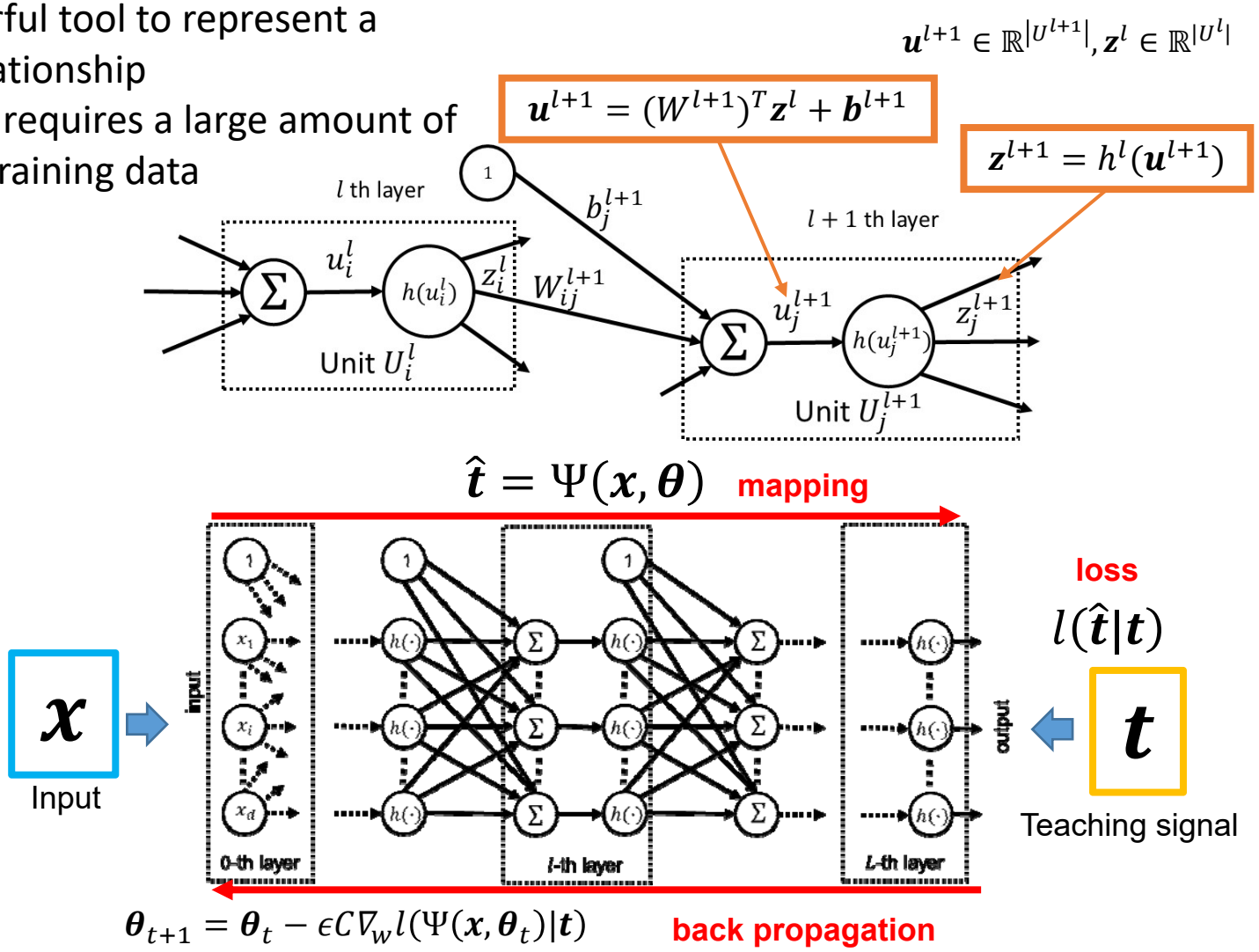
Summarization and Retrieval of Large-Scale Multimedia Data

Tatsuya Harada (The Univ. of Tokyo / RIKEN AIP / NII)
Masashi Sugiyama (RIKEN AIP / The Univ. of Tokyo),
Kazunori Ohno (Tohoku Univ.)
Koji Tsukada (Future Univ. Hakodate),
Masamichi Shimosaka (Tokyo Institute of Technology)

Deep Neural Networks

Pros: a powerful tool to represent a nonlinear relationship

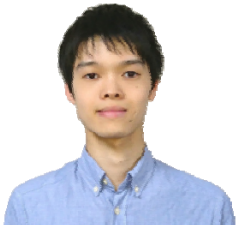
Cons: usually requires a large amount of high-quality training data



Learning from limited training data is one of the intrinsic problems in Deep NN!

