Vol.2022-CVIM-230 No.28
2022/5/12

宝くじ仮説を用いた継続学習における破滅的忘却の抑制

城居 巧[†1,a)]　岩村 雅一[†1,b)]　黄瀬 浩一[†1,c)]

概要：継続学習では，識別器が複数のタスクを順に学習する．識別器にニューラルネットワークを用いる場合には，学習を重ねる度に先に学習したタスクの知識が失われる破滅的忘却が起こる．それを軽減する従来手法が提案されているが、タスクが多い場合は精度が著しく低下する．これは従来手法ではパラメータが不足するためと考えられる．本稿では，宝くじ仮説を用いることで使用するパラメータを削減する．宝くじ仮説により，従来手法の破滅的忘却を抑制する効果を維持しつつ，タスクが多い場合に精度が著しく低下する問題を改善する．50 個のタスクの継続学習の実験で，提案手法は従来手法 HAT の精度を最大 14.70%改善することを確認した．

Lottery Ticket Hypothesis for Overcoming Forgetting on continual Learning

Abstract: Catastrophic forgetting occurs when a neural network loses the information learned in a previous task after training on subsequent tasks. Conventional methods have been proposed to mitigate catastrophic forgetting, but their effectiveness is significantly reduced when the number of tasks is large. In this paper, we improve this problem by using The Lottery Ticket Hypothesis. We show the proposed method improves the accuracy of the conventional method HAT by 14.70% with experiments on continuous learning of 50 tasks.

1. はじめに

継続学習では識別器が複数のタスクを順に学習する．一つのタスクでは一つのデータセットを学習する．つまり，継続学習では識別器が複数のデータセットを順に学習することになる．その過程で，過去に学習したデータセットの情報が識別器から失われ，推論精度が下がる破滅的忘却と呼ばれる現象が起こることが知られている [1, 2].

識別器にニューラルネットワーク（以下、ネットワークと呼称）を用いる場合には，脳神経ネットワークにおける忘却現象において報告されている [3, 4] ように，過去に学習したデータセットの情報が失われるのは，それを保持するネットワークのパラメータ（結合重み）が上書きされることが原因と考えられる [5, 6].

破滅的忘却を抑制する手法の一つに HAT [7] がある．HAT は簡潔ながら他の手法と比べて忘却の抑制に優れ，比較的高い精度を達成できる手法である．HAT では，パラメータの上書きを防ぐために，タスク毎に学習に使うパラメータをランダムに選択する．そして，選択されたパラメータのみでそのタスクを学習する．過去のタスクで学習に使用したパラメータは保護され，次のタスク以降では更新されない．このようなメカニズムにより，破滅的忘却を抑制する．HAT では通常，学習するデータセット数の増加に伴い，ネットワーク中の使用されるパラメータが増加する．そのため，タスクが多い場合には，後に学習されるデータセットでパラメータが不足し，学習の阻害により精度の低下が生じる問題がある．

本稿では，HAT でタスクが多い場合に発生するパラメータ不足を回避するため，タスク毎に割り当てるパラメータを減らすことを考える．ただし，割り当てるパラメータをランダムに削減する単純な方法では，各タスクの学習が十分に行われず，精度の低下が予想される．そこで提案手法では，データセットの学習に適したパラメータを宝くじ仮説 [8] を用いて選択する．

宝くじ仮説は，学習に用いるデータセットとネットワークのパラメータの関係を論じた仮説である．ランダムに初期化したネットワークのパラメータを「くじ」に例えて，与えられたデータセットの学習に適したもの（当たりくじ）

[†1] 現在，大阪公立大学
Presently with Osaka Metropolitan University
[a)] sb22635u@st.omu.ac.jp
[b)] masa.i@omu.ac.jp
[c)] kise@omu.ac.jp

と適さないもの（外れくじ）に分けたとき，同じ初期値とデータセットで学習する限り，何度学習しても「当たりくじ」が学習に使われることが実験により示唆された．そのため，宝くじ仮説の「当たりくじ」を利用すれば，学習に用いるデータセットを効率的に学習できると考えられる．幸い，当たりくじはデータセット毎に異なるため，タスク毎に学習するデータセットが異なる継続学習の問題設定に相応しい．

本稿の貢献は，ニューラルネットワークの継続学習に宝くじ仮説の考え方を導入して，与えられたデータセットの学習に適したパラメータを選択することで，タスクが多い場合の破滅的忘却を抑制した点にある．

2. 関連研究

Delange らの分類 [9] によれば，継続学習に対する破滅的忘却を抑制手法は以下の 3 つのアプローチに大別される．

- 再学習によるアプローチ．(Replay methods)
- 正則化項によるアプローチ．(Regularization-based methods)
- パラメータ分離によるアプローチ．(Parameter isolation methods)

本節では 2.1 節， 2.2 節， 2.3 節でそれぞれ再学習のアプローチ，正則化項によるアプローチ，パラメータ分離によるアプローチの概要を述べる．また 3 節で HAT の概要を述べる．

2.1 再学習によるアプローチ

再学習によるアプローチでは過去に学習したデータセットの一部（サブセット）やデータセットに類似した生成データを再利用する．再学習によるアプローチはさらにリハーサル [10–16] と，疑似データの再学習 [17–21]，制約付きの再学習 [22–24] の 3 つに大別される．リハーサルによる再学習では，過去に学習したデータセットの一部を保存し，新たにデータセットを学習する際に保存したデータセットを同時に学習する．先行研究では，保存するデータの選択方法を改善する手法が提案されている．

