人の作業の代行を目的とした人工知能研究の現状と今後

Current and Future Research of Artificial Intelligence to Automate Human Work

岡谷　貴之[*1]

Takayuki OKATANI

Professor, Graduate School of Information Sciences, Tohoku University /
Team Leader, RIKEN Center for AIP

Abstract

This paper discusses the past, current, and future of deep learning, a core technology of AI that continues to evolve quickly. We first summarize the mainstream approach based on fully supervised learning and its limitations and then explain the recently developed self-supervised learning to overcome them. We then describe how language models developed in natural language processing will alter artificial intelligence in the future, discussing the possibility of developing further advanced AI by incorporating text-based knowledge.

●***Keywords*:** Artificial intelligence, Deep learning, Self-supervised learning, Language model

[*1]東北大学　教授／理化学研究所　チームリーダー

1. はじめに

人工知能（AI）は近年、長足の進歩を遂げた。本稿では、人が行う作業を代替・自動化することを目的としたAIで、特に画像を扱うものについて、その研究開発のこれまでを要約し、今後を占う。この意味でのAIの応用には、自動運転、製造物の完成品検査、インフラ構造物の点検・維持管理など、様々なものがあるが、そこで適用される技術の中身は大きく変わらない。ここでは、そのような技術の汎用的な側面を考える。

さて、ここ10年ほどの間のAIの進展のほとんどは、深層学習、すなわち深層ニューラルネットワーク（deep neural network、以下DNN）を用いた機械学習によって、もたらされている[1)]。深層学習をうまく適用することで、以前は難しいと思われた数々の問題が、実用的な水準、あるいは限りなくそれに近い水準で解決され、または大幅な性能向上を果たすに至っている。コンピュータビジョン（視覚を対象とするAI）の分野における、そのような成功例の一部を表1に示す。

表1　深層学習のコンピュータビジョンへの応用例

物体認識	物体認識（画像分類）	画像１枚からそこに写る物の種類（カテゴリー）を認識. 既知のカテゴリのどれかに分類する。
	物体検出	画像１枚の中にあるすべての物体に対し、画像内の位置（矩形領域で表現）と物体の種類を与える。
	セマンティックセグメンテーション	シーンの画像 1枚を入力に、各画素の位置での物体が何かを認識. さらに物体を１つ１つ区別する問題をインスタンスセグメンテーションと呼ぶ
人・生体認証	顔認識認識	顔画像から人物を同定する。対象者の数（クラス数）が多い一般的なケースでは、開クラス集合のクラス分類として扱う。
	個人再同定(person ReID)	複数の監視カメラがある程度離れた場所に設置されており、同一人物がその視野内に映り込む時、同一人物を同定。
幾何推定	人体(関節物体)ポーズ推定	人体の関節（例えば14 個を選定）が画像のどこに写るかを特定・
	顔向き推定	人の顔の画像１枚から顔の向き（3自由度）を推定。
生成	画像操作・変換・生成	スタイル変換（実物の写真を絵画調に変換するなど）
意味理解	人（物体）と物体の関係性の推定	シーンの画像 1 枚を入力に、そこに写る人が物体に対してどういう関係性を持っているかを認識する。
	動画を用いた認識	動画中で行われる人の行動を分類する。
	画像と言語のマルチモーダルタスク	シーンの画像からシーンの記述を与える image captioning や、画像１枚とその内容に関する質問文の２つを入力に受けとり、質問に回答する visual question answering など；その他多数。
3次元	奥行きの推定	カメラからのシーンの奥行きを画素ごとに求める：１枚の画像で行うものや、ステレオカメラの画像からこれを行うもの、より多くの視点の画像からシーンの３次元形状を求めるもの（multi-view stereo）がある。
運動	オプティカルフロー推定	動画の連続する２画像フレームを入力に、シーン（や運動物体）表面の同一点が、画像フレーム間でどれだけ移動したかを返す；ステレオカメラで同じことをすれば、視差の推定＝奥行きの推定に一致する。
画質改善	ノイズ除去、ボケ・手ブレ除去、超解像	入力画像１枚の品質を向上させる。

