認知科学 第 28 巻 第 2 号 (2021) pp. 299–307 https://doi.org/10.11225/cs.2021.012
Cognitive Studies: Bulletin of the Japanese Cognitive Science Society, Vol. 28, No. 2, pp. 299–307

深層学習と人工知能

松尾 豊 *

東京大学

Deep learning and artificial intelligence

Yutaka Matsuo*

The University of Tokyo

This article tries to position deep learning in the intersection of artificial intelligence and cognitive science, as a long quest toward human intelligence. First, the recent development of huge language models obtained by transformer-based methods such as BERT and GPT-3 is introduced. Then, I explain what these models can do and can not do, and why. Two essential problems, which is embodiment and symbol grounding, are shown. In order to solve these problems, deep reinforcement learning with world models are currently studied. Disentanglement is shown to be an important concept to find factors to control. Lastly, I explain my perspective toward the future advancement, and conclude the paper.

Keywords: deep learning（深層学習），artificial intelligence（人工知能），language model（言語モデル），world models（世界モデル），disentanglement（もつれを解くこと）

Received 16 February 2021

1. はじめに

人工知能は，人間の知能の仕組みを構成論的に解き明かそうとする学問分野である．研究者によってはさまざまに異なる人工知能の定義があるが (松尾, 2016)，概ね，知能とは何かという大きな問いに向かって，1940 年代から 50 年代にかけて，計算機の出現とほぼ時を同じくして成立した．現在までの歴史のなかで認知科学とも大きな重なりがあり，数々の研究者が両分野に関わりながら，人間の知能に関する大きな問いを追いかけてきたと理解している．

人工知能は構成論的なアプローチをとり，そのときどきで副産物として最先端の計算機技術を生み出してきた．近年，注目を集める深層学習 (Goodfellow et al., 2016 岩澤他訳 2018; LeCun et al., 2015) も，2 つの顔をもつ．ひとつは「脳の構造をヒントに」作られたニューラルネットワークを用い知能の探求を行うという真理探究的・学術的な顔，もうひとつは画像認識や音声認識を中心に最先端の計算機技術として実社会に多くの活用事例を作り出してきたという工学的・実用的な顔である．

前者については，当初のニューラルネットワークの研究，例えば，McCulloch & Pitts (1943) は，脳の神経細胞をヒントに形式ニューロンを提案し，Fukushima (1980) らの複雑細胞，単純細胞は視覚野の構造をヒントにしていたが，近年の深層学習の研究は，人間の脳，知能の仕組みとはかなりかけはなれていると理解されている．例えば，活性化関数は，シグモイド関数から ReLU（$f(x) = \max(0,x)$ で表される）へと変化し，さらに最近では Swish 関数 ($f(x) = x \cdot \mathrm{sigmoid}(\beta x)$; Ramachandran et al., 2017) や周期性を捉えるためのへび関数 ($f(x) = x + \sin(x)^2$; Ziyin et al., 2020) など当初の神経細胞の仕組みから乖離した進展が見られる．さらに，ニューラルネットワークのパラメータを推定するための，損失関数の勾配の計算方法も RMSprop，Adam をはじめ，さまざまな方法が用いられている．大規模なニューラルネットワークの最適化に関して，新しく分かって

* E-mail: matsuo@weblab.t.u-tokyo.ac.jp

きた現象も多くあり，例えば，同じように良い局所最適値が多く存在するという特徴をもつこと，モデルのパラメータを増やしていくと汎化性能が最初は良くなるもののいったん悪くなり，その後また良くなるという二重降下という現象があること (Nakkiran et al., 2019)，パラメータが多いことによってたまたま初期値と構造が問題に適合しているという「当たりくじ」を見つけ出しているという，宝くじ仮説という説明がありえること (Frankle & Carbin, 2019) なども分かってきた．しかし，誤差逆伝播を脳でやっているのかという議論が依然として続いている (Bartunov et al., 2018) ことからも分かる通り，深層学習の進展は，人間の脳の仕組みと離れたところで活発に続いている．

一方で後者の工学的な顔については，深層学習の実社会での活用は目覚ましい進展を遂げており，特に画像認識の浸透は目をみはるものがある．例えば，新型コロナウィルスの体表温測定で顔の物体検知は一般人にとってごく自然なものとなったし，成田空港や羽田空港でも入国時の顔認証が導入された．また，深層学習を用いる自動運転の技術は着々と進展し，医療における画像診断も次々と医療機器の承認が降りている．製造業における外観検査も普及が急速な例のひとつである．いずれも深層学習をさまざまな形で用いている．これ以外にも，無数の深層学習の実用化の例があり，また米国や中国でのベンチャー企業等の躍進も顕著で，産業的にもますます拡大を続けている (情報処理推進機構 AI 白書編集委員会, 2020)．