疑似データの再学習では，過去学習したデータセットと疑似的に同じ生成分布を持つデータを新たなデータセットに学習する際に同時に学習する．過去学習したデータセットと疑似的に同じ生成分布を持つデータは GAN や，バックプロパゲーションによる画像の変更によって生成する．

制限付きの再学習では，サブセットは保存せず，さらに疑似的なサブセットの生成も行わずに再学習を行う．データセットや疑似的なデータセット勾配に代表されるように過去のデータセットでの学習時の情報を代替するものを保存し，新たなデータセットの学習時に用いる．

2.2 正則化項によるアプローチ

正則化項によるアプローチでは損失関数に正則化項を導入することで，過去に学習したデータセットにより獲得した情報を参照しつつ新たなデータセットを学習する．正則化項によるアプローチは，過去のネットワークぼパラメータに注目した手法 [25–30] と過去のデータに注目した手法 [31–34] が存在する．

過去のネットワークのパラメータに注目する手法では，新しいデータセットを学習する際に，過去にデータセットを学習した際のネットワークのパラメータから大きく離れないようにネットワークのパラメータを拘束する．拘束は損失関数に正則化項を導入し達成する手法や，過去のデータセットを学習した際のネットワークのパラメータを用いて，新たなデータセットに適応したパラメータを生成する手法が存在する．

過去のデータに注目した手法では，過去に学習したネットワークの蒸留学習などによって，過去のデータセットを学習したネットワークに含まれる過去のデータセットの情報を新たなデータセットを学習するネットワークに転送する．これにより，データに対し，新たなデータセットを学習するネットワークに過去に学習したデータセットの情報が含まれるように学習し，忘却を抑制する．

2.3 パラメータ分離によるアプローチ

パラメータ分離によるアプローチでは各タスクで使用するパラメータを分離する．パラメータ分離によるアプローチはさらにネットワーク中のパラメータを再利用する手法 [7,35–39] と，学習するデータセットの増加に伴ってネットワークを動的に拡大する手法 [40–45] に大別される． 3 節で述べる HAT も，パラメータ分離によるアプローチのネットワーク中のパラメータを再利用する手法に属する．

ネットワーク中のパラメータを再利用する方法では，データセットごとに使用するパラメータを指定し，ネットワーク中の指定されたパラメータのみを用いて学習と推論を行う．過去のデータセットで使用されているパラメータを固定することで忘却を回避する．またネットワーク中に固定されていないパラメータを適宜新たなデータセットの学習に用いる．

ネットワーク動的に拡大する手法では，データセットの増加に伴ってネットワークにパラメータを追加し，またデータセットごとに使用するパラメータを指定することで忘却を回避する．前者の手法とは違い，ネットワーク中にどのデータセットでも使用されていないパラメータを追加できるため，精度が担保しやすいが，データセットの増加とともにネットワークサイズが増大してしまう．

3. HAT

HAT では，図 1 に示すように，データセット毎に使用す

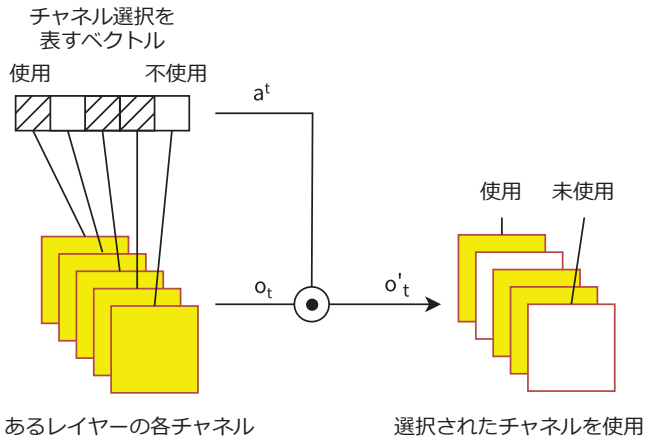

図 1 HAT の概観．データセット毎に使用するチャネルを選択する．一度使用したチャネルのパラメータはその後の学習では保護され，更新しない．

Fig. 1 Overview of HAT.

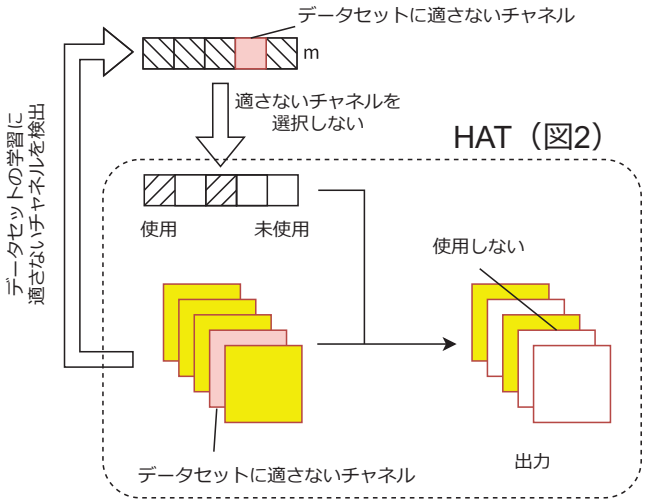

図 2 提案手法の概要．データセットの学習に適さないチャネルを検出し，そのチャネルを使用しないようにチャネル選択を表すベクトルを更新する．

Fig. 2 Overview of the proposed method.

