

依頼論文

深層学習は何をどのように“学習”するのか

How and What Deep Learning *Learns*

松原 崇

Abstract

Deep learning models have already achieved significant results and surpassed other sophisticated models based on features made by experts in various tasks; image processing, sound processing, and natural language processing. They are considered to *automatically learn and extract concepts* such as “cat face” and “human body” from given big dataset. However, what are indeed called concepts, and how they are extracted? This manuscript provides a rough history of neural networks preceding deep learning, explanation of concepts learned by deep learning models, and then future perspective of deep learning study with (cognitive) neuroscience.

1 はじめに

深い構造を持った人工ニューラルネットワーク、いわゆる深層学習は、様々な分野において目覚ましい成果を挙げている。画像処理、音声処理、自然言語処理、ロボット制御などなど。深層学習は大量のデータを与えるだけで、問題を解くために必要な概念を自動的に学習し利用するのだと言われている。過熱する「深層学習」ブームにおいては、あたかも大量のデータさえあれば、あらゆる問題が解けるかのような宣伝を目にすることが少なくない。将来についてうかつなことは言えないが、少なくとも現時点において、深層学習がどのように情報あるいは概念の獲得を行っているかについて議論することは重要であると思われる。またそもそも深層学習の文脈で言う「概念」とは何であるのかについても、考えなければならない。本稿はまず人工ニューラルネットワークの概史を述べ、それを基に、深層学習が獲得した「概念」が他の分野においてどのような事物に対応するのかについて、いくつかの仮説を紹介する。また最後に、人工ニューラルネットワークと認知神経科学分野の関わりと、今後の展望について述べる。

2 人工ニューラルネットワークの概史

2.1 パーセプトロン

ニューラルネットワークの歴史は 1900 年頃にニューロン (神経細胞) 説を唱えたスペインの神経解剖学者ラモン・イ・カハールまで遡る [1]. 神経細胞からは軸索及び樹状突起と呼ばれる突起が伸びており、ある神経細胞の軸索の先端にあるシナプスから、別の神経細胞の樹状突起上にあるスパインに向けて、神経伝達物質が放出される [2]. スパイン上の受容体は神経伝達物質を受け取るとイオンチャネルを開き、イオンが流入流出することで、細胞内外の電位差が変化する. こうして生まれたシナプス後電位は、さらに電位依存イオンチャネルを開き、樹状突起上を伝播していく. この信号をスパイクと呼ぶ. 急上昇直後に急降下する波形から、スパイクとも呼ばれる. 細胞体は静電容量が高いため容易には活動電位が伝わらないが、複数の樹状突起から同時に活動電位が与えられれば、活動電位を生じる. 細胞体に生じた活動電位はさらに軸索からシナプスに伝わり、神経伝達物質の放出に繋がる.

この働きを模擬したものが 1943 年のウォーレン・マカロックとウォルター・ピッツの形式ニューロンである [3]. ある神経細胞 i からスパイク $x_i \in \{0, 1\}$ が与えられたとき、シナプス後細胞が受ける影響の強さを結合強度 (または「荷重」「重み」) と呼び w_i で表す. 細胞体がスパイク $y \in \{0, 1\}$ を生じるために必要な電位の閾値を θ で表すと、神経細胞の働きは次のように定式化される.

$$y = \begin{cases} 1 & \text{if } \sum_i w_i x_i = \mathbf{w}^T \mathbf{x} > \theta \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

これをパーセプトロンと呼ぶ. 入力に従い 0 か 1 を出力するので、入力を 2 グループにグループ分けできる.

神経細胞の出力は二値的だが、生理学的ノイズによって決定論的とはいえない挙動を示す. そのため次のような一般化を考える.

$$y = \sigma(\mathbf{w}^T \mathbf{x} + b). \quad (2)$$

ただし $\sigma(\cdot)$ は活性化関数と呼ばれる 0 から 1 の範囲を取る関数であり、 $b = -\theta$ である. このとき y は神経細胞はスパイクを発生させる確率であると解釈できる. また活性化関数 $\sigma(\cdot)$ には、用途に応じて非線形で実数値全体の集合を値域とする関数を用いることもある. このとき、 y は発生したスパイクの数などと解釈することもできる.

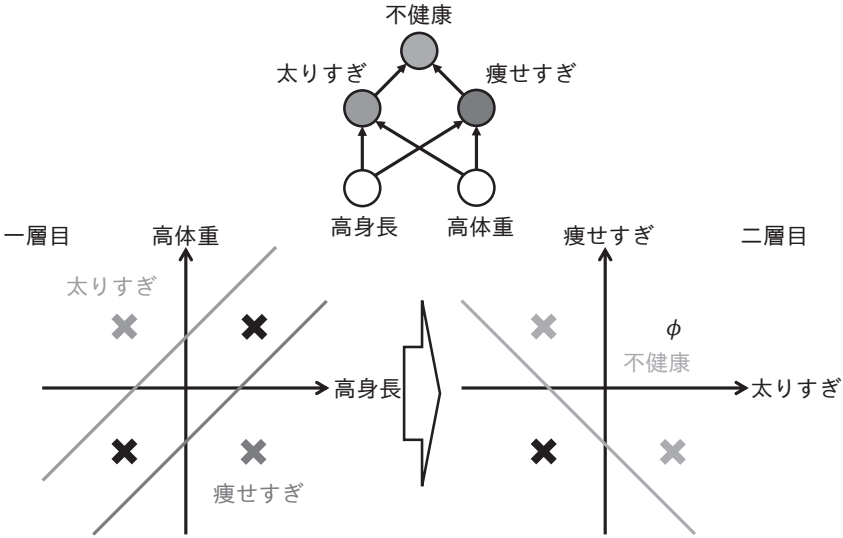


図 1: 身長体重から健康診断するパーセプトロン。

一つの例を考えよう．図 1 は身長体重から健康診断するパーセプトロンである．横軸に身長が低いか高いか ($x_0 \in \{-1, 1\}$)，縦軸に体重が軽いか重いか ($x_1 \in \{-1, 1\}$) を取る．二値を取る変数を考えるとき，このように 0/1 ではなく -1/+1 にしたほうが計算上都合の良いことが多い．身長の高くに太りすぎた人を分類するなら， $\mathbf{w} = (-1, 1)^T$ ， $\theta = 1$ とすると，「身長が低くて体重が重い人」を $y = 1$ のグループに分類でき，それ以外の人 $y = -1$ に分類できる．グループを分ける境界線を決定境界と言い，図中左のグラフの左上の斜線に当たる．身長の高くに痩せすぎた人を分類するなら， $\mathbf{w} = (1, -1)^T$ ， $\theta = 1$ とすると良い．図中左のグラフの右下の斜線に当たる．しかし，太りすぎもしくは痩せすぎた人をまとめて不健康な人として分類したい場合，パーセプトロンでは分類できない．このような問題を XOR 問題と呼ぶ [4]．

