1A5-OS-27c-2

Deep Learningから身体性、シンボルグラウンディングへ

松尾 豊 *1
Yutaka Matsuo

*1東京大学
University of Tokyo

Deep Learning is a recetly-developed set of machine learning techniques which attracts much attention. The most prominent part of the techniques is its ability to recognize things, and typically the representation composes a layered-structure. Various studies on the integration of deep learning with reinforcement learning and image generation are proposed so far. In this paper, we address how the deep learning can realize the long-term dream of AI such as embodiment and symbol grounding. Especially, we discuss on how recent techniques on generative models can adress to the good-old AI research such as SHRDLU, subsumption architecture, and machine translation.

1. はじめに

近年、ディープラーニングが注目を集めている。ディープラーニングとは、深い層を重ねることでその学習精度を上げるように工夫したニューラルネットワークを用いる機械学習技術のことである。2006 年に Geoffrey Hinton らが教師無し学習を反復的に用いることで深い階層のニューラルネットワークで精度を上げることに成功して以来 [8]、さまざまな手法が提案され、2011 年には音声認識のタスクで優勝、2012 年には ILSVRC という一般物体認識のコンテストで圧勝するなど、数多くのコンペティションで成果を収めてきた [1, 11]。2013 年には、Google が DNNresearch という Hinton とその学生らが立ち上げた会社を買収、2014 年初頭には、DeepMind というディープラーニングと強化学習を有する会社を 4 億ドルで買収した。中国検索最大手の Baidu は、スタンフォード大学で GPU を使ったディープラーニングの研究を進めていた Andrew Ng を招き、ディープラーニング研究所を作った。Facebook は、ディープラーニングの主要な研究者であるニューヨーク大の Yann LeCun をトップに据えて人工知能研究所を設立、ニューヨーク、シリコンバレー、パリに拠点を広げている。ICLR*1、NIPS*2、ICML*3 などのディープラーニングと関連する国際会議も、ここ数年は急激にその参加者を増やしている。

ディープラーニングがなぜこのように注目を集めるのだろうか。その立場には大きく 2 つあるように感じる。単なる流行であるという立場と、大きなブレークスルーであるという立場である。前者の立場から見ると、人工知能あるいはニューラルネットワークに関するブームは、歴史的には何回も繰り返されている [24]。現在用いられている手法も、ほとんどが昔からあるものであり、その用途も現在のところ画像認識や音声認識などに限定されている。したがって、今回が真の突破口であると信ずる客観的な理由はないとするものである。後者は、ディープラーニングが今後起こる大きな変化の突破口であると考える。その理由は、人工知能の分野で議論されてきたさまざまな難問において、結局のところはデータをもとにして特徴量を抽出するところに最も大きな困難性があり、それがいま、「現実的な方法で」「実際に」解けるようになっているからである。

私自身の立場は後者である。例えば、情報検索の研究は 1970 年代からあったが、1990 年代後半、インターネットの普及という環境を得て、一気に花開いた。インターネット広告という収益の手段をもち、技術者・研究者を圧倒的に集め、その後の検索エンジンの研究はもはや大手の検索エンジン企業でなければ事実上不可能になった。いったん産業界における収益化の手段と結びついた後の学術研究は（特に米国においては）すさまじい発展を見せる。同じように、ディープラーニングの技術も計算機環境の進展とデータの拡大という環境を得て、一気に花開くのではないだろうか。（その際に、産業界における収益化の手段と学術研究をいかに結びつけるかが日本においては最大の鍵だと感じている。）

私は、今までの人工知能研究は何も間違っていないと思う。古くは、1980 年ごろに福島の提唱したネオコグニトロン [25] が視覚的な認識の仕組みとして結局のところ正しかったように、これまでの人工知能研究の多くは正しいものであると思う。ただし、現実のデータから学習できるだけのデータ量と計算機のパワーがなかった。

本稿では、古くから議論されている身体性、あるいはシンボルグラウンディングに焦点をあて、ディープラーニングを用いることでどのように従来の議論が捉えることができるのかを述べる。そのために、SHRDLU や包摂アーキテクチャ、述語論理などの概念を再訪しながら、その再解釈を試みる。特に、ディープラーニングにおける生成モデルが、最も重要な技術要素であることを述べる。

