[\ell(f(x)^T(\mu_c-\mu_c))\right] = 1$$

$$L_{un}^{\neq}(f) = \underset{\substack{c^+,c^-\sim\rho^2 \\ x,x^+\sim\mathcal{D}_{c^+}^2 \\ x^-\sim\mathcal{D}_{c^-}}}{\mathbb{E}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))\,|\,c^+\neq c^-\right]$$

$$\widehat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{L}_{un}(f)$$

Theorem 4.5. *With probability at least $1-\delta$, $\forall f \in \mathcal{F}$*

$$L_{sup}(\widehat{f}) \leq L_{sup}^{\mu}(\widehat{f}) \leq L_{un}^{\neq}(f) + \boxed{\beta}\, s(f) + \eta\, Gen_M$$

where $\boxed{\beta = c'\frac{\tau}{1-\tau}}$, $\eta = \frac{1}{1-\tau}$ *and c' is a constant.*

and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

The prob: we have
true negative

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau)L_{un}^{\neq}(f) \qquad (8)$$

$$L_{un}^{=}(f) = \mathop{\mathbb{E}}_{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))\right]$$
$$\geq \mathop{\mathbb{E}}_{c\sim\nu, x\sim\mathcal{D}_c}\left[\ell(f(x)^T(\mu_c-\mu_c))\right] = 1$$

$$L_{un}^{\neq}(f)$$
$$= \mathop{\mathbb{E}}_{\substack{c^+,c^-\sim\rho^2 \\ x,x^+\sim\mathcal{D}_{c^+}^2 \\ x^-\sim\mathcal{D}_{c^-}}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))|c^+\neq c^-\right]$$

$$\widehat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{L}_{un}(f)$$

Theorem 4.5. *With probability at least* $1-\delta$, $\forall f \in \mathcal{F}$

$$L_{sup}(\widehat{f}) \leq L_{sup}^{\mu}(\widehat{f}) \leq L_{un}^{\neq}(f) + \boxed{\beta}\, s(f) + \eta\, Gen_M$$

where $\boxed{\beta = c'\frac{\tau}{1-\tau}}$, $\eta = \frac{1}{1-\tau}$ *and* c' *is a constant.*

If tau is very small?

and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Self-supervision: contrastive learning

- The gap between supervised learning and contrastive learning?

The prob: we have true negative

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(\boxed{x^-})}}\right)\right]$$

$$L_{un}(f) = \tau L_{un}^{=}(f) + (1-\tau) L_{un}^{\neq}(f) \qquad (8)$$

$$L_{un}^{=}(f) = \mathop{\mathbb{E}}_{\substack{c\sim\nu \\ x,x^+,x^-\sim\mathcal{D}_c^3}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))\right]$$
$$\geq \mathop{\mathbb{E}}_{c\sim\nu, x\sim\mathcal{D}_c}\left[\ell(f(x)^T(\mu_c-\mu_c))\right] = 1$$

$$L_{un}^{\neq}(f)$$
$$= \mathop{\mathbb{E}}_{\substack{c^+,c^-\sim\rho^2 \\ x,x^+\sim\mathcal{D}_{c^+}^2 \\ x^-\sim\mathcal{D}_{c^-}}}\left[\ell(f(x)^T(f(x^+)-f(x^-)))|c^+\neq c^-\right]$$

$$\boxed{\widehat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{L}_{un}(f)}$$

Theorem 4.5. *With probability at least* $1-\delta$, $\forall f \in \mathcal{F}$

$$L_{sup}(\widehat{f}) \leq L_{sup}^{\mu}(\widehat{f}) \leq L_{un}^{\neq}(f) + \boxed{\beta}\, s(f) + \eta\, Gen_M$$

where $\boxed{\beta = c'\frac{\tau}{1-\tau}}$, $\eta = \frac{1}{1-\tau}$ *and* c' *is a constant.*

If tau is very small?

If we have a large number of classes and we perform uniform random sampling?

and Nikunj Saunshi. "A theoretical analysis of contrastive unsupervised representation learning." *arXiv preprint arXiv:1902.09229* (2019).

Meta-learning

$\boxed{\text{Multi-Task Learning}}$

Solve multiple tasks $\mathscr{T}_1, \cdots, \mathscr{T}_T$ at once.

$$\min_{\theta} \sum_{i=1}^{T} \mathscr{L}_i(\theta, \mathscr{D}_i)$$

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

Multi-Task Learning

Solve multiple tasks $\mathscr{T}_1, \cdots, \mathscr{T}_T$ at once.

$$\min_{\theta} \sum_{i=1}^{T} \mathscr{L}_i(\theta, \mathscr{D}_i)$$

D_a D_b

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

$\boxed{\text{Multi-Task Learning}}$

Solve multiple tasks $\mathcal{T}_1, \cdots, \mathcal{T}_T$ at once.

$$\min_{\theta} \sum_{i=1}^{T} \mathcal{L}_i(\theta, \mathcal{D}_i)$$

$D_a \neq D_b$

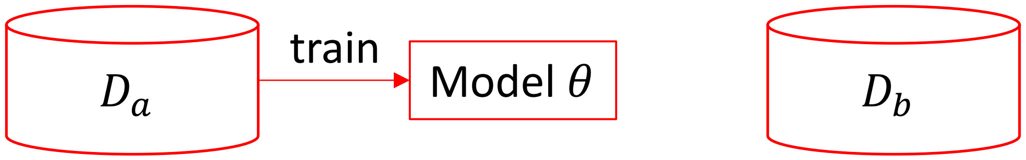
Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

Multi-Task Learning

Solve multiple tasks $\mathscr{T}_1, \cdots, \mathscr{T}_T$ at once.

$$\min_{\theta} \sum_{i=1}^{T} \mathscr{L}_i(\theta, \mathscr{D}_i)$$



$D_a \neq D_b$

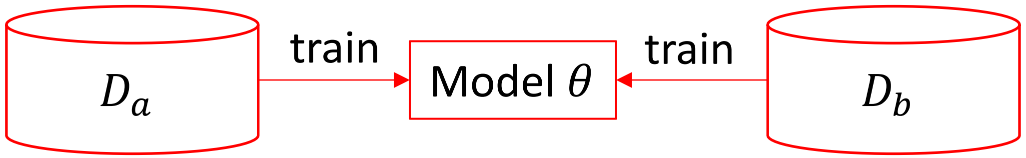
Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

Multi-Task Learning

Solve multiple tasks $\mathcal{T}_1, \cdots, \mathcal{T}_T$ at once.

$$\min_{\theta} \boxed{\sum_{i=1}^{T} \mathcal{L}_i(\theta, \mathcal{D}_i)}$$



$D_a \neq D_b$

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

Transfer Learning

Solve target task $\mathscr{T}_b$ after solving source task $\mathscr{T}_a$
by *transferring* knowledge learned from $\mathscr{T}_a$

Key assumption: Cannot access data $\mathscr{D}_a$ during transfer.

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

Transfer Learning

Solve target task $\mathscr{T}_b$ after solving source task $\mathscr{T}_a$ by *transferring* knowledge learned from $\mathscr{T}_a$

<u>Key assumption</u>: Cannot access data $\mathscr{D}_a$ during transfer.

$D_a \neq D_b$

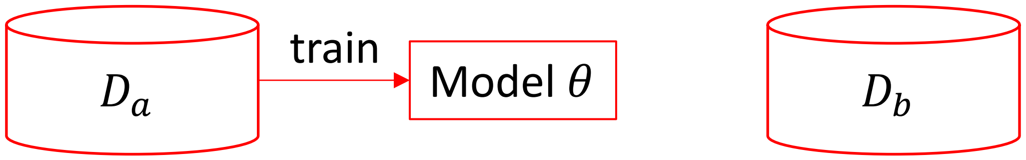
Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

Transfer Learning

Solve target task $\mathcal{T}_b$ after solving source task $\mathcal{T}_a$
by *transferring* knowledge learned from $\mathcal{T}_a$

Key assumption: Cannot access data $\mathcal{D}_a$ during transfer.



$$D_a \neq D_b$$

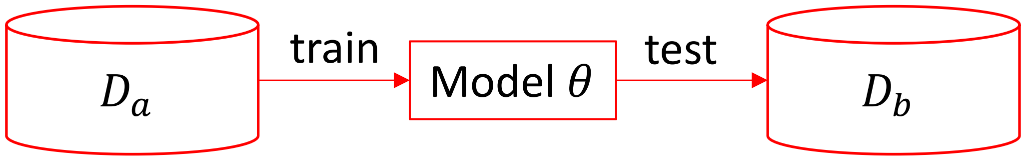
Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

Transfer Learning

Solve target task $\mathcal{T}_b$ after solving source task $\mathcal{T}_a$
by *transferring* knowledge learned from $\mathcal{T}_a$

<u>Key assumption</u>: Cannot access data $\mathcal{D}_a$ during transfer.



$$D_a \neq D_b$$

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

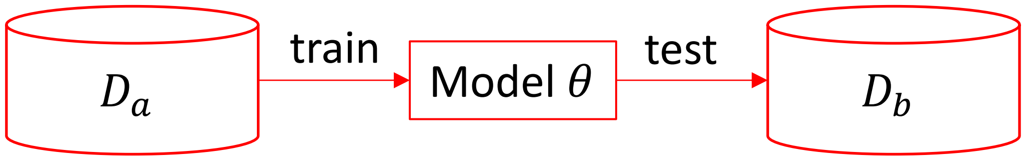
Meta-learning

Q: anything we learned can be regarded as "transfer learning"?

Transfer Learning

Solve target task $\mathscr{T}_b$ after solving source task $\mathscr{T}_a$ by *transferring* knowledge learned from $\mathscr{T}_a$

<u>Key assumption</u>: Cannot access data $\mathscr{D}_a$ during transfer.

