社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

PyramidNet に対する新たな確率的正則化手法 ShakeDrop の提案

山田　良博†　岩村　雅一†　黄瀬　浩一†

† 大阪府立大学大学院工学研究科
〒 599–8531 大阪府堺市中区学園町 1–1
E-mail: yamada@m.cs.osakafu-u.ac.jp, {masa,kise}@cs.osakafu-u.ac.jp

あらまし　高精度な一般物体認識を実現するために，確率的に学習を乱すことで認識精度を向上する，確率的な正則化手法が注目を集めている．確率的な正則化を用いた一般物体認識手法の 1 つである PyramidDrop は，Randam Drop と呼ばれる確率的な正則化を導入することで発表当時の世界最高性能を実現した．しかし，Randam Drop は二値の乱数を用いて各 Residual Unit を使うか使わないかを制御するため，正則化の効果が限定的と考えられる．本稿では，最近提案された実数の乱数を用いる確率的な正則化を PyramidDrop に導入することで，より高い認識精度の手法の実現を目指す．この乱数は，各 Residual Unit をどの程度使うかを表す．一般物体認識データセット CIFAR-10 及び CIFAR-100 を用いた実験の結果，提案手法 ShakeDrop は，現在の state-of-the-art の手法に比べて、エラー率を CIFAR-10 で 0.25%、CIFAR-100 で 3.01%軽減して，現時点の世界最高精度を達成することを確認した．

キーワード　一般物体認識，深層学習，Deep Residual Network，Residual Learning，正則化

1. はじめに

2012 年に開催された一般物体認識コンペティションにおいて従来手法を凌駕する成績を収めて以来，一般物体認識において Convolutional Neural Network (CNN) が本質的な特徴を見出し優れた性能を発揮すると注目を浴びている [1]．CNN は，画像の畳み込みを実現する複数の畳み込み層から成り，畳み込み層が増える程，抽象化した特徴を取り出すことができる．そのため，層数が増えれば，より抽象的で複雑なカテゴリが認識できるようになると考えられている．しかし，これは諸刃の剣であり，多くの畳み込み層を持つ深層ネットワークでは特徴を抽象化し過ぎてしまい，単純な特徴で構成されるカテゴリを上手く表現できず，認識精度が頭打ちになることが報告されている [2]．これは，ネットワークの前段の少数の畳み込みで得られた，比較的単純な特徴が後段の畳み込みによって潰れてしまうからと考えられている．

この従来の CNN の問題を解決したのが ResNet [3] である．ResNet の最大の特徴は，Residual Unit の導入である．Residual Unit は，従来のように畳み込みを行う場合と，この層への入力をそのまま出力して畳み込みを行わない場合の結果を足し合わせる処理機構によって，畳み込みが非常に多い CNN ながら，特徴を潰さず高い精度を実現している．ResNet は図 1 のような Residual Unit を多数含む構造になっている．ResNet の登場以降，Residual Unit の構造をいかに改善するかが CNN における大きな課題となっている．

PyramidDrop [9] は ResNet の改良手法の一つである．PyramidDrop は，ResNet の派生手法である PyramidNet [8] に毎回ランダムに選ばれた一部の畳み込みを学習時に無視する Random Drop と呼ばれる確率的な正則化 [10] を導入することで，ResNet の認識精度を向上させた．PyramidDrop は正則化の導入によって，PyramidNet よりも高い認識精度を示すことが実験的に確認されている．ただし PyramidDrop は二値の乱数に従うシンプルな正則化手法であるため，正則化の効果が限定的となり，認識精度が頭打ちになっている可能性がある．実際に確率的な正則化は二値の乱数にのみ従う Dropout [11] 以降，Shakeout [12]，Whiteout [13] など様々な工夫を含む多くの改良手法が提案され，大幅に認識精度を改善してきた経緯がある．同様に PyramidDrop が様々な工夫を含む正則化手法と組み合

