Weakly- & Self-Supervised Learning

Lantian Li

2020.02.24

Weakly Supervised Learning

Supervised learning

- Concepts
  - learning from a large number of examples
  - each example has its individual label.

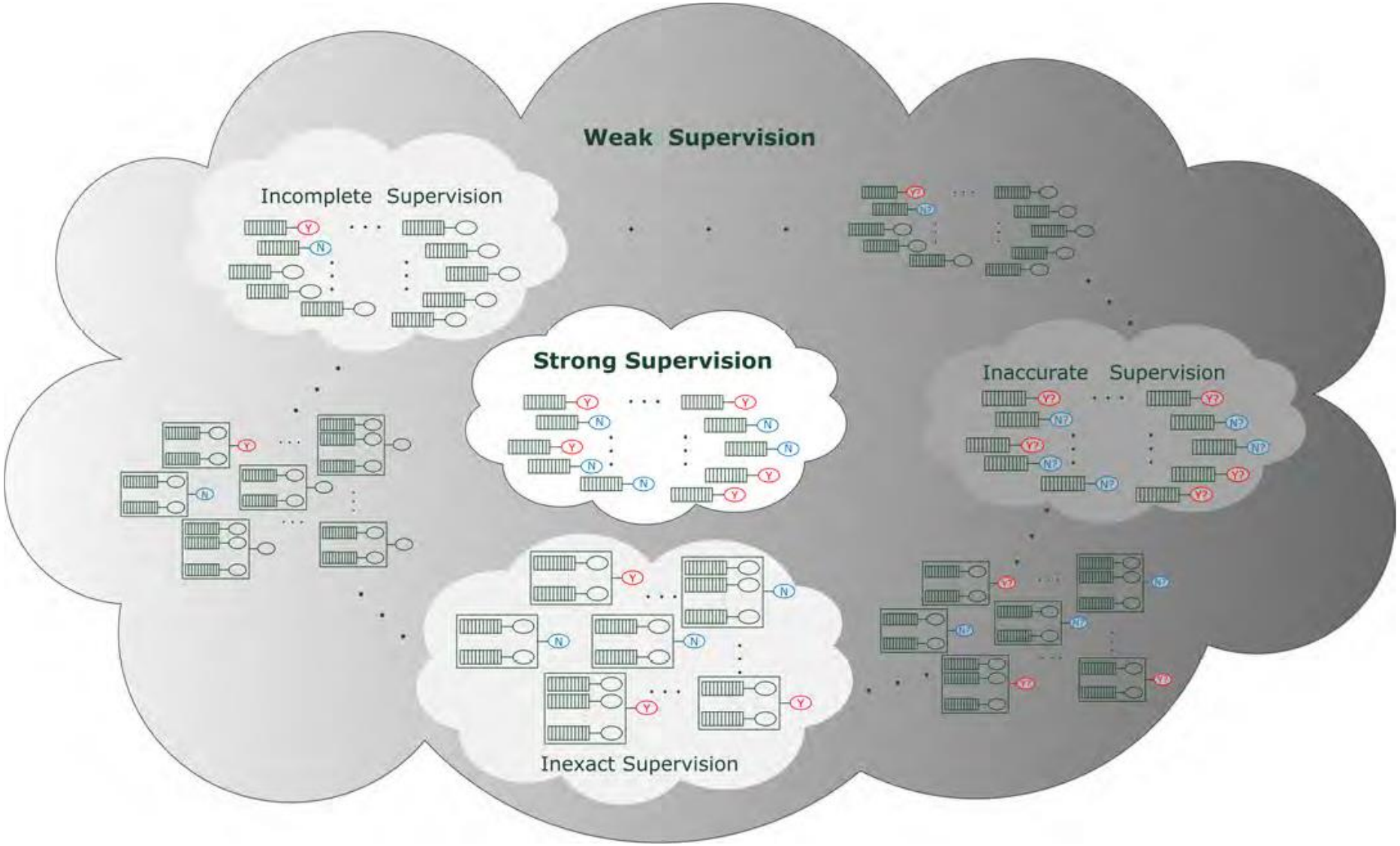
- Pros and Cons
  - task-related, good performance (deep neural networks)
  - high cost of data labeling.

Weakly supervised learning

- Concepts
  - learning with weak supervision.
  - noisy, limited, or imprecise sources

- Three types of weak supervision
  - *incomplete*: speaker / image categorization
  - *inexact*: object in a video / image / doc.
  - *inaccurate*: crowdsourcing