2021 年現在においては，前者の学術的な意義よりも，後者の実用性が重要だと注目されがちであるが（そしてそのことは著者自身は 2015 年ごろからずっと主張してきたことではあるが (松尾 2015)），むしろ前者のほうが重要な段階に入ってきていると感じている．

すなわち，深層学習が知能の仕組みの解明において，重要な役割を果たす可能性がある．なぜなら，多くの発見を通して，人間の知能の仕組みに対しての示唆が多く積み上がってきているからである．本稿では，そうした可能性の中心となる議論について，最新の研究動向を紹介しつつ，著者なりの見解を交えながら解説していきたい．

2. 大規模言語モデル

近年，深層学習を用いた大規模な言語モデルが大きな注目を集めている．

きっかけとなったのは，2017 年の *Attention is all you need* という論文 (Vaswani et al., 2017) で，トランスフォーマ（transformer）という機構がよく機能することが明らかになった．その後，BERT (Devlin et al., 2019)，XLNet，RoBERTa，ALBERT，T5，GPT-3 (Brown et al., 2020) など，次々とトランスフォーマを用いた巨大な言語モデルが発表された．BERT は Google から発表されたもので，GPT-3 は，Elon Musk や Peter Thiel などが 2015 年に設立した OpenAI という非営利の研究団体から発表されたものである．

トランスフォーマとは，もともとは自然言語文において長い距離の依存関係を把握したいということで作られたものであるが，従来からよく用いられてきたリカレントニューラルネットワーク（以下，RNN）ではなく，アテンション（注意）という仕組みを用いる．これは，対象とする層のユニットと，注意に対応するベクトルとの内積をとることで，注意を向けた部分の情報だけを取り出すというものである．そして，この注意のベクトルを，対象とする層自体から取り出してしまうのが自己注意機構である．この自己注意がひとつの層に複数あることをマルチヘッドという．

まとめると，トランスフォーマは，マルチヘッドの自己注意機構を用い，それと通常の多層パーセプトロンの層を交互に多層に重ねたものである．機構は複雑であるが，長い距離の依存関係を含め，入力からの情報を取り出し，加工することが柔軟にできる仕組みである．

BERT や GPT-3 などでは，事前学習（pre-training）で言語モデルを学習し，その後，下流の（downstream）タスクに対して，ファインチューニング（fine-tuning）を行うという点も，トランスフォーマの使用と並んで大きな特徴のひとつである．事前学習として教師なし学習を用いるが，最近は自己教師あり学習（self-supervised learning）と呼ばれる場合が多い．教師なし学習の重要性は人工知能の分野でも古くから指摘されているが，深層学習で 2019 年にチューリング賞を共同受賞した Yann LeCun が，

ケーキの比喩で自己教師あり学習はケーキのスポンジの部分だとしてその重要性を主張したのは有名である[1)]．具体的にいうと，自己教師あり学習では，手元にあるデータから，擬似的に教師あり学習の問題を作成する．自然言語では，文中の一部の語にマスクをかけて隠し，それを当てるという問題や，2つの文が連接しているかどうかを当てる問題，次の語を予測する問題などが自己教師あり学習の問題として用いられる．

GPT-3 は，非常に巨大なモデルを用い，最も大きなバージョンでは，トランスフォーマ＋フィードフォワードの層を 96 層もつ．各層に 12,288 ユニット，ヘッドの数が 96 であり，全体として，1,750 億パラメータ（パラメータとは，重みとして学習可能な変数）を持つ．事前学習としては，次の語を予測するタスクを用いる．データとして，Wikipedia データや子供向けの本に加えて，巨大な Web クロールのデータを用い，全体で 4 兆トークン（近似的に単語と考えてよい）以上のデータを用いている．

GPT-3 の事前学習済みのモデルは，2020 年 7 月に API が限定的に公開された．それを使った印象的なデモがたくさん構築され，あまりにその能力を高く見せてしまうために誇大広告気味であるとして，投資家による注意喚起がなされるほどであった．(もともと，GPT-2 の公開時には，このモデルが危険過ぎるのではないかという開発陣自身の懸念から，公開が延期されたこともあった．)

デモは多岐に及び，翻訳や質問応答などの一般的なものはもちろん，例えば，少しの情報の入力を行うだけで「それっぽい」レジュメを自動で生成するもの，エクセルのデータ（例えば都市と人口）の欠損部分を周りから推測して自動で補っていくもの，また，自然言語文から HTML のコードを生成するもの，アプリのプロトタイプを生成するもの，機械学習のコードを生成するものなども示された．プログラマが不要になるのではといった議論すら起こった．

