Unlabeled and Incomplete Data are Cool Too

Mateusz Pach

Attention-based Deep Multiple Instance Learning

Maximilian Ilse [*1] **Jakub M. Tomczak** [*1] **Max Welling** [1]

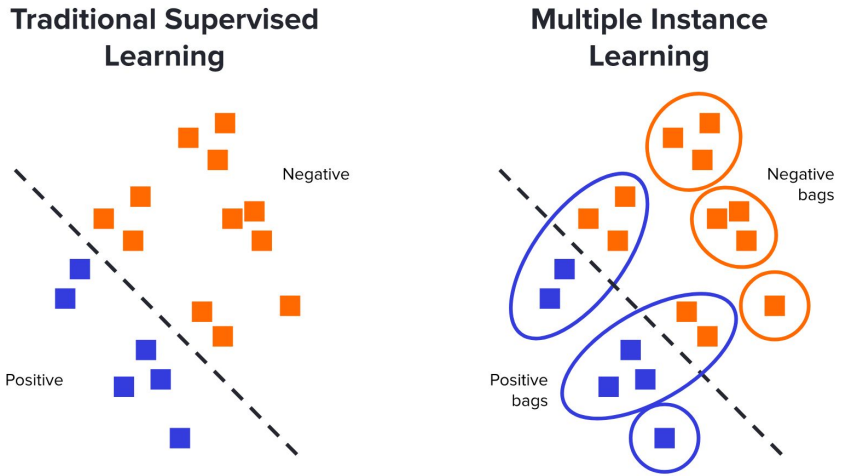
Abstract

Multiple instance learning (MIL) is a variation of supervised learning where a single class label is assigned to a bag of instances. In this paper, we state the MIL problem as learning the Bernoulli distribution of the bag label where the bag label probability is fully parameterized by neural networks. Furthermore, we propose a neural network-based permutation-invariant aggregation operator that corresponds to the attention

model that predicts a bag label, *e.g.*, a medical diagnosis. An additional challenge is to discover *key instances* (Liu et al., 2012), *i.e.*, the instances that trigger the bag label. In the medical domain the latter task is of great interest because of legal issues[1] and its usefulness in clinical practice. In order to solve the primary task of a bag classification different methods are proposed, such as utilizing similarities among bags (Cheplygina et al., 2015b), embedding instances to a compact low-dimensional representation that is further fed to a bag-level classifier (Andrews et al., 2003; Chen

Motivation

- Weakly annotated data is especially common in medical imaging.
- Multiple Instance Learning deals with a bag of instances for which a single class label is assigned.
- Let’s design a more interpretable and flexible method in the MIL setup.