2. SHRDLU をもう一度考える

まず、本稿での議論を、SHRDLU から始めよう。SHRDLU とは、人工知能研究の初期に行われた有名な研究であり、1968 年から 1970 年にかけて Terry Winograd が開発したシステムである [17]。画面の中の「積み木の世界」に、ブロックや円錐、球などが存在し、ユーザからのさまざまな質問に自然言語文で答えることができた。例えば、「円錐は何に支えられているか？」などである。また、自然言語文の命令により動かすことができた。例えば、ユーザは、「緑色の円錐を赤いブロックの上に置け」と指示した後、「その円錐を取り除け」と指示することができた。

Winograd はこの研究の後、人工知能研究をやめ、HCI の研究を行うようになった。次々と先進的な研究を生み出し、研究室からは Google の創業者も輩出したわけだが、Winograd

連絡先: 松尾 豊，東京大学 大学院工学系研究科，東京都文京区弥生 2-11-16，03-5841-7718，matsuo@weblab.t.u-tokyo.ac.jp

∗1 International Conference on Representation Learning
∗2 Annual Conference on Neural Infomation Processing Systems
∗3 International Conference on Machine Learning

はなぜ自然言語処理の研究を辞めてしまったのだろうか。日本でもこうした変化をたどった研究者の方は何人もいるのではないかと思う。おそらくとんでもないほどの絶望感を感じたのだろう。積み木の世界はすべて人間が設計した世界であり、ありとあらゆるお膳立てをして、ようやくコンピュータが少し知的に見える振る舞いをすることができるようになる。裏を分かっている開発者にとっては、こうして実現されるものと人間の知能には呆然とするほどの距離がある。

しかし、私は、この SHRDLU に代表される積み木の世界の研究が、知能の重要な側面を捉えようとしていること自体は何も間違っていないと思う。そして、それがいま、ディープラーニングを突破口に新たな展開を見せつつあるのではないかと考える。

その鍵となるのが、ディープラーニングにおける生成モデルである。通常、機械学習においてクラス分類を解くための手法は、識別モデルと生成モデルに分けられる。識別モデルは、データ X が与えられた時のクラス C の条件付き確率をモデル化し、生成モデルは X と C の結合確率をモデル化する。ディープラーニングで良く使われる畳み込みニューラルネットワークは識別モデルであるが、生成モデルも近年、数多く提案されている。

有名なものには、VAE（Variationoal Auto-Encoder）がある。視点を入れた拡張である DRAW（Deep Recurrent Attention Writer）がある [7]。また、GAN（Generative Adversarial Network）[6] は Goodefellow らが提案したものであり、生成器と識別器から構成され、互いに騙そう、騙されまいとすることにより、精度をあげる。これを拡張した、LAPGAN（Laplacian Pyramid of Generative Adversarial Networks）[3] も有名である。

これらを使うと、物理世界での動きを「予想」することができるようになる。例えば、ボールがはずむ動きなどを予想することができる [16]。また、Fragkiadaki らは、ビリヤードの球の動きを、Convolutional Neural Netowrk (CNN) と Long Short-Term Memory (LSTM) で学習させた。さらに、ある状況で特定の方向に力を加えると、どのようなことが起こるかを「想像」する [5]。これを使って、行動の計画を立てることができるというものである。また、Oh らは、ATARI のゲームを題材に、アクションを挟み込んだオートエンコーダでフレームを学習することにより、特定の行動を行うと次に何ができるかを予測している [15]。それにより、Deep Q-Network (DQN) を使ったゲームのスコアが向上する。

これらが示しているのは、明示的に積み木の世界を作らなくても、ディープラーニングの生成モデルを使うことによって、その世界を描くことができるということである。従来は人間が細部までお膳立てをしないといけなかったという状況なしに、SHRDLU で目指していたような「どういう行動をすれば何が起きるか」をシミュレートすることができ始めているのである。

「生成モデルで世界をシミュレートする」というのは単純なアイディアだが、これをベースにして、さまざまな可能性が広がっていると考える。以降では、身体性への拡張（センサ情報だけではなくアクチュエータの情報も含んだ拡張）を述べ、その上で、自動翻訳（生成モデルで作った世界と言語との結びつきによる言語の意味理解）、さらには、述語論理等による推論（生成モデルで作った世界の記号的な要約）について述べる。これまでの人工知能研究で目指していたものが、生成モデルによる世界シミュレータによって実現が大きく近づくのではないかという本論文の主張である。

