人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-055-6
(11/22)

Between-class Learning for Sound and Image Classification

床爪 佑司 [1]　牛久 祥孝 [1]　原田 達也 [1,2]
Yuji Tokozume[1]　Yoshitaka Ushiku[1]　Tatsuya Harada[1,2]

[1] 東京大学
[1] The University of Tokyo
[2] 理化学研究所
[2] RIKEN

Abstract: We introduce our novel learning method for sound and image classification called between-class learning (*BC learning*). We generate between-class images by mixing two images belonging to different classes with a random ratio. We then input the mixed image to the model and train the model to output the mixing ratio. BC learning has the ability to impose constraints on the shape of the feature distributions, and thus the generalization ability is improved. As a result, classification performance of sounds and images was improved.

1 はじめに [1]

本発表では，実環境理解に関する研究として，ICLR 2018 および CVPR 2018 で提案した深層ニューラルネットワークの新しい教師付学習手法 between-class learning (BC learning) [Tokozume 18a, Tokozume 18b] について紹介する．

音や画像の認識において，深層学習を用いた手法が高い性能を発揮している．深層学習は，線形分離不可能なデータ空間から線形分離可能な特徴空間への関数を学習する．限られた学習データから出来る限り判別的な特徴空間を学習することが，深層学習における重要な課題である．

そこで本研究では，限られた学習データから判別的な特徴空間を学習できる，深層ニューラルネットワークの新しい教師付学習手法を提案する．新しい教師付学習手法には，ネットワーク構造や正則化等の従来の学習技術に影響を与えないこと，限られた学習データを効率的に使えること，判別的な特徴空間を学習できること，の 3 つが求められる．

ここで，判別的な特徴空間とはどのようなものだろうか．まず，クラス間の Fisher's criterion [Fisher 36] が大きい特徴空間は判別的である．Fisher's criterion とは，クラス内分散に対するクラス間距離の比のことであり，2 つのクラスがどの程度判別的であるかを表す指標である．また，各クラスが無相関な特徴空間は判別的である．識別タスクでは各クラスを等価に扱う必要があるため，特徴空間において各クラスが等間隔に並んでいることが望ましい．本研究ではこれら 2 つを判別的な特徴空間の要件とする．

従来の教師付学習では，学習データセットから単一の学習データを選択し，対応するクラスは 1，それ以外は 0 を出力するようにニューラルネットワークを学習していた．このような学習手法では，特徴空間において各クラスが線形分離可能であれば罰則が与えられないので，特徴空間が判別的になる保証は無い．本研究ではこの問題を解決する学習手法を提案する．

[1] 本稿の内容は JSAI2019 における講演予稿 (3E4-OS-12b) の転載 (一部改変) である．

2 Between-class Learning

2.1 概要と効果

本研究では，深層ニューラルネットワークの新しい教師付学習手法として，between-class learning (BC learning) を提案する．BC learning では，以下の手順でモデルを学習する．

- 異なるクラスに属する 2 つのデータを選択する．
- それらをランダムな比率で合成し，モデルに入力する．
- 合成比率を出力するようにモデルを学習する．

BC learning は，従来の学習技術に影響を与えない．また，データの合成によって学習データのパターン数が増えるため，限られた学習データを効率的に使うことができる．さらに，判別的な特徴空間を学習できる効果がある．その理由を以下に示す．

効果 1. Fisher's criterion の増大　図 1 (左) のように，特徴空間においてクラス A, B 間の Fisher's criterion が小さい場合を考える．クラス A, B に属するデータをある比率で合成してモデルに入力した際に，

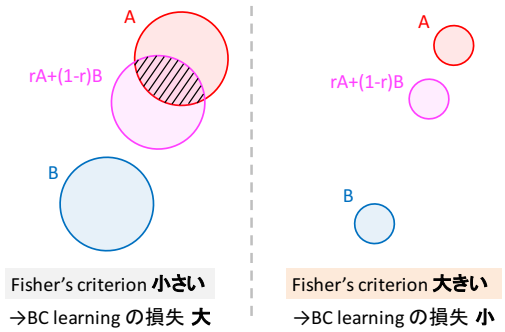

図 1: BC learning による Fisher's criterion の増大.