るチャネルを選択する．データセット t に対する，ニューラルネットワークのあるレイヤーの出力を o_t とする．o_t は画像の縦横とチャネルの 3 つを基底とする 3 階のテンソルである．また，チャネル選択を表すベクトルを a^t とする．a^t の要素は 0 または 1 を取り，0 になったチャネルの出力を抑制する．a^t の要素は，概ね確率 50%で 0 と 1 となるように初期化される[*1]．a^t を o_t と同じ大きさのテンソルに拡張する関数を f としたとき，HAT を用いるときのこのレイヤーの出力 o'_t はアダマール積 $\odot$ を用いて，$o'_t = o_t \odot f(a^t)$ で与えられる．データセットの学習時には，既に使用したチャネルのパラメータは更新しない．また，a^t も誤差逆伝播法で更新される．これらのメカニズムに加え，データセット同士で a^t が類似するように制約条件を加える．これにより，新たに学習するデータセットと既学習のデータセットでパラメータを共有する．HAT は既学習のデータセットで使用されたパラメータを積極的に再利用することで新たなデータセットを効率的に学習する．

HAT の改良手法として CAT [39] が存在する．CAT はパラメータの選択をランダムにするのでは無く，アテンション機構を用いる．その際，データセット同士の関連度を考慮するため，既学習のデータセットと似ているデータセットでは同じパラメータが選ばれる確率が上がる．逆に，既学習のデータセットと類似性の低いデータセットでは異なるパラメータが選ばれる．データセット同士の類似性低い継続学習を考えた場合，HAT ではデータセットの類似度を考慮しないため，HAT より，CAT の方がパラメータ不足が発生しやすいと考えられる．そのため，本研究では従来手法として CAT ではなく，HAT と比較を行う．

[*1] 厳密に言えば，標準正規分布に従って乱数を生成した後，シグモイド関数を掛けて 0〜1 の範囲にマッピングする．

4. 提案手法

提案手法は，宝くじ仮説を用いてデータセットの学習に適さないチャネルを特定し，除外する仕組みを HAT に導入する．図 2 に示すように，チャネル選択を表すベクトル a^t を決める際に，データセットの学習に適さないチャネルが選ばれないようにする．その後の処理は HAT と同じである．データセットの学習に適さないチャネルを特定する処理を以下に示す．

(1) 最初のデータセットの学習時のみ，ネットワークのパラメータをランダムに初期化する．
(2) a^t を 3 節に記載した方法でランダムに初期化する．
(3) ネットワークを j エポック学習し，パラメータを更新する（本稿では $j = 1$）．
(4) 宝くじ仮説を用いてデータセットの学習に適さないチャネルを特定し，マスクベクトル m を生成する．
(5) マスクベクトル m を用いて a^t からデータセットの学習に適さないチャネルを除外する．
(6) （3）〜（5）の処理を学習が終わるまで繰り返す．

ここで（4）の m を決める処理を説明する．（4）の m を決める処理には宝くじ仮説を用いる．宝くじ仮説は枝刈り [46, 47] の一種である．宝くじ仮説では，パラメータ毎に「当たりくじ」か「外れくじ」が決まる．宝くじ仮説に拠れば，絶対値の大きいパラメータが当たりである．我々はチャネル毎に「当たり」か「外れ」かを知りたいので，パラメータ単位の情報をチャネル単位に集約する．あるレイヤーへの入力とたたみ込む重み（パラメータ）を $W_{whc'c}$ と書くことにする．ここで $W_{whc'c}$ は 4 階のテンソルであり，w，h，c'，c はそれぞれ画像の横と縦の大きさ、一つ前のレイヤーのチャネル数，このレイヤーのチャネル数を

表す．このとき，このレイヤーの c 番目のチャネルが，今考えているデータセットの学習に適すかを表す指標 I_c を

$$I_c = \left| \max_{w,h,c'} (W_{whc'c}) \right| \quad (1)$$

とおく．すなわち，c に関係する whc' 個の重みの値の最大値を取る．I_c が小さい程，このデータセットの学習に適さないと考えられる．そこで，この指標に基づいて，適していないチャネルを割合 p だけ選ぶ．もしチャネル数が 20 であり，$p = 0.5$ であれば 10 のチャネルを選ぶことになる．そして，除外したいチャネルを表すマスクベクトル m を定義し，除外したいチャネルに該当する要素を 0 とし，それ以外を 1 とする．

続いて，上記の処理の（ 5）では，$\tilde{a}^t = a^t \odot m$ により $\tilde{a}^t$ を求め，これを HAT の a^t と置き換える．これにより，チャネル選択を表すベクトルからこのデータセットの学習に適さないチャネルを除外して，以後の学習を行う．

なお，宝くじ仮説は枝刈りに使用した例があるものの，我々の知る限り、継続学習に使われた例は無い．

5. 実験

5.1 データセット

既存研究（例えば [48]）に習い，CIFAR-10 と CIFAR-100 [49] をベースに SplitCIFAR-10，SplitCIFAR-100，SplitCIFAR-110 を作成して用いる．CIFAR-10 と 100 はそれぞれ 10，100 クラスからなる画像分類データセットである．それぞれの画像は $32 \times 32 \times 3$ ピクセルで構成される．CIFAR-10 と 100 はいずれも 50,000 枚の訓練用画像と 10,000 枚の検証用画像で構成される．HAT を含め，既存手法の評価には通常 10〜20 タスク程度が用いられる．提案手法の良さはタスクが多いときに発揮される．そのため，最大 50 個のデータセットからなる継続学習の条件で実験を行った．SplitCIFAR-10 は，CIFAR-10 の 10 クラス問題をランダムに 5 個の 2 クラス問題に分割して，5 個のデータセットから成る．同様に，SplitCIFAR-100 は CIFAR-100 を 50 個の 2 クラス問題に分割して，50 個のデータセットから成る．SplitCIFAR-110 は，CIFAR-10 をデータセット 1 とし，CIFAR-100 をランダムに 10 クラスごと分割してデータセット 2 から 11 として構成した．

