Vol.2022-CVIM-230 No.24
2022/5/12

学習済みモデルに任意の出力を促す画像生成とそれを用いた学習

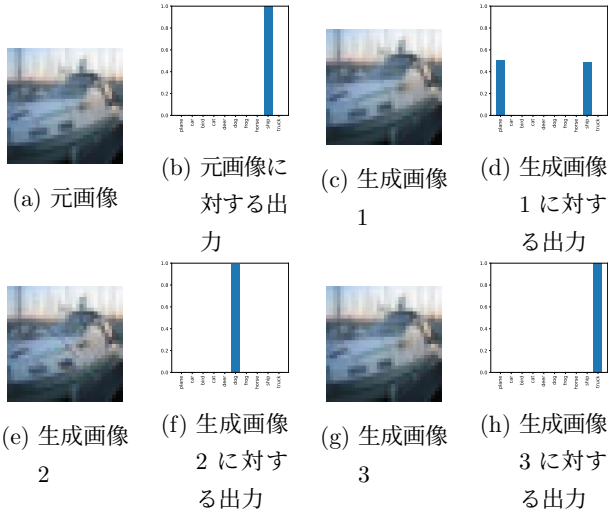
平野 甫[1,a)]　岩村 雅一[1,b)]　黄瀬 浩一[1,c)]

概要：本稿では，画像分類モデルに任意の出力を促す画像生成手法を提案するとともに，提案手法による生成画像を用いた学習を検証する．従来の画像生成手法では画像分類モデルの 1 位の出力のみを改変させるが，提案手法では 1 位の出力のみならず 2 位以降の出力をも任意に改変することができる．提案手法による生成画像の活用例として，生成画像を用いた画像分類モデルの学習を行った．実験の結果，生成画像を用いない学習に比べ，画像分類の精度が向上した．

1. はじめに

近年，ニューラルネットワークは急速に発展してきており，画像分類や物体検出，自然言語処理などの様々な分野での研究が盛んに行われている．中でも，画像分類は，医療分野での病気の診断や工場での検品など多岐にわたる分野で活用されている重要な技術の 1 つである．画像分類技術の急速な発展の要因の 1 つに誤差逆伝播法による重み更新が挙げられる．画像分類モデルは入力として画像を受け取り，出力として入力された画像が各クラスに属する確率を返す．画像分類モデルの学習では，入力画像に対する画像分類モデルの出力が入力画像に割り当てられた教師ラベルに近づくように画像分類モデルの重みを更新する．この画像分類モデルの重み更新に用いられるのが誤差逆伝播法である．

本研究では，この誤差逆伝播法による重み更新を画像生成に用いることができるかを試みた．提案手法による画像生成では，画像分類モデルの出力が予め設定した教師ラベルとなるような画像の生成を目標とする．提案手法では，誤差逆伝播法による重み更新を画像分類モデルの重みにではなく，画像に対して適用することを考えた．そのため，提案手法では，画像を「重み」として扱い，誤差逆伝播法による重み更新によって画像を改変させていく．このとき，生成画像に対する画像分類モデルの出力が予め設定した教師ラベルに近づくように重み更新を行う．これにより，提案手法による画像生成では，**図 1** に示すような，画像分類モデルの出力全体を任意に改変させる画像を生成することができた．



(a) 元画像　(b) 元画像に対する出力　(c) 生成画像 1　(d) 生成画像 1 に対する出力
(e) 生成画像 2　(f) 生成画像 2 に対する出力　(g) 生成画像 3　(h) 生成画像 3 に対する出力

図 1　提案手法による画像生成例

また，従来の画像生成手法では不可能であった複数のクラスの特徴を持つ画像を生成できる．提案手法による画像生成では，重み更新の方法によっては得られる生成画像に大きなノイズが現れ，不自然な画像が生成される場合がある．そこで，本研究では生成画像のノイズを減らし，自然な画像を生成する方法についても実験を行った．