そこでパーセプトロンの上にパーセプトロンを載せて多層化したものが開発され，多層パーセプトロンと呼ばれた．以下のように式で表すことができる．

$$\mathbf{h} = \sigma(W\mathbf{x} + b), \quad y = \sigma(\mathbf{v}^T \mathbf{h} + c). \quad (3)$$

重み W は行列であり， $\mathbf{h}$ は中間層あるいは隠れ層と呼ばれる一つ目の演算の出力ベクトルである．先程と同様に h_0 で太り過ぎた人を， h_1 で痩せ過ぎた人を検出し， y が $h_0 \vee h_1$ になるように重み v と c を調整すれば良い．図中右の

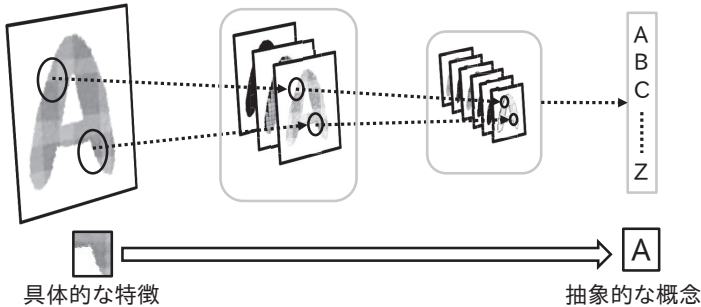


図 2: ネオコグニトロンと畳み込みニューラルネットワークの模式図。

グラフの斜線がそれに当たる．このように，部分的な特徴や分類を入力に近い層で抽出し，出力に近い層で統合して高次の特徴とするのが，多層パーセプトロンの性質である．多層パーセプトロンは特に前向きニューラルネットワークとも呼ばれ，確率モデルとして解釈した場合，データ x が与えられた場合にその分類や特徴量 y の条件付き確率 $p(y|x)$ のモデル化だと考えられる．

余談だが，歴史的に図 1 上部のように積和演算を二回行うものを，入力も含めて三層パーセプトロンと呼ぶが，一回だけ行うものを単層パーセプトロンと呼ぶという不整合がある．今日的には中間層の数（≒積和演算の数）を数えるため，図 1 上部は二層のニューラルネットワークである．

2.2 ネオコグニトロン

脳の中でもほぼ後頭部に位置する視覚野という部位は，その名の通り眼（網膜）で受け取った視覚情報を処理していると考えられている [5]．視神経や第一視覚野は空間的な局所性があり，網膜に写った一部分にしか反応しないことがわかっていた．またデイヴィッド・ヒューベルらは猫の第一視覚野の神経細胞の反応を測定し，特定の方向の直線といった具体的な記号に選択的に反応する神経細胞があることを突き止めた [6]．この仕組みをモデル化した多層パーセプトロンが 1980 年頃に発表されたネオコグニトロンである [7] (図 2 参照)．ネオコグニトロンは無数の小さなパーセプトロンからできている．それぞれの小さなパーセプトロンは同じ重みパラメータを持っており，画像の中のごく一部にしか反応しない．反応する場所はパーセプトロンごとに様々であり，空間のあちこちから同じ種類の具体的な特徴が得られる（例えば直線の向きや黒点の有無）．抽出された特徴は空間中に並べられ特徴マップを作る．こ

の特徴マップから更に特徴を抽出すると、今度は先程得られた具体的な特徴の組み合わせであり、もう少し広い範囲に該当する少し抽象的な特徴が得られる（例えばハネや払いの有無）。これを繰り返すことで、最終的には像全体に対応した抽象的な概念を得ることが期待できる（例えば文字の種類）。通常のパーセプトロンだと画像全体を一度に処理しようとするが、ネオコグニトロンは画像の一部しか計算しないために計算コストが低く、重みパラメータを共有することで学習が速くメモリのにも効率的である。なにより、各パーセプトロンが重みパラメータを共有しているおかげで、画像が少し歪んでいたり横にずれていても、同じような特徴が抽出できるため、画像の変形（例えば人による筆跡の違いなど）に対してとても堅牢である。

2.3 連想記憶モデルと誤差逆伝播法

パーセプトロンやネオコグニトロンは適切な重みが与えれば、様々な情報を処理することができる。しかしどうすれば適切な重みを獲得（つまり学習）することが出来るのだろうか。生物の神経システムの学習については、「パブロフの犬」の名称で知られる実験が有名である（図3を参照）。まず刺激（例えばベルの音）を与え、その後に報酬（例えば餌）を与える。これは2つの信号（ベルの音と餌）に相関性を与えることを意味する。すると犬は刺激が与えられると報酬を予想するように学習する。本来報酬に対する反応（例えば唾を出す）であるはずのものが、刺激に対する反応になることで、学習されたことが確認できる。刺激に対応した神経活動と、報酬に対応した神経活動を同時に励起させることで、両者の関係性が学習されたと考えられる。ドナルド・ヘップは1949年にこのメカニズムをモデル化し、ヘップの法則を提案した。