このように成功例は多岐に渡るが、深層学習の適用のされ方は、それらの間でほとんど同じである。ある入力xと出力yを考え、それらの間に成り立つ写像ϕ:x→yを、1つのDNNでy=f(x;w)と表現し、DNNのパラメータwを学習によって定める。新たな問題に適用するには、まず入力と出力を選定し、合わせて訓練データを収集した上で、目標とする写像を与えるDNNの構造を設計し、その学習を成功させることが必要となる。先述のような様々な参考となる応用事例がある今、成功の鍵は訓練データをどのように収集するかと、それを使った学習をいかにうまく行うかが握っている。

2. これまでの深層学習と、その課題

DNNの学習方法には様々あるが、これまで最も成功してきたのは、オーソドックスな教師あり学習である（図1（a））。そこでは入力の標本xと、あるべき出力yのペア(x, y)を集めた訓練データ$\mathcal{D} = \{(x_n, y_n)\}_{n=1,\cdots,N}$を、学習に用いる。具体的には、$x_n$に対するDNNの予測$\widehat{y_n} = f(x_n; w)$と、その正解$y_n$の差――例えば$\Sigma_n |y_n - f(x_n; w)|^2$――を小さくするように、勾配降下法でパラメータ$w$を定める。

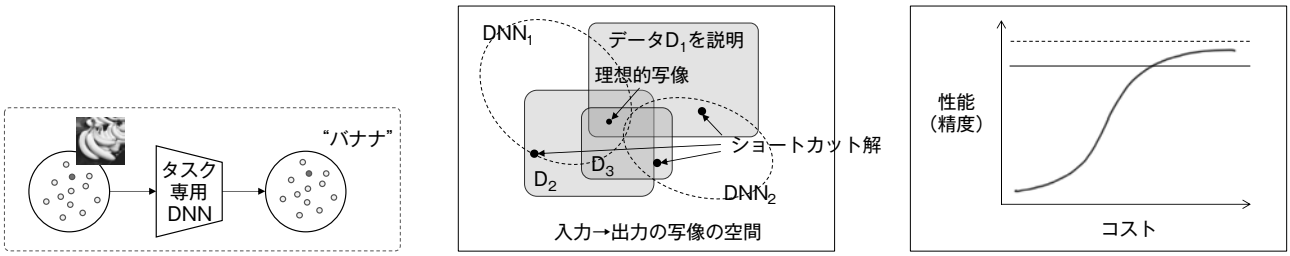

図1　(a) 完全教師あり学習　(b) ショートカット学習概念図　(c) コストと性能のS字カーブ

ただしこの方法は、机上では高い性能を見せても、すんなりとは実用化には至らないことも多い。やっかいなのは、性能評価が一筋縄ではいかないため、現実の性能を正しく評価しづらいことである。あらゆる入力を試せれば理想的だが、通常、入力xの空間が広すぎて、そのごく一部でしかテストを実施できず、品質保証が難しい。しかしより本質的な問題は、深層学習が、机上と現実との間で異なる性能を見せ易いことである。これは訓練データへの過剰な依存から来る性質で、著者らはショートカット学習と呼んでいる。これは、DNNが与えられた訓練データ（+テストデータ）内のみに潜む、統計的な偏りを活用するように学習してしまう現象を指す（図1（b））。そんな形で得られた解は多くの場合、われわれが期待するものではなく、実世界で現実に遭遇するデータに対して、望んだ性能を達成しない。

一例を挙げれば、皮膚病の医用画像診断――皮膚変状の画像1枚から、その悪性良性を診断、つまり「ほくろ」と「皮膚ガン」を判別する問題――において、訓練データDを作る際、変状が悪性のときのみ、「ものさし」を添えて画像を撮影してしまったという事例である。このとき、DNNは病変の悪性良性ではなく、画像内の「ものさし」の有無を判断するように学習してしまう。これは分かり易い例だが、表面的には分かりづらいショートカットも多く、訓練データの作成時、意図しない「ものさしの配置」を生じ、そのことをなかなか突き止められないことが良くある。

もう一つの例として、物体認識で「バナナ」を認識させたい場合を考える。バナナの画像を大量に集めて訓練データDを作って学習に利用することになるが、これは、バナナの概念をデータDで定義しているとも言える。Dが実物のバナナの自然画像のみからなる場合、その集合からこぼれ落ちたバナナの概念、例えば、絵画で描かれたバナナや食品サンプルのバナナを、DNNはうまく認識できないことが多い。