また，近年ではデータ拡張として画像生成を用いる研究も行われている [1–3]．そこで，本研究では，提案手法による画像生成の活用例として提案手法による生成画像を用いた画像分類モデルの学習を行った．以降，2 節で提案手法及び重み更新の違いによる生成画像の違いについて説明

[1] 大阪公立大学大学院情報学研究科基幹情報学専攻
Graduate School of Informatics, Osaka Metropolitan University
[a)] sb22852v@st.omu.ac.jp
[b)] masa.i@omu.ac.jp
[c)] kise@omu.ac.jp

し，3 節で提案手法による生成画像を用いた学習の詳細と結果を示し，4 節で関連研究を紹介し，最後に 5 節で結論を述べる．

2. 提案手法

2.1 提案手法の概要

提案手法による画像生成では，画像分類モデルの出力が予め設定した出力となるように画像を改変していく．そのために用いるのが誤差逆伝播法である．通常，誤差逆伝播法はニューラルネットワークの出力が教師ラベルに近づくようにニューラルネットワークの重みを更新するために用いられる．提案手法ではこの誤差逆伝播法を用いて画像分類モデルの重みではなく，画像を更新する．そのためには画像分類モデルの重みを固定したうえで，画像を「重み」として扱えるようにする必要がある．提案手法では画像を「重み」として扱うために，**図 2(a)** のように画像分類モデルの入力層の前に新たに全結合層 (以降，画像重み層と呼ぶ) を 1 層追加する．一般に，全結合層は入力 $\boldsymbol{x} \in \mathbb{R}^D$ に対して，重み $\boldsymbol{W} \in \mathbb{R}^{d \times D}$，バイアス $\boldsymbol{b} \in \mathbb{R}^d$ をパラメータとして持ち，以下の式で計算される $\boldsymbol{y} \in \mathbb{R}^d$ を出力する．

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \tag{1}$$

画像重み層の役割は，画像を「重み」として保持し，保持している「重み」を画像として出力することである．そこで，提案手法では $D = 1$，$d = c \times h \times w$ とし，常に $\boldsymbol{x} = 1$，$\boldsymbol{b} = \boldsymbol{0}$ とすることで，保持している重みを出力することができる．ただし，c, h, w はそれぞれ画像のチャンネル数，高さ，幅である．画像を画像重み層の「重み」として設定する際には画像の画素値の取りうる値は $[0, 1]$ であるのに対し，「重み」は $(-\infty, +\infty)$ であるため，提案手法では以下の式を用いて画像の画素値を $[0, 1]$ から $(-\infty, +\infty)$ にマッピングしたのちに重みとして設定する．

$$\boldsymbol{x'} = \tan((\boldsymbol{x} - 0.5)\pi) \tag{2}$$

ここで，$\boldsymbol{x} \in \mathbb{R}^{c \times h \times w}$ は元画像，$\boldsymbol{x'} \in \mathbb{R}^{c \times h \times w}$ はマッピング後の値である．したがって，画像重み層の重み $\boldsymbol{W}$ は

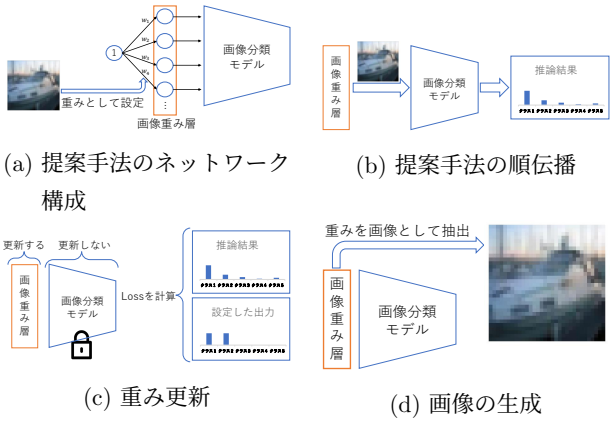
$$\boldsymbol{W} = (x'_{111}, x'_{112}, x'_{113}, \ldots, x'_{chw})^\top \tag{3}$$