なかでも，私が特に面白い，かつ重要だと思うデモのひとつは，素数を列挙するものである（図 1）．最初のいくつかの素数を列挙すると（太字部分．これを下流のタスクとして学習させると），自動的に次々と素数を挙げていくことができる[2)]．

2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41,
43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97, 101, 103, 107, 109, 113, 127, 129*, 131, 137, 139, 149, 151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199, 211, 223, 227, 229, 233, 239, 241, 251, 257, (263*), 269, 271, (277*), 281, 283, 293, 307, 311, 313, 317, 331, 337, 347, 349, 353, 359, 367, 373, 379, 383, 389, 397, 401, 409, 419, 421, 431, 433, (439*), 441*, (443*), 449, 457, 461, 463, 467, 479, 487, 491, 499, 503, …

上記で太字はFew-shot学習として与えたもの。43以降はGPT-3による出力。*は素数ではなく、間違い。括弧つきはGPT-3が出力できなかった素数で、間違い。

図 1　GPT-3 を用いた素数の生成

素数を列挙するということがなぜできるのか？これはもちろん事前学習の効果であり，大量の Web ページをデータとして次の語を予測するタスクによって言語モデルが作られているからである．当然，Web ページのなかには，素数を列挙したページもたくさんあるので，それを学習している，おそらく「丸覚え」しているのではないかと思うかもしれないが，よく見ると，ところどころで間違えている．*をつけたものは一見素数のように見えるが素数ではなく，(*) をつけたものは列挙することができなかったが実際には素数であるものである．丸覚えしているのであれば，間違えはないはずであり，すなわち，「だいたい合っているがたまに間違える素数生成の仕組み」を学習によって獲得しているわけである．つまり，素数の計算の方法が，十分に汎化されない不十分な形で学習されている．

これを支持するような知見として，GPT-3 の論文 (Brown et al., 2020) には，次のような実験について記されている．数学的なタスクがどのくらい学習できるかという実験であり，下流のタスクとして 10 個の算術演算（足し算や引き算，掛け算など）を与えるだけで，どのくらい精度が出るかを調べる．さらに，モデルのパラメータ数を増やすとこの精度がどう変化するかを調べる．それによると，モデルのパラメータを増やしていくと，130 億パラメータに至ると急に 2 桁の足し算や引き算ができるようになる．1,750 億パラメータに至ると，急に 3 桁の足し算や引き算ができるようになり，また，2 桁の掛け算も少しできるようになる．すなわち，算術的な演算を行うアルゴリズムが，多層のトランスフォーマによって言語モデルとして学習されているということになる．

BERT や GPT-3 などのトランスフォーマに基づくモデルは，関数の万能近似器であり，かつ系列を系

1) 例えば https://medium.com/syncedreview/yann-lecun-cake-analogy-2-0-a361da560dae
2) Aravind Srinivas の実験による．

列に変換するアルゴリズムとしてチューリング完全（チューリングマシンと同じ計算能力を持つ）であることが示されている (Yun et al., 2020)．この性質は，多層パーセプトロンや RNN 等でも示されているものであり，それ自体は特筆すべきものではないが，実際に，算術的な演算を行うアルゴリズムが，事前学習によって大量の Web 上のデータから自然に獲得されているということである．素数を計算するというのは単純な計算ではなく，少なくとも，素因数分解と因数の数え上げができなければならない．(あるいは，もっと簡易で学習しやすい方法があるのかもしれないが.）これが意味することは，適切なアーキテクチャをもつ（巨大な）モデルに，適切な学習タスクを与えることによって，(理論的にではなく）現実に目的とするアルゴリズムを学習させ，実行できるようにできるのではないかということである．

なお，深層学習の領域では，2015 年ごろ，任意のアルゴリズムを実行するチューリングマシンに該当するものとして，ニューラルチューリングマシン (NTM; Graves et al., 2014) やメモリーネットワーク (Weston et al., 2014) といった手法が提案されていた．いずれも RNN にメモリへのアクセスを明示的に組み込んだものである．一方で，BERT や GPT-3 などは，明示的なメモリがなくともチューリング完全であることが示されている (Perez et al., 2019).