Contents

- Learning from limited data
- Novel learning method for deep neural networks in supervised setting
  - Between-class learning (BC learning)



Y. Tokozume

Between-class Learning

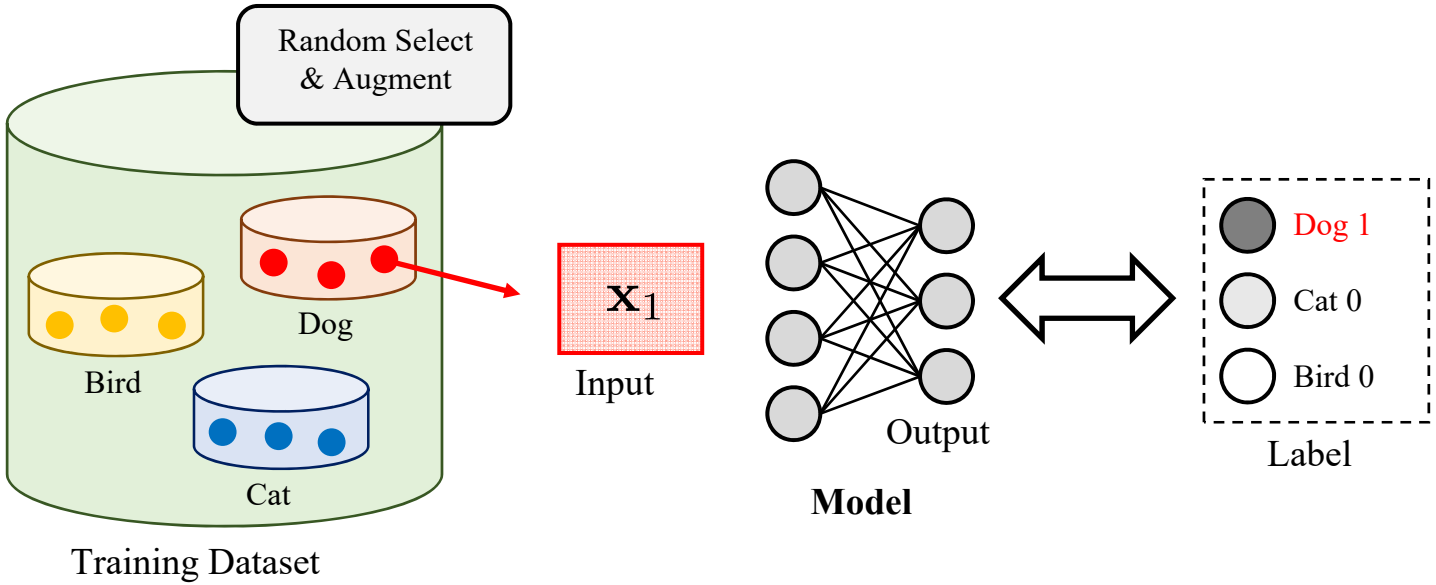
Learning from Between-class Examples for Deep Sound Recognition
Yuji Tokozume, Yoshitaka Ushiku, Tatsuya Harada
ICLR 2018

Between-class Learning for Image Classification
Yuji Tokozume, Yoshitaka Ushiku, Tatsuya Harada
CVPR 2018

https://github.com/mil-tokyo/bc_learning_sound
https://github.com/mil-tokyo/bc_learning_image