https://nilg.ai/202105/an-introduction-to-multiple-instance-learning/

Approaches

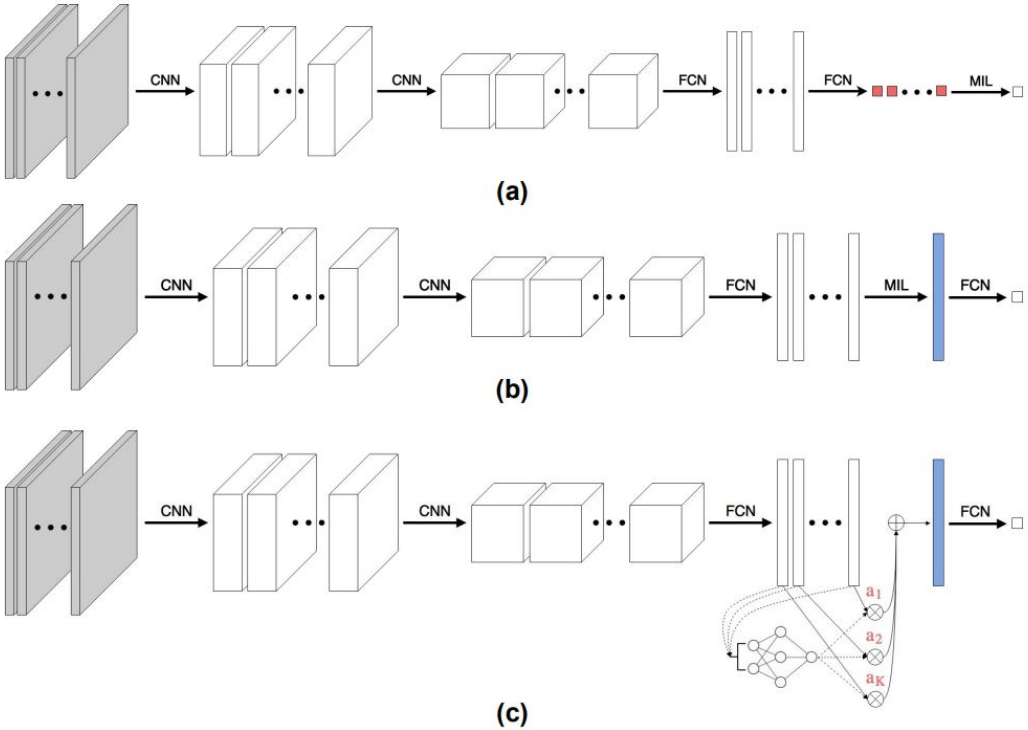


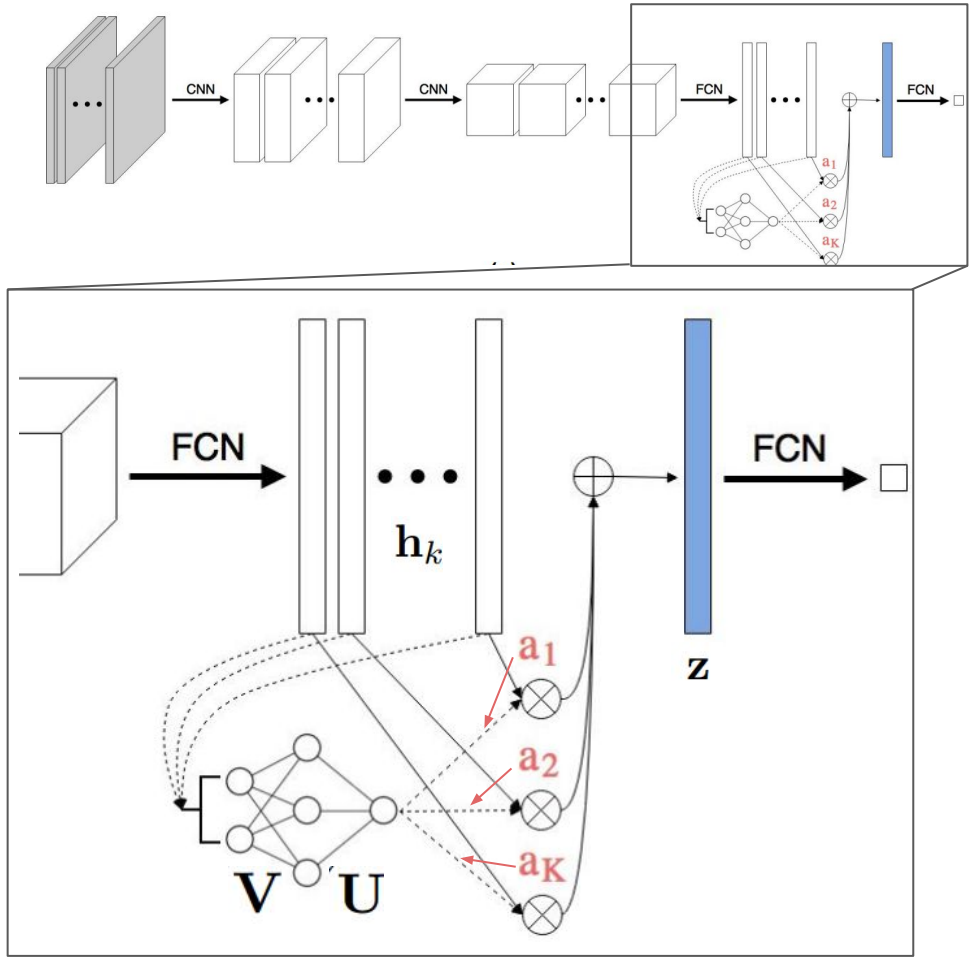
Figure 6. Deep MIL approaches: (a) the instance-based approach, (b) the embedding-based approach, (c) the proposed approach with the attention mechanism as the MIL pooling. Red color corresponds to instance scores, blue color depicts a bag vector representation. *Best viewed in color.*