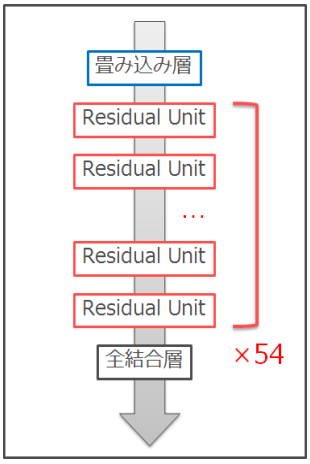


図 1: ResNet の全体図．Residual Unit は 1 つあたり 2 つの畳み込みを含むため，Residual Unit を 54 個持つこの ResNet は，畳み込みを行う処理層と全結合と呼ばれる処理を行う処理層を全て合わせて 110 層の処理層で構成されている．

表 1: 各手法とその特徴および一般物体認識用データセット CIFAR-10 と CIFAR-100 を用いた際のエラー率．表中の 6 倍学習は，ResNet の関連研究の多くは 300Epoch での学習を行うのに対して，Shake-Shake で利用された，その 6 倍となる 1800Epoch での学習を表す．矩形前処理は，Cutout [4] や RandomErasing [5] で効果的なことが示された，ランダムに決定した矩形に応じて学習画像の一部を塗りつぶす学習法を表す．ShakeDrop に 6 倍学習や矩形前処理を組み合わせた場合，エラー率が最も低い従来手法に比べて、エラー率を CIFAR-10 で 0.25%、CIFAR-100 で 3.01%軽減できた．

手法	正則化	6 倍学習	矩形前処理	層数	パラメータ数	CIFAR-10 のエラー率	CIFAR-100 のエラー率
ResNeXt [6,7]	×	○	×	26	26.2M	3.58%	-
				29	34.4M	-	16.34%
Shake-Shake [7]	○	○	×	26	26.2M	2.86%	-
				29	34.4M	-	15.85%
Cutout [4]	○	○	○	26	26.2M	2.56%	-
				29	34.4M	-	15.20%
PyramidNet [8]	×	×	×	272	26.0M	3.31%	16.35%
PyramidDrop [8,9]	○	×	×	272	26.0M	-	15.94%
ShakeDrop (Proposed)	○	×	×	272	26.0M	-	**14.90%**
	○	×	○	272	26.0M	2.89%	**13.85%**
	○	○	○	272	26.0M	**2.31%**	**12.19%**

わされば，認識精度を大幅に改善し得ると考えられる．

確率的な正則化の中でも ResNet の改良手法の一つである Shake-Shake [7] は，発表当時の世界一の認識精度を達成した確率的な正則化における最先端の手法である．Shake-Shake は Residual Unit の畳み込みの分岐を従来の一分岐から二分岐に増やし，特定の値域となる実数の乱数を用いて抽出した特徴を混ぜ合わせて学習を行う．Shka-Shake は特徴を混ぜ合わせることで，物体の色や形状などの見えの変化を吸収し，本質的な特徴の学習を目指した手法である．Shake-Shake はこの工夫によって ResNet の認識精度を飛躍的に向上させた．PyramidDrop と Shake-Shake を組み合わせれば認識精度を大幅に改善出来ると考えられる．しかし Shake-Shake は二分岐の構造を持つ手法でないと導入することができない制限がある．二分岐の構造は同様の条件の一分岐の場合の倍の計算を要することになるため，Shake-Shake を用いた場合，計算時間や学習パラメータ数の都合から多層な CNN にすることが難しい．一方で PyramidDrop で用いられる確率的な正則化 Random Drop は多層な CNN で有効なことが報告されている [10] ため，多層な CNN にすることが難しい Shake-Shake の二分岐の構造と相性が悪く，PyramidDrop と Shake-Shake を単純に組み合わせて正則化の効果を高めることは難しいと考えられる．

PyramidDrop と Shake-Shake 及び本稿において目標とする提案手法の特徴をまとめると表 2 のようになる．PyramidDrop は正則化の効果が限定的で，Shake-Shake は計算時間や学習パラメータ数の都合から多層な構造にすることが難しい．そこで我々は PyramidDrop の一部に Shake-Shake と同様の実数の乱数を用いた高度な正則化と多層構造を両立する新たな手法の実現を目指した．本稿ではこの新しい確率的な正則化手法を ShakeDrop と呼ぶことにする．我々は ShakeDrop が特定のパラメータにおいて従来の正則化を用いた認識精度を大幅に上回ることを実験的に確認した．さらに，ShakeDrop は表 1 に示す通り，最先端の手法 Cutout [4] で利用されている，従来の 6 倍