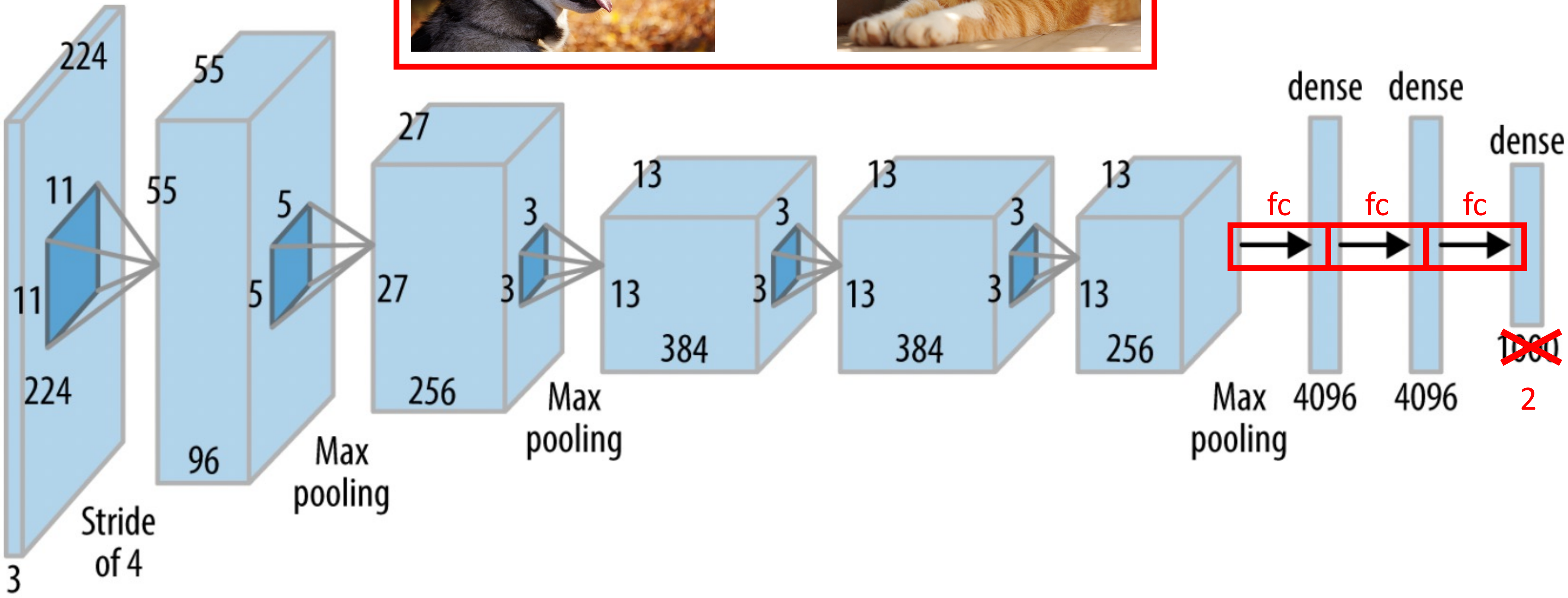


$$D_a \neq D_b$$

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Pre-train?

Two classes



Suppose we have a learned model → weight parameters are determined and fixed

Meta-learning

The Meta-Learning Problem

Given data from $\mathcal{T}_1, \ldots, \mathcal{T}_n$, quickly solve new task $\mathcal{T}_{\text{test}}$

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

The $\boxed{\text{Meta-Learning}}$ Problem

Given data from $\mathscr{T}_1, \ldots, \mathscr{T}_n$, $\boxed{\text{quickly}}$ solve new task $\mathscr{T}_{\text{test}}$

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

The $\boxed{\text{Meta-Learning}}$ Problem

Given data from $\mathcal{T}_1, \ldots, \mathcal{T}_n$, $\boxed{\text{quickly}}$ solve new task $\mathcal{T}_{\text{test}}$

Not directly apply the trained model for prediction

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

The $\boxed{\text{Meta-Learning}}$ Problem

Given data from $\mathcal{T}_1, \ldots, \mathcal{T}_n$, $\boxed{\text{quickly}}$ solve new task $\mathcal{T}_{\text{test}}$

Not directly apply the trained model for prediction

Allow some quick update to fit the testing data

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning

The $\boxed{\text{Meta-Learning}}$ Problem

Given data from $\mathcal{T}_1, \ldots, \mathcal{T}_n$, $\boxed{\text{quickly}}$ solve new task $\mathcal{T}_{\text{test}}$

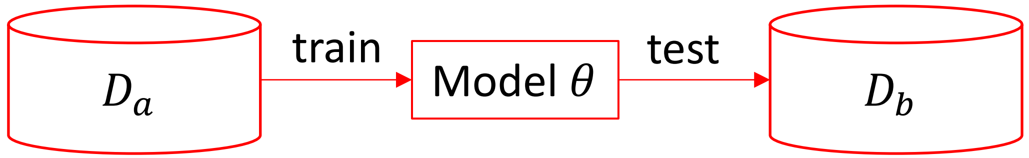
Not directly apply the trained model for prediction

Allow some quick update to fit the testing data

$\mathcal{T}_{\text{test}}$ can be one of many target tasks

Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning



Multi-Task Learning

Solve multiple tasks $\mathcal{T}_1, \cdots, \mathcal{T}_T$ at once.

$$\min_{\theta} \sum_{i=1}^{T} \mathcal{L}_i(\theta, \mathcal{D}_i)$$

Transfer Learning

Solve target task $\mathcal{T}_b$ after solving source task $\mathcal{T}_a$ by *transferring* knowledge learned from $\mathcal{T}_a$

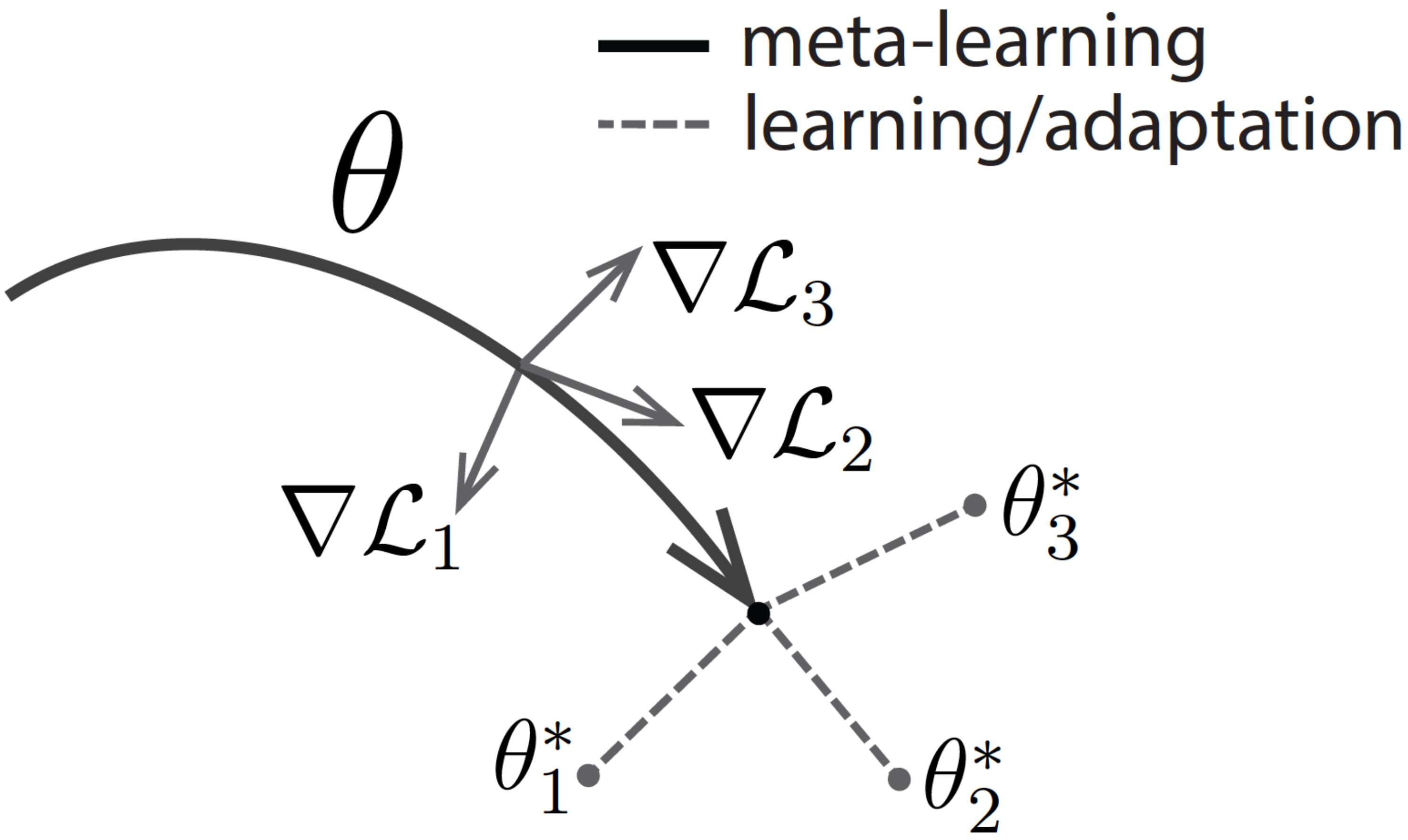
Key assumption: Cannot access data $\mathcal{D}_a$ during transfer.

The Meta-Learning Problem

Given data from $\mathcal{T}_1, \ldots, \mathcal{T}_n$, quickly solve new task $\mathcal{T}_{\text{test}}$

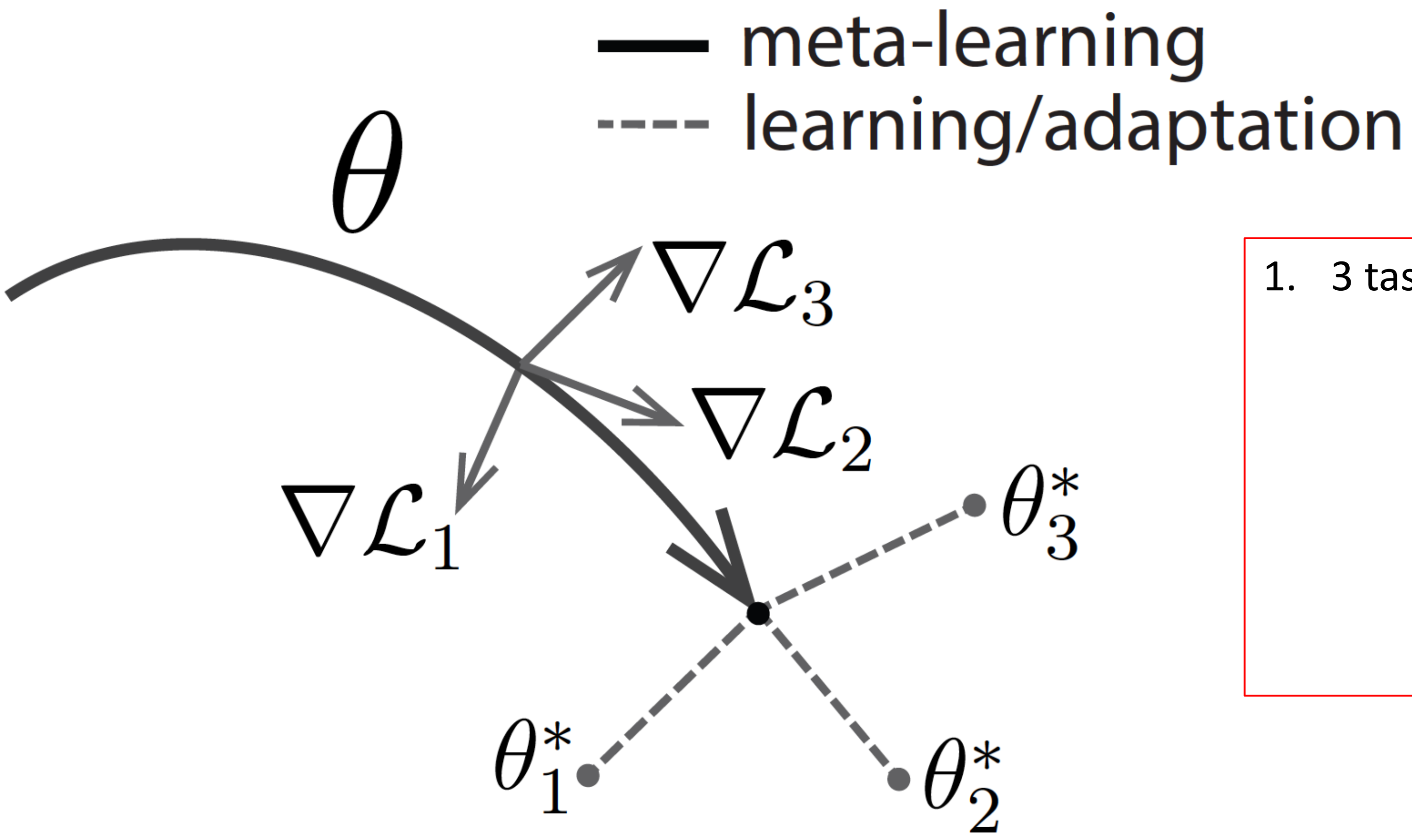
Image credit https://cs330.stanford.edu/slides/cs330_metalearning_bbox_2020.pdf