1970年代には電気生理学的な実験でも確認された[8]。そしてヘップの法則によってネットワークに相関性を覚えさせ、限られた情報から残りの情報を想起・復元できるニューラルネットワークである、連想記憶モデルが提案された[9]。ヘップの法則を数式で書くなら

$$\Delta W_{0,1} = x_0 x_1 \quad (4)$$

となる。ここで、 x_0 と x_1 は刺激と報酬に対応するニューロン、 $W_{0,1}$ はその間の重み、 $\Delta W_{0,1}$ は重みの変化量である。刺激と報酬が同時に与えられる（つまり $x_0 = 1$ かつ $x_1 = 1$ ）とき、重みが増える（ $\Delta W_{0,1} = 1$ ）。ネットワークの学習が終わると、「ベルの音」「餌の発見」「唾を出す」という3つの要素に対応した神経細胞は、正の相関を表す正の重みで結合するようになる（それ

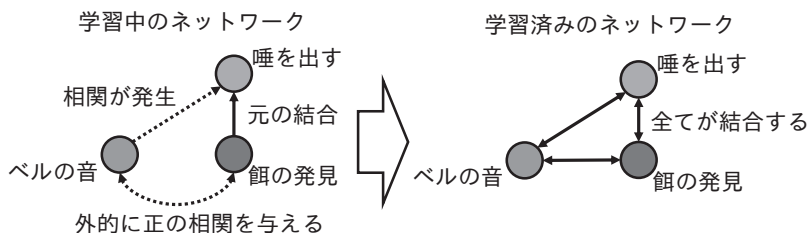


図 3: パブロフの犬を連想記憶型ネットワークに実装した模式図。

それが結合の重みが 1、閾値が 0.5 のパーセプトロンであると考え)。すると、ベルの音を聞いた犬は餌を予想するだけでなく、直接的に唾を出し始める。「唾を出す」「唾を出さない」という 2 パターンの情報は、「両方ある」と「両方ない」という情報にそれぞれ対応したものであり、4 パターンの情報を 2 パターンに圧縮している。これを応用すると、連想記憶型ネットワークは画像の修復などを行うことが出来る。大量の人の顔の画像を与えて、画素ごとの相関性を学習させたあと、新しい顔画像の右目を隠して与える。ネットワークは学習によって「右目は左目が左右反転したもの」と記憶しているため、左目の情報をもとに、隠された右目を予測することができる。ちょうど「ベルの音」から「餌の発見」を予想するのと同じである。

連想記憶モデルは、自己組織化マップ (self-organizing maps; SOM)、ホップフィールドネットワーク (Hopfield network)、そしてボルツマンマシン (Boltzmann machine) へと発展していく [10]。これらのモデルは与えられた情報の同時分布 $p(x, y)$ のモデル化と考えられ、双方向型のニューラルネットワークである。また一度学習が終われば、前向きニューラルネットワークとして使うこともできる。

連想記憶モデルで画像を学習するには全てのピクセルの間で相関性を学習する必要がある、組み合わせは膨大である。また非線形あるいは 3 つ以上の要素が絡むような複雑な相関性を学習するには隠れ層が必要であり、組み合わせが増えるだけでなく、隠れ層をどのように学習すれば良いのかは自明ではなかった。そこで代わりに、多層パーセプトロンの教師あり学習法として導入されたのが、デビッド・ラメルハートらにより 1986 年に提案された誤差逆伝播法である [11]。ただし誤差逆伝播法と類似した考え方は古くから何度も再発見されている。目的値 (教師) と現在の出力との誤差を微分し、どのように重みパラメータを変化させれば誤差が小さくなるかを計算し、それに従って重みパラメータを変化させていく方法である。

2.4 深層学習へ

ニューラルネットワークを用いれば様々なデータを解析できることは経験的に知られていたが、1990 年前後に、少なくとも中間層がひとつあるニューラルネットワークは、任意の有界で連続な関数を近似できることが証明された [12]。しかし実際に任意の関数を近似しようとする、と、入力のみ組み合わせ爆発が起これ、入力次元数の指数オーダーというとても数多くのニューロンが中間層に必要な。一方で、もし中間層の層数が増えていけば、組み合わせの組み合わせを作ることができるため、中間層のユニット数及びそれらの結合パラメータを大幅に削減することができる [13, 14]。そのためニューラルネットワークをいかに深くするか、という研究が注目をあびるようになった。しかしニューラルネットワークの深層化を阻んだのが誤差逆伝播法における勾配消失問題である [15]。入力に近い層の勾配を計算すると、出力に近い層の勾配の積になる。もしここでゼロに近い値が一つでも紛れ込んでしまうと、入力に近い層の勾配がゼロに近くなり、学習が進まなくなってしまう。そしてパーセプトロンに用いる活性化関数 $\sigma(\cdot)$ は一般に単調増加かつ有界なので、十分大きいもしくは小さい入力値に対して、勾配が消失するため、深いニューラルネットワークを学習できない。結局のところ、ニューラルネットワークを現実的な規模の問題を解けるほど大規模かあるいは汎用化することができず、「ニューラルネットワーク冬の時代」を招くことになる。

そのため誤差逆伝播法を諦め、部分ごとに小さな連想記憶モデルとみなし学習することが積極的に行われた [15]。また誤差逆伝播法を使う時も、ある層で計算を行うか素通しするかを層自身が決定し、勾配の消失を防ぐ機構も考案された [16]。ネオコグニトロンはネットワークを使い回しているため、一部で勾配が消失しても、残りの部分で学習が進むため、勾配消失に強かった。これを畳み込みニューラルネットワーク (convolutional neural network) という [17]。他にも再帰ニューラルネットワーク (recursive neural network) と呼ばれる、文章のような木構造を持ったデータに対して、親ノードから子ノードへの接続に同じネットワークを使いまわす方法も提案された [18]。またネットワークを学習することを諦め、ランダムな重みを使った層を学習しない基底とし、そこから必要なものを学習によって抽出するというアプローチも考えられた [19, 20]。とはいえ、どの手法にも何らかの限界や問題点があった。

そんな中で成功を収めたものが「事前学習」と「ファインチューニング (微調整)」という方法である [21]。まず 1000 ニューロンの入力があったとすると、連想記憶モデルを使ってそれを 500 ニューロンに圧縮する (「ベルの音」と「餌の発見」という 2 ニューロンと「唾を出す」という 1 ニューロンに圧

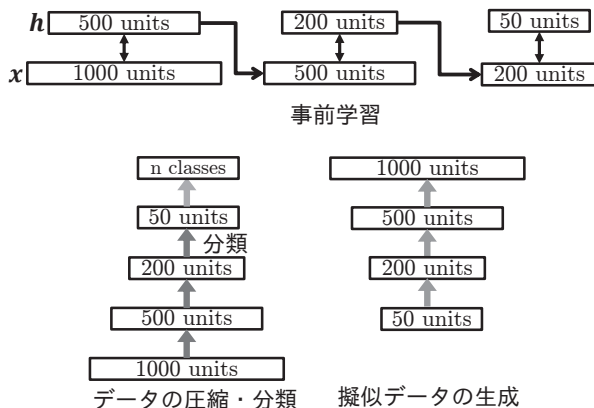


図 4: 自己符号化器と事前学習と出来上がったパーツの使い方.