表 2: PyramidDrop，Shake-Shake，提案手法 ShakeDrop の特徴．PyramidDrop は多層構造が可能だがシンプルな正則化であり，Shake-Shake は実数の乱数を用いた高度な正則化を行うが多層な構造にすることが難しい．

手法	多層な構造	高度な正則化
PyramidDrop	○	×
Shake-Shake	×	○
ShakeDrop (Proposed)	○	○

の更新回数を用いる工夫や，一定の割合で学習に用いる画像の一部を塗りつぶす工夫と組み合わせることで，認識精度が大幅に上昇することを確認した．これらの検討の中で提案手法は一般物体認識用データセット CIFAR-10 および CIFAR-100 [14] を用いた実験で，表 1 に示すように原稿執筆時点 (2017 年 9 月 21 日) においての世界最高の認識精度を達成している．

2. 先 行 研 究

本節では，ResNet，PyramidDrop，Shake-Shake について説明する．

2.1 ResNet

前述のように CNN は畳み込み層が多いと様々な特徴を取り出すことができるが，得られた重要な特徴が多くの畳み込み層の中で消えてしまうことがある．これは畳み込みで恒等写像 $f(x) = x$ を実現することが難しいことに起因する．恒等写像を学習することができれば，得られた重要な特徴が畳み込み層の中で消えてしまいにくくなると考えられる．

そこで恒等写像を扱うため，Residual Unit が提案された．これは畳み込みによる写像と入力を足し合わせて出力とするもので，入力を x としたとき，以下のように表される．ただし $G(x)$ は入力 x に対する Residual Unit 全体の変換であり，$F(x)$ は Residual Unit の畳み込み部分のみの変換である．

$$G(x) = x + F(x) \tag{1}$$

Residual Unit では恒等写像を扱うとき，常に $F(x) = 0$ になる畳み込みを学習することになる．これは $f(x) = x$ よりも簡単に実現できるため，従来の畳み込みに比べて恒等写像の学習が容易になると考えられる．実際に Residual Unit を大量に積み重ねた ResNet は従来の CNN に比べて大きく精度が改善しており，ImageNet [15] を用いた実験では人の平均的な認識精度を超えるまでになっている．

2.2 PyramidDrop

PyramidDrop は ResDrop で提案された処理機構 RandomDrop を PyramidNet に導入した手法である．

ResDrop は確率的な正則化 Random Drop を ResNet に導入した手法である．ResNet は従来より多くの畳み込みを持つため，Residual Unit の導入をもってしてもなお，得られた特徴が学習の途中で失われてしまう効果が無視できない．また，学習時間が長いという問題もある．ResDrop [10] は式 (1) の畳み込み部分 $F(x)$ について，確率的に $F(x) = 0$ とする処理機構 RandomDrop を導入し，学習過程で学習を行わない層を毎回ランダムに決定する．RandomDrop を組み込んだ Residual Unit の式 (1) は以下のように表される．

$$G(x) = x + p_b F(x) \tag{2}$$

ただし p_b は RandomDrop を制御する二値乱数であり，N 個の Residual Unit があるとき，入力から n 番目の Residual Unit において $0.5 * n/N$ の確率で 0 の値を取る．この工夫によってそれぞれの Residual Unit が特徴を補い合い精度の高い特徴を取り出せるようになり，認識精度を高めながら学習時間を削減できる．ResDrop を導入することで ResNet に比べ精度が高くなること，特に ResNet で最も認識精度が高かった 110 層の処理層を持つ CNN から Residual Unit を大幅に増やした 1202 層の学習で精度が向上することが確認されている．