Weak supervision

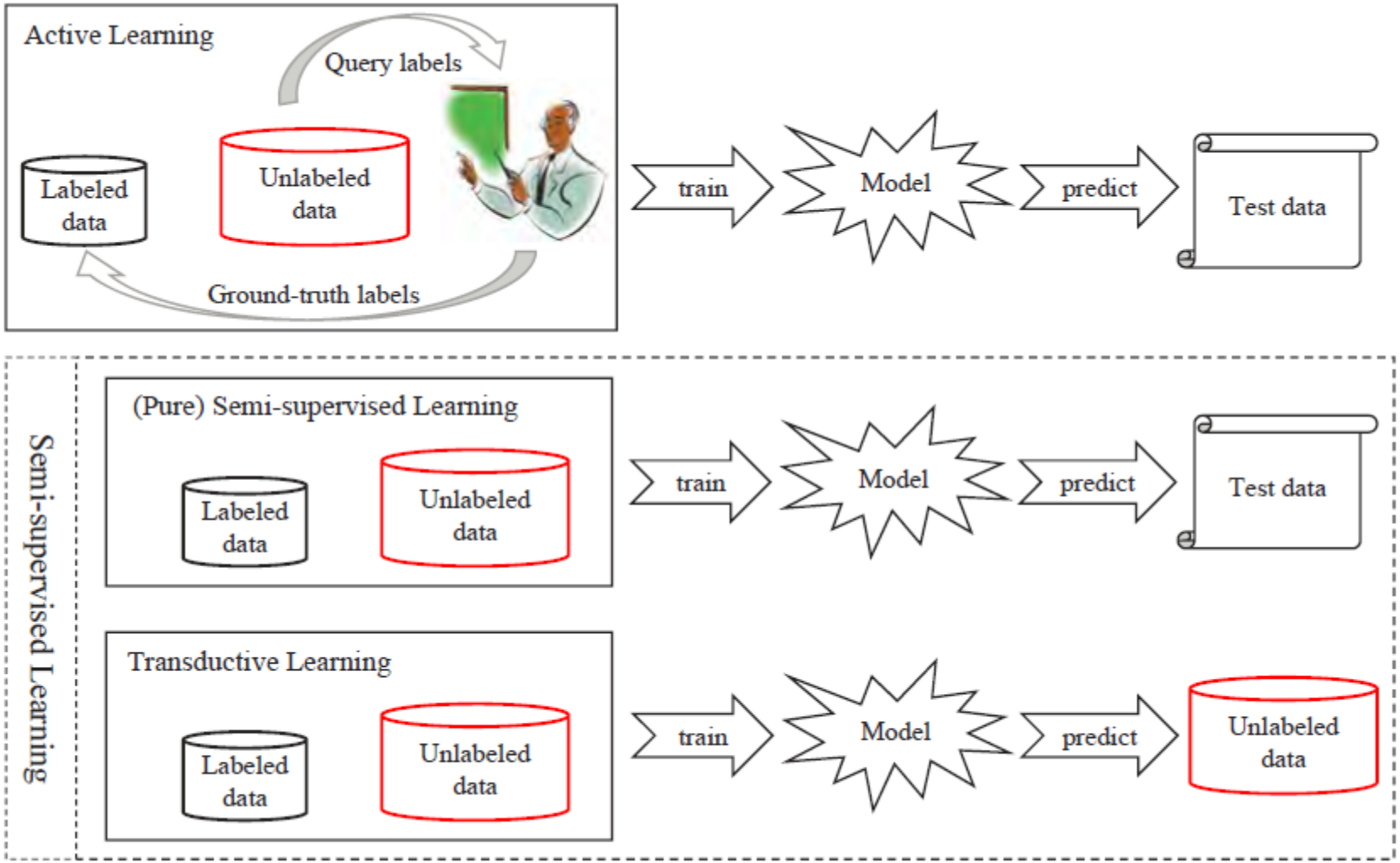


Incomplete Supervision

- Active learning
  - with human intervention
  - labels can be queried from an oracle.

- Semi-supervised learning
  - without human intervention
  - automatically exploit unlabeled data to improve performance

Incomplete Supervision



Active learning

- Goal
  - minimize the number of queries to minimize labeling cost
  - to select valuable unlabeled data

- Selection criteria
  - *informativeness*: uncertainty and entropy (0.55 vs. 0.99; 4:3 vs 6:1)
  - *representativeness*: sampling distribution (clusters)

Semi-supervised learning

- Goal
  - data without labels to help construct models

Explanation by GMM

$$f(\boldsymbol{x}|\Theta) = \sum_{j=1}^{n} \alpha_j f(\boldsymbol{x}|\theta_j) \,, \tag{1}$$

where α_i is the mixture coefficient, $\sum_{i=1}^{n} \alpha_i = 1$, and $\Theta = \{\theta_i\}$ are the model parameters. In this case, label y_i can be considered as a random variable whose distribution $P(y_i|\boldsymbol{x}_i, g_i)$ is determined by the mixture component g_i and the feature vector $\boldsymbol{x}_i$. According to the maximum *a posterior* criterion, we have the model

$$h(\boldsymbol{x}) = \arg\max_{c \in \{Y,N\}} \sum_{j=1}^{n} P\left(y_i = c|g_i = j, \boldsymbol{x}_i\right) P\left(g_i = j|\boldsymbol{x}_i\right) \,, \tag{2}$$

where

$$P\left(g_i = j|\boldsymbol{x}_i\right) = \frac{\alpha_j f(\boldsymbol{x}_i|\theta_j)}{\sum_{k=1}^{n} \alpha_k f(\boldsymbol{x}_i|\theta_k)} \,.$$