Meta-learning



Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." In *International Conference on Machine Learning*, pp. 1126-1135. PMLR, 2017.
http://proceedings.mlr.press/v70/finn17a/finn17a.pdf

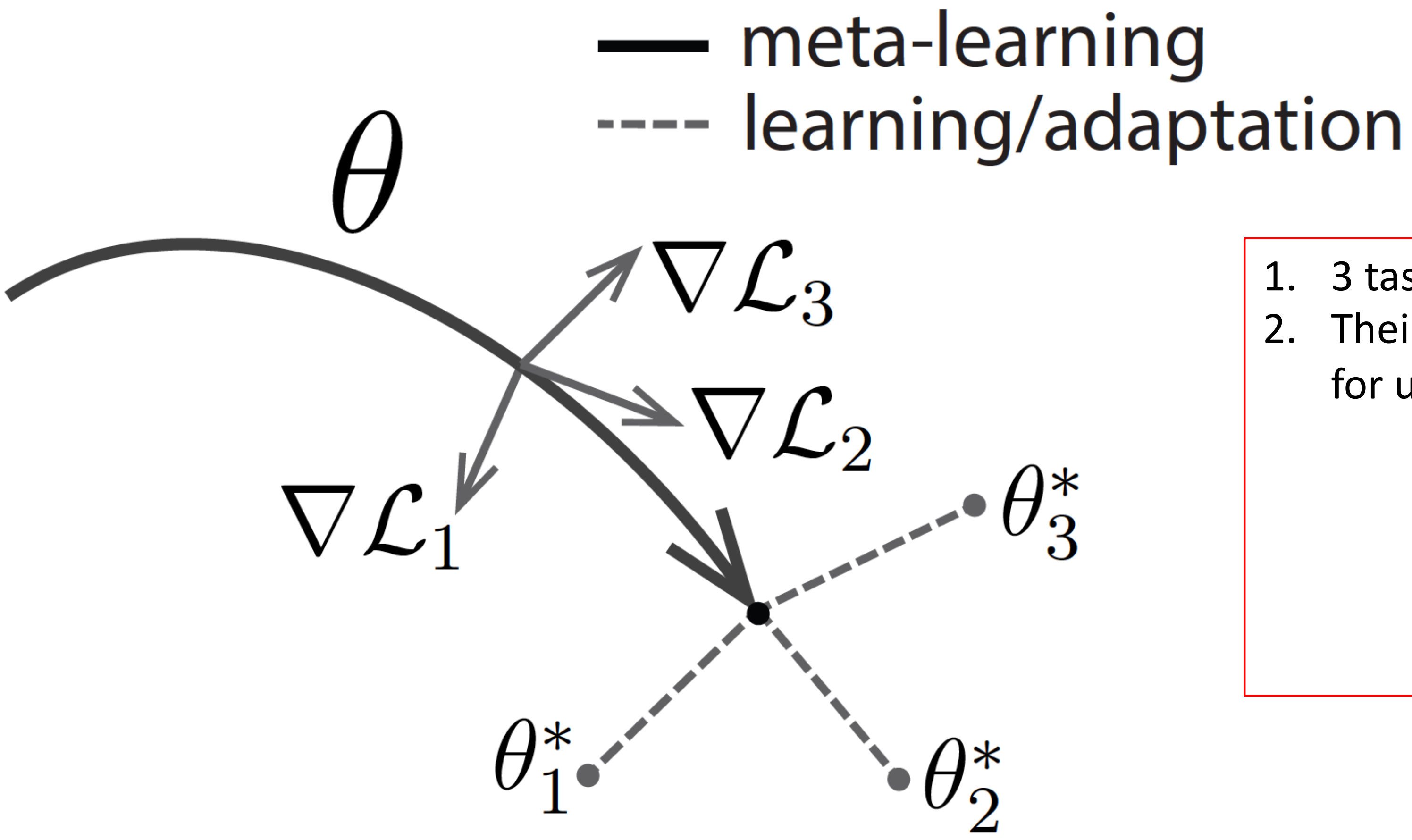
Meta-learning



1. 3 tasks

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." In *International Conference on Machine Learning*, pp. 1126-1135. PMLR, 2017.
http://proceedings.mlr.press/v70/finn17a/finn17a.pdf

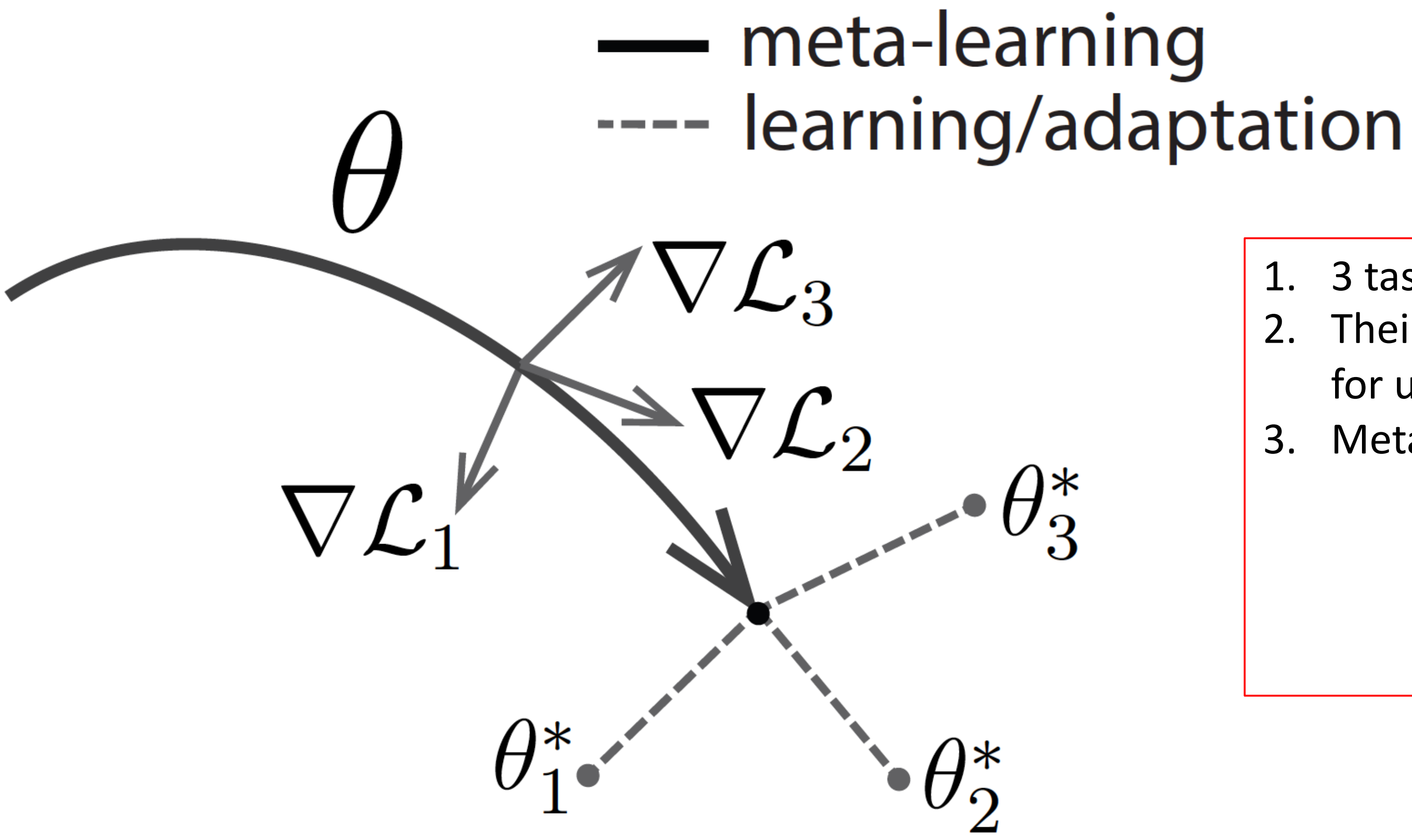
Meta-learning



1. 3 tasks
2. Their datasets may have different directions/gradients for updating current model

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." In *International Conference on Machine Learning*, pp. 1126-1135. PMLR, 2017.
http://proceedings.mlr.press/v70/finn17a/finn17a.pdf