Method

Embedding-level approach with attention pooling

$$\mathbf{z} = \sum_{k=1}^{K} a_k \mathbf{h}_k$$

$$a_k = \frac{\exp\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\mathbf{h}_k^\top\right) \odot \mathrm{sigm}\left(\mathbf{U}\mathbf{h}_k^\top\right)\right)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\mathbf{h}_j^\top\right) \odot \mathrm{sigm}\left(\mathbf{U}\mathbf{h}_j^\top\right)\right)\}}$$



Databases

Musk1, Musk2, Fox, Tiger, Elephant

MNIST-Bags

Breast Cancer, Colon Cancer

Results

Table 1. Results on classical MIL datasets. Experiments were run 5 times and an average of the classification accuracy (± a standard error of a mean) is reported. [1] (Andrews et al., 2003), [2] (Gärtner et al., 2002), [3] (Zhang & Goldman, 2002) [4] (Zhou et al., 2009) [5] (Wei et al., 2017) [6] (Wang et al., 2016)

METHOD	MUSK1	MUSK2	FOX	TIGER	ELEPHANT
mi-SVM [1]	0.874±N/A	0.836±N/A	0.582±N/A	0.784±N/A	0.822±N/A
MI-SVM [1]	0.779±N/A	0.843±N/A	0.578±N/A	0.840±N/A	0.843±N/A
MI-Kernel [2]	**0.880**±0.031	**0.893**±0.015	**0.603**±0.028	0.842±0.010	0.843±0.016
EM-DD [3]	0.849±0.044	**0.869**±0.048	**0.609**±0.045	0.730±0.043	0.771±0.043
mi-Graph [4]	**0.889**±0.033	**0.903**±0.039	**0.620**±0.044	**0.860**±0.037	**0.869**±0.035
miVLAD [5]	**0.871**±0.043	**0.872**±0.042	**0.620**±0.044	0.811±0.039	**0.850**±0.036
miFV [5]	**0.909**±0.040	**0.884**±0.042	**0.621**±0.049	0.813±0.037	**0.852**±0.036
mi-Net [6]	**0.889**±0.039	**0.858**±0.049	**0.613**±0.035	0.824±0.034	**0.858**±0.037
MI-Net [6]	**0.887**±0.041	**0.859**±0.046	**0.622**±0.038	**0.830**±0.032	**0.862**±0.034
MI-Net with DS [6]	**0.894**±0.042	**0.874**±0.043	**0.630**±0.037	**0.845**±0.039	**0.872**±0.032
MI-Net with RC [6]	**0.898**±0.043	**0.873**±0.044	**0.619**±0.047	**0.836**±0.037	**0.857**±0.040
Attention	**0.892**±0.040	**0.858**±0.048	**0.615**±0.043	**0.839**±0.022	**0.868**±0.022
Gated-Attention	**0.900**±0.050	**0.863**±0.042	**0.603**±0.029	**0.845**±0.018	**0.857**±0.027

Table 2. Results on BREAST CANCER. Experiments were run 5 times and an average (±