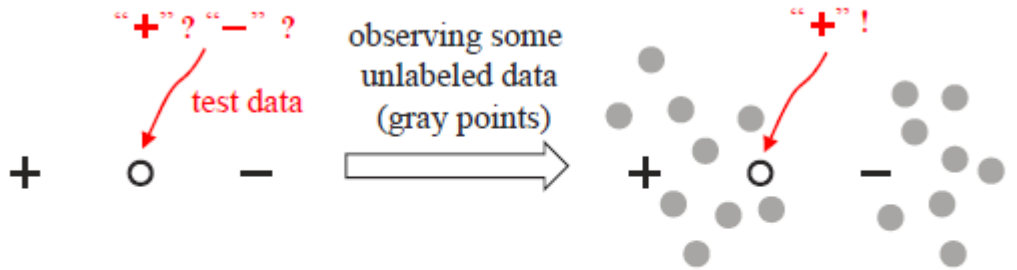


Figure 3. Illustration of the usefulness of unlabeled data

Semi-supervised learning

- Goal
  - data without labels to help construct models

- Data assumptions
  - cluster assumption (the same cluster has the same class)
  - manifold assumption (nearby instances have similar predictions)

Semi-supervised learning

- Categories
  - generative methods (GMMs)
  - graph-based methods (knowledge graph: relation completion)
  - low-density separation methods (S3VMs)

S3VMs vs. SVM

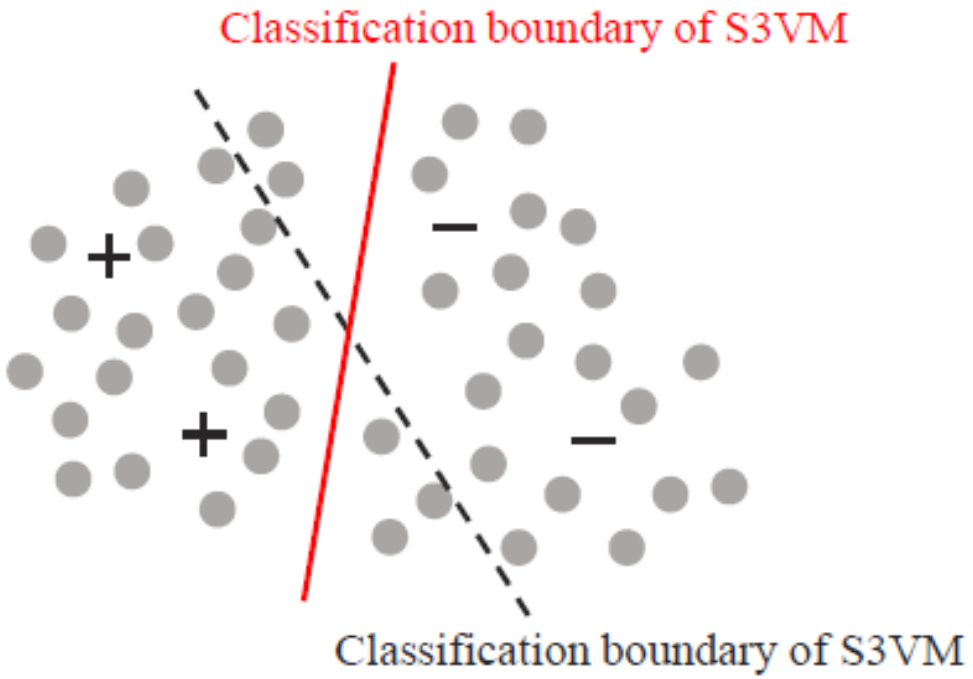


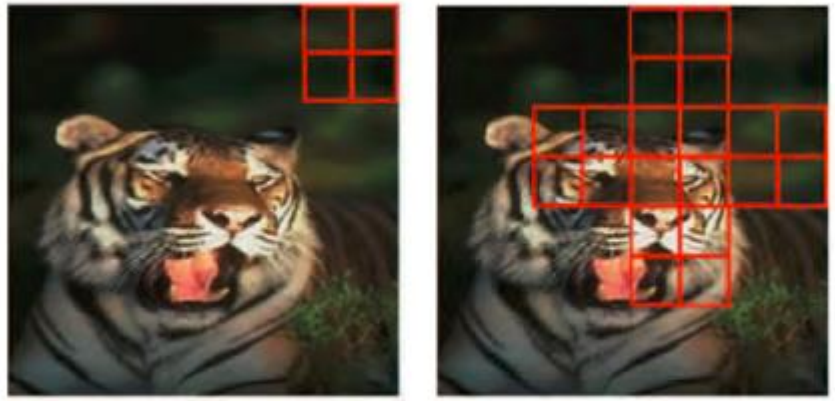
Figure 4. Illustration of the usefulness of un-labeled data