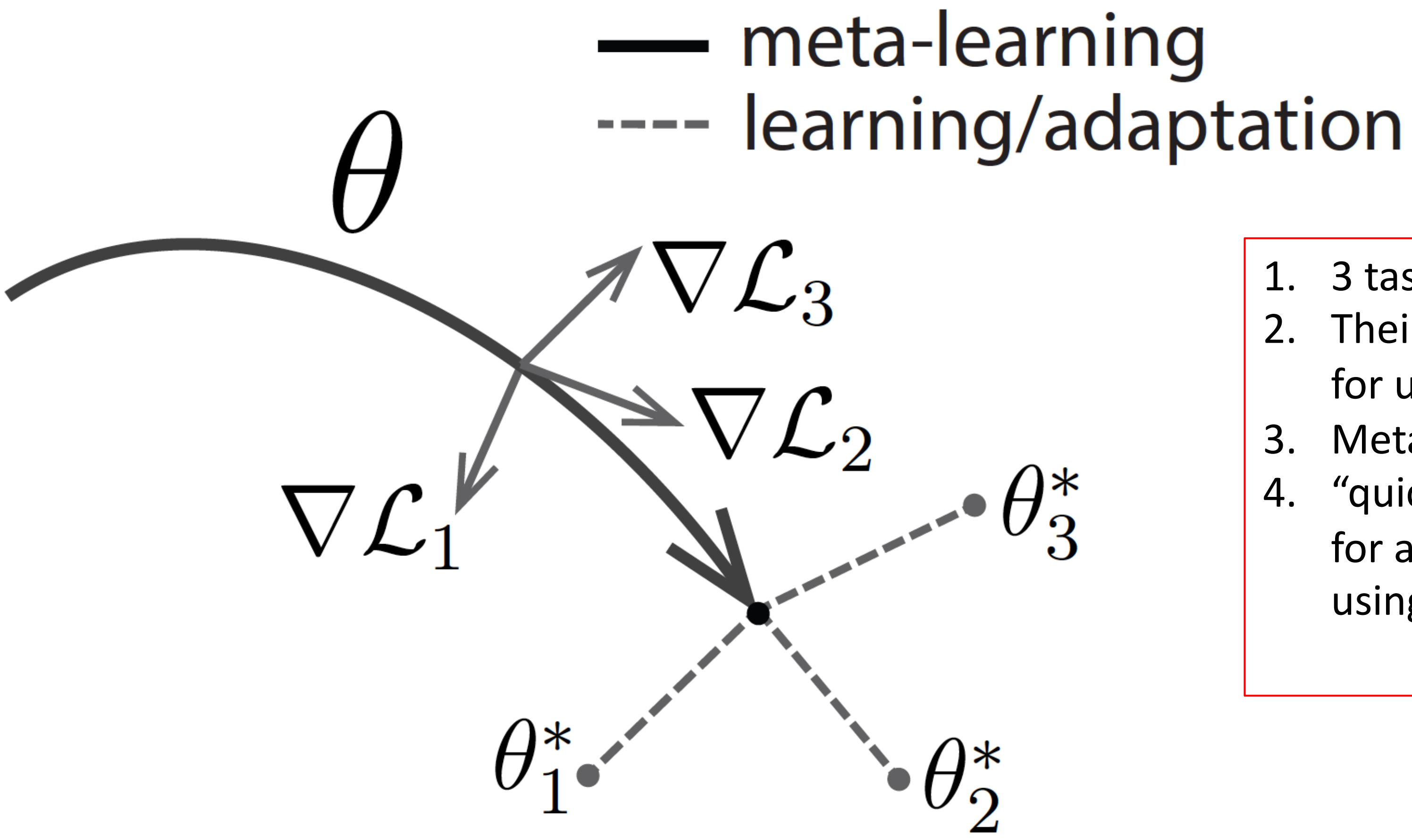
Meta-learning



1. 3 tasks
2. Their datasets may have different directions/gradients for updating current model
3. Meta learning determines one update direction

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." In *International Conference on Machine Learning*, pp. 1126-1135. PMLR, 2017. http://proceedings.mlr.press/v70/finn17a/finn17a.pdf

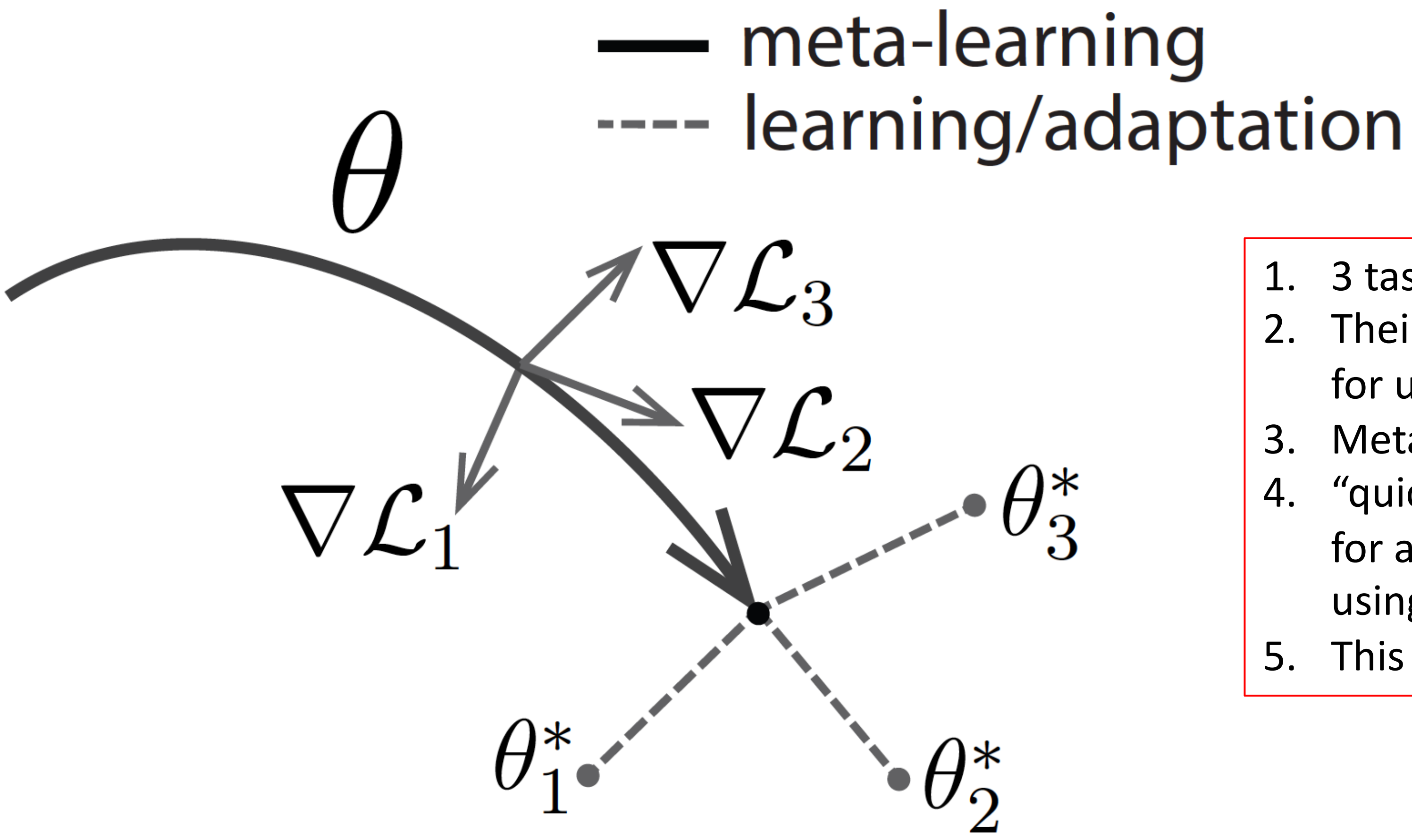
Meta-learning



1. 3 tasks
2. Their datasets may have different directions/gradients for updating current model
3. Meta learning determines one update direction
4. “quickly solve” new tasks: after meta learning update, for any new task, we can simply update a few steps using the data from the corresponding task

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." In *International Conference on Machine Learning*, pp. 1126-1135. PMLR, 2017. http://proceedings.mlr.press/v70/finn17a/finn17a.pdf

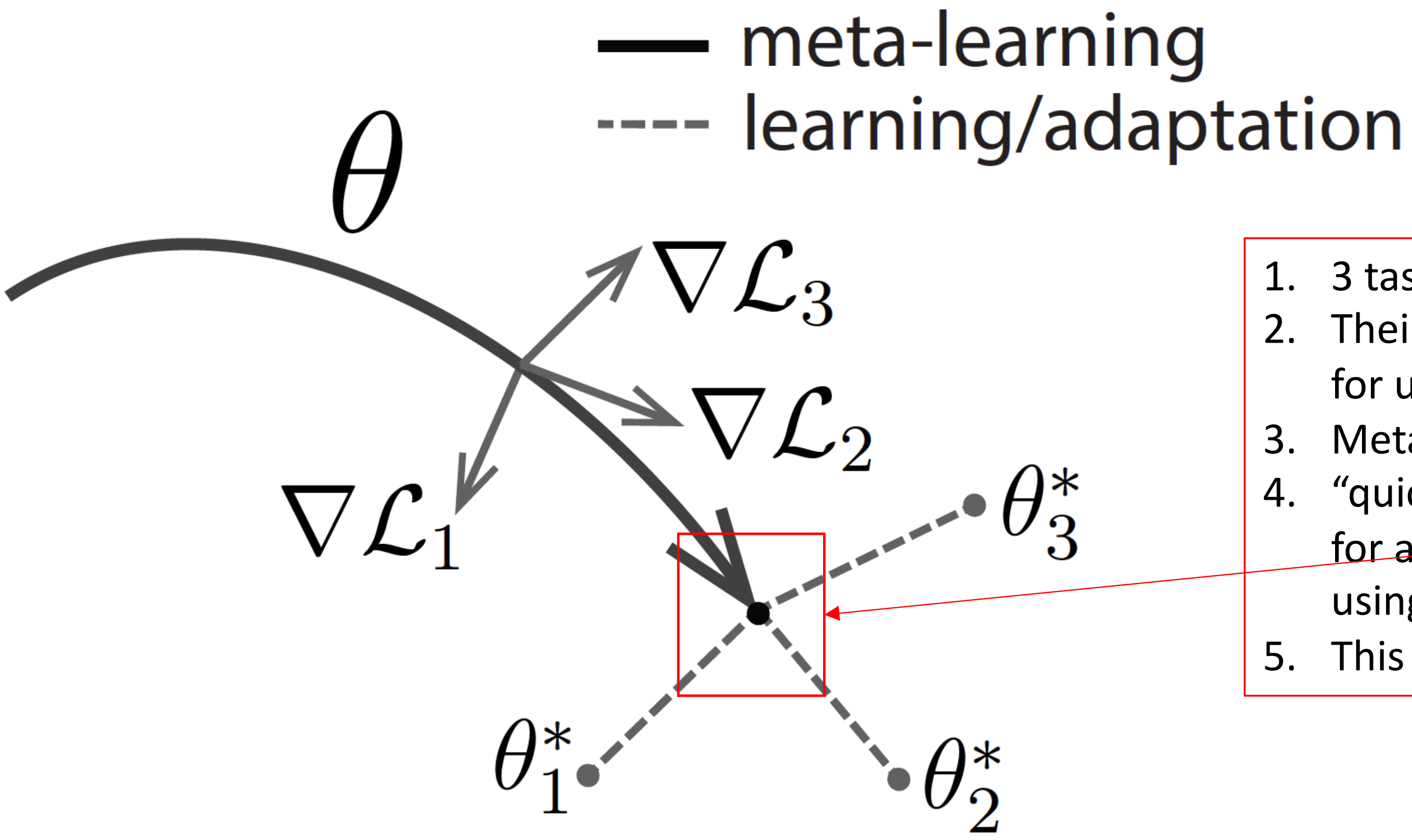
Meta-learning



1. 3 tasks
2. Their datasets may have different directions/gradients for updating current model
3. Meta learning determines one update direction
4. "quickly solve" new tasks: after meta learning update, for any new task, we can simply update a few steps using the data from the corresponding task
5. This final model fits target task well

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." In *International Conference on Machine Learning*, pp. 1126-1135. PMLR, 2017.
http://proceedings.mlr.press/v70/finn17a/finn17a.pdf