縮したのと同じである). さらにその 500 次元を 200 次元に圧縮し, さらに 50 次元に, と繰り返して連想記憶モデルによってデータを圧縮する (図 4 上). こうして作った複数の連想記憶モデルを縦に繋げてやる. すると 1000 次元とデータを徐々に 50 次元に圧縮するネットワークと, 逆に 50 次元のデータを徐々に 1000 次元に復元するようなネットワークが得られる (図 4 下). このようなニューラルネットワークを自己符号化器と呼ぶ. 前半のデータを圧縮する部分を使えば, データをグループ分けしたり, さらにパーセプトロンと組み合わせて分類することができる (図 4 左下). 後半の復元する部分は擬似データを作る事にも使える (図 4 右下). どちらも深いニューラルネットワークとしての力をある程度発揮できる. またこのニューラルネットワークはある程度きれいな情報伝達ができている (=事前学習) ため, 勾配消失に悩まされずに, 誤差逆伝播法によってさらなる学習をすることができる (=ファインチューニング). 事前学習だけでは, 例えばある複数の組み合わせでのみ意味を成す情報などが, 最初の連想記憶モデルで取りこぼされてしまい, 一番圧縮された層まで伝達しないため, 深いニューラルネットワークの能力を最大限に活かされていない. ファインチューニングを行うことで, 深いニューラルネットワークを能力を最大限発揮できるようになった.

データを圧縮する方法であるという点において, 自己符号化器は主成分分析や独立成分分析にとっても良く似ている. 主成分分析は線形基底を用いてデータを圧縮するが, 独立成分分析はデータ内の分散共分散あるいは非線形相関も考慮に入れ, 自己符号化器はニューラルネットワークを基底に用いるとい

う点異なる。

なお、2014 年頃から深いニューラルネットワークを一度に学習させる方法が発展したため、教師あり学習のために事前学習が用いられることはもはや多くない [22, 23]。

3 表現学習

3.1 神経科学と表象

「表象 (representation)」という語はそもそも哲学分野における用語であるそうだが、異なる物について「表現する」あるいは「代わりをする」ものを指すと聞く。認知科学の分野においては、外界からの情報を符号化し処理するシステムを表象の語を用いて呼ぶ [24]。例えば新生児は人の顔を他と区別し興味を示すような行動を取る。人の顔のパーツをバラバラにした画像や、周波数スペクトル等が人の顔の画像と一致した擬似画像にはそのような行動をあまり示さない。このことは人間の脳が生後間もなく人間の顔を区別するシステムを獲得することを示唆している。この脳活動こそが人の獲得した表象であると考えることができる。同じようにニューラルネットワークが入力を符号化したものも “representation” と呼ばれてきた [11, 25, 18]。

3.2 グーグルの猫と表現学習

しかし、視覚的にも明確な表象の構築が確認できたのは、2000 年代も後半になってからである [26]。ボルツマンマシンによって事前学習した深いニューラルネットワークにおいて、多くの画像が共通して持つような特徴、つまり黒丸や斜線と行ったエッジを検出するようなニューロンが自動的に構築されたのである。さらに第二層においては、第一層で得られた特徴を組み合わせ、空間的な位置が異なっても同種のエッジに反応するようなニューロンが確認できた。このような特徴は電気生理学的な実験で確認された、第一視覚野と第二視覚野の関係に酷似していた [27]。それだけではない。続く研究では畳み込みニューラルネットワークが用いられた [28]。2.2 節で述べた通り、畳み込みニューラルネットワークはその構造上、入力に近い層であれば近傍の狭い範囲からのみ情報を受け取るため、局所的な情報の成分に対応する。出力に近い層であれば、ほぼ画像全域の情報の成分に対応する。顔画像に限定したデータセットを与えると、画像 x に近い層では黒丸や斜線のような具体的

な特徴が、さらに中間の層では目や鼻や口に見える特徴が、 y に近い層ではほぼ顔そのものが特徴が得られるという、三段階以上もの抽象化が確認できた。このように高次の抽象的な特徴が低次の具体的な特徴の組み合わせでできていること、そしてその関係が学習によって自動的に獲得できることが示された。これは 2.2 節で述べたネオコグニトロンの仮説の実証でもある。

自己符号化器の形に組んだ畳み込みニューラルネットワークを学習したときに確認されたのが、いわゆる「グーグルの猫」である [29]。この実験は動画投稿サイト YouTube に投稿された 1000 万本の動画から画像を取り出したものをデータセットに用いた。YouTube の動画は一般人などが投稿した物が多く、ほとんどは人に関するものや、犬猫のような身近な動物に関するものである。そこで、動画から大量に人の顔が写ったシーンを集め、顔画像が入力されると、大きな値を出力するが、それ以外の画像には小さな値しか出力しないニューロンを見つけた。そしてまたそのニューロンが最も反応するような画像を、誤差逆伝播法を用いて作ると、ぼやっとした人の顔が浮かび上がった。こうして見つかったニューロンは「人の顔という概念を学習したニューロン」と考えられた。同じようにネットワークは「猫の顔という概念」や「人の体という概念」も学習していることが確認され、大きなセンセーションを巻き起こした。

これらの経緯から深層学習とは表現学習 (representation learning) である、と明確に主張されるようになった [30]¹。しかしこの表現とは、より具体的に何であろうか。