5.2 評価基準：精度と忘却，パラメータ使用率

継続学習では，あるデータセット t の精度（ACC_t）と忘却（FGT_t）を精度の指標として用いる．ACC_t は全データセットの学習後に測定される通常の物体認識の精度であり，FGT_t はあるデータセットの学習直後の精度が全データセットの学習後にどの程度下がるかを表す．データセット数を T とすれば，$\mathrm{FGT}_t = \mathrm{ACC}_{t,t} - \mathrm{ACC}_{t,T}$ である（$1 \leq t \leq T$）．ここで $\mathrm{ACC}_{t,t}$ はデータセット t 学習直後のデータセット t の精度，$\mathrm{ACC}_{t,T}$ はデータセット T 学習直後のデータセット t の精度を表す．さらに，いずれかのデータセットに割り当てられたパラメータの割合（パラメータ使用率）を $\frac{|\bigcup_{1 \leq t \leq T} \theta_t|}{|\theta|}$ とする．ここで，θ はネットワーク中のパラメータ集合を表し，θ_t はデータセット t に対し使用されるパラメータ集合を表す．

5.3 実装

従来手法である HAT の実験と同様に，実験には AlexNet [50] をベースとしたネットワークを用いた．ネットワークは 64, 128, 256 フィルタのサイズとカーネルサイズがそれぞれ 4×4, 3×3, 2×2 である畳み込み層と，2048 ユニットで構成される 2 層の全結合層とマルチヘッド全結合層から構成される．また，畳み込みの後に 2×2 の最大プーリングを行い，レートが 0.2 であるドロップアウトを初めの 2 層に，レートが 0.5 であるドロップアウトを残りの層に追加した．損失関数としてクロスエントロピーを用いた．学習率の初期値は 0.05 で，検証損失が改善されない場合，1/3 倍に減衰させながら、誤差逆伝播法を用いて最適化した．それぞれのデータセットは SplitCIFAR-10，SplitCIFAR-100 で 100 エポック，SplitCIFAR-110 で 200 エポック学習した．また，実験では機械学習のフレームワークである Pytorch [51] と，継続学習の Pytorch 拡張フレームワークである Avalanche [52] を用いて実装した．

従来手法 HAT と提案手法との比較のために，HAT の簡単な拡張として，使用するパラメータをランダムに削減する方法を実装した．この手法を「ランダム」と呼ぶ．直感的には，HAT の a^t に 0 が出る確率を 50%より増やしたものである．a^t の作成方法の制約により，以下のように実装した．図 2 のマスクベクトル m のように，ランダムに 0 か 1 の値を要素として持つマスクベクトル m' を作成する．このとき，0 が b%の確率で生成されるようにする．そして，$\tilde{a}^t = a^t \odot m'$ により $\tilde{a}^t$ を求め，これを HAT の a^t と置き換える．ランダムと提案手法の比較により，パラメータ削減の方法として宝くじ仮説が有効であることを確認する．

5.4 結果と考察

まず，表 1 に 3 つのデータセットでの従来手法 HAT と提案の精度を示す．SplitCIFAR-10，SplitCIFAR-100 については 10 回の平均値，SplitCIFAR-110 については 5 回の平均値である．表から，SplitCIFAR-10 や 110 のようにデータセットの数が少ない場合には，提案手法の精度の向上幅が小さいが，データセットの数が多い SplitCIFAR-100 においては精度の大幅な向上が見られた．また，パラメータ使用率に注目すると，データセットの数が 50 個の場合，HAT はネットワーク中のパラメータをすべて使い切ってしまったが，提案手法はすべて使い切っていない．同様に，従来手法より，提案手法の方が使用中のパラメータが少ないことがわかる．以上の結果から，提案手法はネットワー

表 1　データセット毎の精度とパラメータ使用率.
Table 1　Overview of average accuracy and using parameter rate of HAT and the proposed method.

データセット	タスク数	HAT	提案手法	精度の向上	HAT のパラメータ使用率	提案手法のパラメータ使用率
SplitCIFAR-10	5	0.8010	0.8436	+4.26	0.9649	0.7533
SplitCIFAR-110	11	0.6040	0.6268	+2.28	0.9996	0.8993
SplitCIFAR-100	50	0.5745	0.7215	+14.70	1.0000	0.8837

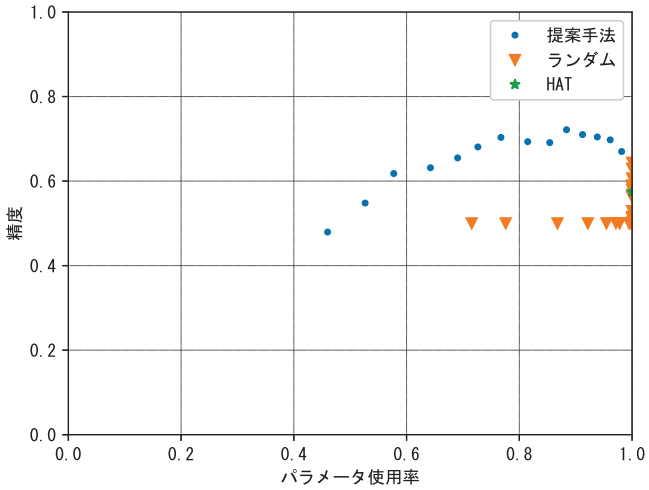

図 3　SplitCIFAR-100 の精度とパラメータ使用率の関係．丸の点，星の点，逆三角形の点はそれぞれ宝くじ仮説を用いたパラメータ削減の方法（提案手法），HAT，ランダムにパラメータを削減する方法を表す．提案手法ではパラメータ使用率の減少とともに HAT や比較手法ランダムより高い精度が得られた場合があるものの，比較手法ランダムではパラメータ使用率の減少とともに HAT より精度が低下した．この結果からパラメータを削減する手法として宝くじ仮説が有効であることがわかる．

Fig. 3　Average accuracy and using parameter rate obtained by experiment using SplitCIFAR-100.