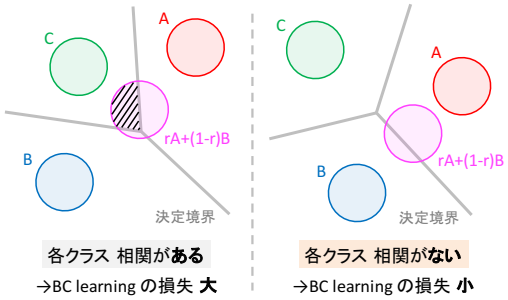

図 2: BC learning による各クラスの無相関化.

その特徴量分布 (桃色) はクラス A, B のいずれかの特徴量分布と重複することが予想される．このとき，合成するデータの組み合わせによっては，合成したデータがいずれかのクラスに分類されてしまい，モデルが合成比率を出力することができない．そのため，BC learning を行った場合の損失が大きい．一方，図 1 (右) のように Fisher's criterion が大きい場合，重複が発生しないため，BC learning による損失が小さい．学習は損失が小さくなる方向に進むため，BC learning によって図 1 (右) のような Fisher's criterion が大きい特徴空間が学習される．

効果 2. 各クラスの無相関化　図 2 (左) のように特徴空間において各クラスに相関がある場合，クラス A, B の合成物がクラス C に分類されるケースが発生するため，BC learning の損失が大きい．一方，図 2 (右) のように各クラスに相関がない場合，クラス A, B の合成物がクラス C に分類されないため，BC learning の損失が小さい．よって，BC learning によって図 2 (右) のような各クラスが無相関な特徴空間が学習される．

2.2　環境音識別への適用

音はデータ同士を合成しても音として成り立つため，BC learning が有効であると考えられる．選択された 2 つの学習データをそれぞれ $\mathbf{x}_1, \mathbf{x}_2$ とし，それらの one-hot ラベルをそれぞれ $\mathbf{t}_1, \mathbf{t}_2$ とする．また，合成比率 r を一様分布 $U(0, 1)$ から生成する．ラベルの合成は単純に $r\,\mathbf{t}_1 + (1-r)\,\mathbf{t}_2$ とする．一方，データの合成は，

表 1: 環境音データセットにおける実験結果.

モデル	学習手法	誤識別率 (%) ESC-50	ESC-10	US8K
EnvNet-v2	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**

表 2: 一般物体画像データセットにおける実験結果.

モデル	学習手法	誤識別率 (%) CIFAR-10	CIFAR-100
11 層 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	Standard	4.24 ± 0.06	20.18 ± 0.07
	BC (ours)	3.75 ± 0.04	19.56 ± 0.10
	BC+ (ours)	$\mathbf{3.55 \pm 0.03}$	$\mathbf{19.41 \pm 0.07}$
Shake-Shake	Standard	2.86	**15.85**
	BC (ours)	2.38 ± 0.04	15.90 ± 0.06
	BC+ (ours)	$\mathbf{2.26 \pm 0.01}$	16.00 ± 0.10

同様に $r\,\mathbf{x}_1 + (1-r)\,\mathbf{x}_2$ とするのが単純であるが，$\mathbf{x}_1$, $\mathbf{x}_2$ それぞれの音圧レベル G_1, G_2 (dBA) の差を考慮した以下の合成式を提案する．