どういった事前学習のタスクを課せば，どのようなアルゴリズムが学習され得るのか，それがどういったアーキテクチャや機構で効率化され得るのかというのは，今後徐々に明らかになっていくだろうが，大変興味深い課題である．さらに，誤解を恐れずに言えば，大規模なニューラルネットワークに適切なタスクを課すことが，人間の持つ万能性，特に言語による学習の可能性や教育の意義と関係があるように思えてならない．

3. 大規模言語モデルの課題

BERT などのトランスフォーマに基づく手法で得られた大規模な言語モデルに何が学習されているかについては，さまざまな研究がある．BERTology という論文 (Rogers, 2020) がよくまとまっており少し紹介したい．

この論文では，なぜ BERT がうまくいくのか，そのモデルが何を学習しているのかについて，150 以上の論文からの知見が整理してまとめられている．まず，BERT の表現は階層的であり，形態素，構文チャンクや意味役割についての情報が，階層的に符号化されている．統語構造は，自己注意の重みには直接は符号化されていないが，表現から統語情報を再現することはできる．BERT で学習されている統語情報は，我々が言語学的にアノテーションするようなものとは異なるようで，例えば，BERT は NPI（Negative Polarity Item のこと．*ever* など）や，その使用を許すような言葉（*whether* など）を見つけるのがうまい．BERT は，否定を「理解」していない．おかしな入力を与えられてもそれを判定できずに動いてしまう．以上の意味で，BERT の統語知識は不完全であると言えるが，これは与えられた事前学習のタスクを解くのに必要ないからとも言うことができる．

BERT は，意味処理についての知識も持っている．エンティティのタイプ，関係，意味役割なども符号化している．しかし，数の表現には苦労している．浮動小数点の良い表現も得られていない．また，BERT は，固有表現（named entity）の置換に対して驚くほどもろい．固有名詞が同じ実体を指すかどうかを判定する共参照のタスクでその固有名詞を別の固有名詞に置き換えると 85%のケースで予測が変化する．つまり，固有表現の一般的な意味概念を作り出していない．

BERT は，実践的な推論やイベントに関する知識も得意ではない．概念の抽象的な属性や，視覚的，感覚的な属性についても難点があり，世界知識に基づいて推論を行うことができない．例えば，人間が家に歩いて入ることができ，家は大きいことは知っているが，人間が家より大きいかは推論することができない，などである．

何が得意でないかについて，さらに深めた研究 (Talmor et al., 2020) では，以下のことを示している．Always-Never 質問（例えば，鳥は角を（決して持たない | 時々もつ | いつも持つ）という質問），年齢比較質問（21 際の男は，私が 35 歳であれば私より（若い | 年寄り）だという質問）など 7 種類の質問に対して，BERT 等の手法がどのくらい正当できるかを比較している．それによると，Always-Never 質問はほとんどのケースでうまくいかない．年齢比較質問はそこそこであるが，これが 3 つ以上の比較になるタスク（Multi-hop Composition）では急に

精度が落ちる．また，辞書を引いた上で回答しないといけないような問題（例えばジョン・レノンが演奏したバンドはいつ結成されたか?という問題であり，ジョン・レノンが演奏したバンドをまずビートルズに置き換えてから，ビートルズがいつ結成されたかを回答しなければならない）も苦手である．

以上をまとめると，現在の大規模言語モデルで苦手であることとして，(他にもあるかもしれないが) 大きく次の 2 点が挙げられる．

［課題 1］ 大規模言語モデルは，複数の行為からなる系列に基づく処理が苦手である．(否定の扱い，名前の置き換え，3 つ以上の数の比較，辞書を引いて回答するなど.)

［課題 2］ 大規模言語モデルは，実世界の経験や行動に基づく知識の処理が苦手である．(Always-Never 質問や，世界知識に基づく推論など.)

課題 1 は，多層のトランスフォーマのモデルは，基本的に 1 回の系列の入力から 1 回の系列の出力の変換を学習しており，(その内部で複数の処理が学習されていることはあり得るとしても）そうした変換を組み合わせて用いることはできない．これは，「言語の空間における身体性」と言ってもよいのかもしれないが，対象と相互作用することによって結果を導き出すという視点が欠けている．

課題 2 は，人工知能の分野で従来から指摘されているシンボルグラウンディング問題と言われるものと同じであるが，言語がそれを指す具体的な画像や映像，相互作用などの経験と紐付いていないため，それらを必要とする種類の処理には対応できない．(あるいは言語的なパターンだけから学習することは甚だ非効率である.)

以上のふたつの課題について，次の節ではどういった研究が行われているのかを見ていきたい．

4. 世界モデルと「もつれを解くこと」

実世界を扱うための画像認識や音声認識は，深層学習によってこの 10 年で大きく進展した．また，深層学習を実世界知能につなげようとする研究，すなわち深層強化学習やロボットへの応用研究も大きく進展した．

トランスフォーマと自己教師あり学習の成功は，画像処理の分野にも波及した．OpenAI が出した iGPT という手法 (Chen et al., 2020) では，画像の解像度を粗くし，また 2 次元を 1 次元にしベクトル化する．すると，基本的に GPT のアルゴリズムがそのまま適用できる．マスクしたピクセルの予測，あるいは次のピクセルを予測という事前学習を行い，タスクに応じてファインチューニングを行うことで，トップレベルの教師あり学習の精度に並ぶ．

他にも，画像分類，物体検知，セグメンテーションなどで，ViT，DETR，IPT，Max-DeepLab などの手法が提案されている．いずれも，事前学習と簡単な線形分類器で，教師あり学習と同程度の精度を出すことが次々と示されている (Han et al., 2020).