3. 身体性をもう一度考える

人工知能を実現する上で、身体性が必要かどうかは、長らく論争の的である。MIT の人工知能研究所所長であった Rodney Brooks は、表象なき知能（Intelligence without representation）という考え方を提案した [2]。知的能力は、「生存と生殖を最低限保持するのに十分なほど周囲を知覚し、動的な環境世界を動き回ることのできる能力」を土台として、その上に築かれるべきだとした。そして、昆虫型ロボット等によって、環境とインタラクションするだけで、十分に知的に見える振る舞いが出現することを示した。

ディープラーニングによる認識の能力が飛躍的に向上することで、まさに Brooks の言うような身体性（embodiment）の研究が重要になると思う。ここでいう身体性は、必ずしも現実の身体を意味するわけではなく、記号システムの上になりたつ身体性（例えば、囲碁や将棋における記号創発のための身体性）であっても良いだろう。ただし、注意しなければならないのは、現実は驚くほど複雑で非線形であるということである。Marvin Minsky は、心の社会という本のなかで、紙を組み合わせて置くだけでなぜ「閉じ込め」という機能が発生することについて議論したが、この例は世界の不思議さを最も端的に表していると思う。それぞれの紙には「閉じ込め」という機能はない。しかし、それが組み合わされて四方を囲んだ途端に、「閉じ込め」という機能が出現するのである。人工物の機能、都市、生物、組織、どれひとつをとっても非線形な機能の創造に満ちあふれている。それを考えると、現象をモデル化し、その上での身体性を考えることも可能ではあるが、モデル化によって現実の非線形性が大いに削られてしまう。現実を対象にしない身体性は、どうしても「ちゃちな」ものになってしまうため、現実的には、人間の身体と同じように、現実世界と相互作用できる装置が必要だろう。

ディープラーニングとロボットを組み合わせた研究はすでに始まっている。UC Berkeley では、強化学習＋ディープラーニングをロボットに対して適用する、Sensorimotor ディープラーニング（感覚運動ディープラーニング）のプロジェクトが進められている [12]。組み立てロボットがさまざまなタスクを試行錯誤により学習し、おもちゃの飛行機を組み合わせたり、レゴのブロックを組み立てたり、木の輪をくいにはめ込んだりすることができる。日本でも、Preferred Networks 社や、早稲田大学の尾形 [14] らがディープラーニングとロボットを組み合わせた研究を行っている。さらに言えば、今回のディープラーニングのブームのずっと前から、谷らは、センソリモータ・フローの分節の問題を長らく研究している [22]。浅田、國吉らは、認知発達ロボティクス [20] というテーマで、また谷口らは、記号創発ロボティクスというテーマで研究を行っている [21]。川人らは、脳科学の知見を活かしながらロボットで強化学習を行うさまざまな研究を 1990 年代から行っている [10]。

ディープラーニングによる認識の精度が急激に向上したことは、こうした研究にも相当な追い風になると考えられるが、課題もいくつかある。ひとつは、多数回の試行錯誤にも頑健に耐え得るハードウェアの構築である。将来的には、アルゴリズムとともに、ハードウェア自体も進化をさせていく必要があるだろう。また、ハードウェアの頑健性と裏表であるが、学習に必要なサンプル数をいかに減らしていくかである。すでにさまざまなアプローチが議論されている。i) 教師あり学習を使う（人間を教師とした見まね学習）、ii) 学習結果を共有する、iii) 階層タスクネットワークを使う、iv) 他のドメインから転移する、v) ソーシャルメディア（YouTube 等）からの見まね

学習を使う [19]、などが主なアプローチであろう。このうち、階層タスクネットワークを使う方法は、まさに、Brooks の包摂アーキテクチャを、ディープラーニングという現代的な武器を得て、実現する方法なのかもしれない。