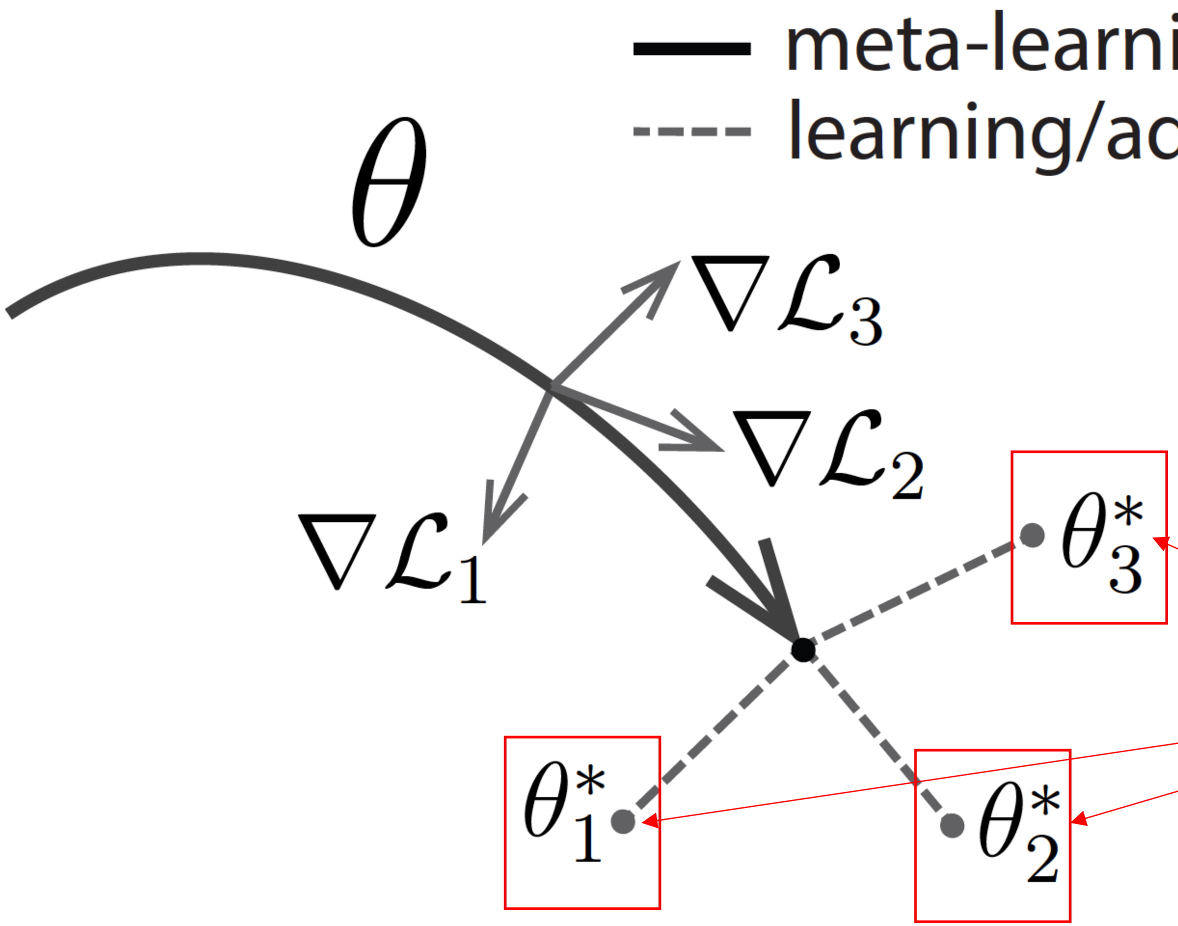
Meta-learning



1. 3 tasks
2. Their datasets may have different directions/gradients for updating current model
3. Meta learning determines one update direction
4. “quickly solve” new tasks: after meta learning update, for any new task, we can simply update a few steps using the data from the corresponding task
5. This final model fits target task well

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." In *International Conference on Machine Learning*, pp. 1126-1135. PMLR, 2017. http://proceedings.mlr.press/v70/finn17a/finn17a.pdf

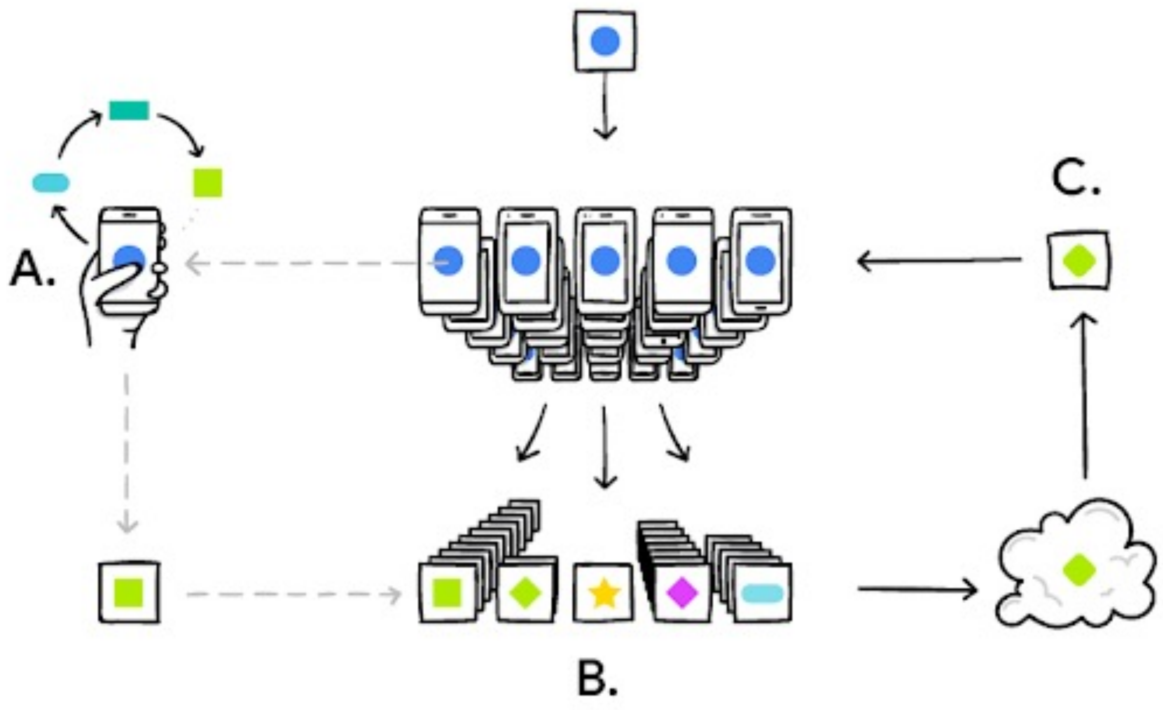
Meta-learning



1. 3 tasks
2. Their datasets may have different directions/gradients for updating current model
3. Meta learning determines one update direction
4. “quickly solve” new tasks: after meta learning update, for any new task, we can simply update a few steps using the data from the corresponding task
5. This final model fits target task well

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." In *International Conference on Machine Learning*, pp. 1126-1135. PMLR, 2017. http://proceedings.mlr.press/v70/finn17a/finn17a.pdf