画像に限らず，自己教師あり学習によって，エージェントが自己をとりまく環境をモデル化しようという研究が，2018 年ごろから盛んになっている.「世界モデル」（world models）と呼ばれる研究であり (Ha & Schmidhuber, 2018)，従来から人工知能あるいはロボットの分野で，内部モデル，あるいはメンタルモデルと呼ばれていたものに近い．DeepMind の Generative Query Network（GQN）という研究では，ある環境（積み木の世界）において，ある視点を固定し，そこからの「見え」を予測する (Eslami et al., 2018)．この問題を多数解かせて学習することで，空間の 3 次元構造を表すような内部表現を獲得することができる．

こうしたモデルは，深層強化学習，あるいはそのロボットへの応用という文脈で大きな意味を持つ．強化学習では，従来から，状態の明示的な遷移をモデルとして持っているモデルベース強化学習と，それを持っていないモデルフリー強化学習の 2 つがあり，それぞれ別のものとして研究が進んできた．例えば，囲碁は状態の遷移が明示的に定まっているため，アルファ碁はモデルベース強化学習であり，DeepMind が 2014 年ごろに行った ATARI のゲームの学習を行う手法は，Deep Q-Learning（DQN）と呼ばれ，状態の遷移を明示的にもたないモデルフリー強化学習である．一般に，モデルベース強化学習は，サンプル数が少なくとも効率的に学習できるが，事前知識が必要であり，モデルフリー強化学習は，事前知識がいらないがサンプル数が多く必要で，ゲームやオンライン上のシミュレータなど高速に実行できる環境でないとうまくいかない場合が多い．

ところが，世界モデルを獲得するということは，状態遷移モデルが与えられていない前提で，データから状態表現および状態遷移モデルを獲得し，モデルベース強化学習を行うことに相当する．した

がって，こうした手法が，深層強化学習，あるいはロボットへの応用で成果を挙げることが期待されている．

さて，こうした世界モデルと言語を結びつけるにはどうしたらよいだろうか？

その際には，画像と言語をつなぐ研究が参考になる．言語と画像のアラインメントはさまざまに研究が進んでいる．例えば，Microsoft COCO や Visual Genome，VQA（Visual Question Answering）を改良した GQA などのデータセットを用い，文と画像のアライメントを学習させることができる．すると，画像から文を生成する，あるいは逆に，文から画像を生成することができる (Mansimov et al., 2015; Reed et al., 2016)．例えば，*Stop sign flying in the sky* という文を入力し，空を飛ぶ止まれ標識を画像として生成することができる．よりきれいな画像を生成するために，StackGAN では，粗い画像をまず生成し，次に高解像度の画像を生成するという 2 段階で行う (Zhang et al., 2017)．BERT を用いた，画像と文のアライメントを行う手法（VisualBERT, VL-BERT，LXMERT など）も提案されている．

画像の生成は，深層生成モデルと呼ばれる技術によって大きく進展した．代表的な 2 つの手法が，Generative Adversarial Network（GAN，敵対的生成ネットワーク）と Variational Auto-Encoder（VAE，変分オートエンコーダ）である．GAN は，識別器と分類器が互いに競うことで精度を上げ，VAE は，データ自身を再構成する（自己教師あり学習としての）オートエンコーダのモデルを学習することによって精度を上げる．従来は，VAE よりも GAN のほうがきれいな画像が生成できるとされてきたが，最近では VAE も潜在変数を離散化する，多層にするなどによって，GAN と同じようにきれいな画像を生成できるようになってきた (Razavi et al., 2019)．

ここで重要なのは「もつれを解くこと」（disentanglement）という概念である．自己教師あり学習の最終層のひとつ手前では，うまく学習すると disentangle された表現が得られている．また，VAE の潜在表現も，うまく学習すると disentangle された表現が得られる．disentangle された表現とは，世界の変形的な性質に注目し，いくつかの独立な部分空間に分解された表現であり，それぞれの部分空間が他の部分空間のアクションから影響を受けないというものである (Higgins et al., 2018)．