Semi-supervised learning

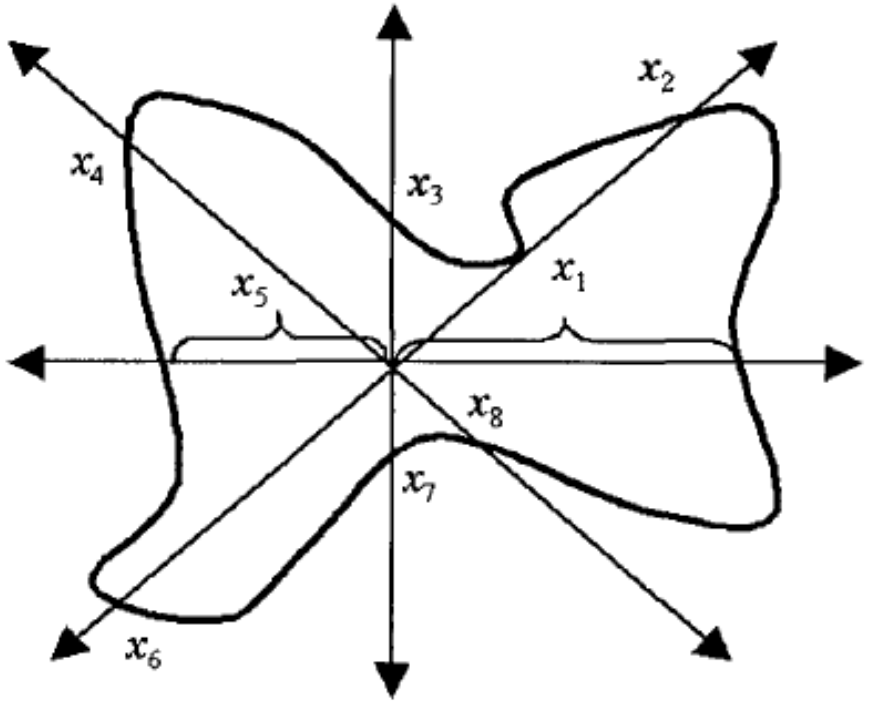
- Categories
  - generative methods (GMMs)
  - graph-based methods (knowledge graph: relation completion)
  - low-density separation methods (S3VMs)
  - disagreement-based methods (co-training)

Inexact Supervision

- Multi-instance learning
  - object in a video / image / doc.
  - Bag generates instances based on concept.



(a) SB (b) SBN



Inaccurate Supervision

- Learning with label noise
  - add error rate in the cost function
  - data editing

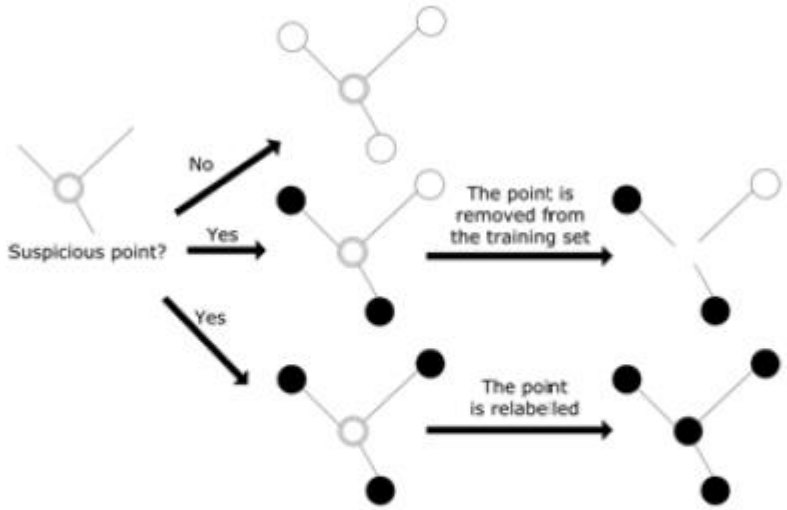


Figure 6. Identify and remove/relabel suspicious points

Inaccurate Supervision

- Learning with label noise
  - add error rate in the cost function
  - data editing
  - crowdsourcing
    - ensemble methods with voting
    - spammer elimination
    - combine with economics (Nash equilibrium)