PyramidNet [8] は ResDrop と同様に ResNet の認識精度を向上させる手法である．CNN は特徴を抽出する過程で，画像が持つ高さや幅とは異なるもう一つの次元，channel に対して処理を行う．ResNet ではいくつかの Residual Unit でこの channel に関する出力が大きく増加する．ResNet はどれか 1 つの Residual Unit の畳み込みで出力される特徴を使わない場合も，全ての Residual Unit を使った場合と認識精度がほぼ変わらないことが示されているが，この channel が急激に増加する Residual Unit の畳み込みで出力される特徴を使わない場合は，認識精度が大幅に低下することが示されている [16]．ResDrop は channel が急激に増加する Residual Unit の畳み込みで出力される特徴を使わない場合も認識精度が大きく変わらず，ResDrop は認識精度を大きく向上させているため，channel が急激に増加する Residual Unit が認識精度の向上を妨げていると考えられる [16]．したがって channel の次元数は大きく変化しないことが望ましい．ただし channel 数が少なければ充分な特徴を抽出できず，多過ぎるとメモリ容量が足りなくなるため，入力付近の channel は少なく，出力付近の channel が多くなるよう調整する必要がある．そこで PyramidNet はいくつかの Residual Unit で channel を急激に増加させるのではなく，各 Residual Unit で channel を徐々に増加させることで，急激な channel の増加に関する問題の解決を図っている．PyramidNet はこの工夫によって，一般物体認識データセット CIFAR-10 および CIFAR-100 において高い認識精度を実現している．

ResDrop と PyramidNet は互いに異なる工夫で ResNet の改良を実現している．ResDrop は channel が急激に増加する Residual Unit への対策を持たず，PyramidNet は非常に多くの畳み込みを持つことへの対策を持たない．そこで両者の弱点を補い合うように両者を組み合わせたのが PyramidDrop である [9]。PyramidDrop のアイディアは PyramidNet の論文中で言及はされているものの，具体的な実験結果は報告されていなかった．山田らは，PyramidDrop の層数が増加した場合や，並列学習を行った際に認識精度が改善することを実験で確認し，当時の世界一の認識精度を達成した [9]．

2.3 Shake-Shake

多層 CNN は従来の手法に比べて圧倒的に学習対象となるパラメータ数が多く，一般に学習に用いるデータ数よりもパラメータ数の方が多くなる [17]．学習データ数よりもパラメータ数が十分多い場合，あらゆる学習データと正解ラベルの多対一の対応を学習することができるため，学習データと正解ラベルの対応関係を学習してしまい，学習データに含まれないデータをうまく扱えなくなる現象が起こる．このような現象は過学習と呼ばれる．実際に正解ラベルをランダムに決定した場合に，容易に対応関係を学習してしまうことが知られており [17, 18]，過学習は CNN の学習において大きな課題となっている．ResNet の改良手法の一つである Shake-Shake は特徴を確率的に混ぜることによって，対応関係を学習してしまうことを避け，過学習の解決を図ったものである．Shake-Shake の入力 x に対する Residual Unit 全体の変換 $G(x)$ は以下のように表される．

$$G(x) = x + \alpha F_1(x) + (1 - \alpha) F_2(x) \tag{3}$$

ただし，$F_1(x)$，$F_2(x)$ は Residual Unit の畳み込み部分のみの変換である．α は変換において用いられるパラメータであり，学習時には 0 ～ 1 の実数を取る一様乱数であり，テスト時には 0.5 の値を取る．式 (3) は α に基づいて $F_1(x)$ と $F_2(x)$ で得られる特徴を混ぜ合わせることになる．$F_1(x)$ と $F_2(x)$ はそれぞれ異なる畳み込みを学習し，異なる特徴を抽出するため，これらの特徴が混ざることによってより汎化性の高い特徴を抽出することができると考えられる．Shake-Shake は畳み込みパラメータの更新時にも α の代わりに一様乱数 β を用いて，計算結果を混ぜ合わせる．これらの「学習を阻害する」工夫によって，より汎化性の高い学習を実現し精度を向上させた．

3. 提 案 手 法

PyramidDrop は高度な正則化処理を持たず，Shake-Shake は二分岐でなければならない．二分岐の構造は同様の条件の一分岐の場合の倍の計算を要することになるため，Shake-Shake

を用いた場合，計算時間や学習パラメータ数の都合から多層なCNN にすることが難しい．PyramidDrop で用いられる Random Drop は多層構造で有効であることが確認されているため，多層かつ高度な正則化処理を実現することが望ましい．ここで実数の乱数で特徴を融合させることが Shake-Shake の目的であり，特徴が融合できれば分岐構造は問わない点に着目する．

