# 深層学習で意味の深みへ

東京大学
松尾 豊

概要 : 本稿では、深層学習のこれまでの進展を踏まえた上で、意味をどのように扱うことができるのかを議論する。そのために、深層生成モデルによる画像の生成をベースとし、そこからの技術進展を考えることを提案する。次に、深層学習の関連研究を紹介し、どこまでが現在の技術で可能なのかを述べる。最後に、記号が生まれた進化的な意義からの考察を加え、人間の知能に特有な部分がアルゴリズム的にはどこに相当するのかを述べる。

## 1. 意味とは何か

本稿では、以下の仮説をベースに議論を進める。

0. 原始的な意味とは、ある記号を見たときに、脳神経回路内に再現される「絵」である。

意味に関する研究の長い歴史を踏まえると、乱暴な仮説であることは承知の上で、ここから議論をスタートしたい。りんごの意味は、りんごの絵であり、馬の意味は、馬の絵である。意味を理解するとは、脳内に絵を描けることである。ここでの絵は構成性原理を満たしており、馬の絵にりんごが加わると、馬がりんごを食べる絵にもなる。したがって、りんごは馬にとっての餌という意味を取り出すこともできる。

「絵」というのは最も単純化した表現であるので、これにいろいろな拡張が加わる。

1. まず、時間的な概念が加わる。馬の意味は馬の絵ではなく、時間発展を伴う馬の絵、つまりは馬の「映像」である。
2. 次に、視覚以外のモダリティが加わる。視覚的な馬の映像だけに留まらず、馬のいななく声や、馬のたてがみの感触も含む。つまり、馬に関しての「センサ情報の集合」である。
3. そして、アクチュエータも加わる。手綱をもって自分が馬をひく、そのときの手足の動き。それに対して、馬が従ったり暴れたりする感じ。馬に跨るときの手足の感覚、馬が手綱や足の指示にしたがって反応する様子などである。センサだけでなくアクチュエータの情報も含む集合であり、つまりは「体験」である。
4. さらには、抽象度が上がる。馬にのったときの感じと車にのったときの感じは似ている面があり、抽象化して「乗り物にのった感じ」になる。馬と牛を飼育するさまは似ており、「家畜を育てる感じ」になる。そして、「絵」はもはや視覚的に見える「絵」ではなくなる。

以下では、これを深層学習の視点から実装することを考える。

## 2. 深層学習で実装する

深層学習により、「絵」を描く技術はすでに存在する。深層生成モデルという。特に、Generative Adversarial Network（GAN, 敵対的生成ネットワーク）[Goodfellow14]、Variational AutoEncoder（VAE, 変分オートエンコーダ）[Kingma13]がその代表例である。DCGAN、LAPGAN、BigGAN など、さまざまな進化系が知られており、最近では本物の写真と全く区別のつかない、高解像度で非常にきれいな絵を生成することもできる。また、CycleGAN やその進化系では、写真のなかの馬をシマウマに変えたり、昼を夜に変えたりすることができる[Zhu17]。GAN では、ノイズをもとに画像を生成する生成器と、本物の画像か生成した画像かを識別する識別器が、互いに競い合うことによって、より本物に近い画像が描けるようになる。また、VAE では、画像からいったん低次元の潜在空間に縮退させ、そこから画像を生成する。その際の再構成誤差が少なくなるように学習を行うことで画像を生成できるようになる。

本物に近い絵を描けるだけでなく、さまざまな条件をつけて絵を描くこともできる。そのために、GAN や VAE に対して潜在変数でその出力を条件づける Conditional GAN[Mirza14]や Conditional VAE[Kingma14]、およびその多数の進化系が知られている。潜在変数をいじることで例えば、人物の顔をより若くしたり、女性よりにしたり、メガネをかけさせたりできる。こうした潜在変数と、「メガネ」などの言葉が結びつけば、基本的には、「メガネをかけた女の人」などの絵を描くことができる。フレーズあるいは文から画像や映像を生成する技術も進んでいる。Microsoft COCO や Visual Genome、VQA（Visual Question Answering）を改良した GQA[Hudson19]などのデータセットを用いると、文と画像のアライメントを学習あせることができる。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[Mansimov15, Reed16]。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段階で行う手法を提案している[Zhang17]。最新の BERT という手法を用いた、画像と文のアライメントの手法も提案されている[Li19]。

動画から次のシーンを予測することもできる。映像予測（video prediction）という技術であり、GAN や RNN（Recurrent Neural Network）を使ったりすることで短い将来（例えば１秒後）のフレームを予測する[Mathieu15]。

マルチモーダルな情報に関しては、画像と結びつけた音の生成[Owens16]などの研究があるが、映像と音声などを結びつけた予測は、まだ活発ではない。尾形らは、映像と音声など、複数のセンサ入力を同時に扱う研究をしている[Sasaki16]。著者らは、マルチモーダルな VAE のモデルを提案した[Suzuki16]。