3.3 多様体学習としての表現学習

ヨシュア・ベンジオらはレビュー論文 [30] において、表現学習とは多様体学習であると述べている。多様体とは幾何学の用語である。

世界は 3 次元（以上）であるが、私たちは基本的に地球の上を歩いている。ビルに登っても、飛行機に乗っても、地球の大きさから考えれば誤差のようなものである。私たちの位置を寄せ集めれば、そこには緯度と経度で表せる 2 次元の空間が広がる。このような空間を多様体という。一般に多様体は曲がっている。日本からアメリカまで移動するとき、地表に沿った動きをすることになる (図 5 右の車を参照)。もし 3 次元空間中でまっすぐ歩いてしまっ

¹representation の訳として表象ではなく表現が定着しているところに、(少なくとも日本における) 機械学習分野と認知神経科学分野の断絶を感じることができるかもしれない。

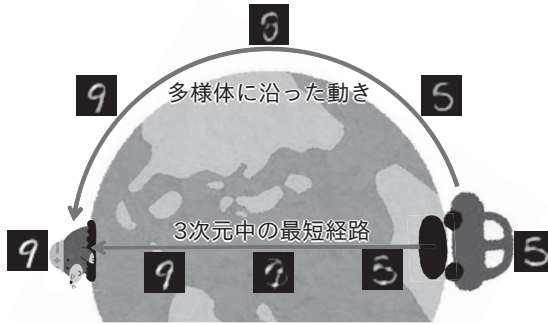


図 5: 多様体のイメージ図.

たら、地殻を突き抜けマントルを経由することになってしまう (図 5 左のもぐらを参照).

ここで MNIST[31] と呼ばれる手書き数字データセットを例にして考えてみる (図 5). これは縦横ともに 28 ピクセルのモノクロ画像であり、0~9 の数字がひとつ書いてある. それぞれのピクセルを白と黒の二値で表してみると、ピクセルが $28 \times 28 = 784$ 個あるので、 2^{784} 通りの画像が存在する. しかし 2^{784} 通りもの手書き数字があるわけではない. このサイズの画像であれば、例えばアルファベットや漢字を書くことも可能だろうし、簡単な絵を書いたり、ただのノイズのような画像も考えられる. 実際に手書き数字とみなせる組み合わせは、画像が持ちうる 2^{784} 通りよりも遥かに少ない. つまり全ての手書き数字は、全ての画像を含む 784 次元空間の部分集合であり、より低次元の多様体の中にあると考えられる.

しかし緯度と経度で表現できる地球上の座標と違って、手書き数字を表現できる変数は自明ではない. MNIST から適当な 2 枚の画像を選び図 5 の右と左に表示する. この 2 枚の画像を画像空間中で少しずつ変化させたものを、下部に 3 枚並べた. その 3 枚は確かに 784 次元空間中に存在しているが、手書き文字の画像とは言い難い. これは 2 枚の画像の間を「まっすぐ」移動したために、手書き数字の多様体から逸脱してしまったからである [32]. 日本からアメリカまでをまっすぐ歩いたらマントルに突っ込んでしまった状態に相当する.

そこで自己符号化器 (の一種) 使って、左右の画像から MNIST から表現を抽出してみる [33]. 符号化器は 784 次元の画像を受け取り、より少ない次元 (例えば 10 次元) の表現を出力する. 次に 2 枚の画像から得られた表現を、表

現空間でまっすぐ変化させる．こうして得られた表現を復号化器に入れ、疑似画像を生成し、図5の上部に3枚並べた．この3枚は右から、5、8、9の数字に見える．復号化器は10次元分の表現しか与えられていないので、生成する画像は784次元のデータ空間全てを網羅できず、学習した手書き数字の多様体中の画像に限定される．表現空間中の座標は多様体上の座標であり、それこそが表現であると考えられる．すると生成した画像は、常に手書き数字として成立するような見た目になる．このように多様体を学習することが自己符号化器の特徴であり、またそれによってデータから意味のある情報を取り出せると考えられる．さらに言えば、表現の各次元には独立性が仮定されることがあり、すると各次元が「長さ」「太さ」「傾き」といった人間に解釈可能な何らかの意味をもっているように見えることも多い．

3.4 特徴抽出としての表現学習

もっと単純な解釈も考えられる．表現は自己符号化器やボルツマンマシンのような教師なし学習だけではなく、教師あり学習でも確認できる．2.1節で述べた「不健康な人」を分類する多層パーセプトロンでは、中間層に「身長が低くて体重が重い人」という表現を明示的に作っている．しかし、「不健康な人」を分類したいという目的さえ与えて誤差逆伝播法で学習すれば、それが達せられるように前半のニューラルネットワークが自動的に、必要な表現を構築していくと考えられる．前半のニューラルネットワークは特徴抽出器であり、出力は特徴量である．その特徴量を後半のニューラルネットワークが分類する．

データセットが大規模で様々なデータを含んでおり、ニューラルネットワークを様々な目的で訓練した場合、訓練済みのニューラルネットワークはさまざまな特徴量を出力することが期待される．この特徴量はHOG特徴量[34]のような従来画像認識に使われてきた特徴量と同じように使うことができる．そのため、後半のニューラルネットワークを入れ替えてしまえば、異なるタスクに流用することができる．これを特に転移学習という[35]．これは流用先(転移先)のデータセットが小規模な場合に有効である．小規模なデータセットが対象では、ニューラルネットワークがデータセットをまる覚えしてしまい、汎用的な特徴量を抽出することできず、意味のない分類をしてしまう．後半のニューラルネットワークは、ニューラルネットワークではなくサポートベクターマシンやトピックモデルでも良い．とはいえ、ニューラルネットワークがどのような特徴量を学習するのかは初期値やアルゴリズムに依存

する．例えば図 2.1 では「身長が低く体重も軽くて健康な人」と「身長が高く体重も重くて健康な人」という表現を獲得しても，最終的な分類が成功する．

明示的に表現を構築する方法としてはマルチタスク学習がある [30]．これは例えば画像を与えた場合に，同時に「物体の分類」「物体の領域分け」「物体の境界線の抽出」「顕著性の検出」といった似ているが異なるタスクを同時に行うことで，それぞれの精度が向上する．