Quick summary

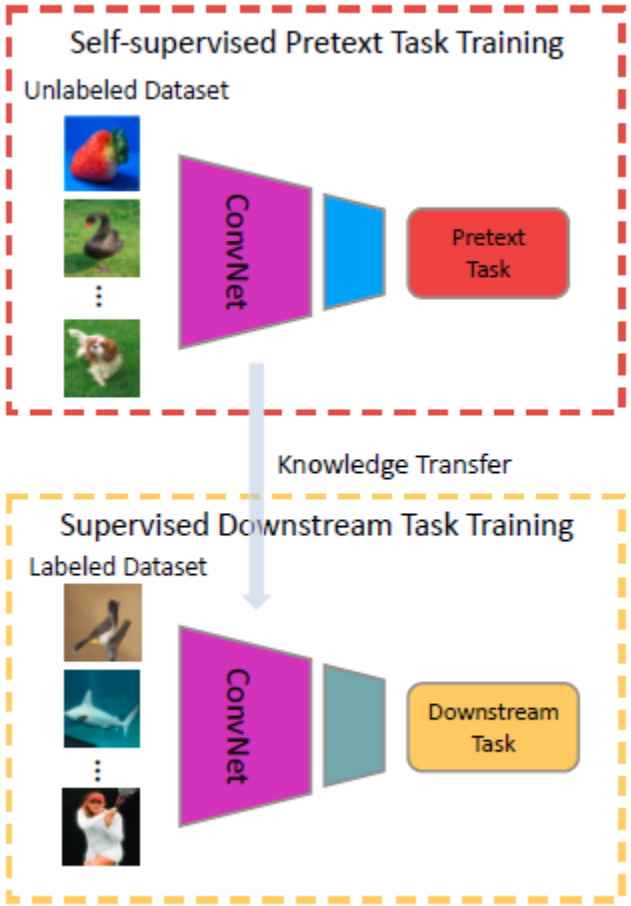
- Weakly supervised learning

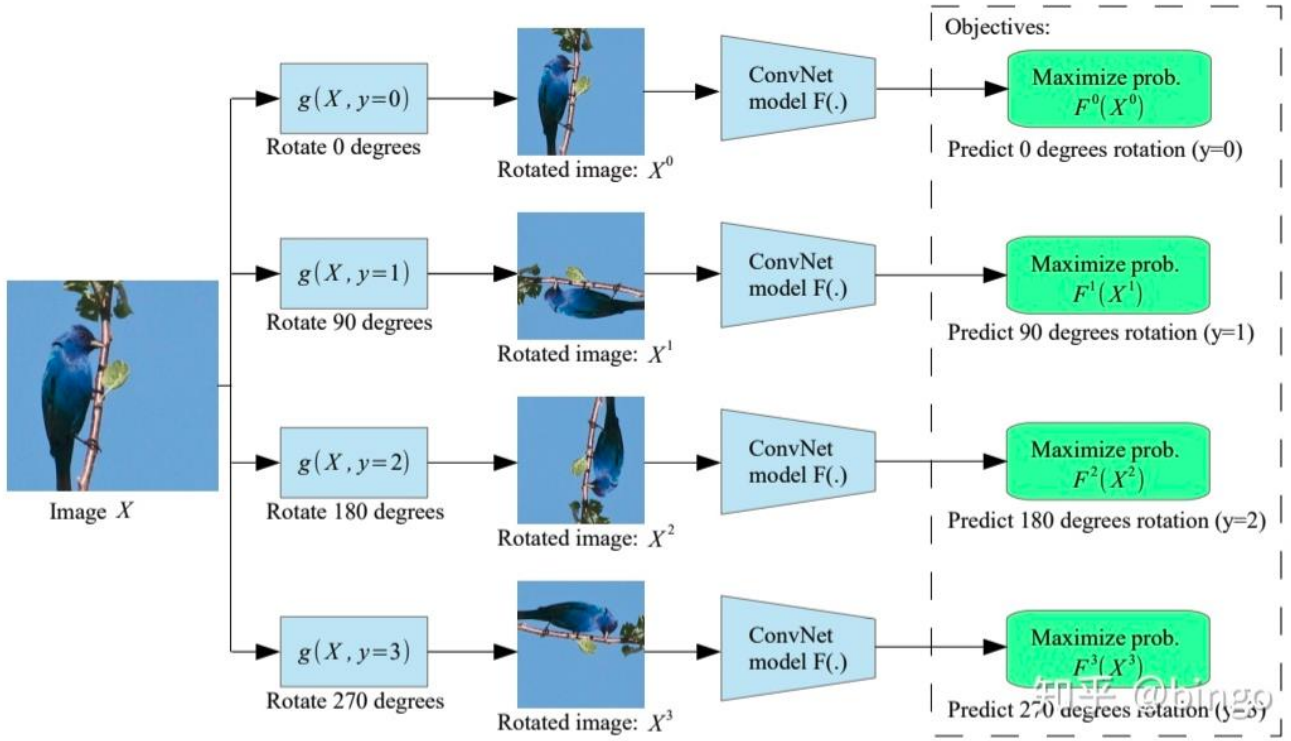
- Weak supervision
  - Semi-supervised learning

Self-Supervised Learning

Self-supervised learning

- Self-supervised learning is supervised learning without human-annotated labels.



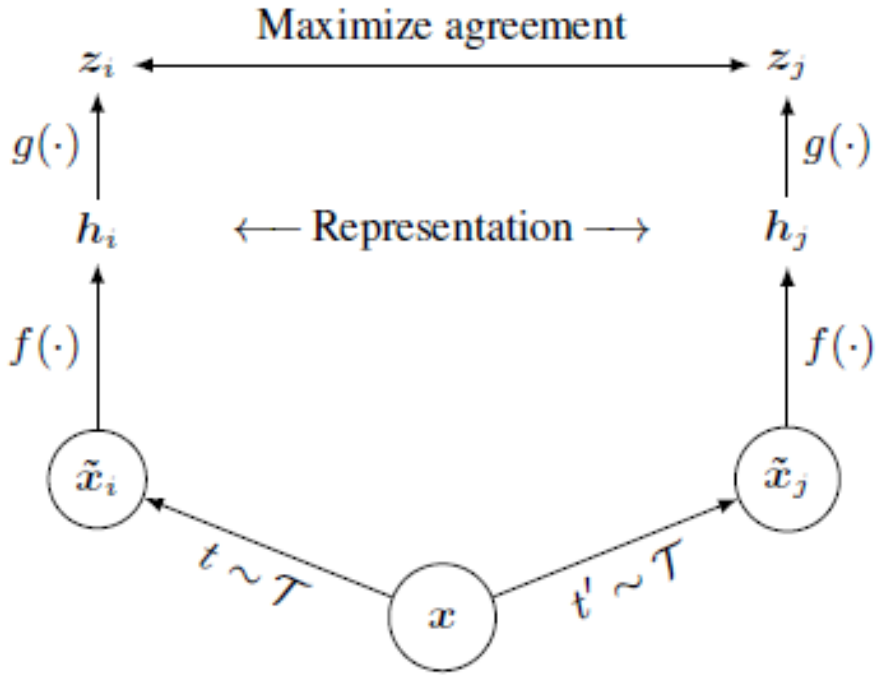


Learning representation without supervision

- Generative model
  - AE, VAE, PixRNNs …
- Discriminative model
  - objective function is the same as supervised learning.
  - perform on pretext task.
  - but inputs and labels are derived from an unlabeled dataset.