Veit らの研究 [16] において，ResNet はどれか 1 つの Residual Unit の畳み込みで出力される特徴を使わない場合も，全ての Residual Unit を使った場合と認識精度がほぼ変わらないことが示されている．PyramidDrop の元となった ResDrop や PyramidNet は更にこの性質が強いことが実験的に示されている [8, 16]．これは Residual Unit の畳み込みで出力される特徴がほぼ対等になることを示しており，最終的には各 Residual Unit の特徴が融合していると言える．この性質によって，PyramidDrop の各 Residual Unit で実数の乱数による正則化を用いれば，一分岐でも二分岐と同様に特徴が融合でき，Shake-Shake と同様の効果を得ることができると考えられる．そこで前述のパラメータ p_b 及び α，β を用いて，条件を満たす正則化 ShakeDrop を実現する．

ShakeDrop の Residual Unit は以下の式で表される．

$$G(x) = x + (p_b + \alpha - p_b\alpha)F(x) \tag{4}$$

式 (4) は p_b の値に応じて以下に変化する．

$$G(x) = \begin{cases} x + F(x), & \text{if } p_b = 1 \\ x + \alpha F(x), & \text{otherwise.} \end{cases} \tag{5}$$

これは $p_b = 1$ のとき式 (1) と同じ値を，$p_b = 0$ のとき式 (3) の $F_2(x) = 0$ の場合と同じ値を出力することになる．また畳み込みパラメータの更新時にも α の代わりに実数を取る一様乱数 β を用いて，計算結果を混ぜ合わせる．これらの工夫によって，多層な構造と高度な正則化を両立し，認識精度の向上を目指した．

4. 実　　験

それぞれの手法について比較を行い，提案手法の有効性を検証した．実験 1 では ShakeDrop に適したパラメータを探索した．実験 2 では ShakeDrop の実数の乱数を Batch 単位，Image 単位，Channel 単位，Pixel 単位で決定するよう変化させたときに，認識精度にどのような影響が出るのかを確認した．実験 3 では実験 1，2 の結果を踏まえて，より高い精度で画像が可能か検討した．

4.1 実　験　1

ShakeDrop において有効なパラメータを探索した．

データセットは CIFAR-100 を用いた．ResDrop，及び提案手法の Random Drop における死亡率 p_b は，最初の Residual Unit から一定で増加し最後の Residual Unit で 0.5 となるよう設定した．提案手法については以下の条件に従った．層数は 110 で basic block を用い，最終層の channel は 286，Epoch は 300，BatchSize は 128，重み減衰は 0.0001，モメンタムは 0.9，Nesterov の加速法を用い，初期学習率を 0.5 とした．学習率は Epoch が半分進んだ時点で 0.05，4 分の 3 進んだ時点で 0.005 となるように設定した．α と β はモデル毎の各 Residual Unit で固定で，学習を早く終えるためにモデルを複数にコピーした上で，並列に特徴抽出と勾配計算を行う学習を行った．この並列学習には 4 つモデルを使用した．

表 3: 層数 110 の PyramidNet，PyramidDrop，ShakeDrop における最終 Epoch のエラー率．

手法	α の範囲	β の範囲	CIFAR-100 のエラー率
PyramidNet [8]	1	1	17.87%
PyramidDrop [9]	0	0	17.78%
ShakeDrop (Proposed)	$0 \sim 1$	$0 \sim 1$	19.02%
	$-1 \sim 1$	0	17.81%
	$-1 \sim 1$	$-1 \sim 1$	18.18%
	$-1 \sim 1$	$0 \sim 1$	**16.03%**

表 4: α および β を Batch 単位，画像単位，Channel 単位，Pixel 単位で変化させた場合の層数 110 の ShakeDrop における最終 Epoch のエラー率．

手法	α の範囲	β の範囲	単位	CIFAR-100 のエラー率
ShakeDrop (Proposed)	$-1 \sim 1$	$0 \sim 1$	Batch	16.03%
	$-1 \sim 1$	$0 \sim 1$	Image	16.18%
	$-1 \sim 1$	$0 \sim 1$	Channel	**15.39%**
	$-1 \sim 1$	$0 \sim 1$	Pixel	15.80%

それぞれの手法の結果を表 3 に示す．$\alpha = \beta = 1$ のとき ShakeDrop と PyramidNet は一致する．$\alpha = \beta = 0$ のとき ShakeDrop と PyramidDrop は一致する．これらの条件については元の手法名で併記した．また PyramidNet と PyramidDrop の実験結果については，同条件の [9] のものを表記した．