METHOD	ACCURACY	PRECISION	RECALL
Instance+max	0.614±0.020	0.585±0.03	0.477±0.087
Instance+mean	0.672±0.026	0.672±0.034	0.515±0.056
Embedding+max	0.607±0.015	0.558±0.013	0.546±0.070
Embedding+mean	**0.741**±0.023	**0.741**±0.023	0.654±0.054
Attention	**0.745**±0.018	0.718±0.021	**0.715**±0.046
Gated-Attention	**0.755**±0.016	**0.728**±0.016	**0.731**±0.042

Table 3. Results on COLON CANCER. Experiments were run 5 times and an average (±

METHOD	ACCURACY	PRECISION	RECALL	F-SCORE	AUC
Instance+max	0.842 ± 0.021	0.866 ± 0.017	0.816 ± 0.031	0.839 ± 0.023	0.914 ± 0.010
Instance+mean	0.772 ± 0.012	0.821 ± 0.011	0.710 ± 0.031	0.759 ± 0.017	0.866 ± 0.008
Embedding+max	0.824 ± 0.015	0.884 ± 0.014	0.753 ± 0.020	0.813 ± 0.017	0.918 ± 0.010
Embedding+mean	0.860 ± 0.014	0.911 ± 0.011	0.804 ± 0.027	0.853 ± 0.016	0.940 ± 0.010
Attention	**0.904** ± 0.011	**0.953** ± 0.014	**0.855** ± 0.017	**0.901** ± 0.011	**0.968** ± 0.009
Gated-Attention	**0.898** ± 0.020	**0.944** ± 0.016	**0.851** ± 0.035	**0.893** ± 0.022	**0.968** ± 0.010

Results, but in pictures

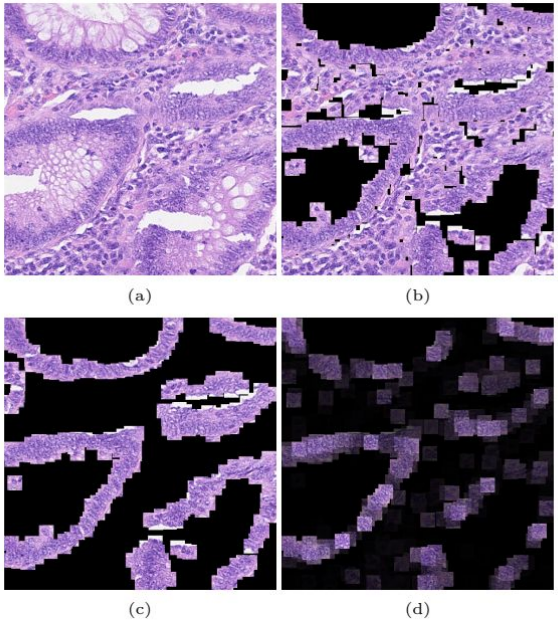


Figure 5. (a) H&E stained histology image. (b) 27×27 patches centered around all marked nuclei. (c) Ground truth: Patches that belong to the class epithelial. (d) Heatmap: Every patch from (b) multiplied by its corresponding attention weight, we rescaled the attention weights using $a'_k = (a_k - \min(\mathbf{a}))/(\max(\mathbf{a}) - \min(\mathbf{a}))$.

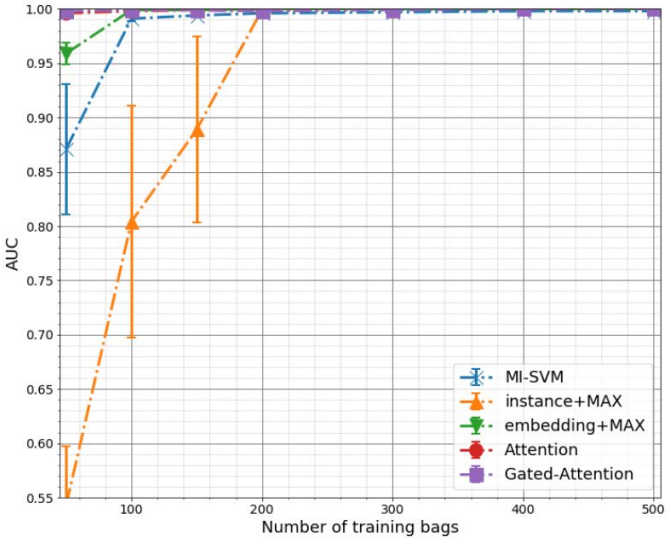


Figure 3. The test AUC for MNIST-BAGS with on average 100 instances per bag.