SimCLR

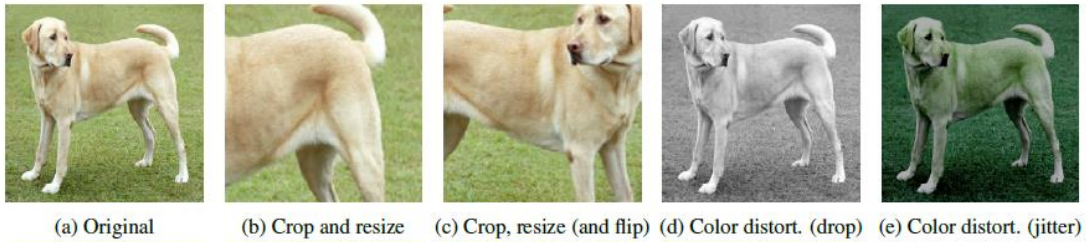
- data augmentation operations
- projection head
- contrastive loss function



Algorithm 1 SimCLR's main learning algorithm.

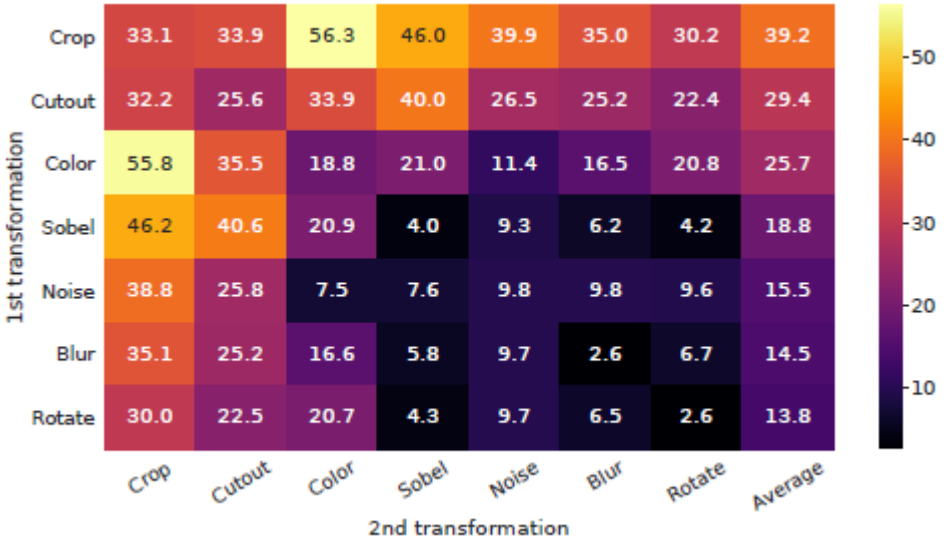
input: batch size N, temperature τ, structure of $f, g, \mathcal{T}$.
for sampled minibatch $\{x_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}, t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{x}_{2k-1} = t(x_k)$
    $h_{2k-1} = f(\tilde{x}_{2k-1})$ # representation
    $z_{2k-1} = g(h_{2k-1})$ # projection
    # the second augmentation
    $\tilde{x}_{2k} = t'(x_k)$
    $h_{2k} = f(\tilde{x}_{2k})$ # representation
    $z_{2k} = g(h_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\tau \|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i, j)$ **as** $\ell(i, j) = -\log \frac{\exp(s_{i,j})}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k})}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network f

Data augmentation



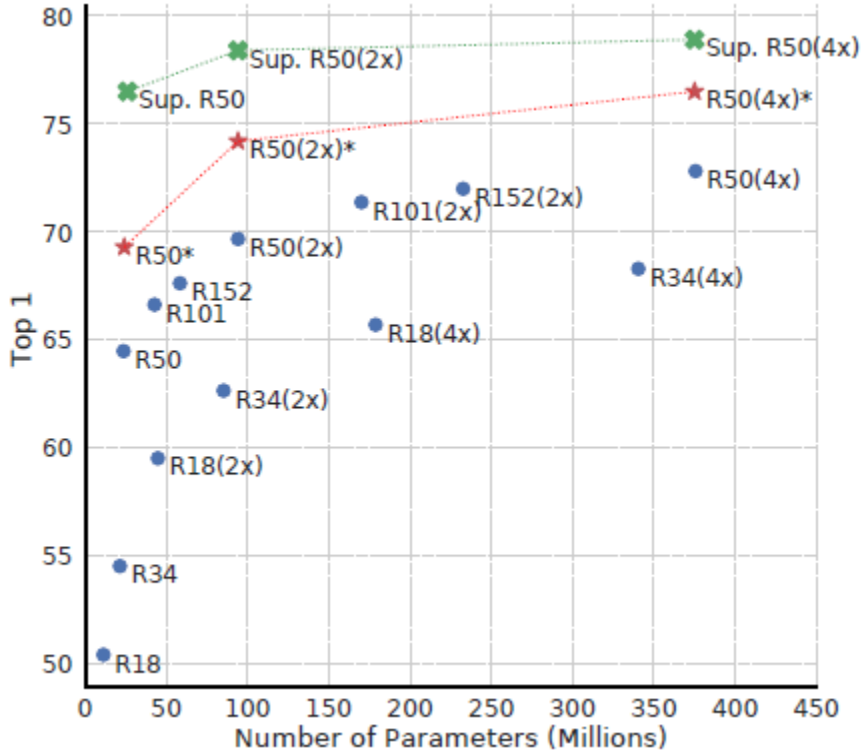
(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering



Methods	Color distortion strength 1/8	1/4	1/2	1	1 (+Blur)	AutoAug
SimCLR	59.6	61.0	62.6	63.2	64.5	61.1
Supervised	77.0	76.7	76.5	75.7	75.4	77.1

Projection head



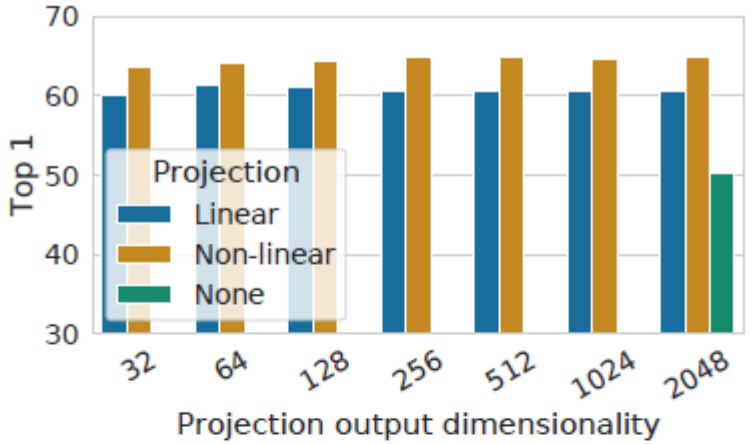


Figure 8. Linear evaluation of representations with different projection heads $g(\cdot)$ and various dimensions of $z = g(h)$. The representation h (before projection) is 2048-dimensional here.

What to predict?	Random guess	Representation h	Representation $g(h)$
Color vs grayscale	80	99.3	97.4
Rotation	25	67.6	25.6
Orig. vs corrupted	50	99.5	59.6
Orig. vs Sobel filtered	50	96.6	56.3

Table 3. Accuracy of training additional MLPs on different representations to predict the transformation applied. Other than crop and color augmentation, we additionally and independently add rotation (one of $\{0°, 90°, 180°, 270°\}$), Gaussian noise, and Sobel filtering transformation during the pretraining for the last three rows. Both h and $g(h)$ are of the same dimensionality, i.e. 2048.

Contrastive loss

Name	Negative loss function	Gradient w.r.t. $\boldsymbol{u}$
NT-Xent	$\boldsymbol{u}^T\boldsymbol{v}^+/\tau - \log \sum_{v\in\{v^+,v^-\}} \exp(\boldsymbol{u}^T\boldsymbol{v}/\tau)$	$(1 - \frac{\exp(\boldsymbol{u}^T\boldsymbol{v}^+/\tau)}{Z(\boldsymbol{u})})/\tau\boldsymbol{v}^+ - \sum_{v\in\{v^+,v^-\}} \frac{\exp(\boldsymbol{u}^T\boldsymbol{v}/\tau)}{Z(\boldsymbol{u})}/\tau\boldsymbol{v}$
NT-Logistic	$\log\sigma(\boldsymbol{u}^T\boldsymbol{v}^+/\tau) + \log\sigma(-\boldsymbol{u}^T\boldsymbol{v}^-/\tau)$	$(\sigma(-\boldsymbol{u}^T\boldsymbol{v}^+/\tau))/\tau\boldsymbol{v}^+ - \sigma(\boldsymbol{u}^T\boldsymbol{v}^-/\tau)/\tau\boldsymbol{v}^-$
Margin Triplet	$-\max(\boldsymbol{u}^T\boldsymbol{v}^- - \boldsymbol{u}^T\boldsymbol{v}^+ + m, 0)$	$\boldsymbol{v}^+ - \boldsymbol{v}^-$ if $\boldsymbol{u}^T\boldsymbol{v}^+ - \boldsymbol{u}^T\boldsymbol{v}^- < m$ else $\mathbf{0}$

Table 2. Negative loss functions and their gradients. All input vectors, i.e. $\boldsymbol{u}, \boldsymbol{v}^+, \boldsymbol{v}^-$, are ℓ_2 normalized. NT-Xent is an abbreviation for "Normalized Temperature-scaled Cross Entropy". Different loss functions impose different weightings of positive and negative examples.