Shake-Shake 中で有効とされた $\alpha = 0 \sim 1$ の一様乱数，$\beta = 0 \sim 1$ の一様乱数は効果がなかった．実験の中で $\alpha = -1 \sim 1$ の一様乱数，$\beta = 0 \sim 1$ の一様乱数が最も認識精度が高くなった．また $\alpha = -1 \sim 1$ の一様乱数，$\beta = 0 \sim 1$ の一様乱数を用いた場合は，PyramidNet や PyramidDrop よりも精度が高くなった．

4.2 実　験　2

α および β を各モデル中の mini-batch 単位で決定するのではなく，画像単位，Channel 単位，Pixel 単位で変化させた場合に，認識精度にどのような影響が現れるのかを確認した．実験 1 と同様の条件で，$\alpha = -1 \sim 1$ の一様乱数，$\beta = 0 \sim 1$ の一様乱数を用いた．

結果を表 4 に示す．大きな差は見られなかったが，Channel 単位での結果が最も優れていた．

4.3 実　験　3

最先端の手法では学習の際に様々な工夫を用いている．ResNet の関連研究の多くでは 300Epoch での学習を行うのに対して，Shake-Shake ではその 6 倍となる 1800Epoch での学習を行い，高い精度を達成した．Cutout [4] や RandomErasing [5] では学習の際にランダムに決定した矩形に応じて学習

画像のデータを塗りつぶす学習法が効果的なことを示した．これらの工夫を本稿ではそれぞれ 6 倍学習と矩形前処理と呼ぶことにする．この実験では 6 倍学習や矩形前処理を用いた際の認識精度を確認した．実験 1，2 と同様の条件で，channel 単位の $\alpha = -1 \sim 1$ の一様乱数，$\beta = 0 \sim 1$ の一様乱数を用いた．PyramidNet，PyramidDrop，ShakeDrop 層数は 272，bottleneck，最終層の channel は 864 とした．6 倍学習は Shake-Shake の中で使用された 1800Epoch の cosine 学習率に基づく学習法であり，ResNet 以降によく用いられる 300Epoch の 6 倍の時間を学習に費やす。ShakeDrop における初期値を 0.5 とした．ただし矩形前処理として Cutout は Cutout [4] を用い，ShakeDrop は RandomErasing [5] を使用している．

結果を表 1 に示す．ただし ResNeXt，Shake-Shake，Cutout，PyramidNet はそれぞれの論文中の値である．提案手法は PyramidNet や PyramidDrop よりも高い認識精度を示した．ShakeDrop は 6 倍学習や矩形前処理を行わずに最先端の認識精度を達成した．また，6 倍学習や矩形前処理を行った場合，従来手法を大幅に上回る認識精度を達成した．

4.4 考　　察

実験 1〜3 では提案手法 ShakeDrop が優れた認識精度を示すことを確認した．ここでは従来の手法に比べて ShakeDrop において何故学習が上手くいくのか，ShakeDrop における学習は PyramidNet や PyramidDrop とどのような差異が存在するのかを調べる．実験 1 と同様の条件で，PyramidNet，PyramidDrop，ShakeDrop の学習中の loss と勾配の平均，勾配の分散の推移を記録した．ShakeDrop は Batch 単位のものを使用した．loss は学習中の各 Epoch の平均値を使用した．勾配の平均，勾配の分散については，ネットワーク中に 54 個存在する Residual Unit の入力側から 1 個目，27 個目，54 個目の Residual Unit の 2 つ目の畳み込みの勾配の各 Iteration の平均，分散を用いた．これらをそれぞれ first，middle，final と呼称する．

学習中の loss の推移を図 2 に示す．PyramidNet に比べ，確率的正則化を導入した PyramidDrop と ShakeDrop は loss の低下が鈍い．しかし PyramidDrop と ShakeDrop に大きな差は見られない．

学習に用いられる勾配では大きな差が見られた．勾配の平均の推移を図 3，勾配の分散の推移を図 4 に示す．loss では PyramidDrop と ShakeDrop が同程度だったにも関わらず，勾配では PyramidDrop と ShakeDrop が大幅に異なる値を示していることが分かる．特に first での差が激しく，middle においても差が大きい．