このように書くと理解が難しいが，例えば，顔画像であれば「目の大きさ」「ひげがあるか」「若いか年寄りか」「男性的か女性的か」などは disentangle されたものであり，それぞれの要素を独立に変化させることができる．（通常，男性的であることとひげがあることは相関しており，絡み合った概念である．）画像にしても世界モデルにしても，要素分解することで，いかにこの disentangle された表現を獲得するかが重要な鍵となる．disentangle された表現を得ることさえできれば，あとは言語で条件づけた深層生成モデルを用いて，さまざまなデータを生成できるということになる．これは，簡単にいえば「想像する」ということであり，知能における想像の重要性は，DeepMind の Demis Hassabis らによってたびたび指摘されている．

OpenAI が 2021 年 1 月に公開した DALL-E というモデルは，文と画像のペアを事前学習したもの (CLIP と呼ばれるコントラスティブな事前学習; Radford et al., 2021) を用いて，画像の生成を行うものである [3]．自然言語に対してのトランスフォーマと，画像に対してのトランスフォーマが用いられている．例えば，*an armchair in the shape of an avocado* という文を入力すると，アボカドの形をした椅子が画像として生成される．驚くような例がいくつも紹介されている．

5. 今後の研究の方向性

前節で述べたような自然言語からの生成は，多くの場合，画像に限った話であるが，こうした自然言語からのデータの生成を実世界で獲得した世界モデルに対して行えば，前節で述べた課題 2，すなわちシンボルグラウンディングの問題に大きくアプローチすることができるだろう．例えば，鳥は角を（決して持たない | 時々もつ | いつも持つ）という Always-Never 質問に対して，あるいは，人間が家より大きいかという質問に対して，言語から画像あるいは実世界に紐づく世界モデルを生成し，それに基づいて答えを出力することができれば，原理的には答えることができるようになると考えられる．（ロボットのセンサ・モータ情報の disentanglement とその言語との関連付けは，尾形らの一連の研究，Zhong et al., 2019, にその端緒を見ることができる．）

では，課題 1 に対してはどうだろうか？上記の課

3) https://openai.com/blog/dall-e/ 論文は今後公開されるとのことである．

題 2 に関連する「画像あるいは世界モデルを生成し，それに基づいて答えを出す」ということは，一連の行動の系列である．こうした行動の系列を用いることが，現状の大規模言語モデルで苦手であるということは，3 節の課題 1 で述べた通りである．

これを解くには，やはり行動の系列を直接扱う，深層強化学習の研究を参照する必要がある．深層強化学習の研究は，例えば，カリフォルニア大学バークレー校やスタンフォード大学，DeepMind などで進展している．自己教師あり学習の手法等も取り入れているが，これまでのところ飛躍的な精度の向上や，従来にない印象的なデモの実例があったかというとそうではない．

ここからは著者の仮説になるが，その理由はおそらく，時間の取り扱いに関しての技術が未成熟，あるいは方向が誤っているからであろう．

現在の強化学習では，この分野の大家である Richard Sutton の大きな貢献以降，伝統的に状態空間表現（state space representation）が用いられている．どの教科書でも，はじめにこの説明がでてきた上で，MDP（マルコフ決定過程），あるいは POMDP（部分観測マルコフ決定過程）などの説明が続く．

この際，状態間の時間のステップ幅を決め，状態を遷移する際の行動を定義する．しかし，我々が日常的に感じるように，時間の扱いにおいて，行動ごとに対象となる時間スケールは異なるし，その影響がステップごとに同一の扱いを必要とするというのは考えてみれば不自然である．そのため，階層的プランニング等の手法によって行動を階層的にすることが試みられるが，それでも大きなブレークスルーはこれまでのところ得られていない．

こうした状況はどう理解すればよいだろうか．私は，次のように考える．深層学習の画像認識におけるブレークスルーで明らかになったものは，特徴量の生成そのものが大きな問題であったということである．画像に関しては，この特徴量の生成の問題を，畳み込みニューラルネットワーク（CNN）等の深層学習，あるいは最近ではトランスフォーマで解決したわけである．深層強化学習が抱えている問題も同様であり，時間方向に広がりをもった時空間の情報に関して，適切な特徴量を抽出できていないという点にあるのではないか．つまり，従来の画像認識と相似の問題を抱えているように思う．

仮にそうであれば，さまざまな特徴生成の既存手法，具体的には時空間方向に延ばした CNN やトランスフォーマを用いればよいということになるが，事態はそれほど簡単ではない．画像は，ピクセルというある程度情報の密度がそろったものを扱っているし，自然言語は，同様に情報の密度がそろった語（あるいは文字や音素）という単位を扱っている．したがって，遠距離の影響も同じオーダの想定の範囲内にある．ところが，我々の実世界における時間的な影響は，例えば，水の入ったコップを把持するときの，少し傾いたから腕の筋肉を補正するといった数百ミリ秒の制御から，部屋の隅にあるかばんを取ろうという数秒に渡る身体全体の大域的な制御から，電車にのって友人と落ち合うといった数十分単位のかなり高次の行動まで，オーダの異なるさまざまな粒度がある．しかも，多くの場合，行動を制御するためのフィードバック系のループが独立に構成されており，それらが協調しながら全体としての行動を達成している．