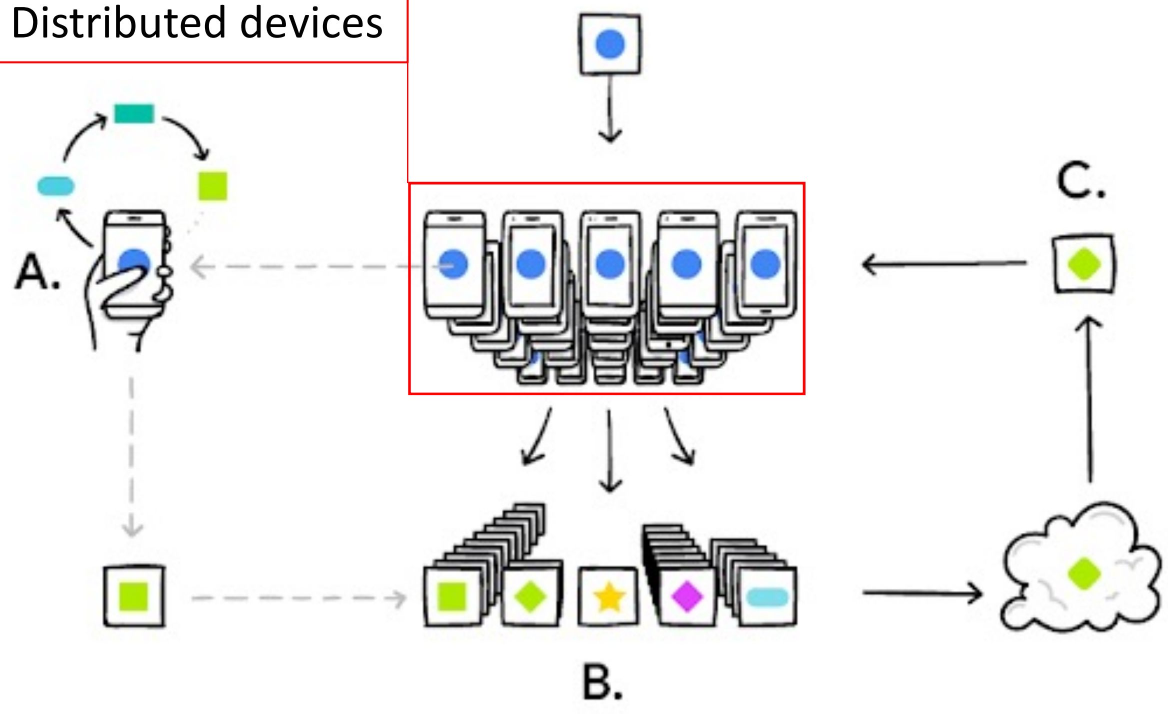
Federated learning



https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Federated learning

1. Distributed data
2. Distributed devices

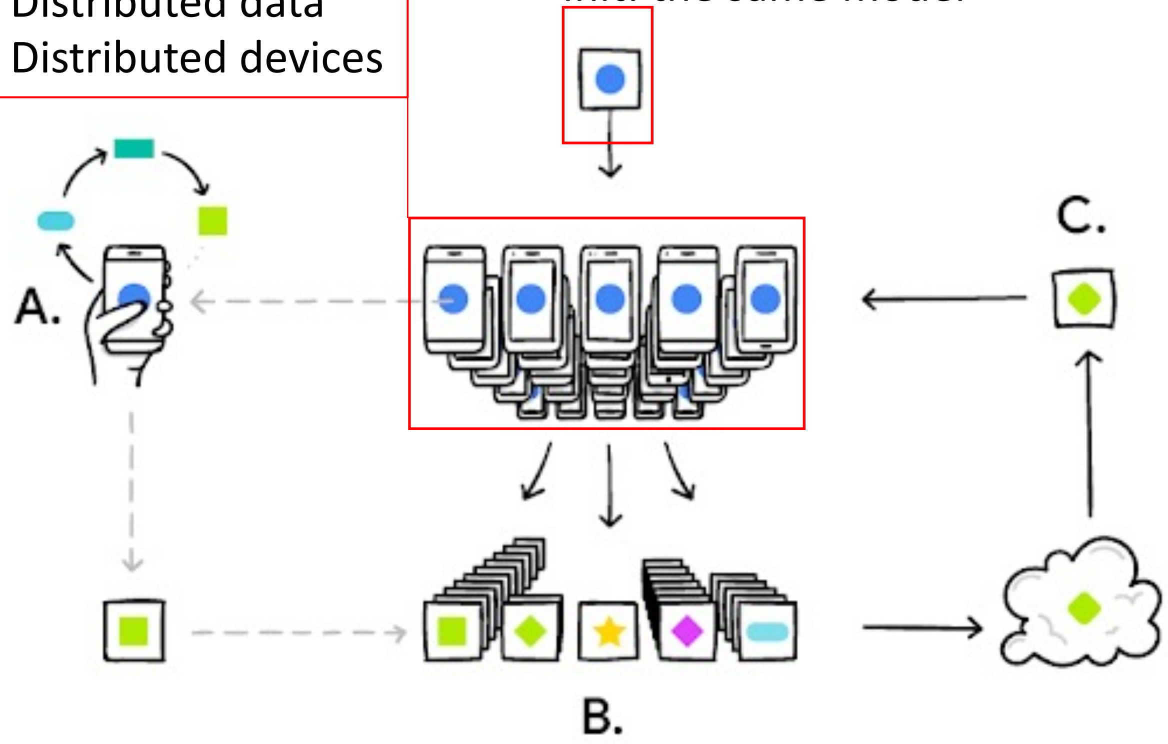


https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Federated learning

1. Distributed data
2. Distributed devices

Init: the same model



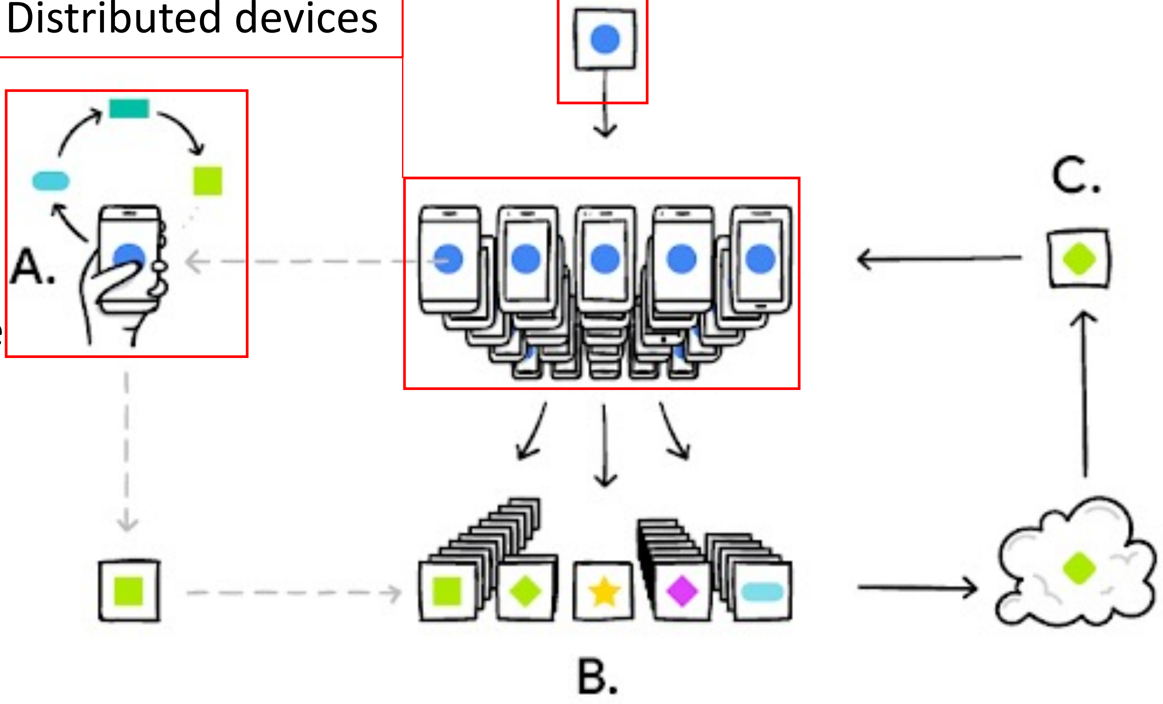
https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Federated learning

1. Distributed data
2. Distributed devices

Init: the same model

Local data → local update



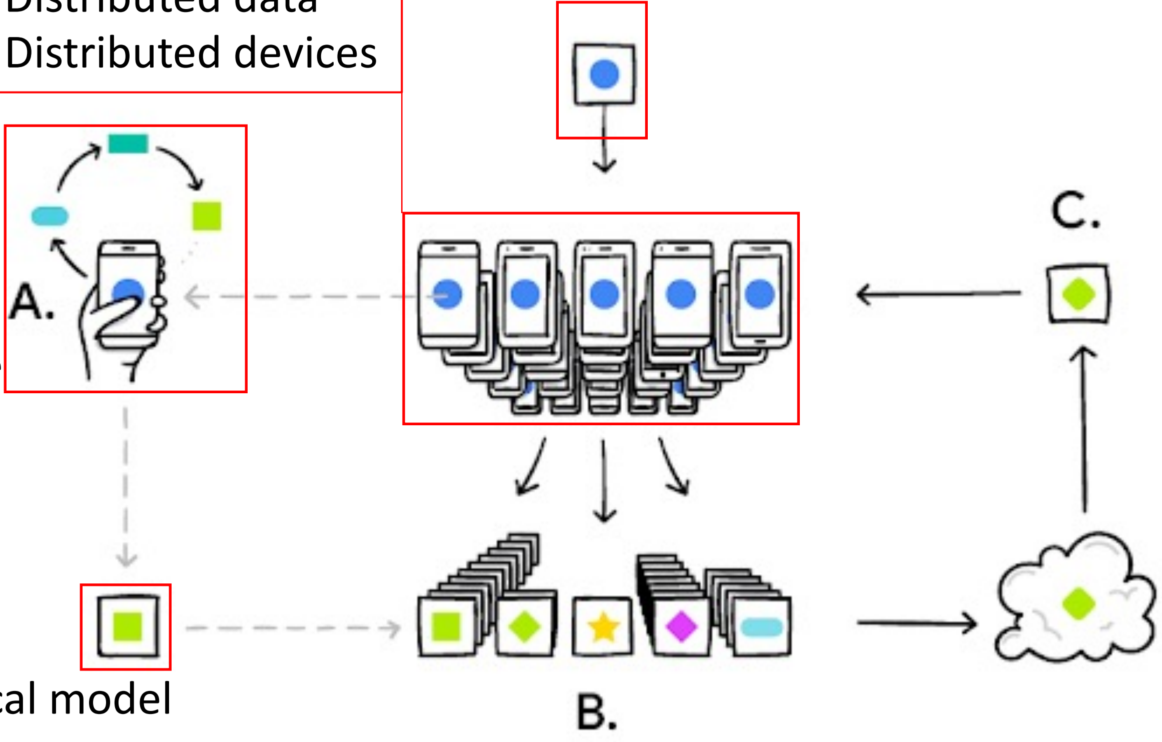
https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Federated learning

1. Distributed data
2. Distributed devices

Init: the same model

Local data → local update



Only upload local model
(no data)

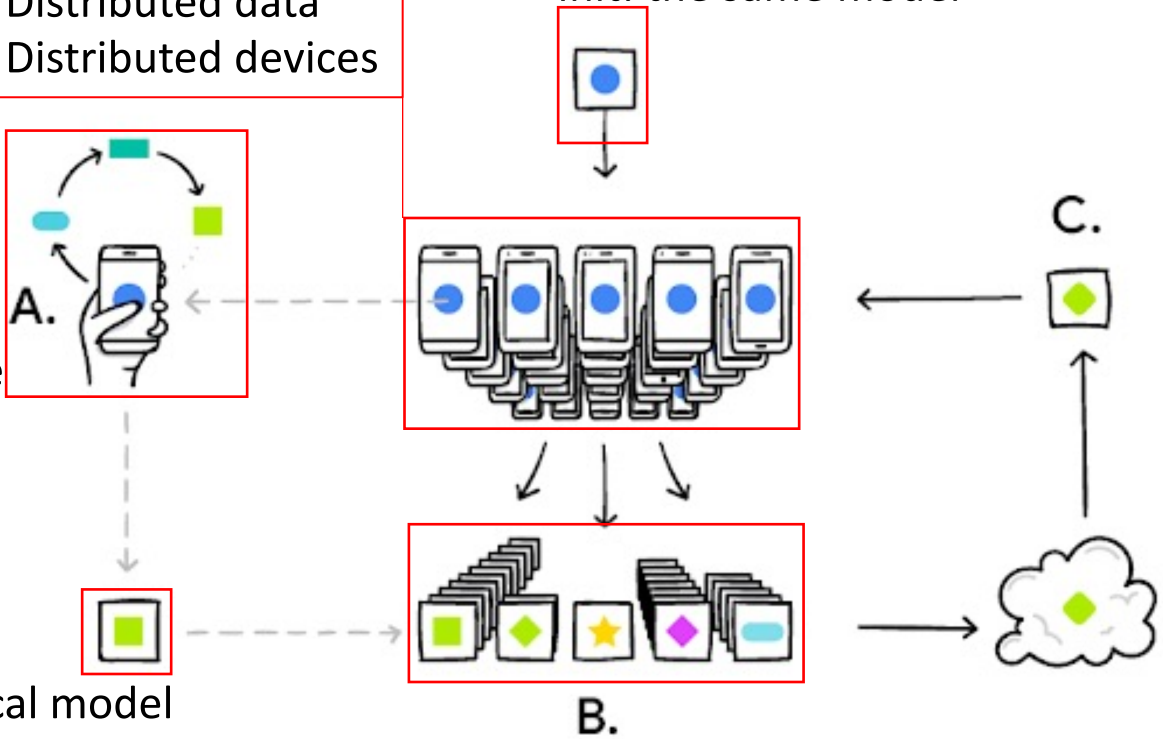
https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Federated learning

1. Distributed data
2. Distributed devices

Init: the same model

Local data → local update



Only upload local model
(no data)

Central server → aggregate

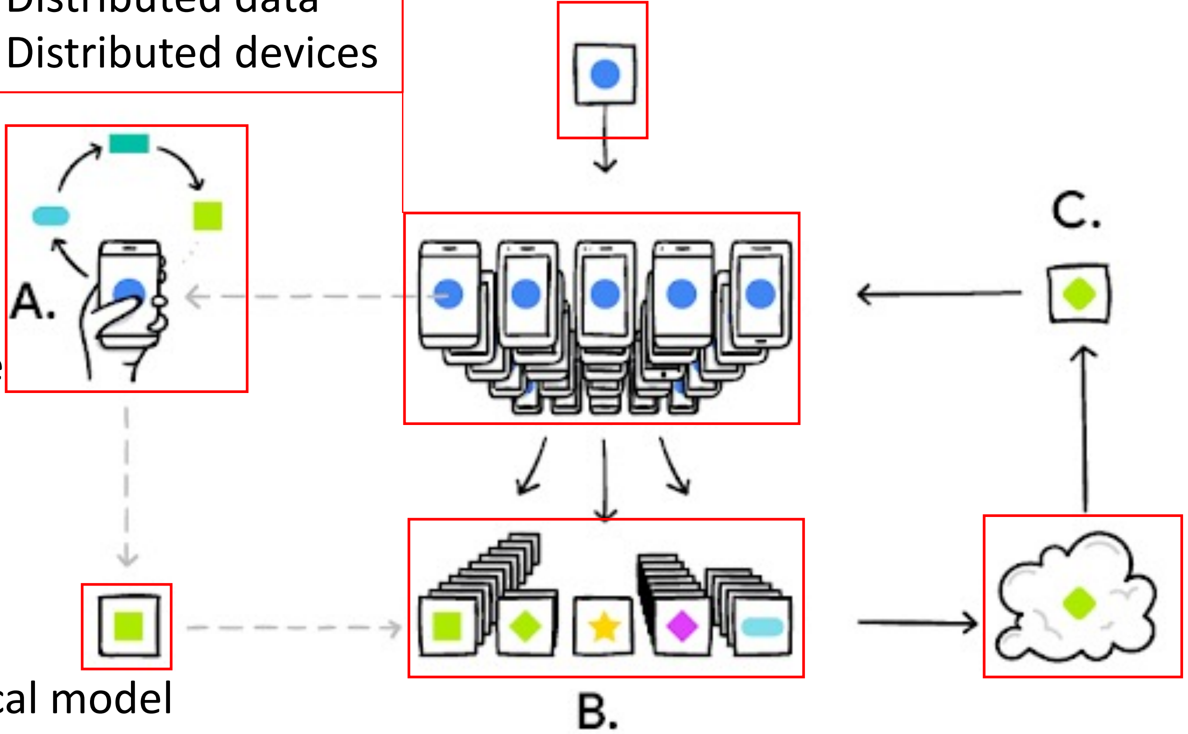
https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Federated learning