となる．ただし，x_{ijk} は画像の ijk 成分を表す．画像重み層からの出力 $\boldsymbol{y}$ に対しては式 (2) の逆関数

$$\boldsymbol{y'} = \frac{\arctan(\boldsymbol{y})}{\pi} + 0.5 \tag{4}$$

によって値を $(-\infty, +\infty)$ から $[0, 1]$ にマッピングしたのち，画像となるように並べ替えたものを画像分類モデルに入力する．すなわち，画像分類モデルへの入力 $\boldsymbol{z} \in \mathbb{R}^{c \times h \times w}$ は

$$z_{ijk} = y'_{(i-1)hw+(j-1)w+k} \tag{5}$$



(a) 提案手法のネットワーク構成　(b) 提案手法の順伝播

(c) 重み更新　(d) 画像の生成

図 2　提案手法の流れ

となる．そのため，**図 2(b)** のように画像分類モデルに入力されるものは画像となり，画像分類モデルは通常通り，推論結果を出力する．得られた推論結果と予め設定した出力との Loss を計算し重み更新を行うことで画像を改変させていく．上記の重み更新を Loss が十分小さくなるまで繰り返し，最後に画像重み層の重みを画像に変換することで画像を生成する．

2.2 重み更新の方法の違いによる生成画像の違い

提案手法による画像生成では，重み更新の方法によって得られる生成画像が大きく異なる．本小節ではいくつかの重み更新手法によって得られた生成画像例について紹介する．

2.2.1 SGD を用いた画像生成

重み更新に SGD を用いた場合の生成画像の例を**図 3** に示す．SGD の更新式は以下の通りである．

$$\boldsymbol{W}^{t+1} \leftarrow \boldsymbol{W}^t - \gamma \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}^t) - \gamma \lambda \boldsymbol{W}^t \tag{6}$$

ここで，$\boldsymbol{W}$ は画像重み層の重み，γ は学習率，$\mathcal{L}$ は損失関数，λ は重み減衰を表す．画像生成時の SGD のパラメータの設定は $\gamma = 0.05, \lambda = 5 \times 10^{-4}$ である．図 3 から分かるように，ノイズがほぼなく，色彩にも不自然な箇所のない画像を生成できた．

2.2.2 Momentum SGD を用いた画像生成

重み更新に Momentum SGD [4] を用いた場合の生成画像の例を**図 4** に示す．Momentum SGD の更新式は以下の通りである．

$$\boldsymbol{v}^{t+1} \leftarrow \mu v^t + \gamma \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}^t) + \gamma \lambda \boldsymbol{W}^t \tag{7}$$

$$\boldsymbol{W}^{t+1} \leftarrow \boldsymbol{W}^t - \gamma \boldsymbol{v}^{t+1} \tag{8}$$

ここで，$\boldsymbol{W}$ は画像重み層の重み，γ は学習率，μ はモメンタム，$\mathcal{L}$ は損失関数，λ は重み減衰を表す．Momentum SGD のパラメータの設定は $\gamma = 0.05$，$\mu = 0.9$，$\lambda = 5 \times 10^{-4}$ である．図 4 から分かるように，重み更新に Momentum SGD を用いた場合，ノイズが目立つ画像が生成された．そ