したがって，画像の場合の CNN で仮定したような空間的近接性のプライア（事前知識），あるいは，自然言語処理の場合に BERT や GPT-3 で仮定したような「文」という単位での情報の固まりといったプライアに相当するものが時空間の処理においても必要であり，それは，おそらく，センサ−アクチュエータのフィードバックループを基本にした構造ということではないだろうか．実世界での機械やロボットの制御には，未だに PID 制御（出力値に目標値を一致させるためにその差分，差分の微分や積分を用いて制御を行う手法）がよく使われる．PID 制御で仮定されているのは，

(i) フィードバックループであること
(ii) 時定数があること
(iii) 目標値と出力値から（簡単な計算で）算出される操作量（つまり行動に関しての因子）が制御されること

ということである．これらは，どれも，実際の機械やロボットの制御では重要であるが，現在の強化学習の文脈では明示的に扱われないものである．人間の脳や身体において，さまざまな時定数のさまざまなフィードバックループが存在することを考えても，こうしたフィードバックループを基礎とすることは自然な仮定であると思われる．

そして，これを実現するためには，おそらく，「何が目標値か」「なにが出力値か」，そして，何が行

動の因子かを，深層学習における特徴抽出の結果として見つけなければならない．(逆に，これさえ見つけてしまえば，あとは少自由度の簡単な制御の問題になる．) そして，これらはすべて，時空間のデータを対象とした自己教師あり学習により disentangle された結果として得られるものであるはずである．

そのためには，自己教師あり学習の問題設定を大きく変える必要があるかもしれない．ここで，少し脳科学の話にはいるが，ヒトの大脳皮質は，どこも 6 層構造をしており，可塑性がある．そして，扱うデータにあわせて異なる構造が構成される．視覚的な入力を扱う視覚野は，空間的な隣接性を重視したカラム構造になるし，聴覚を扱う部分，運動野や前頭前野などもそれぞれ異なった構造になる．これは，いまの深層学習が仮定しているような，何らかのネットワークのアーキテクチャを最初において，そこから重みを学習するという方法ではなく，ニューロンがより柔軟にデータの特性にあわせて組み合わさるうちに，データに応じた構造ができあがるということではないだろうか．

そこまで考えると，実は，CNN やトランスフォーマといったアーキテクチャ自体を作り出す仕組み自体を再設計しないといけなくなるので，かなり大掛かりな変更を必要とすることになる．これが私の見立てでの現在の深層学習の最も根本的な課題である．仮にここでの議論が正しいとすると，そうしたアーキテクチャの大きな変更をした先に，時空間の特徴量を自己教師あり学習で適切に取得することができ，フィードバックループを仮定した行動の学習をすることができ，実空間での行動制御が格段に向上するとともに，課題 1 で述べたような，言語空間での知識処理に関しても大きくその性能があがることにつながるのではないか．

以上が，現時点で，著者が持っている見通しであるが，その真偽はさておき，こうした視点での技術の進展が予想されるからこそ，冒頭に述べたように，人間の知能（あるいは脳科学の知見）と人工知能の研究が近づいている段階に入っていると予感するのである．

6. おわりに

本稿では，まず，深層学習による大規模言語モデルの進展を紹介し，その可能性とともにその課題を述べた．そして，その課題の根本的な原因となる 2 つの課題について説明した．その課題が，実世界知能に関連していること，そして，課題解決の糸口に対しての私見を述べた．

前半部分はできるだけ論文等の客観的な知見になるように心がけたつもりであるが，後半はやや一面的，恣意的な部分もあったかもしれない．しかし，いずれにしても，深層学習の急速な進展が，知能の仕組みに関して大きなヒントを与えてくれている段階に入っていることを感じ取っていただければ幸いである．そういった中で，再度，認知科学と人工知能の両分野は刺激しあい，融合していく必要があるのかもしれない．

最後に，本特集において，著者自身の考えの整理にもなる貴重な執筆の機会をいただいたことに感謝し，結語としたい．

文　献

Bartunov, S., Santoro, A., Richards, B., Marris, L., Hinton, G., & Lillicrap, T. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems*, 9390–9400.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. https://arxiv.org/abs/2005.14165

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Dhariwal, P., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. *Proceedings of the 37th International Conference on Machine Learning, PMLR, 119*, 1671–1703.