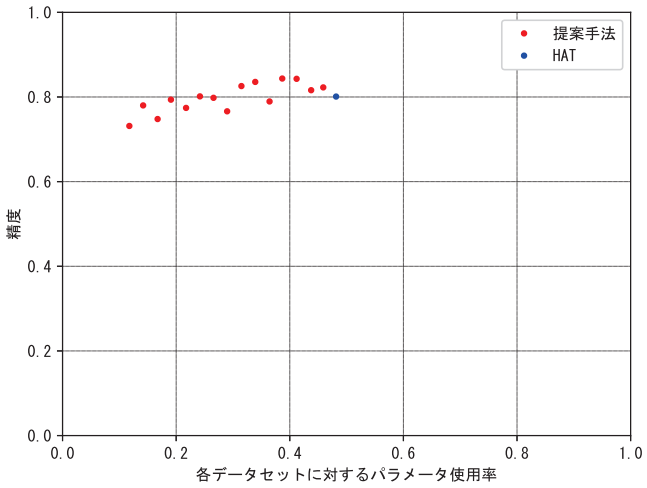

図 4　SplitCIFAR-10（データセット数 5）を用いた実験のグラフ．横軸がネットワークに対する各データセットに割り当てられるパラメータの割合を表し，縦軸が精度を表す．青の点が HAT を表し，赤の点が提案手法を表す．各データセットに対するパラメータ使用率が低下すると精度が低下する傾向がみられる．データセット数が 5 の場合では HAT でのパラメータ不足が深刻でないため，HAT と提案手法で同程度の精度が得られた．

Fig. 4　Average accuracy and using parameter rate obtained by experiment using SplitCIFAR-10.

ク中のパラメータをすべて使い切る現象を軽減し，データセットの数が多い場合に精度を向上が確認できた．

図 3 に精度とパラメータ使用率の HAT と提案手法，ランダムにパラメータの削減する方法の比較を示す．精度，パラメータ使用率は各データセット $t(1 \leq t \leq$ データセット数) における平均をとった．実験結果は SplitCIFAR-10, SplitCIFAR-100 においては 10 回試行，SplitCIFAR-110 においては 5 回試行の平均値である．

HAT では，学習の過程でネットワーク中のパラメータが全てデータセットに割り当てられ，パラメータ不足が発生した．比較手法の「ランダム」では，ハイパーパラメータ b を変化させて，選択するチャネル数を徐々に減らすと，最初はパラメータ使用率は HAT のときと同様にほぼ 1 であり，精度だけが変化した．その際，一時的に精度が向上し HAT より高い精度が得られたものの，提案手法で一番高い精度と比べ低い精度であった．さらに選択するチャネル数を減らすと，選択するチャネル数の減少に伴い精度は低下した．そして，精度がチャンスレートである 50%に下がった後，その精度のままパラメータ使用率が低下した．このことから，ランダムな削減ではパラメータ不足を軽減するために，提案手法と比べ各データセットに割り当てられるパラメータを大幅に削減する必要があることが分かった．また，パラメータを大幅に削減することは精度がチャンスレートである 50%まで低下することにつながる．それに対して提案手法は，ハイパーパラメータである p を調整することで，ネットワークのパラメータを使い切らず，かつ HAT より高い精度を達成した．そのため，宝くじ仮説がパラメータを削減する前と比べ精度を維持したままパラメータを削減する方法として有効であるといえる．

図 4 から 6 にそれぞれ SplitCIFAR-10, SplitCIFAR-100, SplitCIFAR-110 を用いた実験における，各データセットに割り当てられるパラメータの割合と，精度を示す．図 4 から 6 で示すパラメータ使用率は図 3 で示した全データセットを通したパラメータ使用率とは異なり，それぞれのデータセットごとのパラメータ使用率である． また，表 2 と 3 に SplitCIFAR100 を用いた実験において提案手法のハイパーパラメータである p，比較手法ランダムのハイパーパ

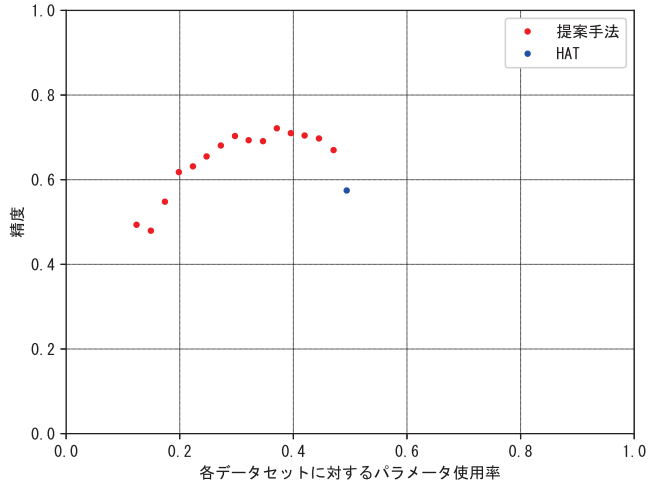

図 5 SplitCIFAR-100（データセット数 50）を用いた実験のグラフ．横軸がネットワークに対する各データセットに割り当てられるパラメータの割合を表し，縦軸が精度を表す．青の点が HAT を表し，赤の点が提案手法を表す．各データセットに対するパラメータ使用率が低い場合で精度が低下する傾向がみられる．SplitCIFAR-100 を用いたデータセット数が 50 の場合の実験では HAT で深刻なパラメータ不足が発生しているため，提案手法の精度のピークと比べ HAT や HAT のパラメータ使用率に近い提案手法で精度の低下が見られる．

Fig. 5 Average accuracy and using parameter rate obtained by experiment using SplitCIFAR-100.