Standard Supervised Learning

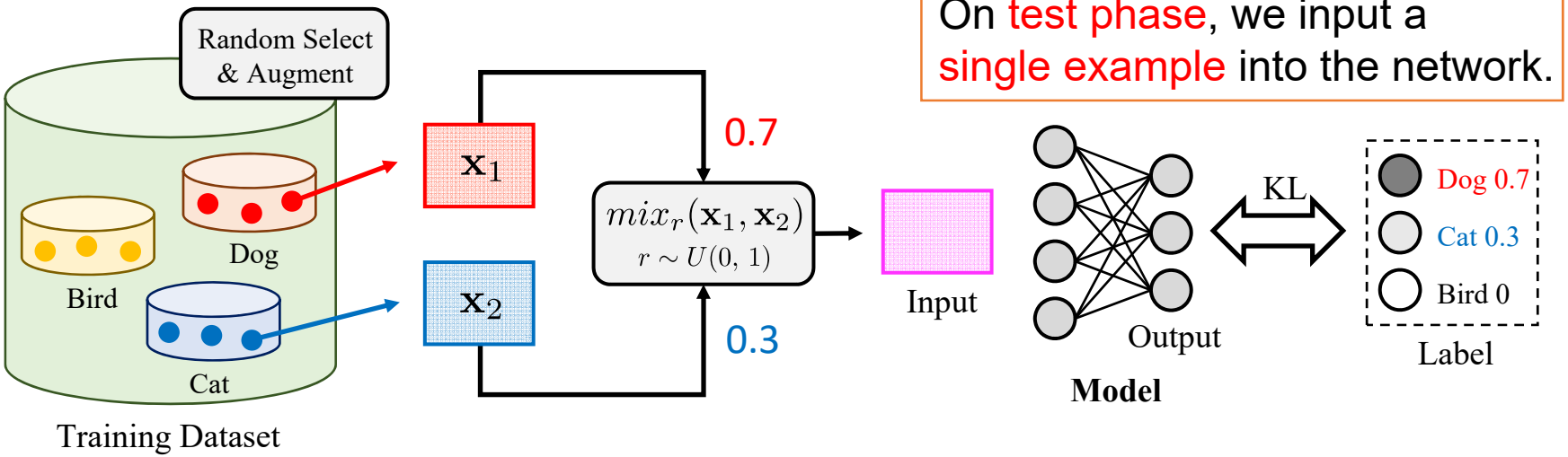
1. Select one example from training dataset
2. Train the model to output 1 for the corresponding class and 0 for the other classes



Between-class (BC) Learning

Proposed method

1. Select two training examples from different classes
2. Mix those examples with a random ratio
3. Train the model to output the mixing ratio and mixing classes

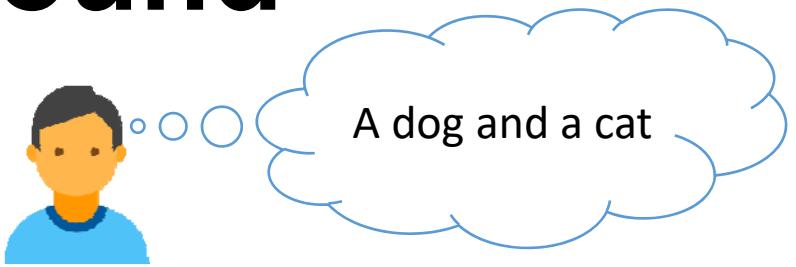


On test phase, we input a single example into the network.

Merits

- Generate infinite training data from limited data
- Learn more discriminative feature space than standard learning

BC Learning for Sound



- Two training examples $(\mathbf{x}_1, \mathbf{t}_1), (\mathbf{x}_2, \mathbf{t}_2)$
- Random ratio $r \sim U(0, 1)$

labels

- Dog: 1
- Cat: 0
- Bird: 0

$\mathbf{t}_1$

- Dog: 0
- Cat: 1
- Bird: 0

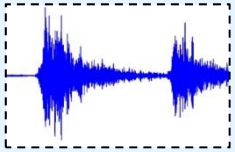
$\mathbf{t}_2$

⇨

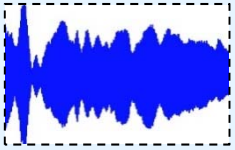
- Dog: r
- Cat: $1 - r$
- Bird: 0

$r\,\mathbf{t}_1 + (1 - r)\,\mathbf{t}_2$

sounds



$\mathbf{x}_1$



$\mathbf{x}_2$

⇨ $mix_r(\mathbf{x}_1, \mathbf{x}_2) = \dfrac{p\,\mathbf{x}_1 + (1 - p)\,\mathbf{x}_2}{\sqrt{p^2 + (1 - p)^2}}$ where $p = \dfrac{1}{1 + 10^{\frac{G_1 - G_2}{20}} \cdot \frac{1 - r}{r}}$