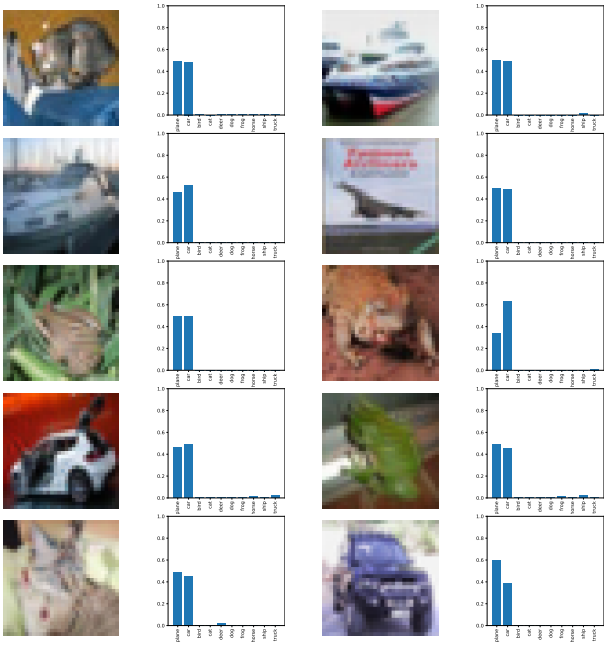


図 3　SGD を用いた画像生成

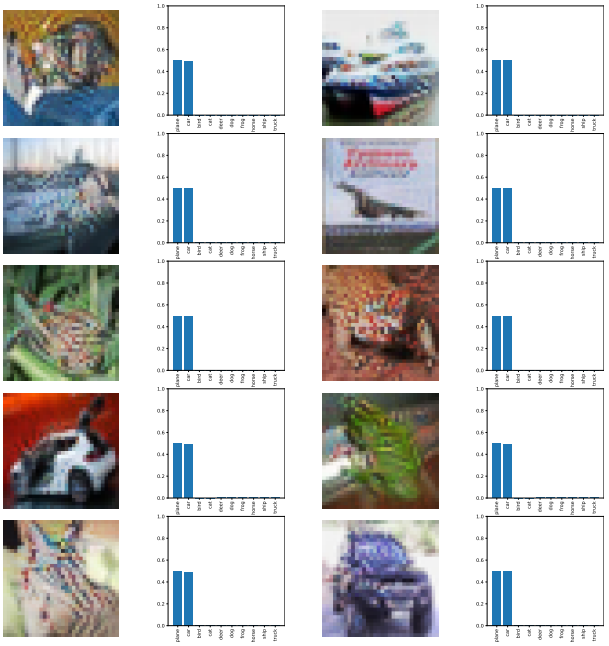


図 4　Momentum SGD を用いた画像生成

の要因としては，SGD を用いた場合には重み更新の際に過去の影響を受けないため，その時点での最適な更新を行うことができる一方，Momentum SGD を用いた場合には重み更新の際に過去の更新の影響を受けてその時点での最適な更新方向に向かうことができないためと考えられる．

2.2.3　ローパスフィルタを用いた画像生成

重み更新に Momentum SGD を用いた場合にノイズが目立つ画像が生成されたことを受けて，本研究では画像のノイズ除去によく用いられるローパスフィルタを適用した．ローパスフィルタは画像重み層から出力される画像に適用され，画像分類モデルに入力される．また，重み更新

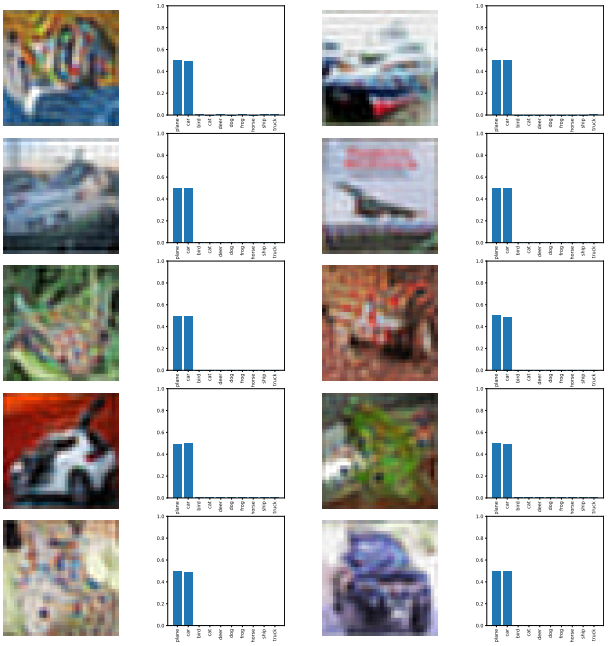


図 5　ローパスフィルタを用いた画像生成

終了後，画像重み層から画像を取り出す際にも適用される．なお，ローパスフィルタは画像の各チャンネルごとにそれぞれ適用される．ローパスフィルタを用いた場合の生成画像の例を**図 5** に示す．図 5 では，32×32 の画像に対し，半径 10 のローパスフィルタを適用した．重み更新には Momentum SGD を使用し，各種パラメータの設定は Momentum SGD を用いた画像生成と同じである．図 5 から分かるように，Momentum SGD を用いた画像生成の際に見られたノイズを除去することができている一方で，生成画像の色彩に不自然な虹色が見られる結果となった．