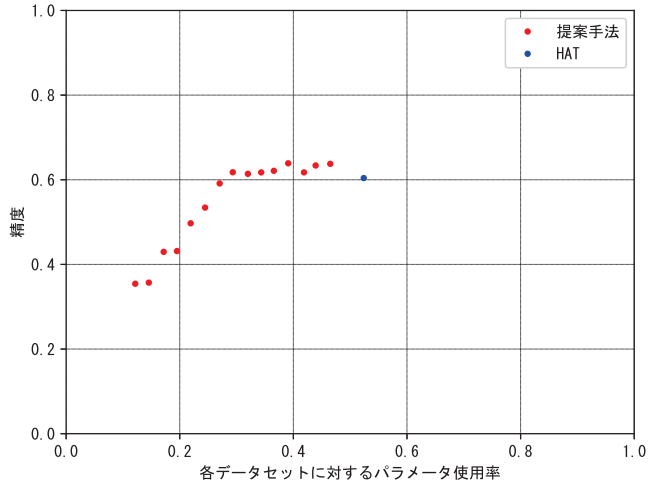

図 6 SplitCIFAR-110（データセット数 11）を用いた実験のグラフ．横軸がネットワークに対する各データセットに割り当てられるパラメータの割合を表し，縦軸が精度を表す．青の点が HAT を表し，赤の点が提案手法を表す．各データセットに対するパラメータ使用率が低い場合で精度が低下する傾向がみられる．SplitCIFAR-110 を用いたデータセット数が 11 の場合の実験では SplitCIFAR-100 を用いたデータセット数が 50 個の実験より HAT でのパラメータ不足が深刻でないため，提案手法の精度のピークと比べた HAT での精度低下は．SplitCIFAR-100 を用いたデータセット数が 50 個の実験より小さい．

Fig. 6 Average accuracy and using parameter rate obtained by experiment using SplitCIFAR-110.

表 2 提案手法のデータセット毎の精度とパラメータ使用率．p は HAT では HAT，提案手法についてはハイパーパラメータである p を表す．各パラメータ使用率は各データセットごとのパラメータ使用率を表す．

Table 2 Average accuracy and using parameter rate obtained by experiment using SplitCIFAR-100 with the proposed method.

p	精度	各パラメータ使用率	パラメータ使用率
HAT	0.5745	0.4947	1.0000
0.05	0.6700	0.4708	0.9815
0.10	0.6975	0.4451	0.9612
0.15	0.7044	0.4198	0.9384
0.20	0.7100	0.3955	0.9121
0.25	0.7215	0.3709	0.8837
0.30	0.6910	0.3466	0.8540
0.35	0.6933	0.3213	0.8148
0.40	0.7032	0.2972	0.7674
0.45	0.6808	0.2723	0.7265
0.50	0.6549	0.2472	0.6906
0.55	0.6315	0.2232	0.6423
0.60	0.6178	0.1984	0.5772
0.65	0.5480	0.1739	0.5265
0.70	0.4795	0.1492	0.4599
0.75	0.4935	0.1239	0.4029

ラメータ b を変化させたときの精度と，各データセットに割り当てられるパラメータの割合，パラメータ使用率を示す．図 4 と 6 からデータセット数の多い，SplitCIFAR-100，SplitCIFAR-110 を用いた実験では，各データセットに割り当てられるパラメータの割合が少ない場合で，各データセットに割り当てられるパラメータの割合が多い場合と比べ精度の低下が顕著に見られた．また，図 4 からデータセット数の少ない SplitCIFAR-10 を用いた実験においても SplitCIFAR-100，SplitCIFAR-110 を用いた実験ほど顕著でないものの，各データセットに割り当てられるパラメータの割合が少ないほど，精度の低下する傾向が見られた．そのため，提案手法を用いた場合でもそれぞれのデータセットに割り当てられるパラメータが少なすぎると精度が低下すると考えられる．提案手法はデータセットを多く学習する場合の精度を向上させるために，それぞれのデータセットに割り当てられるパラメータを削減する必要がある．そのため，提案手法では，今回実験を行った 50 個のデータセットを大幅に超えるようなデータセットを学習させる場合に精度が低下することが予想される．今後さらに精度を保ちつつパラメータを多く削減できる手法が望まれる．

6. 結論

ランダムにデータセットにパラメータを割り当てる従来手法では，データセットの数が多い際に精度が低下する問題があり，これはパラメータ不足によって発生することが示唆された．宝くじ仮説に基づきデータセットにパラメー

表 3 比較手法ランダムデータセット毎の精度とパラメータ使用率. b は比較手法ランダムのハイパーパラメータである b を表す. 各パラメータ使用率は各データセットごとのパラメータ使用率を表す.