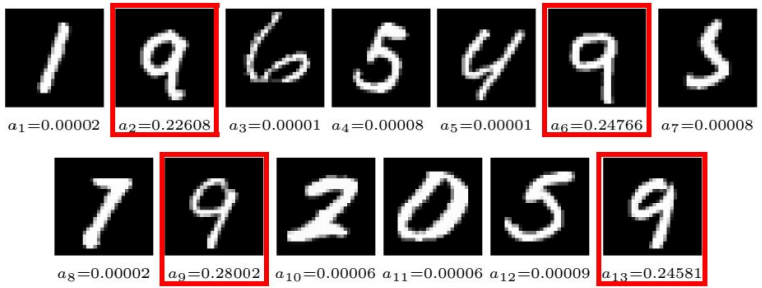


Figure 4. Example of attention weights for a positive bag.

Conclusions

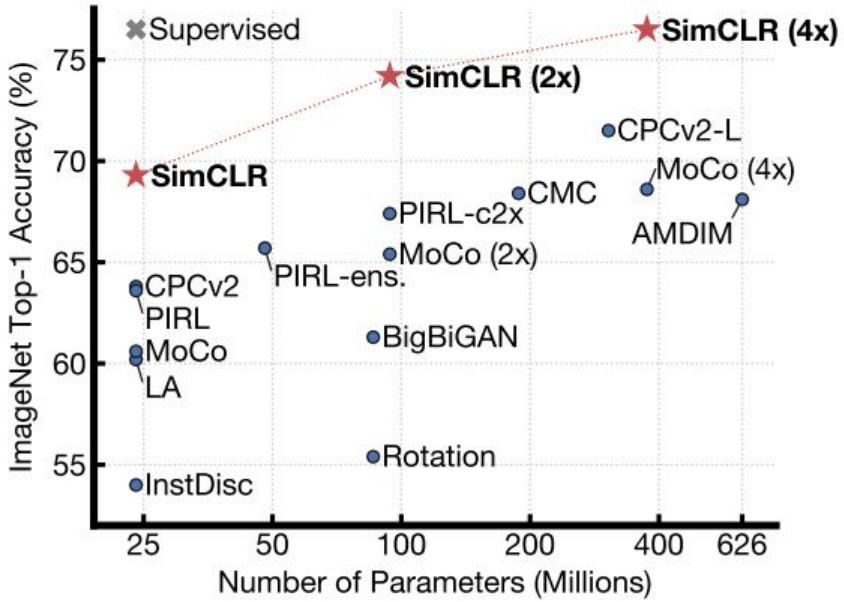
- The paper proposes a flexible and interpretable approach for Multiple Instance Learning (MIL) that uses neural networks.
- It uses deep learning to model bag score function and trainable MIL pooling with attention mechanism.
- It outperforms or performs on par with best-performing methods on various datasets at the same time providing the decision interpretation.

A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen [1] **Simon Kornblith** [1] **Mohammad Norouzi** [1] **Geoffrey Hinton** [1]