ShakeDrop は特徴抽出時に α を用いる．$-1 \sim 1$ の値を取る α を用いた場合，本来抽出するはずだった特徴を打ち消すような，負の特徴を抽出することがある．一方で更新時には $0 \sim 1$ の値を取る β を用いる．β は正の値を取るため，負の特徴を抽出した場合でも正の特徴を取り出していたと仮定して学習を進める．上記の検証で確認された first や middle における勾配の違いは，ShakeDrop の勾配における α と β による勾配計算の齟齬が逆伝播の中で蓄積されたためと考えられる．

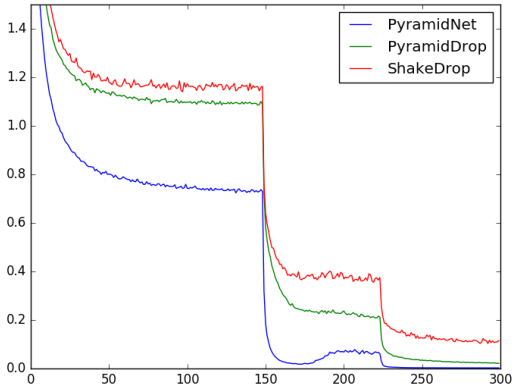


図 2: PyramidNet，PyramidDrop，ShakeDrop の loss の推移．縦軸が loss の値であり，横軸が Epoch 数である．学習率が 1/10 になる 150Epoch と 225Epoch で大幅に loss が下落する．3 つの手法の中でも正則化手法を含む PyramidDrop と ShakeDrop は似た推移となる．

ShakeDrop における勾配の変化にどのような効果があるのか．テストデータにおいて認識精度が高くなる CNN は，あらゆるデータに対して loss が小さくなり，勾配が小さくなる．特に最適なパラメータ付近では勾配が小さくなり，平坦な状態になると考えられており [19]，実験的に平坦さと認識精度の関係が示されている [20]．このような平坦な状態では勾配が小さくなるが，CNN による学習は勾配を用いるため，最適なパラメータ付近では学習が進まなくなる．一方 ShakeDrop は擬似的に勾配を大きくする効果によって，本来の勾配が小さくなっても学習をすすめることができるため，平坦な状態を効率的に探索したことで，従来手法を大きく上回る認識精度を達成したのではないかと考えられる．一方で平坦さと認識精度の関係については懐疑的な見解も存在している [21] ため，ShakeDrop の効果については理論，実験の両面から更なる検討が必要である．

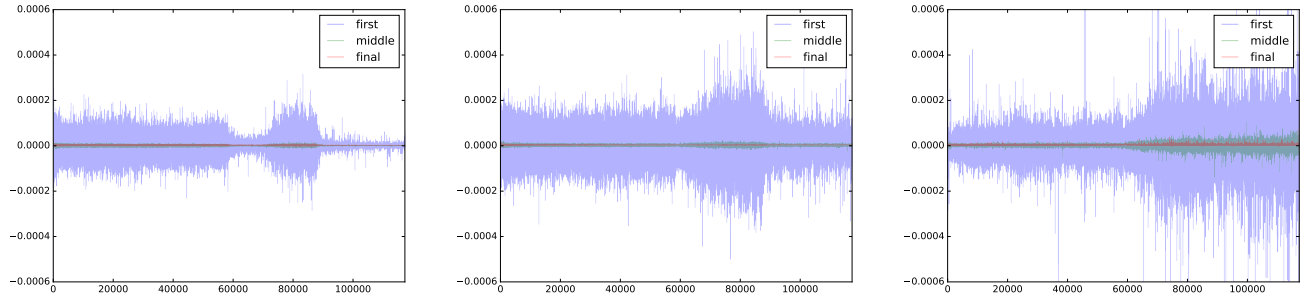
5. まとめと今後の課題

本稿では PyramidDrop と Shake-Shake をもとに，多層かつ高度な正則化を行う新たな確率的正則化手法 ShakeDrop を提案し，その効果を実験的に検討した．その結果，表 1 に示すように，提案手法 ShakeDrop は一般物体認識用データセット CIFAR-10 及び CIFAR-100 において，世界最高の認識精度を達成した．特に CIFAR-100 において最先端の手法から最大で認識精度を 3%程度大幅に改善した．今後の目標として，CIFAR-10 の検証を進めること，大規模な条件に関する検証を進めること，更なるエラー率の低下を目指したパラメータ検証を行うことを目標とすることが挙げられる．