G_1, G_2: sound pressure level of $\boldsymbol{x}_1, \boldsymbol{x}_2$[dB]

Results of Sound Recognition

② Various datasets

① Various models

Model	Learning	Error rate (%) on ESC-50	ESC-10	UrbanSound8K
EnvNet (Tokozume & Harada, 2017)	Standard	29.2 ± 0.1	12.8 ± 0.4	33.7
	BC (ours)	24.1 ± 0.2	11.3 ± 0.6	28.9
SoundNet5 (Aytar et al., 2016)	Standard	33.8 ± 0.2	16.4 ± 0.8	33.3
	BC (ours)	27.4 ± 0.3	13.9 ± 0.4	30.2
M18 (Dai et al., 2017)	Standard	31.5 ± 0.5	18.2 ± 0.5	28.8
	BC (ours)	26.7 ± 0.1	14.2 ± 0.9	26.5
Logmel-CNN (Piczak, 2015a) + BN	Standard	27.6 ± 0.2	13.2 ± 0.4	25.3
	BC (ours)	23.1 ± 0.3	$\mathbf{9.4 \pm 0.4}$	23.5
EnvNet-v2 (ours)	Standard	25.6 ± 0.3	14.2 ± 0.8	30.9
	BC (ours)	$\mathbf{18.2 \pm 0.2}$	10.6 ± 0.6	**23.4**
EnvNet-v2 (ours) + strong augment	Standard	21.2 ± 0.3	10.9 ± 0.6	24.9
	BC (ours)	$\mathbf{15.1 \pm 0.2}$	$\mathbf{8.6 \pm 0.1}$	**21.7**
SoundNet8 + Linear SVM (Aytar et al., 2016)		25.8	7.8	-
Human (Piczak, 2015b)		18.7	4.3	-

③ Compatible with strong data augmentation

④ Surpass the human level

We can improve recognition performance for any sound networks, if we apply the BC learning.

BC Learning for Image

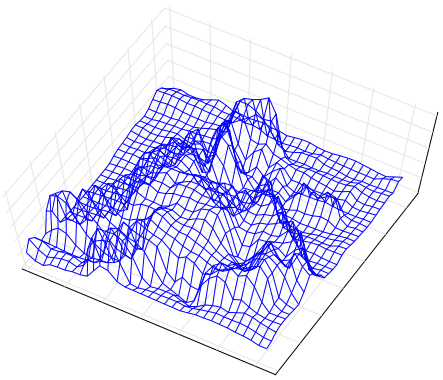
Images as waveforms

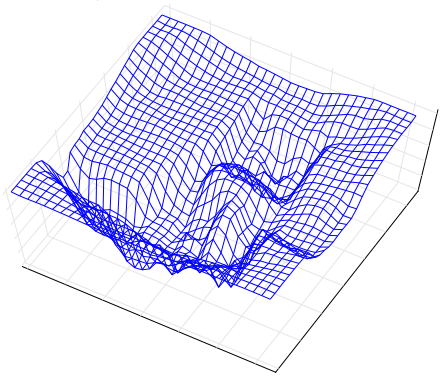
would not be important or even have a bad effect if CNNs treat input data as waveforms

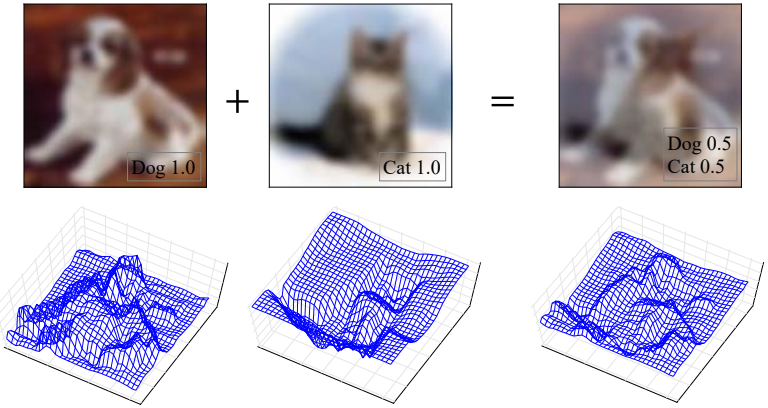
Proposal 1 (vanilla BC)

$$r\,\mathbf{x}_1 + (1-r)\,\mathbf{x}_2 \quad = \quad \{r\,\mu_1 + (1-r)\,\mu_2\} + \{r\,\mathbf{d}_1 + (1-r)\,\mathbf{d}_2\}$$

static component — wave component







$$\mathbf{x}_1 = \mu_1 + \mathbf{d}_1 \qquad \mathbf{x}_2 = \mu_2 + \mathbf{d}_2$$