4. 記号処理を再び考える

Minsky は、Brooks の表象なき知能という主張を痛烈に批判した。動物程度の知能は実現できても、人間の知能にとって記号操作は必要不可欠だという立場であった。私は、両者の主張はどちらとも正しいと思う。環境世界を知覚する仕組みの上に、世界を予測する生成モデルが築かれ、その上にさらに記号の操作が実現されるということではないだろうか。

現在のディープラーニングの研究で、これの端緒にあたるものは、2つある。ひとつは、画像から文を生成する自動キャプションづけである [9]。画像を与えると例えば「ピンクの服を来た女の子が芝生の上でジャンプしている」などの文を生成するものである。[18] では、画像中で最も注目すべき点に焦点を当てて、文を生成する手法を提案している。もうひとつは、画像の生成であり、文が入力されると画像を生成するようなモデルを学習することができる。例えば、[13] では DRAW を使って、「飛行機が空を飛んでいる」「像が砂漠を歩いている」などの文を入れると、該当する画像を描くことができる。「止まれ標識が空を飛んでいる」などの普通あり得ない文を入れても生成できるところが興味深い。これは、すなわち、自然言語文から画像を生成することができ、さらには画像から自然言語文を生成することができるということである。再度、SHRDLU の自然言語による操作と同じことが実現できるわけである。自然言語文の「意味を理解」していると言えるかもしれない。

さて、人工知能の分野では、意味理解についての議論も古くからある。Alan Turing は、チューリングテストを提案し、「計算機は考えることができるか」という問いを、「模倣ゲームをうまく行うことのできるような想像上の計算機は存在するか」という問いに置き換えた。それに対して、John Searl が 1980 年に示した思考実験が「中国語の部屋」である。仮にチューリングテストに合格する機械ができたとしても、操作している対象の「意味が分かっていない」というものである。

文の意味が分かるとはどういうことだろうか。私は文の意味を理解するとは、文から画像を生成することができることだと思う。ここで、画像というのは、視覚的な情報を分かりやすく表現した言い方であり、実際には、センサとアクチュエータの複合的な時系列情報であるので、体験という言葉のほうが適切であろう。つまり、意味理解ができるというのは、文から体験を生成し、あるいは体験から文から生成できる相互変換能力ではないだろうか。

ディープラーニングの生成モデルを使えば、例えば、日本語の文から体験を生成し、それを英語の文に変換するということができるかもしれない。すなわち、自動翻訳、しかも意味理解を伴う自動翻訳が可能になるかもしれない。

もちろん、本格的な自動翻訳を実現するには、たくさんの課題が思い浮かぶ。

- 抽象的な概念をどのように扱うのか。人間の意味理解が、視覚情報や視覚的な処理機構をベースにしているのは確かそうであるとしても（抽象的概念でも空間的な扱いをするものが多い）、映像として再現することは難しい概念もたくさんある。センサ・アクチュエータの高次の特徴量が復元されるということでいいのだろうか。
- 感情や本能等に関わるものをどのように扱うのか。例えば、美しい、おいしいといった感覚は、学習はできるにしても、人間と同じような報酬系が実現されているわけではない。その設計をしなくとも（例えば納豆が嫌いでほとんど食べたことがない人でも、納豆を食べる人を観察して学習できるように）ある程度何とかなるのだろうか。
- 人間と同じセンサ・アクチュエータ系がないと、人間と近い（あるいは理解し得る）概念を生成することは難しいのか。

もうひとつ、考えなければならないのは、記号処理により、いかに見えてないことを予測するかである。人工知能で長らく研究されてきた命題論理や述語論理、あるいは様相論理などによる推論 [23] は、与えられた知識や事実から、最初からは見えていない帰結を導き出すための仕組みであった。

その点では、最近、注目を集めたアルファ碁の研究も意義深い [4]。過去の棋譜データや自己対戦のデータを用いながら、CNN を用いた上で、policy network, value network を構成していく。それによって、先読みを大幅に効率化している。そもそも、こういった思考ゲームがコンピュータに扱いやすいのは、世界モデルの構築なしに、シンプルなルールを記述しておくだけで、未来の状態を展開することができたところにあるのだろう。（つまり生成モデルによる世界のシミュレートをさぼることができたわけである。）ところが囲碁においては、一手一手の操作があまりにもプリミティブすぎて、結局は世界モデルの構築が重要な鍵であった。それに対して、アルファ碁では、CNN による盤面の認識に加えて、強化学習による policy network を構成することで、探索する範囲をかなり絞ることに成功した。