謝辞　本研究は，JST CREST #JPMJCR16E1，JSPS 科研費#25240028 と#17H01803，AWS Cloud Credits for Research Program の補助による．

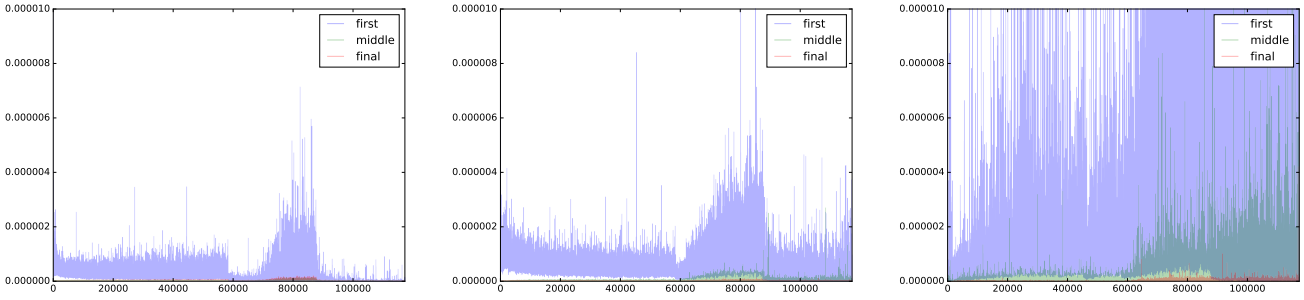
文　　献

[1] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “ImageNet



(a) PyramidNet (b) PyramidDrop (c) ShakeDrop

図 3: 各手法の平均の推移



(a) PyramidNet (b) PyramidDrop (c) ShakeDrop

図 4: 各手法の分散の推移

classification with deep convolutional neural networks," Advances in Neural Information Processing Systems 25, 2012.

[2] K. He and J. Sun, "Convolutional neural networks at constrained time cost," Proc. CVPR, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. CVPR, 2016.

[4] T. DeVries and G.W. Taylor, "Improved regularization of convolutional neural networks with cutout," arXiv:1708.04552 [cs.CV], 2017.

[5] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," arXiv:1708.04896 [cs.CV], 2017.

[6] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," Proc. CVPR, 2017.

[7] X. Gastaldi, "Shake-shake regularization," arXiv:1705.07485v2 [cs.LG], 2017.

[8] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," arXiv:1610.02915v4 [cs.CV], 2016.

[9] 山田良博，岩村雅一，黄瀬浩一，"Pyramidnet における確率的な正則化の効果の検証，" 信学技報，第 116 巻，pp.35–40，2017.

[10] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," arXiv:1603.09382v3 [cs.LG], 2016.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol.15, no.1, pp.1929–1958, Jan. 2014.

[12] G. Kang, J. Li, and D. Tao, "Shakeout: A new regularized deep neural network training scheme.," Proc. AAAI, pp.1751–1757, 2016.

[13] Y. Li and F. Liu, "Whiteout: Gaussian Adaptive Noise Regularization in FeedForward Neural Networks," arXiv:1612.01490 [stat.ML], 2016.

[14] A. Krizhevsky, "Learning multiple layers of features from tiny images," Technical report, University of Toronto, 2009.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," Proc. CVPR, 2009.

[16] A. Veit, M.J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," Advances in Neural Information Processing Systems 29, 2016.

[17] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," arXiv:1611.03530 [cs.LG], 2016.

[18] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M.S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," arXiv:1706.05394 [stat.ML], 2017.

[19] S. Hochreiter and J. Schmidhuber, "Flat minima," Neural Computation, vol.9, no.1, pp.1–42, Jan. 1997.

[20] N.S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P.T.P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," Proc. ICLR, 2017.

[21] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio, "Sharp Minima Can Generalize For Deep Nets," arXiv:1703.04933 [cs.LG], 2017.