3.　提案手法による生成画像を用いた学習

本節では，提案手法による生成画像を用いた画像分類モデルの学習の概要とその結果について述べる．

3.1　実験の流れと実験条件

実験の流れを**図 6** に示す．まず，学習済みの画像分類モデルを用いて学習用データセットの各画像をもとに提案手法による画像生成を行った．このとき，予め設定する出力は「元画像のクラス：0.9，ランダムに選んだクラス：0.1」とした．画像生成時の重み更新には SGD を用いた．また，画像生成時の重み更新の回数は 100 回とした．続いて，学習データセットの画像の一部を生成画像に置き換えることで新規データセットを作成した．このとき，生成画像と置き換えるデータセットの画像は，その生成画像のもととなった画像であることに注意されたい．最後に，作成した新規データセットを用いて画像分類モデルをスクラッチから学習した．学習時の Loss の計算には以下の式を用いた．

$$\mathcal{L} = \begin{cases} 0.5 \times D_{KL}(\boldsymbol{p}, \boldsymbol{q}) \\ \quad + 0.5 \times \mathcal{L}_c(\boldsymbol{p}, \boldsymbol{t}) & \text{(入力が生成画像の場合)} \\ 0.5 \times D_{KL}(\boldsymbol{p}, \boldsymbol{t}) \\ \quad + 0.5 \times \mathcal{L}_c(\boldsymbol{p}, \boldsymbol{t}) & \text{(入力が元画像の場合)} \end{cases} \tag{9}$$

ここで，$\boldsymbol{p}$ は学習中の画像分類モデルの出力，$\boldsymbol{q}$ は提案手法による画像生成で用いた学習済み画像分類モデルの生成画像に対する出力，$\boldsymbol{t}$ は生成画像の元画像の正解ラベルである．D_{KL} は Kullback-leibler divergence (KL divergence) [5] であり，2 つの離散確率分布 $\boldsymbol{p}, \boldsymbol{q}$ に対して以下のように定義される．

$$D_{KL}(\boldsymbol{p}, \boldsymbol{q}) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{10}$$

$\mathcal{L}_c$ は交差エントロピー誤差であり．2 つの離散確率分布 $\boldsymbol{p}, \boldsymbol{q}$ に対して次のように定義される．

$$\mathcal{L}_c(\boldsymbol{p}, \boldsymbol{q}) = -\sum_x q(x) \log(p(x)) \tag{11}$$

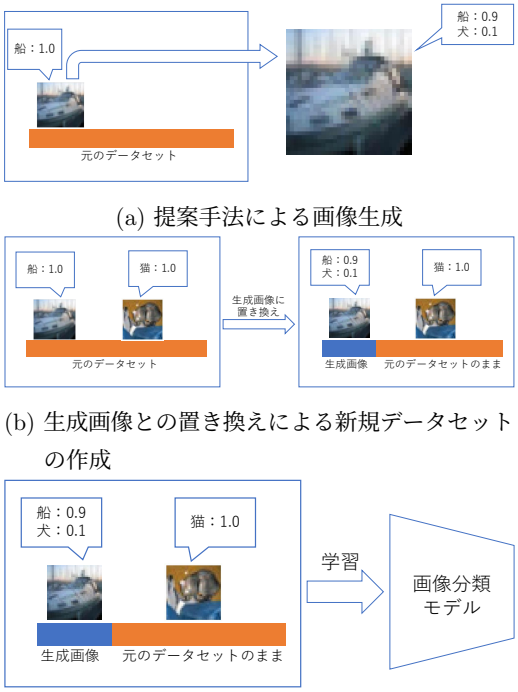
なお，生成画像への置き換えを行わなかった場合には Loss の計算は以下の式を用いた．