Proposal 2 (BC+)

$$mix_r(\mathbf{x}_1, \mathbf{x}_2) = \frac{p\,(\mathbf{x}_1 - \mu_1) + (1-p)\,(\mathbf{x}_2 - \mu_2)}{\sqrt{p^2 + (1-p)^2}}, \text{ where } p = \frac{1}{1 + \frac{\sigma_1}{\sigma_2} \cdot \frac{1-r}{r}}$$

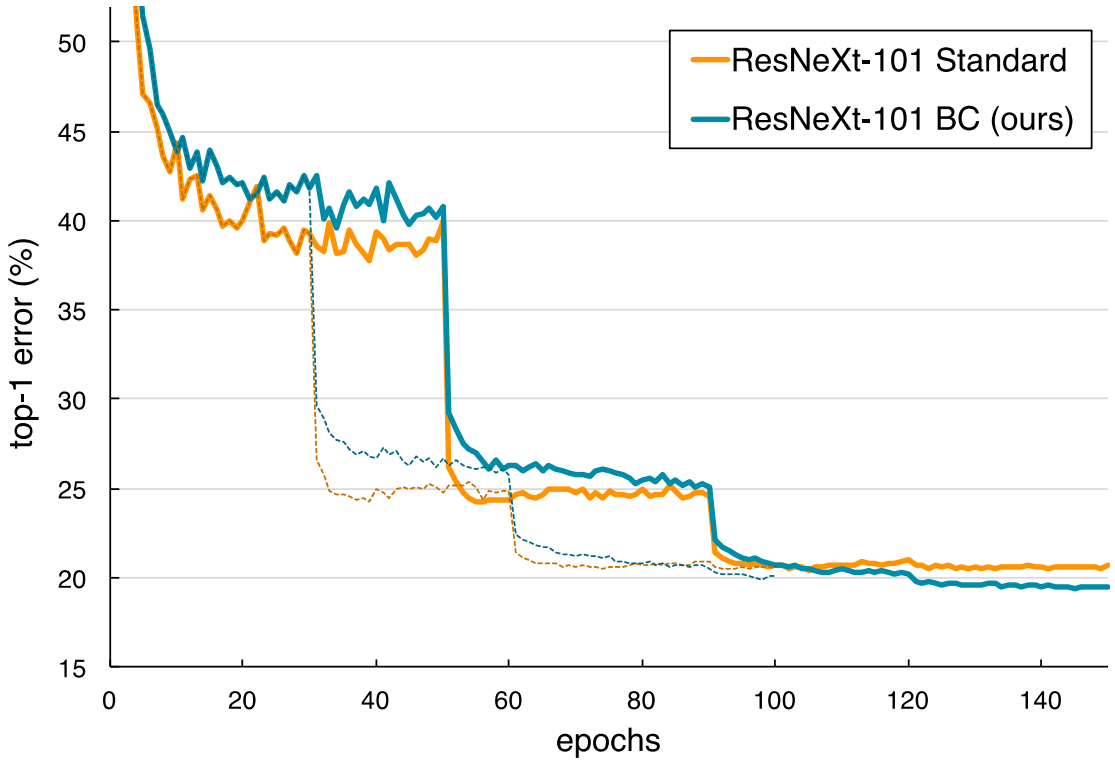
Results on CIFAR

Our preliminary results were presented in ILSVRC2017 on July 26, 2017.