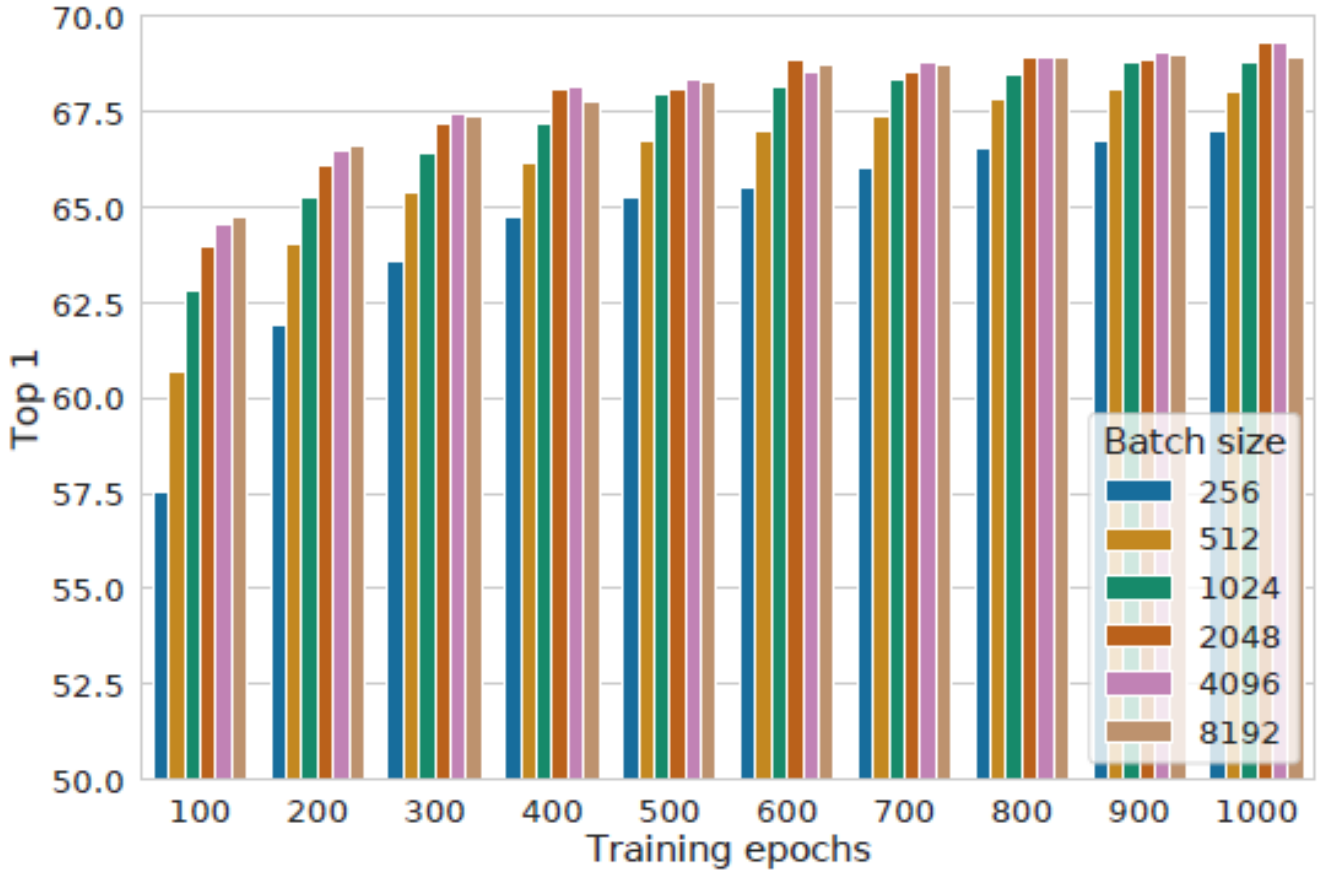
NT-Xent -> 1:1 + 1:2N-1
NT-Logistic -> 1:1
Margin Triple -> 1:1 + 1:1

Margin	NT-Logi.	Margin (sh)	NT-Logi.(sh)	NT-Xent
50.9	51.6	57.5	57.9	63.9

Table 4. Linear evaluation (top-1) for models trained with different loss functions. "sh" means using semi-hard negative mining.

Contrastive loss

Larger batch sizes and longer training compared with supervised learning



Self-supervised methods

Method	Architecture	Param.	Top 1	Top 5
Methods using ResNet-50:				
Local Agg.	ResNet-50	24	60.2	-
MoCo	ResNet-50	24	60.6	-
PIRL	ResNet-50	24	63.6	-
CPC v2	ResNet-50	24	63.8	85.3
SimCLR (ours)	ResNet-50	24	**69.3**	**89.0**
Methods using other architectures:				
Rotation	RevNet-50 (4×)	86	55.4	-
BigBiGAN	RevNet-50 (4×)	86	61.3	81.9
AMDIM	Custom-ResNet	626	68.1	-
CMC	ResNet-50 (2×)	188	68.4	88.2
MoCo	ResNet-50 (4×)	375	68.6	-
CPC v2	ResNet-161 (∗)	305	71.5	90.1
SimCLR (ours)	ResNet-50 (2×)	94	74.2	92.0
SimCLR (ours)	ResNet-50 (4×)	375	**76.5**	**93.2**

Table 6. ImageNet accuracies of linear classifiers trained on representations learned with different self-supervised methods.

Semi-supervised methods

Method	Architecture	Label fraction 1%	Label fraction 10%
		Top 5	Top 5
Methods using other label-propagation:			
Pseudo-label	ResNet50	51.6	82.4
VAT+Entropy Min.	ResNet50	47.0	83.4
UDA (w. RandAug)	ResNet50	-	88.5
FixMatch (w. RandAug)	ResNet50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(∗)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

Table 7. ImageNet accuracy of models trained with few labels.

Transfer learning

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

Table 8. Comparison of transfer learning performance of our self-supervised approach with supervised baselines across 12 natural image classification datasets, for ResNet-50 $(4\times)$ models pretrained on ImageNet. Results not significantly worse than the best ($p > 0.05$, permutation test) are shown in bold. See Appendix B.6 for experimental details and results with standard ResNet-50.

Quick summary

- Speaker representation