$$\mathcal{L} = \mathcal{L}_c(\boldsymbol{p}, \boldsymbol{t}) \tag{12}$$

画像生成に用いる画像分類モデルには学習済みの VGG16 [6]（Top-1 Accuracy：92.27%）を用いた．学習する画像分類モデルには VGG16，データセットとして CIFAR-10 [7] を使用した．また，データセットのうち，10%，20%，30%，40%，50%を生成画像に置き換えた場合を実験した．各学習のエポック数は全て 300 とし，300 エポック終了時の精度を測定した．記載する実験結果は 5 回平均である．

3.2 実験結果と考察

実験結果を**表 1** に示す．データセットの 30%を生成画像に置き換えた場合に Top-1 Accuracy が 92.560%となり，置き換えを行わなかった場合（ベースライン）に比べ，0.084pt 上昇した．しかし，その他の場合ではいずれもベースラインよりも低い結果となった．この原因として，画像分類モデルの学習中に画像生成を行わなかったことが考えられる．図 1 から分かるように，提案手法による画像生成では画像を大域的に変化させることはなく，局所的に変化させる．このような微小な変化は，幾何学的な変換に対して脆弱であると考えられる．今回の実験では，学習前に予め生成画像を含むデータセットを作成し，学習時にそのデータセットから画像を取り出す際には，0.5 の確率で画像が左右反転される RandomHorizontalFlip，画像の外側を 4 ピクセルだけゼロパディングした後 32×32 の領域を切り出す RandomCrop が適用される．そのため，生成画像に対するラベルが不適切である場合が多かったと考えられる．



(a) 提案手法による画像生成

(b) 生成画像との置き換えによる新規データセットの作成

(c) 新規データセットを用いた画像分類モデルの学習

図 6　実験の流れ

4. 関連研究

4.1 Conditional Generative Adversarial Network

Conditional Generative Adversarial Network（CGAN）[8] は，ノイズとラベル情報から画像生成を行う生成モデルである．CGAN には，**図 7** のような Generator と Discriminator と呼ばれる 2 つのネットワークが登場する．Generator は入力としてノイズとラベル情報を受け取り，画像を生成する．Discriminator は入力として画像とラベル情報を受け取り，入力された画像が Generator によって生成された画像か訓練データの画像であるかを判別する．学習時には Generator は Discriminator を騙せるように学習し，Discriminator は騙されないように学習する．このように，お互いが敵対する関係の中で学習を進めることで，Generator は訓練データに似た自然な画像を生成できるようになる．しかし，Discriminator はあくまで生成画像か否かを判別するのみであるため，生成画像を画像分類モデルに入力した際の出力には関心がない．そのため，提案手法のような画像分類モデルに任意の出力を促す画像生成は不可能である．

4.2 Adversarial Attack

Adversarial Attack は**図 8** のように入力画像に微小な摂動を加えることで，摂動を加える前の画像との違いは人間にはほとんど分からないにもかかわらず，画像分類モデルに誤分類を起こさせる攻撃である．Adversarial Attack に

表 1　生成画像を用いた学習

-	置き換えなし	10%	20%	30%	40%	50%
Top-1 Accuracy(%)	92.476	92.364	92.348	**92.560**	92.346	92.222