これまで、このような課題への対策は、訓練データを増やすことしかなかった。問題によっては、コストを費やせば着実にデータ量を増やせる場合があり、現に表1の応用事例では、このアプローチが採られてきた。しかしその場合でも、性能が費やしたコストに比例して向上するわけではない。両者の関係は一般に、図1（c）のようなS字カーブを描くと言える。データの増加が性能向上に素直に反映されるのは、最初のうち（データ量が少ない場合）だけであり、ある程度性能が向上した後では、その効果は減弱する。性能が上がれば上がるほど、より稀なサンプルでなければ、性能向上に貢献しないからである。

以上を分かった上で問題となるのは、実際に求められる性能を、許容できるコストの範囲内で達成できるかどうかである。問題（や

条件）次第でこれは真である場合も、また偽である場合もある。医療応用や自動運転など求められる性能が高いほど、当然ながら達成できない（=コストが高すぎる）可能性が高い。またそもそも、コストに関係なく十分な量のデータを集められないこともよくある（稀な病気は、サンプルを集めること自体が難しい）。

3. 深層学習の新たな潮流：自己教師学習

以上のような課題を克服すべく、研究が行われてきた。現時点での答えの1つは、自己教師学習と呼ばれるアプローチである。自己教師学習とは、プレテキスト（pretext=うわべの）タスクと呼ばれる擬似的なタスクを、目的とする本命のタスクを学習する前に学習（事前学習）しておいて、その結果を目的タスクに転移（転移学習）する方法である（図2）。後半の転移学習には、これまで同様目的タスクの訓練データを要するが、一般に比較的少数で済む。

自己教師学習の狙いは、プレテキストタスクの学習によって、DNNに入力xの良い内部表現を獲得させることにあり、これは「表現学習」とも呼ばれている。この考え方自体は、従来からの基本的な転移学習のそれと同じだが、自己教師学習の肝は、事前学習するプレテキストタスクが、正解がコストフリーで手に入るタスクであることにある。つまりデータおよび学習の規模を好きなだけ大きくすることができ、その恩恵を受けられる。

物体認識への適用では、1,000クラスの物体を認識するのに、従来は約100万枚のラベル付き画像を教師あり学習していたところ、自己教師学習を用いると、それと少なくとも同等以上の認識性能を、1万枚程度のラベル付き画像（つまり1/100のデータ量）の学習で達成できている[2]。（なお、画像分類でのプレテキストタスクには、1つの物体が写る1枚の画像から2つの部分画像を切り出し、それぞれに異なる（内容が変わらないと見做せる範囲の）画像変換を施した時、その画像をDNNに入れて取り出した特徴ベクトルが不変であることを要請する（距離計量学習の一種）ものが選ばれる。）

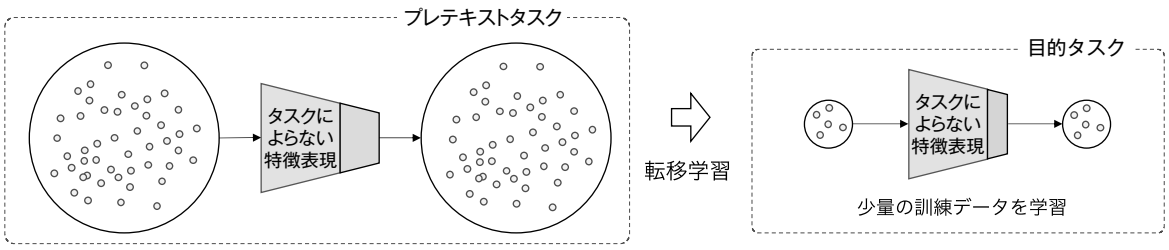

図2　自己教師学習の概要。正解がコストフリーで手に入るプレテキスト（pretext＝うわべの）タスクを事前学習し、入力のリッチな特徴表現を獲得することで、目的とするタスクを少ないデータの学習で解決する。

4. テキストベースの知識を導入した高度な推論へ

自己教師学習は、自然言語、音声認識、画像と、AIのあらゆる分野で有効性が確認されている。中でも、自然言語の分野での「言語モデル」の登場（あるいは再発見）は、AIのフロンティアを大きく前進させた[3]。