Model	Learning	Error rate (%) on CIFAR-10	Error rate (%) on CIFAR-100
11-layer CNN	Standard	6.07 ± 0.04	26.68 ± 0.09
	BC (ours)	5.40 ± 0.07	24.28 ± 0.11
	BC+ (ours)	$\mathbf{5.22 \pm 0.04}$	$\mathbf{23.68 \pm 0.10}$
ResNet-29† [28]	Standard	4.24 ± 0.06 / 4.39 [28]	20.18 ± 0.07
	BC (ours)	3.75 ± 0.04	19.56 ± 0.10
	BC+ (ours)	$\mathbf{3.55 \pm 0.03}$	$\mathbf{19.41 \pm 0.07}$
ResNeXt-29 (16×64d)† [28]	Standard	3.54 ± 0.04 / 3.58 [28]	$\mathbf{16.99 \pm 0.06}$ / 17.31 [28]
	BC (ours)	$\mathbf{2.79 \pm 0.06}$	18.21 ± 0.12
	BC+ (ours)	2.81 ± 0.06	17.93 ± 0.09
DenseNet-BC ($k = 40$)† [13]	Standard	3.61 ± 0.10 / 3.46 [13]	17.28 ± 0.12 / 17.18 [13]
	BC (ours)	2.68 ± 0.03	16.36 ± 0.10
	BC+ (ours)	$\mathbf{2.57 \pm 0.06}$	$\mathbf{16.23 \pm 0.07}$
Shake-Shake Regularization [9]	Standard	2.86 [9]	**15.85** [9]
	BC (ours)	2.38 ± 0.04	15.90 ± 0.06
	BC+ (ours)	$\mathbf{2.26 \pm 0.01}$	16.00 ± 0.10

Results on ImageNet-1K

Our preliminary results were presented in ILSVRC2017 on July 26, 2017.