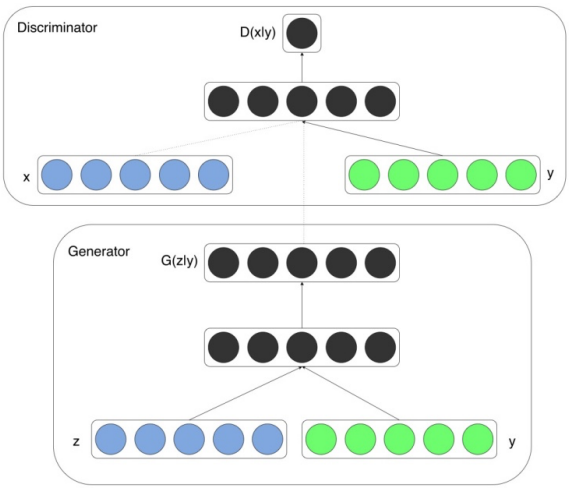


図 7　CGAN のネットワーク構成
Mirza, M. and Osindero, S. (2014) [8] の p.3 の図 1 より転載

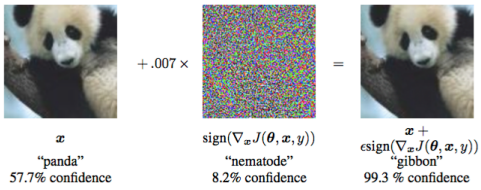


図 8　Adversarial Attack の例
Goodfellow, I. J. et al. (2014) [12] の p.3 の図 1 より転載

は Targeted attack [9–11] と Non-targeted attack [12–14] の 2 つがある．Targeted Attack は画像分類モデルの出力を特定のクラスへと変化させる攻撃であり，Non-targeted attack は入力画像の正解クラスの出力を下げる攻撃である．どちらの攻撃手法も 1 つのクラスの出力にのみ注目しており，提案手法のような画像分類モデルの出力全体を対象とした画像生成は不可能である．

4.3　Mnemonics Training

提案手法と同じように重み更新によって画像を変化させていく手法に Mnemonics Training [15] がある．Mnemonics Training は Continual Learning（Lifelong Learning）[16,17] における破滅的忘却 [18] を回避する手法である．Continual Learning では，モデルは複数のデータセットを順番に学習していく．このとき，あるデータセットを学習した後では，その以前に学習したデータセットに対する精度が著しく低下してしまうという現象が起こる．この現象が破滅的忘却である．

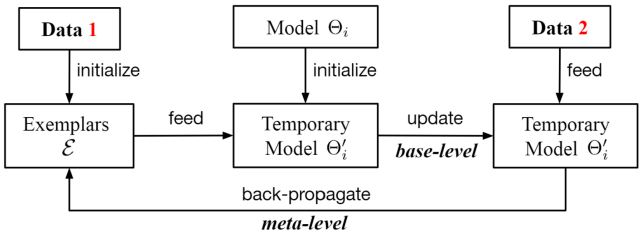


図 9　Mnemonics Training
Liu, Y. et al. (2020) [15] の p.4 の図 3(b) より転載

Mnemonics Training では，この破滅的忘却を回避するために学習中に各データセットごとに Exemplars と呼ばれる，そのデータセットを代表するような画像を複数枚保存する．そして，新しいデータセットを学習する際に Exemplars を使って以前のデータセットを再学習することで破滅的忘却を回避する．Mnemonics Training の特徴として，今回の提案手法のように重み更新によって Exemplars を変化させていくことが挙げられる．具体的には**図 9** の流れの中で以下の式を用いて Exemplars を更新していく．

$$\varepsilon_i \leftarrow \varepsilon_i - \beta_1 \nabla_\varepsilon \mathcal{L}_c(\Theta_i'(\varepsilon_i); D_i) \tag{13}$$

ここで，ε_i は i 番目のデータセットの Exemplars，β_1 は学習率，$\Theta_i'(\varepsilon_i)$ は i 番目のデータセットを学習したモデルを ε_i で再学習した後のパラメータ，D_i は i 番目のデータセット，$\mathcal{L}_c$ は交差エントロピー誤差を求める関数である．

提案手法との違いとしては，Loss の計算方法の違いがある．Mnemonics Training では，D_i に対する Loss を求めており，その Loss が小さくなるように ε_i を更新している．そのため，Mnemonics Training では ε_i の各画像をモデルに入力したときに出力がどのようなものになるのかについては関心がない．一方，提案手法では予め設定した出力との Loss を計算しており，生成画像に対するモデルの出力に主眼を置いている．