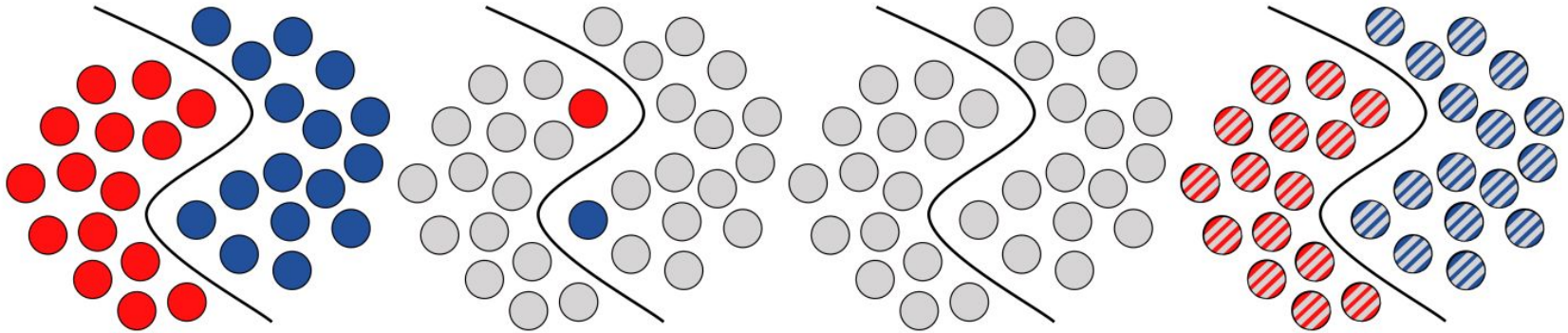
Abstract

This paper presents *SimCLR*: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the repre-



Motivation

- DNNs require large amounts of labeled data to perform well.
- Labeling data is expensive and time-consuming.
- Contrastive learning is a promising self-supervised learning method.
- Let’s learn visual representations with improved contrastive approach.

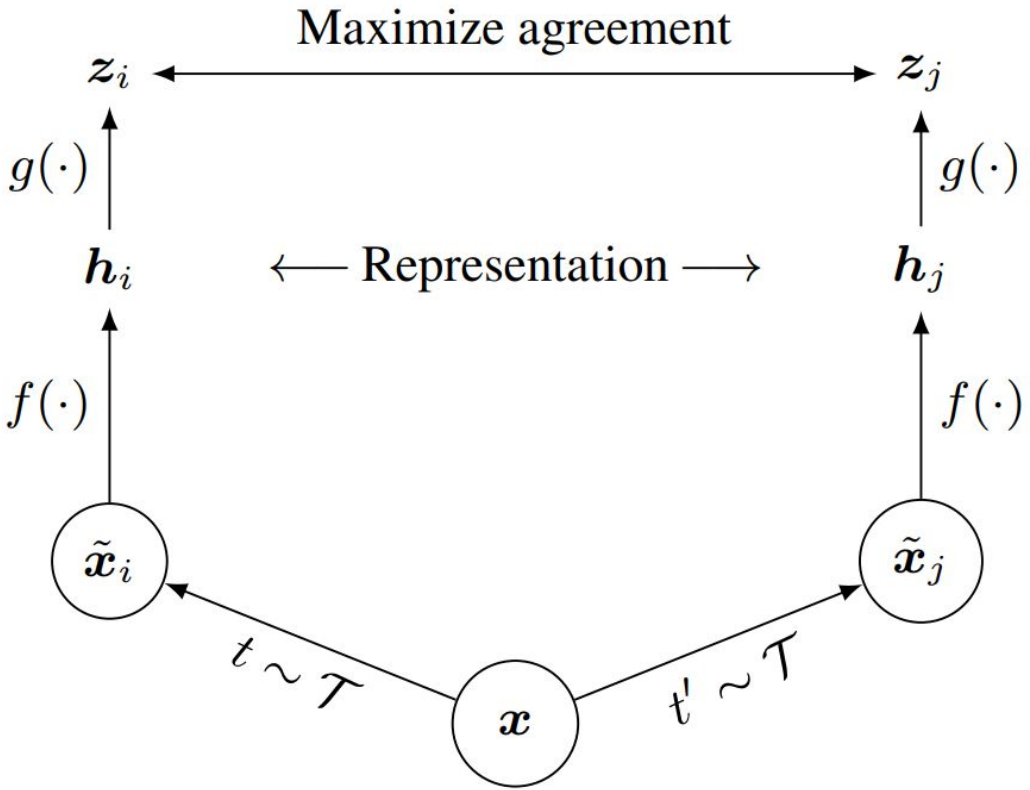


https://arxiv.org/pdf/2002.08721.pdf

Method

Key elements:

- Strong data augmentation
- Non-linear layer g
- Normalized embeddings and adjusted temperature
- Larger batches and longer training



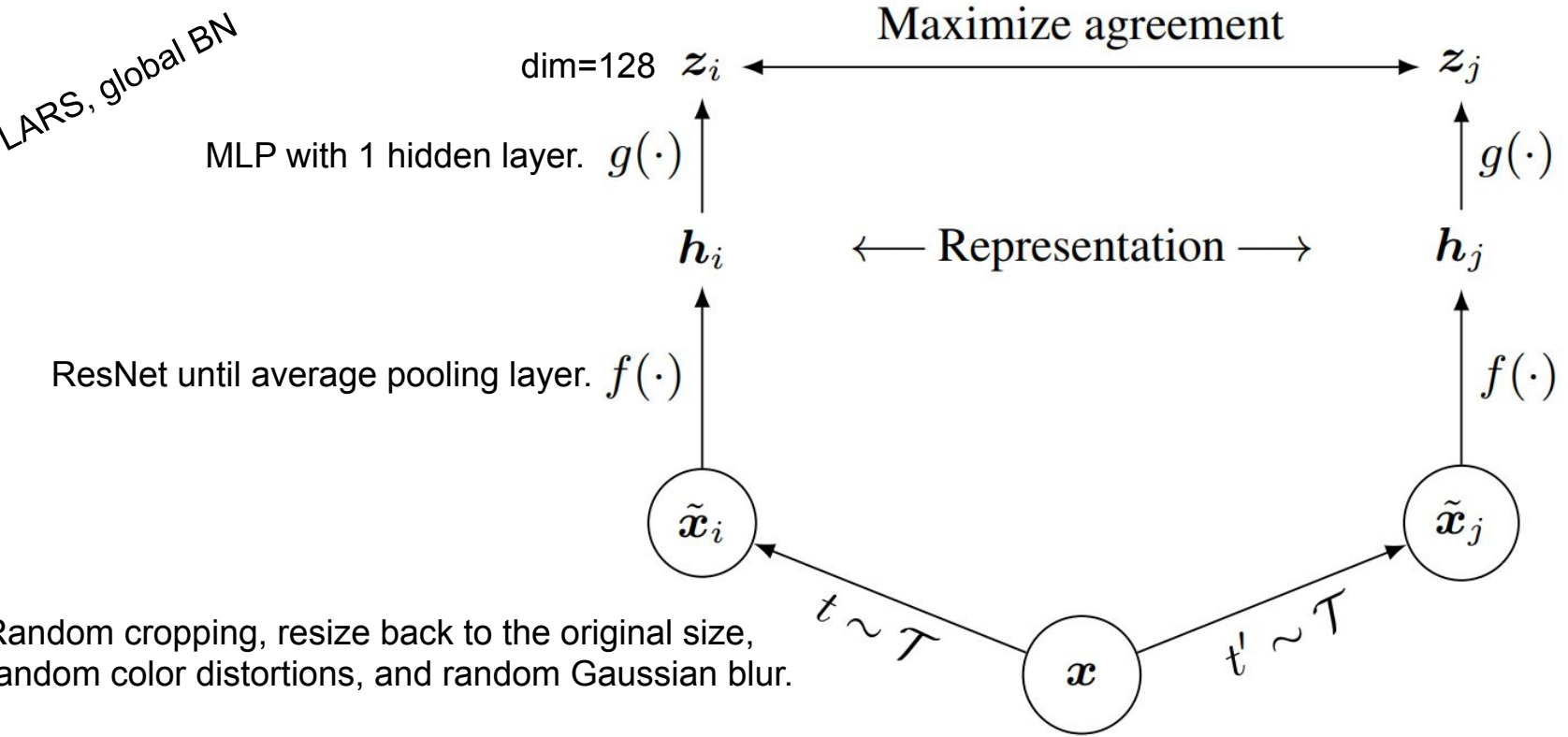
Method details

$$\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$$

NT-Xent

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

4096bs, LARS, global BN



Random cropping, resize back to the original size,
random color distortions, and random Gaussian blur.