多様体学習と違い，この解釈では表現が与えられたデータ全ての説明できない [36]．「不健康な人」の表現は「太りすぎ」と「痩せすぎ」を合わせたものなので，そのどちらであるかという情報は欠落している．ある「犬の画像」を入力とすると，ニューラルネットワークはその画像を「犬」という抽象的な概念の表現に対応させる．しかし「犬」の表現は特定の犬の画像とは対応せず，あらゆる犬の画像の集合に対応する．いわば不可逆圧縮になっている．

3.5 生成モデルとしての表現学習

このような関係は生成モデルを用いることで説明できる．この解釈では自己符号化器のうち図 4 右下にある復号化器が主役になる．生成モデルは確率モデルの一種であり，自然言語分野における情報抽出において活躍してきた．

例えば人物画を考えてみる．様々な画家が「座っている」「中年の」「男性を」想像で描こうと決める．モチーフを統一しても，完成する絵は多種多様である．逆に観客は，見た絵画からその絵画が描かれた背景やモチーフを推測する．結果的に画家や流派，モチーフなどで絵画を分類する．このようにあるデータがモチーフやテーマ（トピック）に基づいて作成されていると考える生成モデルをトピックモデルという [37]．

自己符号化器の符号化器は絵画からトピックを推論する部分に当たり，復号化器はトピックから絵画を生成する部分に当たる．このときトピックは表現に当たり，潜在変数と呼ばれ，データの背後にある直接観測できない変数と考えられる．

なお，通常の自己符号化器の復号化器は決定論的であり，一対一の関係しかモデル化できない．そのため変分自己符号化器 [33] や，深層ボルツマンマシン，深層暗黙的生成モデル [38, 39] など，一対多の関係を扱えるモデルを使う必要がある [36]．このとき不可逆圧縮で失われた情報は，分布からのサンプルとして確率的に補完される．

3.6 どのように学習するのか

本章ではニューラルネットワークが学習したものの解釈についていくつか紹介した。しかし、ニューラルネットワークがそれらの機能を有するように、どのように自動的に学習するのかについてはいまだ議論がまとまっていない。

ニューラルネットワークの奇妙な特徴の一つが汎化性能の高さである [40]。つまり、なぜかニューラルネットワークは、未知のデータに対しても非常に高い精度で分類や情報抽出が可能である。今日的なニューラルネットワークは、内部に 100 万から 10 億のパラメータを持っている [23]。一方で、学習に使う訓練データはベンチマークの識別タスクで 5 万枚、多くて 100 万枚である。一般に説明変数と被説明変数が一つずつの場合、 n 個のデータ対があれば、 $(n-1)$ 次の多項式で全てのデータ対を説明できる。実際のデータにはノイズがあるため、全てのデータ対を完全に説明するようなモデルは、無意味なノイズにまで無理矢理学習してしまっていると考えられ、未知のデータを説明できない。しかし、ニューラルネットワークはデータ対の 100 倍あるいは 1 万倍のパラメータがありながら、過学習せずに、未知のデータに対しても高い識別精度が出せる。一番簡単な解釈は、ニューラルネットワークは多項式ほど柔軟ではなく、そもそも過学習しにくいのだと考えることだ。しかし試しに訓練データの正解ラベルをシャッフルし、意味のないデータ対を作ってみても、ニューラルネットワークはそれらのデータを 100% 正しく識別できる。当然だが検証データは正しく識別できない。つまりニューラルネットワークは過学習する能力をすでに持っているのにも関わらず、過学習しないのである。誤差逆伝播法ではデータ点ごとの勾配を平均化して学習させるため、各データが持つ共通部分のようなものを優先的に学習し、それが得てして人間が欲している情報と一致するため、過学習する前に識別に必要な学習が進むのかもしれない [40]。しかし明確な説明はまだなされていない。

もう少し外部的なレベルであれば、目的となる表現を獲得出来るような拘束条件やネットワーク構造を作ることが可能である。例えば畳み込みニューラルネットワークは空間的な制約を持っており、画像に近い層でいきなり大局的な特徴が抽出されることはない。自己符号化器を用いた生成モデルであれば、複数ネットワークを予想される要素間の従属関係や独立関係に従って組み合わせることで、データ構造を明示的に与えることができる [41]。例えば筆者らの研究では、医療用画像が「疾患特徴」「個人特徴」「環境特徴」そして「ノイズ」という各要素から成り立っているという事前知識を自己符号化器に与え、画像から疾患を診断する研究を実施している [42]。同一人物の複数の画像間には「個人特徴」を共有し、同一疾患の患者間では「疾患特徴」を共有

していることになる。そこから疾患の特徴が何であるかを符号化することで、疾患の有無や疾患に関係している部位をより高い精度で特定することができた。また経済ニュースからそのニュースが書かれたイベントとその好悪を推定することで、株価の騰落を予測する研究も実施している [43]。このような構造的・階層的なニューラルネットワークは、内部で「何を」モデル化し抽出しているかが比較的分かりやすく、性能や実用性が高い。しかし「どのように」モデル化しているかの部分についての説明には必ずしも貢献できない。

4 神経科学と人工知能は再び邂逅するか

人工ニューラルネットワークは神経細胞の数理モデルから始まり、連想記憶型は神経可塑性であるヘップの法則から、畳み込みニューラルネットワークは視覚野の構造からヒントを得ている。しかし畳み込みニューラルネットワークの生みの親であり、今日も深層学習における巨人の一人に数えられるヤン・ルカン¹は、神経科学的な知見を深層学習に持ち込むことには否定的である。この立場から語られる決まり文句として、鳥と飛行機の喩えがある [44]。空を飛ぼうとした人々はまず鳥を真似て、翼を付けて揚力を得、羽ばたくことで推進力を得ようとした。しかしライト兄弟から今日に至るまで、飛行機は羽ばたくことはせず、翼による揚力とプロペラやジェットエンジンによる推進力を別々に持つことで、飛行している。これは自然界にあるものが最適解であるという素朴な考えに対する警鐘である。連想記憶モデルの学習は、誤差逆伝播法に比べて時間がかかりすぎ、有用な状況はかなり限られている。畳み込みニューラルネットワークでは、画像の変形に堅牢にするために視覚野に存在する複雑型細胞 (complex cell) の機能を模したマックスプーリング (max-pooling) という処理が行われていた。しかし現在では、あまり性能に寄与しないと考えられ、用いられない事が多い [23]。畳み込みニューラルネットワークは音声処理の分野でも成果を上げており、そうすると視覚野の構造を模している意味がよく分からなくなる [45]。