5.　おわりに

本稿では，学習済みの画像分類モデルに任意の出力を促す画像生成手法を提案した．提案手法では画像を「重み」として扱うことで誤差逆伝播法による重み更新で画像を改変することを可能にし，従来手法では不可能であった，画像分類モデルの出力全体を任意に改変する画像生成ができる

ことを説明した．また，提案手法による生成画像の活用例として，生成画像を用いた画像分類モデルの学習を実験した．実験の結果，生成画像を用いない学習に比べ，0.084pt画像分類の精度が向上した．

今後の課題として，提案手法による画像生成は誤差逆伝播法による重み更新を用いているため，画像生成に時間がかかってしまうという課題がある．そのため，提案手法による画像生成を高速化する方法を検討する必要がある．また，本稿では提案手法による生成画像の活用先として画像分類モデルの学習を検証したが，その他の分野への活用も検討していきたい．

参考文献

[1] Sundaram, S. and Hulkund, N.: GAN-based Data Augmentation for Chest X-ray Classification, *arXiv preprint arXiv:2107.02970* (2021).
[2] Wickramaratne, S. D. and Mahmud, M.: Conditional-GAN Based Data Augmentation for Deep Learning Task Classifier Improvement Using fNIRS Data, *Frontiers in Big Data*, Vol. 4 (online), DOI: 10.3389/fdata.2021.659146 (2021).
[3] HAN, C., MURAO, K., Shin'ichiSATOH, NAKAYAMA, H.: Learning More with Less: GAN-based Medical Image Augmentation, *Medical Imaging Technology*, Vol. 37, No. 3, pp. 137–142 (online), DOI: 10.11409/mit.37.137 (2019).
[4] Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning representations by back-propagating errors, *nature*, Vol. 323, No. 6088, pp. 533–536 (1986).
[5] Kullback, S. and Leibler, R. A.: On Information and Sufficiency, *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79–86 (online), available from ⟨http://www.jstor.org/stable/2236703⟩ (1951).
[6] Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).
[7] Krizhevsky, A.: Learning multiple layers of features from tiny images, Technical report, University of Toronto (2009).
[8] Mirza, M. and Osindero, S.: Conditional generative adversarial nets, *arXiv preprint arXiv:1411.1784* (2014).
[9] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R.: Intriguing properties of neural networks, *arXiv preprint arXiv:1312.6199* (2013).
[10] Huang, R., Xu, B., Schuurmans, D. and Szepesvári, C.: Learning with a strong adversary, *arXiv preprint arXiv:1511.03034* (2015).
[11] Khrulkov, V. and Oseledets, I.: Art of singular vectors and universal adversarial perturbations, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8562–8570 (2018).
[12] Goodfellow, I. J., Shlens, J. and Szegedy, C.: Explaining and harnessing adversarial examples, *arXiv preprint arXiv:1412.6572* (2014).
[13] Moosavi-Dezfooli, S.-M., Fawzi, A. and Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582 (2016).
[14] Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O. and Frossard, P.: Universal adversarial perturbations, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773 (2017).
[15] Liu, Y., Su, Y., Liu, A.-A., Schiele, B. and Sun, Q.: Mnemonics Training: Multi-Class Incremental Learning Without Forgetting, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
[16] Davidson, G. and Mozer, M. C.: Sequential Mastery of Multiple Visual Tasks: Networks Naturally Learn to Learn and Forget to Forget, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
[17] Parisi, G., Kemker, R., Part, J., Kanan, C. and Wermter, S.: Continual Lifelong Learning with Neural Networks: A Review, *Neural Networks*, Vol. 113, pp. 54–71 (online), DOI: 10.1016/j.neunet.2019.01.012 (2019).
[18] Hassabis, D., Kumaran, D., Summerfield, C. and Botvinick, M.: Neuroscience-Inspired Artificial Intelligence, *Neuron*, Vol. 95, pp. 245–258 (online), DOI: 10.1016/j.neuron.2017.06.011 (2017).