Table 3 Average accuracy and using parameter rate obtained by experiment using SplitCIFAR-100 with the random method.

b	精度	各パラメータ使用率	パラメータ使用率
0.05	0.6435	0.4758	1.0000
0.10	0.6292	0.4498	1.0000
0.15	0.6062	0.4255	1.0000
0.20	0.5898	0.4004	1.0000
0.25	0.5858	0.3755	1.0000
0.30	0.5793	0.3511	1.0000
0.35	0.5623	0.3256	1.0000
0.40	0.5675	0.3006	1.0000
0.45	0.5289	0.2753	1.0000
0.50	0.5156	0.2506	1.0000
0.55	0.5117	0.2256	1.0000
0.60	0.5021	0.2002	1.0000
0.65	0.5009	0.1756	0.9999
0.70	0.5001	0.1501	0.9997
0.75	0.5005	0.1256	0.9986
0.80	0.5000	0.1007	0.9950
0.85	0.5000	0.0754	0.9779
0.86	0.5005	0.0705	0.9713
0.88	0.5000	0.0607	0.9545
0.90	0.5000	0.0505	0.9216
0.92	0.5000	0.0404	0.8676
0.94	0.5000	0.0305	0.7758
0.95	0.5000	0.0254	0.7155

タを割り当てる提案手法では，従来手法の破滅的忘却を抑制する効果を維持しつつ使用するパラメータを削減し，50個のデータセットを学習した実験でパラメータ使用率を11.63%軽減し，精度を 14.70%向上した.

参考文献

[1] Ratcliff, R.: Connectionist models of recognition memory: constraints imposed by learning and forgetting functions., *Psychological review*, Vol. 97 2, pp. 285–308 (1990).
[2] French, R.: Catastrophic forgetting in connectionist networks, *Trends in cognitive sciences*, Vol. 3, pp. 128–135 (online), DOI: 10.1016/S1364-6613(99)01294-2 (1999).
[3] Cichon, J. and Gan, W.-B.: Branch-specific dendritic Ca2+ spikes cause persistent synaptic plasticity, *Nature*, Vol. 520 (online), DOI: 10.1038/nature14251 (2015).
[4] Yang, G., Pan, F. and Gan, W.-B.: Stably maintained dendritic spines are associated with lifelong memories, *Nature*, Vol. 462, pp. 920–924 (2009).
[5] Benna, M. and Fusi, S.: Computational principles of synaptic memory consolidation, *Nature Neuroscience*, Vol. 19 (online), DOI: 10.1038/nn.4401 (2016).
[6] Abraham, W. C. and Robins, A.: Memory retention and weight plasticity in ANN simulations, *Trends in Neurosciences*, Vol. 2, No. 28, pp. 73–78 (2005).
[7] Serrà, J., Surís, D., Miron, M. and Karatzoglou, A.: Overcoming catastrophic forgetting with hard attention to the task, *Proc. ICML 2018*, pp. 4548–4557 (2018).
[8] Frankle, J. and Carbin, M.: The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks, *Proc. ICLR* (2019).
[9] Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G. and Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1 (online), DOI: 10.1109/TPAMI.2021.3057446 (2021).
[10] Rebuffi, S.-A., Kolesnikov, A., Sperl, G. and Lampert, C. H.: iCaRL: Incremental Classifier and Representation Learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
[11] Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T. and Wayne, G.: Experience Replay for Continual Learning, *Advances in Neural Information Processing Systems* (Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R., eds.), Vol. 32, Curran Associates, Inc. (2019).
[12] Isele, D. and Cosgun, A.: Selective Experience Replay for Lifelong Learning, *AAAI* (2018).
[13] Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P., Torr, P. and Ranzato, M.: Continual learning with tiny episodic memories (2019).
[14] De Lange, M. and Tuytelaars, T.: Continual Prototype Evolution: Learning Online From Non-Stationary Data Streams, *Proc. ICCV*, pp. 8250–8259 (2021).
[15] Chaudhry, A., Gordo, A., Dokania, P. K., Torr, P. H. S. and Lopez-Paz, D.: Using Hindsight to Anchor Past Knowledge in Continual Learning, *AAAI* (2021).
[16] Yin, H., peng yang and Li, P.: Mitigating Forgetting in Online Continual Learning with Neuron Calibration, *Advances in Neural Information Processing Systems* (Beygelzimer, A., Dauphin, Y., Liang, P. and Vaughan, J. W., eds.) (2021).
[17] Shin, H., Lee, J. K., Kim, J. and Kim, J.: Continual Learning with Deep Generative Replay, *NIPS* (2017).
[18] Atkinson, C., McCane, B., Szymanski, L. and Robins, A. V.: Pseudo-rehearsal: Achieving deep reinforcement learning without catastrophic forgetting, *Neurocomputing*, Vol. 428, pp. 291–307 (online), DOI: 10.1016/j.neucom.2020.11.050 (2021).
[19] Lavda, F., Ramapuram, J., Gregorova, M. and Kalousis, A.: Continual Classification Learning Using Generative Models (2018).
[20] Ramapuram, J., Gregorova, M. and Kalousis, A.: Lifelong generative modeling, *Neurocomputing*, Vol. 404, pp. 381–400 (2020).
[21] Liu, Y., Schiele, B. and Sun, Q.: Adaptive Aggregation Networks for Class-Incremental Learning, *Proc. CVPR*, pp. 2544–2553 (2021).
[22] Lopez-Paz, D. and Ranzato, M.: Gradient Episodic Memory for Continual Learning, *NIPS* (2017).
[23] Chaudhry, A., Marc' AurelioRanzato, Rohrbach, M., Elhoseiny, M.: Efficient Lifelong Learning with A-GEM, *ICLR* (2019).
[24] Aljundi, R., Lin, M., Goujaud, B. and Bengio, Y.: Online continual learning with no task boundaries, *ArXiv*, Vol. abs/1903.08671 (2019).
[25] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A. et al.: Overcoming

catastrophic forgetting in neural networks, *Proceedings of the national academy of sciences*, Vol. 114, No. 13, pp. 3521–3526 (2017).