1. Distributed data
2. Distributed devices

Init: the same model

Local data → local update



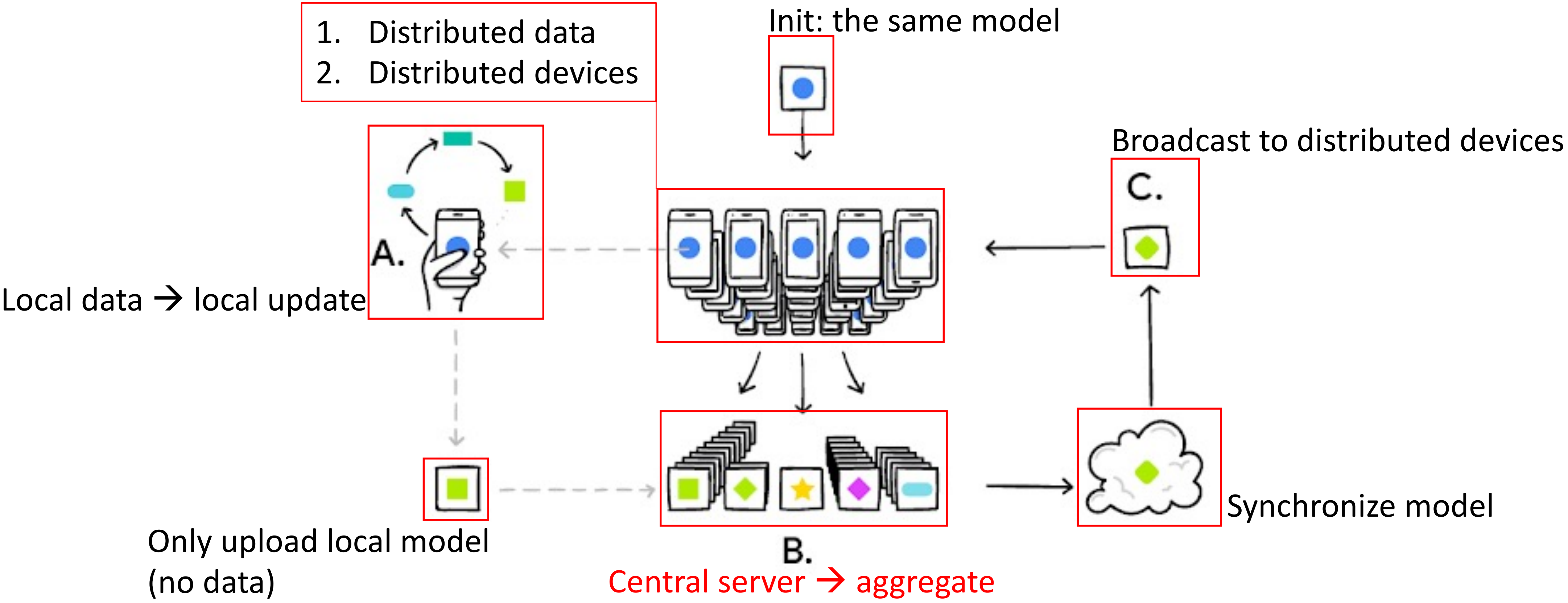
Synchronize model

Only upload local model
(no data)

Central server → aggregate

https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Federated learning



https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Federated learning via meta learning

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i(\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})).$$

Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

n devices/tasks

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \boxed{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i}(\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})).$$

Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

n devices/tasks

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \boxed{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i}(\boxed{\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})}).$$

Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

n devices/tasks

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \boxed{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i}\big(\boxed{\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})}\big).$$

Local model ← using local data for updating (a single step)

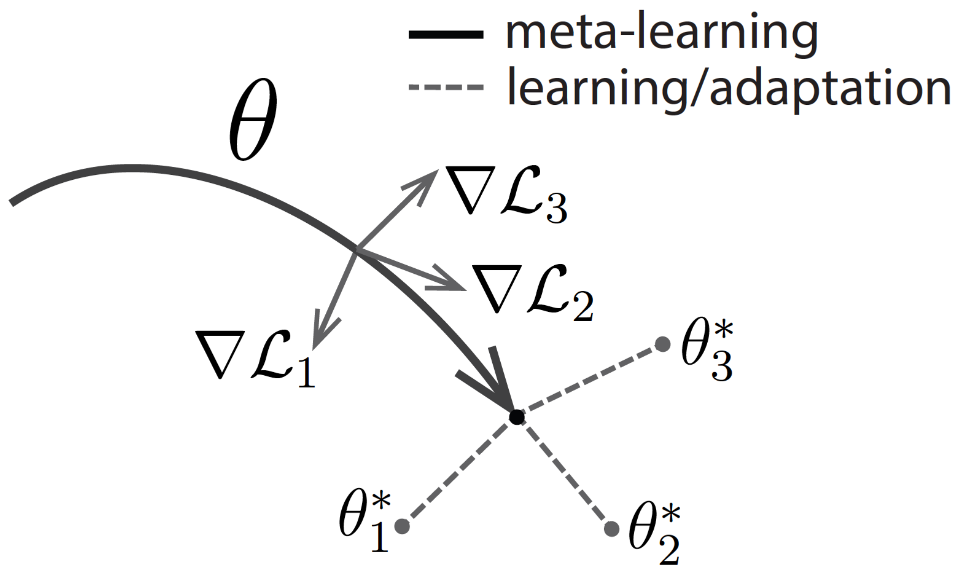
Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

n devices/tasks

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \boxed{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i}(\boxed{\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})}).$$

Local model 🡐 using local data for updating (a single step)



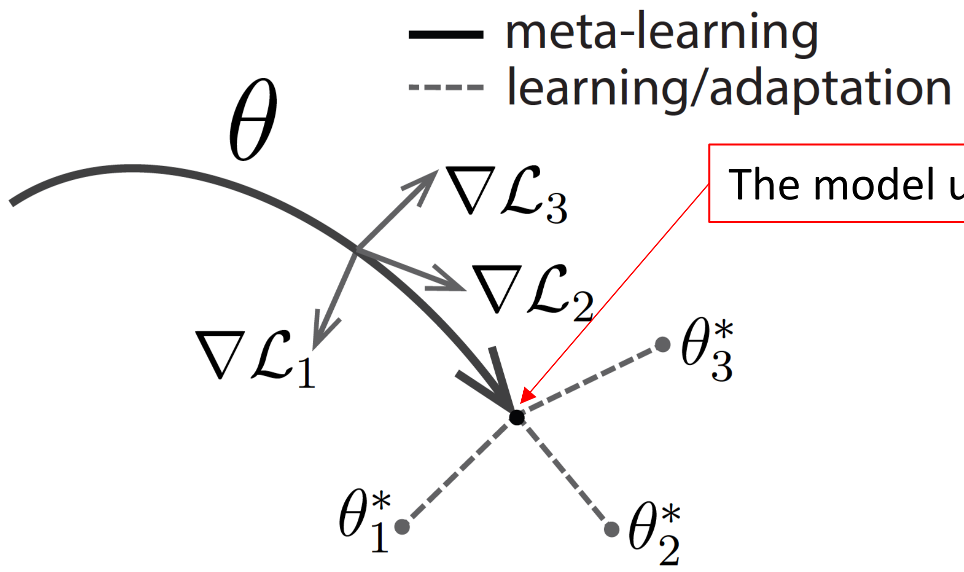
Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

n devices/tasks

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \boxed{\frac{1}{n}\sum_{i=1}^{n}} \mathcal{L}_i(\boxed{\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})}).$$

Local model 🡐 using local data for updating (a single step)



The model used for broadcast

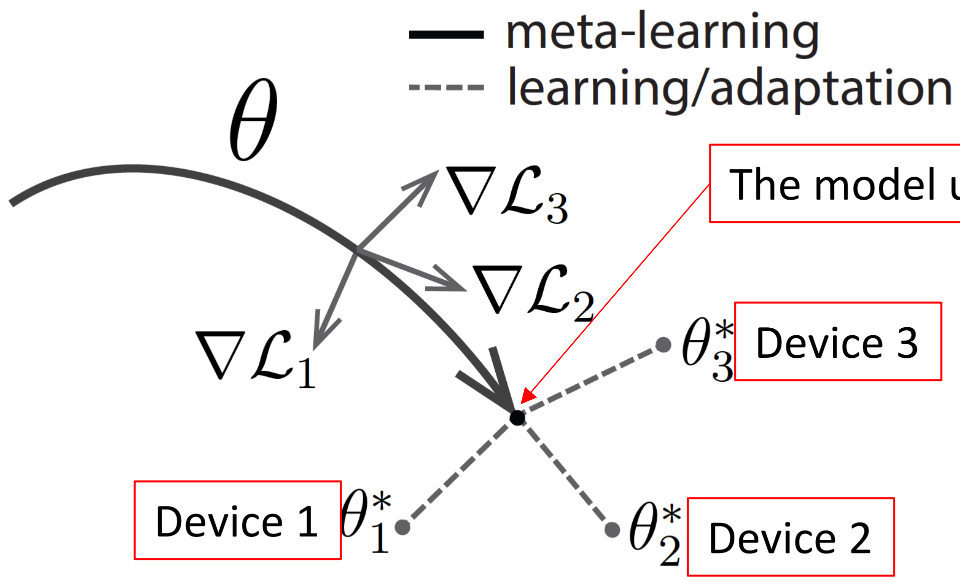
Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

n devices/tasks

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \boxed{\frac{1}{n}\sum_{i=1}^{n}} \mathcal{L}_i(\boxed{\mathbf{w}-\alpha\nabla\mathcal{L}_i(\mathbf{w})}).$$

Local model 🡐 using local data for updating (a single step)