一方、人工ニューラルネットワークにおけるもう一人の巨人ヨシュア・ベンジオ²は、深層学習で得られた知見を計算論的神経科学の分野に持ち込むこと、あるいはその逆について熱心に研究している [46]。例えば、ある種のヘップの法則 (スパイク発火時刻依存可塑性) が誤差逆伝播法に近い働きをすることを理論とシミュレーションで確認している [47]。認知科学の分野において、変分ベイズ法を認知メカニズムの数理モデルとした自由エネルギー原理 (free-energy principle) [48] で知られるカール・ Friston³が、近年モデルの

階層化・深層化に取り組んでいることは深層学習の発展と無関係ではあるまい [49].

アルファ碁 [50] を作り上げた DeepMind の創業者デミス・ハサビスは認知神経科学の学位を有しており、海馬が記憶だけでなく未来を想像することにも関係するという展望的記憶に関する論文を著している [51]. DeepMind は今日に至るまで認知科学および神経科学に関する研究に熱心である [52]. 例えば、誤差逆伝播法や連想記憶モデルの学習は、前向きのネットワークと同じ強度で結合した逆向きのネットワークの存在を仮定する. 電気生理学的な実験で部分的に確かめられているものの、やや強すぎる仮定である [53]. しかし近年、逆向きネットワークは結合が同じ強度でなくとも、あるいは構造が異なっても学習できることが証明された [54, 55]. このことは人工ニューラルネットワークに、より計算量やメモリ消費量の点で効率的に学習アルゴリズムを提供する. また、それに留まらず、一般に逆向きネットワークの存在が期待できない生物の神経系において、誤差逆伝播法のようなアルゴリズムの存在を期待させる. 他にも例えば、人間が物を判別したいときは、そちらに意識を向けて注視する. このように画像を部分的に見る機構をニューラルネットワークに組み込むことで、物体の検出や認識の精度を高めようとも試みている [56].

ほかにも強化学習の世界では、今日でも人間の認知システムに触発された研究が少なくない. 例えばある研究では人の失敗から学ぶ能力をモデル化した [57]. ビリヤードで球を弾いて左にそれたとき、単に失敗したと咎めるのではなく、もっと球が右にあれば今の弾き方でよかったと考えることができる. そのような失敗からも学ぶ機構を組み込むことで、強化学習の学習速度や純粋な性能を高めることができた. ただ、これら人間の認知システムに触発された研究は、あくまで説明のためのアナロジーである場合もあり、実際に研究者らがどれだけその説明を信じているかは分からない.

このように、個々の神経細胞の学習から生物の認知システム全体までのあらゆるレベルで、生物の情報処理メカニズムの機械学習的解釈や、逆に生物システムの機械学習応用は、今日も続けられている.

5 まとめ

人工ニューラルネットワークは自動的に「概念」を学習する. それは認知神経科学において「表象」と呼ばれるものであり、データが持つ形状から考えれば「多様体 (中の座標)」とそこへの「写像」であり、画像処理の分野では

「特徴量」とその「抽出器」であり、確率モデルでは「潜在変数」と「生成モデル」である。これらは与えられたタスクを解く上で必要な要素であり、多くの場合誤差逆伝播法によって学習される。しかしどのように人間が解釈可能にも思える「概念」が学習されていくのかについてはまだまだ明らかではない。とはいえ、ネットワーク構造を工夫したり拘束条件を課すことで、人間が解釈可能な「概念」を明示的に与えることはできる。

人工ニューラルネットワークの研究は神経科学とともに始まったが、その発展は必ずしも足並みを揃えたものではなかった。深層学習の登場以降、解析手法、学習アルゴリズム、あるいは解釈といった様々な面で共通の問題が見つかりつつある。もし人工知能がすでに翼を得ているというなら、鳥などお手本とせずプロペラの開発を急ぐべきである。しかしもしまだ人工知能が翼すら見つけていないなら、神経科学や認知科学はその発展の重要な鍵となっていくだろう。

文献

- [1] A. Nemri, “Santiago Ramón y Cajal,” *Scholarpedia*, vol. 5, no. 12, p. 8577, 2010.
- [2] R. Llinas, “Neuron,” *Scholarpedia*, vol. 3, no. 8, p. 1490, 2008.
- [3] W. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.
- [4] M. Minsky and S. A. Papert, *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, 1969.
- [5] N. Swindale, “Visual map,” *Scholarpedia*, vol. 3, no. 6, p. 4607, 2008.
- [6] D. H. Hubel and T. N. Wiesel, “Receptive fields of single neurones in the cat’s striate cortex,” *The Journal of Physiology*, vol. 148, no. 3, pp. 574–91, 1959.
- [7] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.
- [8] H. Shouval, “Models of synaptic plasticity,” *Scholarpedia*, vol. 2, no. 7, p. 1605, 2007.
- [9] K. Nakano, “Associatron - A model of associative memory,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 2, no. 3, pp. 380–388, 1972.
- [10] G. E. Hinton, “Boltzmann machine,” *Scholarpedia*, vol. 2, no. 5, p. 1668, 2007.