[26] Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W. and Zhang, B.-T.: Overcoming Catastrophic Forgetting by Incremental Moment Matching, *Advances in Neural Information Processing Systems* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Vol. 30, Curran Associates, Inc. (2017).

[27] Zenke, F., Poole, B. and Ganguli, S.: Continual learning through synaptic intelligence, *International Conference on Machine Learning*, PMLR, pp. 3987–3995 (2017).

[28] Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M. and Tuytelaars, T.: Memory Aware Synapses: Learning what (not) to forget, *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[29] Chaudhry, A., Dokania, P. K., Ajanthan, T. and Torr, P. H. S.: Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence, *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[30] Mirzadeh, S. I., Farajtabar, M., Gorur, D., Pascanu, R. and Ghasemzadeh, H.: Linear Mode Connectivity in Multitask and Continual Learning, *International Conference on Learning Representations* (2021).

[31] Li, Z. and Hoiem, D.: Learning Without Forgetting, *Computer Vision – ECCV 2016* (Leibe, B., Matas, J., Sebe, N. and Welling, M., eds.), Cham, Springer International Publishing, pp. 614–629 (2016).

[32] Oren, G. and Wolf, L.: In Defense of the Learning Without Forgetting for Task Incremental Learning, *Proc. ICCV*, pp. 2209–2218 (2021).

[33] Jung, H., Ju, J., Jung, M. and Kim, J.: Less-forgetting Learning in Deep Neural Networks, *CoRR*, Vol. abs/1607.00122 (2016).

[34] Zhang, J., Zhang, J., Ghosh, S., Li, D., Tasci, S., Heck, L., Zhang, H. and Kuo, C. J.: Class-incremental Learning via Deep Model Consolidation, *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Los Alamitos, CA, USA, IEEE Computer Society, pp. 1120–1129 (online), DOI: 10.1109/WACV45572.2020.9093365 (2020).

[35] Mallya, A. and Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning, *Proc. CVPR*, pp. 7765–7773 (2018).

[36] Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A. A., Pritzel, A. and Wierstra, D.: PathNet: Evolution Channels Gradient Descent in Super Neural Networks, *CoRR*, Vol. abs/1701.08734 (2017).

[37] Mallya, A., Davis, D. and Lazebnik, S.: Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights, *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[38] Ke, Z., Liu, B., Ma, N., Xu, H. and Shu, L.: Achieving Forgetting Prevention and Knowledge Transfer in Continual Learning, *Advances in Neural Information Processing Systems* (Beygelzimer, A., Dauphin, Y., Liang, P. and Vaughan, J. W., eds.) (2021).

[39] Ke, Z., Liu, B. and Huang, X.: Continual Learning of a Mixed Sequence of Similar and Dissimilar Tasks, *Advances in Neural Information Processing Systems* (Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. and Lin, H., eds.), Vol. 33, Curran Associates, Inc., pp. 18493–18504 (2020).

[40] Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R. and Hadsell, R.: Progressive Neural Networks, *CoRR*, Vol. abs/1606.04671 (2016).

[41] Aljundi, R., Chakravarty, P. and Tuytelaars, T.: Expert gate: Lifelong learning with a network of experts, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3366–3375 (2017).

[42] Xu, J. and Zhu, Z.: Reinforced Continual Learning, *NeurIPS* (2018).

[43] Rosenfeld, A. and Tsotsos, J. K.: Incremental Learning Through Deep Adaptation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 3, pp. 651–663 (online), DOI: 10.1109/TPAMI.2018.2884462 (2020).

[44] Wang, L., Zhang, M., Jia, Z., Li, Q., Bao, C., Ma, K., Zhu, J. and Zhong, Y.: AFEC: Active Forgetting of Negative Transfer in Continual Learning, *ArXiv*, Vol. abs/2110.12187 (2021).

[45] Raghavan, K. and Balaprakash, P.: Formalizing the Generalization-Forgetting Trade-off in Continual Learning, *Advances in Neural Information Processing Systems* (Beygelzimer, A., Dauphin, Y., Liang, P. and Vaughan, J. W., eds.) (2021).

[46] LeCun, Y., Denker, J. and Solla, S.: Optimal Brain Damage, *Advances in Neural Information Processing Systems* (Touretzky, D., ed.), Vol. 2, Morgan-Kaufmann (1990).

[47] Han, S., Pool, J., Tran, J. and Dally, W. J.: Learning both weights and connections for efficient neural networks, *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pp. 1135–1143 (2015).

[48] Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images (2009).

[49] Krizhevsky, A.: Learning multiple layers of features from tiny images, Technical report, University of Tronto (2009).

[50] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems*, Vol. 25 (2012).

[51] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, *Advances in Neural Information Processing Systems 32* (Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R., eds.), Curran Associates, Inc., pp. 8024–8035 (2019).

[52] Lomonaco, V., Pellegrini, L., Cossu, A., Carta, A., Graffieti, G., Hayes, T. L., Lange, M. D., Masana, M., Pomponi, J., van de Ven, G., Mundt, M., She, Q., Cooper, K., Forest, J., Belouadah, E., Calderara, S., Parisi, G. I., Cuzzolin, F., Tolias, A., Scardapane, S., Antiga, L., Amhad, S., Popescu, A., Kanan, C., van de Weijer, J., Tuytelaars, T., Bacciu, D. and Maltoni, D.: Avalanche: an End-to-End Library for Continual Learning, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2nd Continual Learning in Computer Vision Workshop (2021).