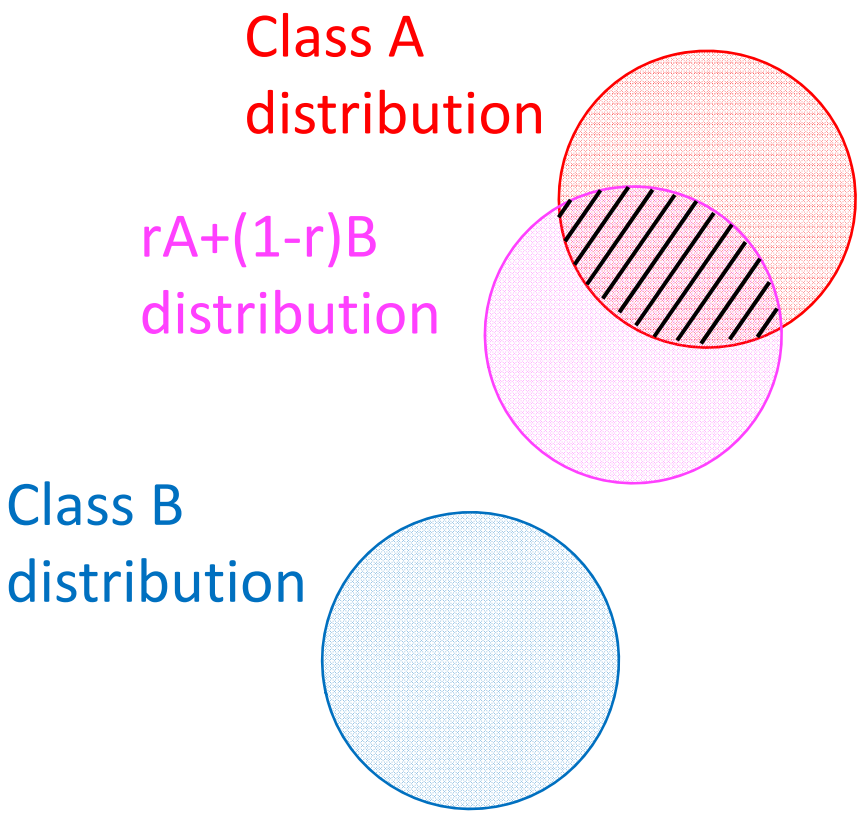


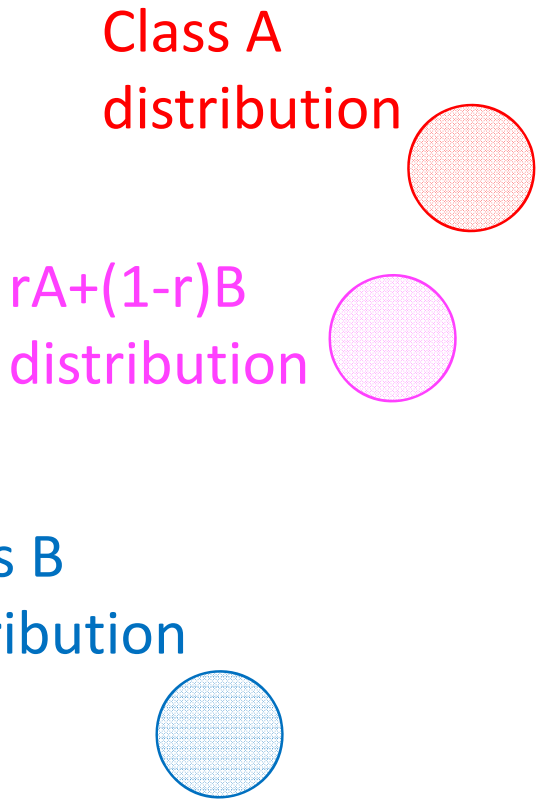
		top-1/top-5 val. error
100 epochs	Standard	20.4/5.3 [28]
	BC (ours)	19.92/4.91
150 epochs	Standard	20.44/5.25
	BC (ours)	**19.43/4.80**