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

2.3　画像識別への適用

画像を合成することは直感に反するが，画像データは x 軸と y 軸に沿った波であると考えられるので，環境音と同様に BC learning が有効であると考えられる．先程と同様に $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{t}_1$, $\mathbf{t}_2$, r を定義する．ラベルの合成は単純に $r\,\mathbf{t}_1 + (1-r)\,\mathbf{t}_2$ とする．データの合成は，同様に $r\,\mathbf{x}_1 + (1-r)\,\mathbf{x}_2$ とするのが単純であるが，$\mathbf{x}_1$, $\mathbf{x}_2$ からそれぞれの平均値 μ_1, μ_2 を引いてゼロ平均にしたのちに，環境音と同様に合成することを提案する．音圧レベルの代わりに各画像の標準偏差 σ_1, σ_2 を用いた以下の合成式を提案する．前者の単純な合成方法を BC，後者を BC+と呼ぶことにする．

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

3　実験

環境音データセット ESC-50, ESC-10, UrbanSound8K, および一般物体画像データセット CIFAR-10, CIFAR-100 を用いて様々なモデルの学習・評価を行った．その結果の一部を表 1 および表 2 に示す．多くの条件において BC learning および BC+によって識別性能が向上した．特に CIFAR-10 において 2018 年 1 月現在の世界最高性能 2.26%を達成した．

次に大規模画像データセット ImageNet-1K を用いて実験を行った．その結果を図 3 に示す．BC learning に

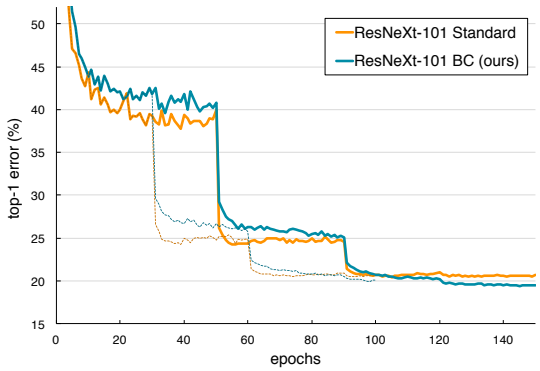

図 3: ImageNet-1K における実験結果．破線は 100 epoch，実線は 150 epoch での実験結果．

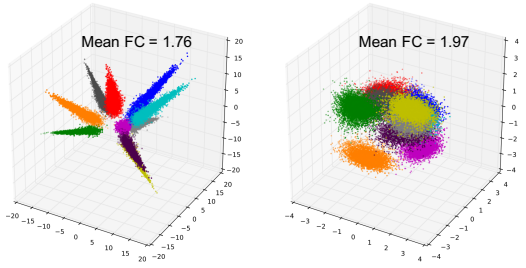

図 4: BC learning によって学習された特徴空間の可視化．

よって最終的な誤識別率が 20.4%から 19.4%へ約 1%向上した．BC learning は大規模データセットに対しても有効であることが示された．

CIFAR-10 で学習した 11 層 CNN の特徴空間（第 10 層）を PCA を用いて可視化した結果を図 4 に示す．BC learning によって学習された特徴空間は，各クラスが球状にまとまっていることが分かる．また，2 クラス間の Fisher's criterion の平均値も，BC learning の方が大きかった．BC learning によって判別的な特徴空間が学習されたといえる．

4 結論と今後の展望

本研究では，between-class (BC) learning という深層ニューラルネットワークの新しい教師付学習手法を提案した．実験の結果，BC learning によって音と画像の識別性能が大きく向上することが示された．BC learning は，音や画像以外のモダリティのデータの識別や，識別以外のタスクにも応用が期待される，非常に汎用性の高い技術である．また，考え方がシンプルで実装も容易であり，実用性も高い．さらに，理論的考察の余地もあり，今後さらなる研究がなされると考えられる．

参考文献

[Tokozume 18a] Y. Tokozume, Y. Ushiku, and T. Harada. Learning from between-class examples for deep sound recognition. In *ICLR*, 2018.

[Tokozume 18b] Y. Tokozume, Y. Ushiku, and T. Harada. Between-class Learning for Image Classification. In *CVPR*, 2018.

[Fisher 36] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, Vol. 7, No. 2, pp. 179–188, 1936.