Databases

Pre-training:

- ImageNet ILSVRC-2012
- CIFAR10

Evaluation:

- Food
- CIFAR10
- CIFAR100
- Birdsnap
- SUN397
- Cars
- Aircraft
- VOC2007
- DTD
- Pets
- Caltech-101
- Flowers
- ImageNet ILSVRC-2012

Results

	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

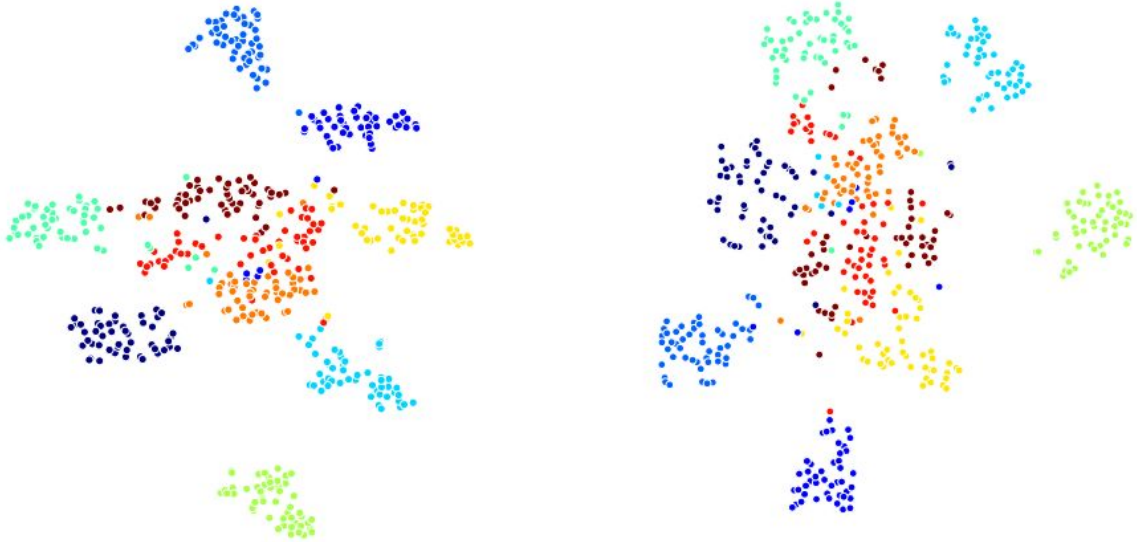
Method	Architecture	Param (M)	Top 1	Top 5
Methods using ResNet-50:				
Local Agg.	ResNet-50	24	60.2	-
MoCo	ResNet-50	24	60.6	-
PIRL	ResNet-50	24	63.6	-
CPC v2	ResNet-50	24	63.8	85.3
SimCLR (ours)	ResNet-50	24	**69.3**	**89.0**
Methods using other architectures:				
Rotation	RevNet-50 (4×)	86	55.4	-
BigBiGAN	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.

Method	Architecture	Label fraction (Top 5) 1%	10%
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	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.

Results, but in pictures



(a) $\boldsymbol{h}$ (b) $\boldsymbol{z} = g(\boldsymbol{h})$

Figure B.4. t-SNE visualizations of hidden vectors of images from a randomly selected 10 classes in the validation set.

Conclusions

- The work presents a framework for contrastive visual representation learning, which improves self-supervised, semi-supervised, and transfer learning methods considerably.
- The framework uses data augmentation, a nonlinear head, and a different loss function compared to standard supervised learning on ImageNet.
- The success of this simple framework highlights the potential of self-supervised learning, which may be undervalued despite recent interest in the field.