- [11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [12] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Mathematics of control, signals and systems*, pp. 303–314, 1989.
- [13] J. Håstad and M. Goldmann, “On the power of small-depth threshold circuits,” *Computational Complexity*, vol. 1, no. 2, pp. 113–129, 1991.
- [14] R. Eldan and O. Shamir, “The Power of Depth for Feedforward Neural Networks,” *Conference on Learning Theory (COLT)*, pp. 1–34, 2016.
- [15] J. Schmidhuber, “Learning Complex, Extended Sequences Using the Principle of History Compression,” *Neural Computation*, vol. 4, no. 2, pp. 234–242, 1992.
- [16] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [17] Y. LeCun *et al.*, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.
- [18] C. Goller and A. Kuchler, “Learning task-dependent distributed representations by backpropagation through structure,” in *International Conference on Neural Networks (ICNN)*, vol. 1. IEEE, 1996, pp. 347–352.
- [19] H. Jaeger, “The “echo state” approach to analysing and training recurrent neural networks,” *GMD Report*, vol. 148, no. 148, pp. 1–47, 2001.
- [20] G.-B. Huang, Q.-y. Zhu, and C.-k. Siew, “Extreme learning machine: Theory and applications,” *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.
- [21] G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [22] D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *International Conference on Learning Representations (ICLR)*, pp. 1–15, 2015.
- [23] K. He *et al.*, “Deep Residual Learning for Image Recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [24] M. H. Johnson and M. de Haan, *Developmental Cognitive Neuroscience: An Introduction*, 4th ed. Wiley-Blackwell, 2015.
- [25] G. Hinton and J. McClelland, “Learning representations by recirculation,” *Neural information processing ...*, pp. 358–366, 1988.
- [26] H. Lee, C. Ekanadham, and A. Ng, “Sparse deep belief net model for visual area V2,” *Advances In Neural Information Processing Systems (NIPS)*, pp. 1–8, 2008.
- [27] M. Ito, “Representation of Angles Embedded within Contour Stimuli in Area V2 of Macaque Monkeys,” *Journal of Neuroscience*, vol. 24, no. 13, pp. 3313–3324, 2004.
- [28] H. Lee *et al.*, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” in *International Conference on Machine Learning (ICML)*. New York, New York, USA: ACM Press, 2009, pp. 609–616.

- [29] Q. V. Le *et al.*, “Building high-level features using large scale unsupervised learning,” in *International Conference on Machine Learning (ICML)*, 2012, pp. 507–514.
- [30] Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [31] “THE MNIST DATABASE of handwritten digits.” url: <http://yann.lecun.com/exdb/mnist/>
- [32] S. Rifai *et al.*, “The Manifold Tangent Classifier,” *Advances In Neural Information Processing Systems (NIPS)*, pp. 1–9, 2011.
- [33] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *International Conference on Learning Representations (ICLR)*, no. M1, pp. 1–14, 2014.
- [34] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. I, pp. 886–893, 2005.
- [35] J. Yosinski *et al.*, “How transferable are features in deep neural networks ?” *Advances In Neural Information Processing Systems (NIPS)*, vol. 27, pp. 1–9, 2014.
- [36] S. Zhao, J. Song, and S. Ermon, “Learning Hierarchical Features from Generative Models,” in *International Conference on Machine Learning (ICML)*, vol. 70, feb 2017, pp. 4091–4099.
- [37] D. M. Blei *et al.*, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [38] S. Mohamed and B. Lakshminarayanan, “Learning in Implicit Generative Models,” *arXiv*, no. 1, pp. 1–8, 2016.
- [39] D. Tran, R. Ranganath, and D. M. Blei, “Deep and Hierarchical Implicit Models,” 2017.
- [40] C. Zhang *et al.*, “Understanding deep learning requires rethinking generalization,” *International Conference on Learning Representations (ICLR)*, pp. 1–15, nov.
- [41] L. Maaløe *et al.*, “Auxiliary Deep Generative Models,” in *International Conference on Machine Learning (ICML)*, vol. 48, 2015, pp. 1–5.
- [42] T. Tashiro, T. Matsubara, and K. Uehara, “Deep Neural Generative Model for fMRI Image Based Diagnosis of Mental Disorder,” in *International Symposium on Nonlinear Theory and its Applications (NOLTA)*, 2017 (accepted).
- [43] T. Matsubara, R. Akita, and K. Uehara, “Stock Price Prediction by Deep Neural Generative Model of News Articles,” *IEICE Transactions on Information and Systems*, 2018 (accepted).
- [44] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 1995, vol. 9, no. 2.
- [45] A. van den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3437–3440, 2016.

- [46] Y. Bengio, “Towards Biologically Plausible Deep Learning,” *arXiv*, p. 18, 2015.
- [47] Y. Bengio *et al.*, “STDP-Compatible Approximation of Backpropagation in an Energy-Based Model,” *Neural Computation*, vol. 29, no. 3, pp. 555–577, 2017.
- [48] K. J. Friston, J. Kilner, and L. Harrison, “A free energy principle for the brain,” *Journal of Physiology Paris*, vol. 100, no. 1-3, pp. 70–87, 2006.
- [49] K. J. Friston *et al.*, “Deep temporal models and active inference,” *Neuroscience and Biobehavioral Reviews*, vol. 77, no. April, pp. 388–402, 2017.
- [50] D. Silver *et al.*, “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [51] D. Hassabis *et al.*, “Patients with hippocampal amnesia cannot imagine new experiences,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 5, pp. 1726–1731, 2007.
- [52] K. J. Miller, M. M. Botvinick, and C. D. Brody, “Dorsal hippocampus contributes to model-based planning,” *Nature Neuroscience*, no. June, 2017.
- [53] S. Song *et al.*, “Highly nonrandom features of synaptic connectivity in local cortical circuits,” *PLoS Biology*, vol. 3, no. 3, p. e68, 2005.
- [54] T. P. Lillicrap *et al.*, “Random synaptic feedback weights support error backpropagation for deep learning,” *Nature Communications*, vol. 7, p. 13276, 2016.
- [55] A. Noekland, “Direct Feedback Alignment Provides Learning in Deep Neural Networks,” *Advances In Neural Information Processing Systems (NIPS)*, vol. 23, no. 3, pp. 540–552, 2016.
- [56] V. Mnih, N. Heess, and A. Graves, “Recurrent models of visual attention,” *Advances In Neural Information Processing Systems (NIPS)*, pp. 1–9, 2014.
- [57] M. Andrychowicz *et al.*, “Hindsight Experience Replay,” *arXiv*, 2017.

(神戸大学)