おそらく人間の場合は、視覚的な生成モデルをベースにしながら、こういうときにはこうなるという関係性を記号レベルの連接関係でも学習していく。すると、いちいち重い処理が走らずとも、簡略化して思考を先に走らせることができる。この記号の想起と、視覚的な生成モデルの組み合わせが、思考の過程であり、それを（生成モデルによる世界のシミュレータがないがゆえに）シンボルの想起だけに限定したものが、従来の述語論理や様相論理による推論ということができるのではないだろうか。

5. おわりに

本稿では、ディープラーニングが認識を高いレベルで可能にしたこと、特に生成モデルにより画像や映像の生成が可能になったことにより、これまで人工知能分野で古くから議論されてきた研究が再度、重要な意義をもつことについて議論した。生成モデルによる世界シミュレーションの構築が、従来の人工知能に欠けていたものであり、それとシンボルを組み合わせる処理こそが、これまで長らく研究されてきた人工知能の主要なテーマでもあり、そして Minsky の述べた人間の知能に必要不可欠な記号操作なのではないだろうか。

なお、本稿では十分に議論することができなかったが、こうした仕組みの上に、さらに、社会的な言語、コミュニティや群知能、他者理解、意識等の議論も行うことができるだろう。いずれにしても、ディープラーニングにおける生成モデルと、これまでに長い蓄積のある人工知能研究の融合こそが、今後の技術の進展の鍵となるのではないだろうか。

参考文献

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 2013.

[2] R. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1-3):139–159, 1991.

[3] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Proc. NIPS2015*, 2015.

[4] D. Silver et. al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[5] K. Fragkiadaki, P. Agrawal, S. Levine, and J. Malik. Learning visual models of physics for playing billiards. In *Proc. ICLR2016*, 2016.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *Proc. NIPS2014*, 2014.

[7] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *Proc. ICML2015*, 2015.

[8] G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006, 2006.

[9] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *arXiv*, 2015.

[10] M. Kawato. From "Understanding the brain by creating the brain" toward Manipulative Neuroscience. *Philosophical Transactions of the Royal Society B*, 2007.

[11] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[12] S. Levine, N. Wagnener, and P. Abbeel. Learning contact-rich manipulation skills with guided policy search. In *Proc. IEEE International Conference on Robotics and Automation (ICRA) 2015*, 2015.

[13] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov. Generating images from captions with attention. In *Proc. ICLR2016*, 2016.

[14] K. Noda, H. Arie, Y. Suga, and T. Ogata. Multimodal integration learning of robot behavior using deep neural networks. In *Proc. IROS2013*, 2013.

[15] J. Oh, X. Guo, H. Lee, S. Singh, and R. Lewis. Action-conditional video prediction using deep networks in atari games. In *Proc. NIPS2015*, 2015.

[16] M. Vincent, R. Memisevic, and K. Konda. Modeling deep temporal dependencies with recurrent grammar cells. In *Proc. NIPS2014*, 2014.

[17] T. Winograd. Procedures as a representation for data in a computer program for understanding natural language. In *MIT AI Technical Report 235*, 1971.

[18] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML2015*, 2015.

[19] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos. Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In *Proc. AAAI-15*, 2015.

[20] 國吉 康夫. 身体が脳をつくるーロボットを題材とした構成論的科学のアプローチー. **認知神経科学**, 11(1):17–22, 2009.

[21] 谷口 忠大. **記号創発ロボティクス 知能のメカニズム入門**. 講談社, 2014.

[22] 谷 淳. ロボットで「科学」する記号の問題. **日本ロボット学会誌**, 28(4), 2010.

[23] 石塚 満. **知識の表現と高速推論**. 丸善, 1996.

[24] 麻生 英樹, 安田 宗樹, 前田 新一, 岡野原 大輔, 岡谷 貴之, 久保 陽太郎, and ボレガラ ダヌシカ. **深層学習**. 近代科学社, 2015.

[25] 福島 邦彦. 位置ずれに影響されないパターン認識機構の神経回路のモデル— ネオコグニトロン —. **電子通信学会論文誌** *A*, J62-A(10):658–665, 1979.