Devlin, J., Chang, M. -W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Eslami, S. M. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., Ruderman, A., Rusu, A. A., Danihelka, I., Gregor, K., Reichert, D. P., Buesing, L., Weber, T., Vinyals, O., Rosenbaum, D., Rabinowitz, N., King, H., Hillier, C., Botvinick, M., ... Hassabis, D. (2018). Neural scene representation and rendering. *Science*, *360* (6394), 1204–1210. http://dx.doi.org/10.1126/science.aar6170

Frankle, J., & Carbin, M. (2019, May 6–9). *The lottery ticket hypothesis: Finding sparse, trainable neural networks.* ICLR 2019: Seventh International Conference on Learning Representations. New Orleans, LA, United States.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recog-

nition unaffected by shift in position. *Biological Cybernetics*, *36* (4), 193–202. https://doi.org/10.1007/BF00344251

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (グッドフェロー, I.・ベンジオ, Y.・クールヴィル, A.　岩澤 有祐・鈴木 雅大・中山 浩太郎・松尾 豊 (監訳) (2018). 深層学習　KADOKAWA)

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing machines. arXiv. https://arxiv.org/abs/1410.5401

Ha, D., & Schmidhuber, J. (2018). World models. arXiv. https://arxiv.org/abs/1803.10122

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2020). A survey on visual transformer. arXiv. https://arxiv.org/abs/2012.12556

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., & Lerchner, A. (2018). Towards a definition of disentangled representations. arXiv. https://arxiv.org/abs/1812.02230

情報処理推進機構 AI 白書編集委員会 (編) (2020). AI 白書 2020：広がる AI 化格差と 5 年先を見据えた企業戦略 角川アスキー総合研究所

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. http://dx.doi.org/10.1038/nature14539

Mansimov, E., Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2015). Generating images from captions with attention. arXiv. https://arxiv.org/abs/1511.02793

松尾 豊 (2015). 人工知能は人間を越えるか：ディープラーニングの先にあるもの　KADOKAWA

松尾 豊 (編著) (2016). 人工知能とは　近代科学社

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5* (4), 115–133. https://doi.org/10.1007/BF02478259

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2019). Deep double descent: Where bigger models and more data hurt. arXiv. https://arxiv.org/abs/1912.02292

Perez, J., Marinkovic, J., & Barcelo, P. (2019, May 6–9). *On the turing completeness of modern neural network architectures*. ICLR 2019: Seventh International Conference on Learning Representations. New Orleans, LA. United States.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv. https://arxiv.org/abs/2103.00020

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. arXiv. https://arxiv.org/abs/1710.05941

Razavi, A., van den Oord, A., & Vinyals, O. (2019). Generating diverse high-fidelity images with VQ-VAE-2. arXiv. https://arxiv.org/abs/1906.00446

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. *Proceedings of the 33rd International Conference on Machine Learning, PMLR*, *48*, 1060–1069.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Talmor, A., Elazar, Y., Goldberg, Y., & Berant, J. (2020). oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, *8*, 743–758. https://doi.org/10.1162/tacl_a_00342

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Kaiser, Ł. (2017). Attention is all you need. *Proceedings of the 31st Conference on Neural Information Processing Systems*, 6000–6010.

Weston, J., Chopra, S., & Bordes, A. (2014). Memory networks. arXiv. https://arxiv.org/abs/1410.3916

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., & Kumar, S. (2020, Apr 26–May 1). *Are transformers universal approximators of sequence-to-sequence functions?* ICLR 2020: Eighth International Conference on Learning Representations. Virtual Conference.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. *Proceedings of 2017 IEEE International Conference on Computer Vision*, 5907–5915. http://dx.doi.org/10.1109/ICCV.2017.629

Zhong, J., Ogata, T., Cangelosi, A., & Yang, C. (2019). Disentanglement in conceptual space during sensorimotor interaction. *Cognitive Computation and Systems*, *1* (4), 103–112. http://dx.doi.org/10.1049/ccs.2019.0007

Ziyin, L., Hartwig, T., & Ueda, M. (2020). Neural networks fail to learn periodic functions and how to fix it. arXiv. https://arxiv.org/abs/2006.08195

松尾　豊

1997 年 東京大学工学部電子情報工学科卒業．2002 年 同大学院博士課程修了．博士（工学）．産業技術総合研究所，スタンフォード大学を経て，2007 年より，東京大学大学院工学系研究科准教授．2019 年より，同教授．専門分野は，人工知能，深層学習，ウェブマイニング．人工知能学会において，2012 年から編集委員長・理事，2014 年から倫理委員長，2019 年より理事．2017 年より日本ディープラーニング協会理事長を務める．