The model used for broadcast

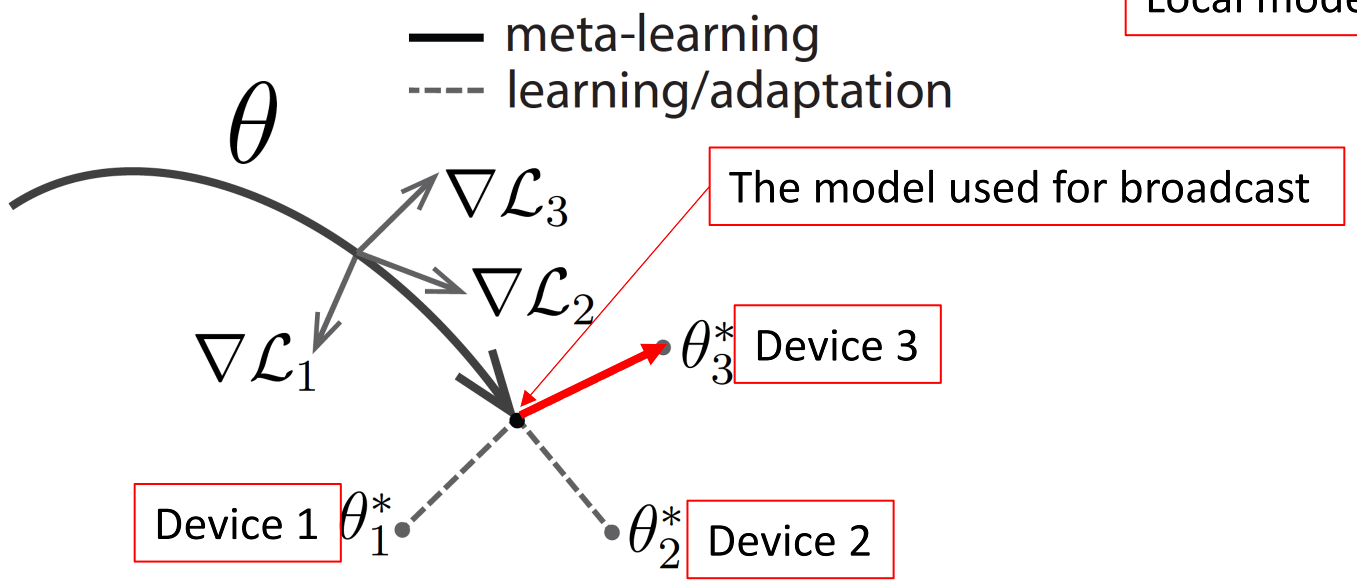
Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

n devices/tasks

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \boxed{\frac{1}{n}\sum_{i=1}^{n}} \mathcal{L}_i(\boxed{\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})}).$$

Local model 🡐 using local data for updating (a single step)



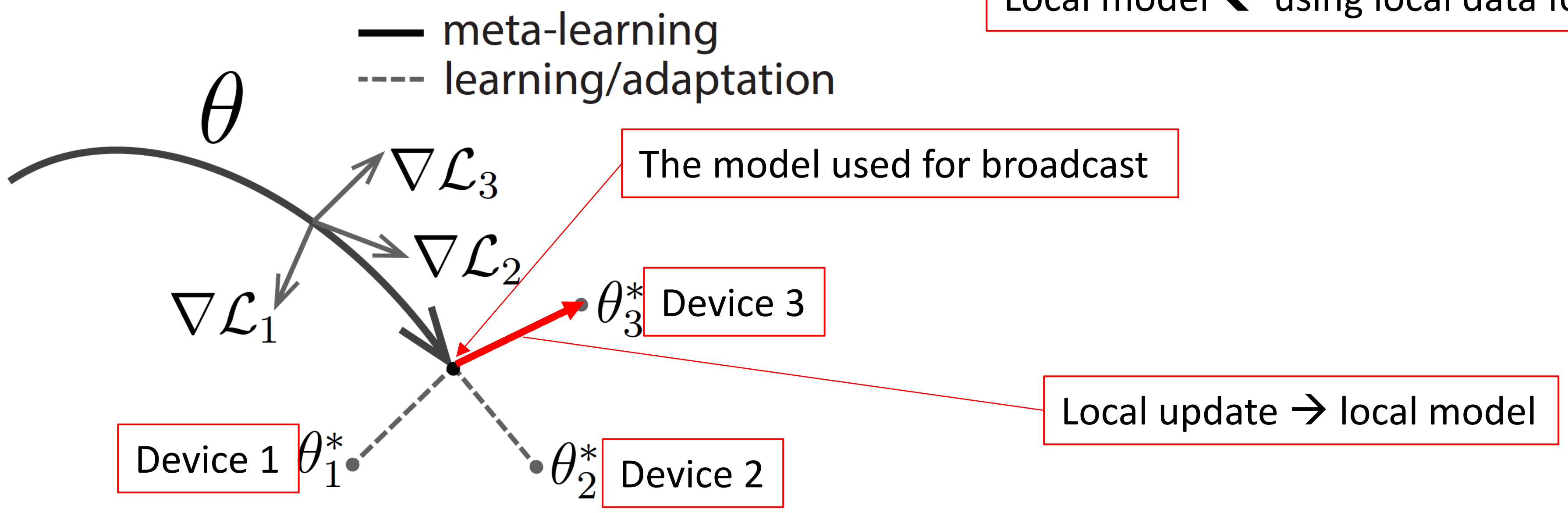
Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

n devices/tasks

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \boxed{\frac{1}{n}\sum_{i=1}^{n}} \mathcal{L}_i(\boxed{\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})}).$$

Local model 🡐 using local data for updating (a single step)



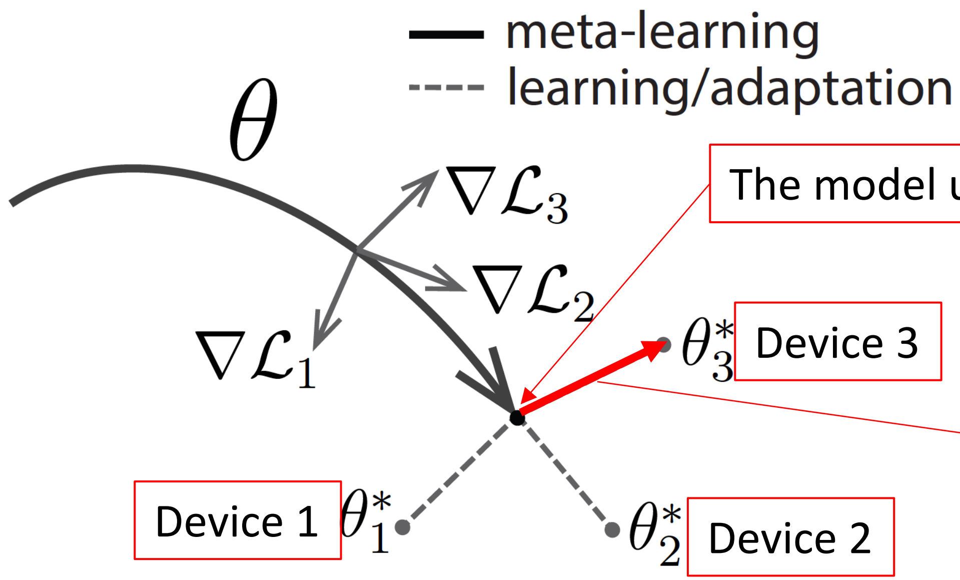
Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

n devices/tasks

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \boxed{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i}\left(\boxed{\mathbf{w} - \alpha\nabla\mathcal{L}_i(\mathbf{w})}\right).$$

Local model ← using local data for updating (a single step)



The model used for broadcast

Local update → local model

Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

Personalized Federated Learning

Algorithm	Client Sampling	Sample Complexity	Communication Complexity	Avg. (K)	#Data points Per Iteration
Per-FedAvg (Fallah et al., 2020b)	✗	$\mathcal{O}\left(n\epsilon^{-6}\right)$	$\mathcal{O}(\epsilon^{-3})$		$\mathcal{O}(\epsilon^{-2})$
Per-FedAvg **(This work)**	✓(3)	$\mathcal{O}(\epsilon^{-7})$	$\mathcal{O}(\epsilon^{-4})$		$\mathcal{O}(\epsilon^{-2})$
LocalMOML (This work)	✗	$\boldsymbol{\mathcal{O}\left(n\epsilon^{-5}\right)}$	$\mathcal{O}(\epsilon^{-3})$		$\boldsymbol{\mathcal{O}(1)}$
LocalMOML (This work)	✓	$\boldsymbol{\mathcal{O}\left(\epsilon^{-6}\right)}$	$\mathcal{O}(\epsilon^{-4})$		$\boldsymbol{\mathcal{O}(1)}$

Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

Personalized Federated Learning

Algorithm	Client Sampling	Sample Complexity	Communication Complexity	Avg. #Data points (K) Per Iteration
Per-FedAvg (Fallah et al., 2020b)	✗	$\mathcal{O}\left(n\epsilon^{-6}\right)$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$
Per-FedAvg **(This work)**	✓(3)	$\mathcal{O}(\epsilon^{-7})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$
LocalMOML (This work)	✗	$\boldsymbol{\mathcal{O}\left(n\epsilon^{-5}\right)}$	$\mathcal{O}(\epsilon^{-3})$	$\boldsymbol{\mathcal{O}(1)}$
LocalMOML (This work)	✓	$\boldsymbol{\mathcal{O}\left(\epsilon^{-6}\right)}$	$\mathcal{O}(\epsilon^{-4})$	$\boldsymbol{\mathcal{O}(1)}$

Number of data samples (on local devices)

Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

Personalized Federated Learning

Algorithm	Client Sampling	Sample Complexity	Communication Complexity	Avg. #Data points (K) Per Iteration
Per-FedAvg (Fallah et al., 2020b)	✗	$\mathcal{O}\left(n\epsilon^{-6}\right)$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$
Per-FedAvg **(This work)**	✓[(3)]	$\mathcal{O}(\epsilon^{-7})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$
LocalMOML (This work)	✗	$\boldsymbol{\mathcal{O}\left(n\epsilon^{-5}\right)}$	$\mathcal{O}(\epsilon^{-3})$	$\boldsymbol{\mathcal{O}(1)}$
LocalMOML (This work)	✓	$\boldsymbol{\mathcal{O}\left(\epsilon^{-6}\right)}$	$\mathcal{O}(\epsilon^{-4})$	$\boldsymbol{\mathcal{O}(1)}$

Number of data samples (on local devices)

Number of broadcasts

Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).

Federated learning via meta learning

1. drop the dependence on n
2. reduce order of \epsilon

Personalized Federated Learning

Algorithm	Client Sampling	Sample Complexity	Communication Complexity	Avg. #Data points (K) Per Iteration
Per-FedAvg (Fallah et al., 2020b)	✗	$\mathcal{O}\left(n\epsilon^{-6}\right)$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$
Per-FedAvg **(This work)**	✓(3)	$\mathcal{O}(\epsilon^{-7})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$
LocalMOML (This work)	✗	$\boldsymbol{\mathcal{O}\left(n\epsilon^{-5}\right)}$	$\mathcal{O}(\epsilon^{-3})$	$\boldsymbol{\mathcal{O}(1)}$
LocalMOML (This work)	✓	$\boldsymbol{\mathcal{O}\left(\epsilon^{-6}\right)}$	$\mathcal{O}(\epsilon^{-4})$	$\boldsymbol{\mathcal{O}(1)}$

Number of data samples (on local devices)

Number of broadcasts

Wang, Bokun, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. "Memory-based Optimization Methods for Model-Agnostic Meta-Learning." *arXiv preprint arXiv:2106.04911* (2021).