読者が今、読んでいるような自然言語の文は、単語の並び（系列）として表現できる。言語モデルとは、そのような文を構成する単語の（統計的な）並び方を捉えて表現する方法である。言語モデルを使うと、例えば文章をその途中まで与えたとき、その次の単語（候補が複数あるなら、それぞれの確率）を予測することができる。予測される単語は、それまでの単語の並びと文法的にも意味的にも、整合するものでなければならないが、ウェブ上の大量のニュース記事や、Wikipediaの記事などのテキストデータを片っ端から自己教師学習すれば、言語モデルはそんな条件を満たすようになる。

こうして作られた言語モデルは、新たな文を生成することができる。文を途中まで入力し、次の単語を予測させた後、それを元の文に追加し、再び言語モデルに入力すれば、さらに1つ先の単語を予測できる。これを延々繰り返せば、長い文章を紡ぎ出すこともできる。数千億オーダのパラメータを持つ巨大な言語モデル（GPT-33）など）は、この文を生成（途中までの文を補完）する能力が極めて高く、その能力だけで、人と高度な「対話」を行えることが示されている。さらにこの能力は、従来の機械学習の考え方を大きく変えるものであることも分かってきた。言語を使う様々な問題――Q&A（クイズ）や試験問題の回答（自由回答や選択肢を選ぶもの

まで）など――が、対話の形式に落とし込むだけで、解くことができると分かってきたのである。これは、先述の完全教師あり学習を基本とする従来のAIのあり方とは大きく異なる、全く新しいパラダイムである。

今後、この言語モデルの能力が、言語で閉じた世界を超えて、他のモダリティ、特に画像と融合するのは時間の問題である。これまでにも、画像と言語の境界領域にある諸問題が活発に研究されてきた。例えば、ある情景の画像を見てそれをテキストで記述する問題（image captioning）や、情景の画像とそれに関する質問文（テキスト）の2つを入力に受け取り、質問に回答する問題（visual question answering）などである。ただしこれまでは、これらの問題に対しても、従来の完全教師あり学習が適用されていた。今後は、言語モデルが何らかの形で取り入れられ、重要な役割を果たすと予想される。

この新しい研究のトレンドが示唆する未来は次のようなものである。世の中の知識の多くは、テキストデータとして表現されている。言語モデルは、そんなテキストデータを学習する（一種の記憶とも言える）ことで、そこに記された知識を、多様な問題解決に活用できるようになるだろう。これによって、画像などを対象とした推論に、テキストベースの知識を導入できる可能性が生まれる。応用の視点で言えば、例えば、インフラ構造物の維持管理において、橋梁の健全度を診断する問題などへの応用が可能かも知れない。これまでのこの種の問題へのAIの適用は、今となっては単純とも言えるパタン認識的アプローチであり、例えば画像から何種類かの変状（異常）を検出するだけであった。今後は、橋梁の構造や工法、さらには橋の点検に関する知識を取り込むことで、より高度な診断を行うAIを実現できるようになるかもしれない。

5. おわりに

これまでのAIの応用の成功例は、完全教師あり学習、すなわち、厳密に入力と出力を指定し、その間の写像を表現するDNNを考え、これを入出力の正解ペアを与えて学習するという方法に基づいていた。この方法は訓練データへの依存度が高く、そのことから様々な限界が生じ、問題によっては現実世界への適用が難しいこともあった。これに対し近年、自己教師学習という新たな方法が生み出され、従来の課題を完全に解決するにはまだ至らないが、AIの研究を確実に前へと進めた。

また、完全教師あり学習に基づく「従来のAI」は、人の認知的情報処理の一切片を切り取った「認知過程の自動化（cognitive automation）」というべきものであり、本当の意味での知能（intelligence）とは、かなり遠いものであった。これに対し、最近登場した巨大言語モデルは、自然言語の分野の問題解決の方法論を一変させるとともに、AIを、より本物の知能に近付けるものと言えそうである。少なくとも応用の観点から言えば、社会でこれまで蓄積されてきたテキストベースの膨大な知識が、言語モデルを通じてAIによる問題解決に様々な形で活用できるようになりそうである。その結果として、これまでは難しかった高度な人の作業を代行するAIの実現が期待される。

参考文献

1）岡谷貴之、深層学習（機械学習プロフェッショナルシリーズ）第2版、講談社サイエンティフィク(2022).

2）Kaiming He Haoqi, Fan, Yuxin Wu, Saining Xie, and Ross Girshick, Momentum contrast for unsupervised visual representation learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 9729–9738.

3）Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla, Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).