around 1% gain in top-1 error

How BC Learning Works

Less discriminative ⟷ More discriminative



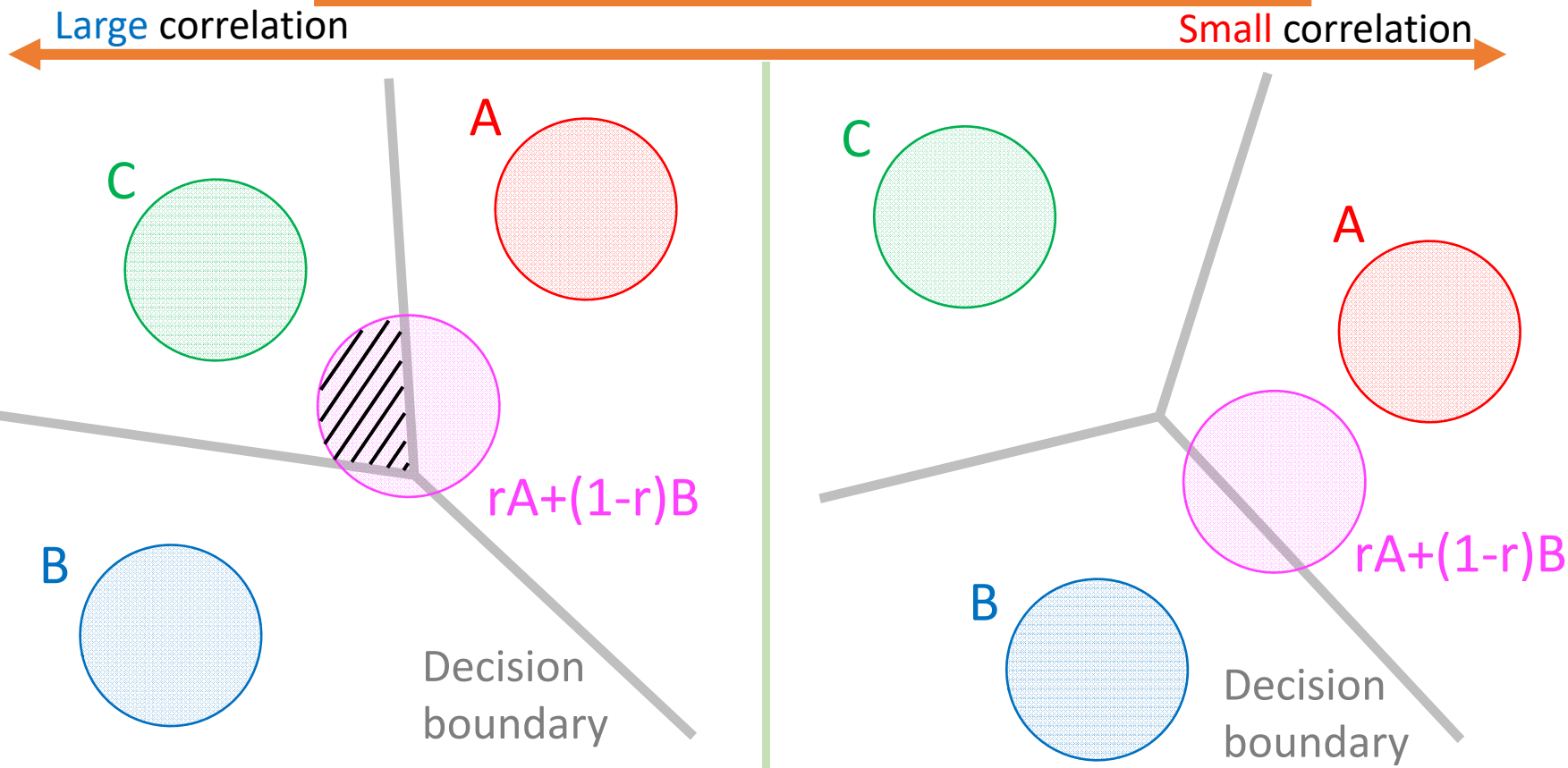


Small Fisher's criterion
→ Overlap among distributions
→ Large BC learning loss

Large Fisher's criterion
→ No overlap among distributions
→ Small BC learning loss

How BC Learning Works

In the classification, the distributions must be uncorrelated because the teaching signal is discrete,.



Large correlation among classes

→ Mixing class of A and B may be classified into class C.

→ Large BC learning loss

Small correlation among classes

→ Mixing class of A and B is not classified into class C.

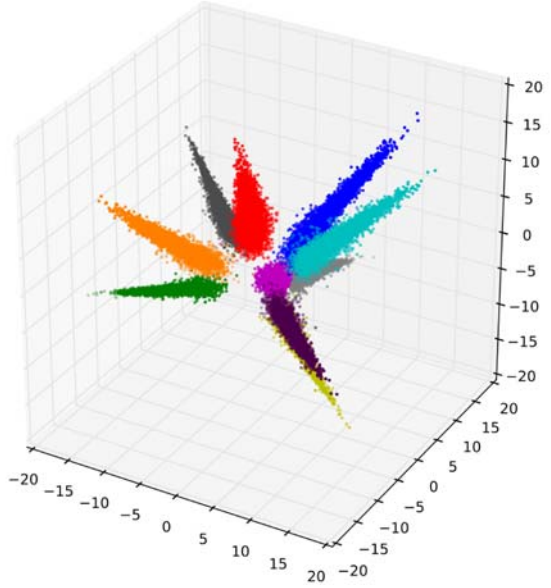
→ Small BC learning loss

Visualization using PCA

Activations of

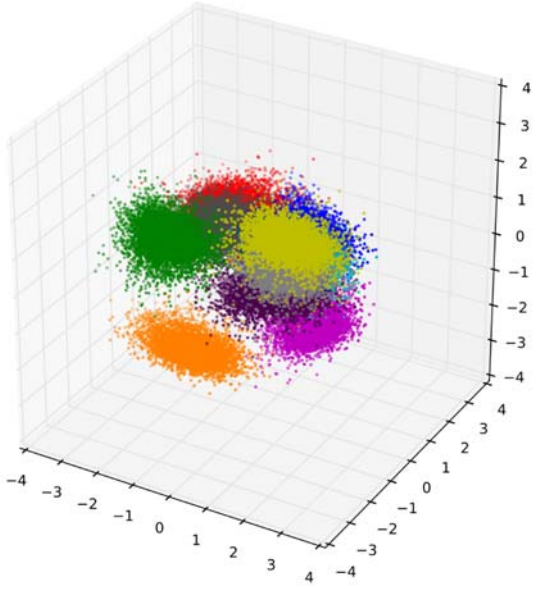
- 10-th layer of 11-layers CNN
- trained on CIFAR-10

Fisher's criterion: 1.76



Standard learning

Fisher's criterion: 1.97



BC learning (ours)

- Distributions are more compact than those from standard learning.
- Distributions are spherical.
- Larger Fisher's criterion than that of standard learning

Take Home Messages

- Learning from limited training data is one of the intrinsic problems in deep neural networks.

- Between-class learning (BC learning)
  - A novel learning method for deep neural networks in supervised setting
  - Mix two training examples with a random ratio
  - Train the model to output the mixing ratio
  - Simple and easy to implement
  - Can be introduced independently from previous techniques: network architectures, data augmentation schemes, optimizers, etc.