センサ情報だけでなく、アクチュエータに広げるとどうだろうか。ロボットを使って深層強化学習を使って、例えば把持や移動等の動作を行わせる研究は盛んである。自己教示学習という、ある種の教師なし学習が重要であることも指摘されている。シミュレータと実機との融合、sim-to-real と呼ばれる研究や、ドメイン適応と呼ばれる研究が、深層学習の領域で急激に進んでいる。しかし、センサとアクチュエータの情報、つまり「体験」を言葉から再現するとなると、研究としてはほとんどない。尾形らの研究[Zhong19]がこれに該当する数少ない例のひとつである。

さらにセンサとアクチュエータの複合体を抽象化するような試みになると、現在のところ皆無である。

したがって、1節で述べたような意味理解の仕組みは、まだ道半ばであり、0から1の段階が研究として行われており、2や3の段階はまだごく少数という状況である。しかし、昨今の急激な進展を考えると、このまま順調に推移すると近いうちに、かなり4に近づいてくるのではと思われる。

## 3. 進化的な意義

そもそも生物は、自己の保存、再生産を目的としている。(というより、そうしたものが生き残ったのでそのような目的を持っているように見える。) 環境が変化しやすいときには、環境への対応を学習で、そうでなければハードワイヤードな生得的な仕組みで実装したほうがよい[Hinton87]。

学習においてサンプル効率というのは大変重要な概念で、できるだけサンプル数が少なく、できるだけ自由度の高いモデルを同定したい。サンプル数が少なくて済むほうが、生存上の危険を冒さずに済むし、自由度の高いほうが、柔軟に「世界モデル」を獲得でき、賢い振る舞いができる。

進化的に考えると、原始的な記号のそもそもの成立は、プランニングに資するためではないか。環境中の情報を知覚し、それに応じて適応的な行動を取る知覚運動系は、さまざまな動物の基本的な行動を司る。これはX1という感覚入力に対してY1という行動出力を行うことを考える。そして、Y1という行動の結果Z1が予測できるとき、Y1という行動がよいのか、はたまたY2やY3という行動がよいのかを比べて、最もよい帰結Zをもたらすであろう行動(例えばY2)を選択することができる。このとき、頭の中では、Y2という行動をとったとした条件のもとでのシミュレータが、深層生成モデルによって動いている。

さらには、Y2という行動が帰結Z2をもたらし、その次にさらにY4という行動をとれば、Z4と結果が得られるというふうに、シミュレータを前向きに動かしていくことで、より長期の予測ができる。そしてよい結果になると分かったときにだけ、「あっ、これだ!」と行動すればいいのである。

そういうわけで、一部の動物(高等な哺乳類)には、こうしたシミュレータが備わっているように思える。これを深層学習の分野では、最近、「世界モデル」(world models)と呼んでいる。従来、内部モデル、メンタルモデルと言われていたものと近い。この世界モデルに関する研究は、深層学習の分野でも2018年以降、盛んである[Ha18, Eslami18]。ほとんど何も仮定していない状態から、VAE等の技術で、環境の要因をdisentangle(もつれを紐解く)し、そうして得られた要素の遷移モデルをRNN等でモデル化し、学習する。このように環境やエージェントの状態を表す状態表現や、その遷移をデータから獲得することの重要性は最近でも指摘されており[Mahadevan18, Hamrick19]、DeepMindのCEOであるデミス・ハサビスは、知能の研究において最も重要なのは想像(imagination)とプランニングであると述べている[Hassabis17]。

ところが、人間の場合、最も重要な進化的な変化は、X1という感覚入力に対して上記のプロセスが発動するのではなく、L1という言葉の入力に対して、X1という感覚入力が生起し、プロセスが発動するということではないか。これによって、現在、X1という状況にないにも関わらず、L1という言葉によってX1という感覚入力(あるいはX1の感覚とY1の行動)を疑似体験することになる。これは、ハラリのいう「虚構」を共有することに相当し、「集団としての結束を高めた」作用があったのではないか[ハラリ 16]。言語による神などの宗教的概念、敵や味方などの概念を持つことはプラスの作用をもたらし

た。

そして、次の大きな変化が、「L1 は L2 だ」というような文を伝えることである。L1 という感覚入力と、L2 という感覚入力を同時に生起させる。（例えば、「馬から落ちると、めちゃ痛い。」）これは、深層生成モデルによって、擬似的にデータを生成していることに相当し、学習データに組み入れることができる。

そして三段階目の変化は、「L1 は L2 だ」とか「L2 は L3 だ」のような記号的な運用規則の適用そのものを、「疑似体験」として概念化し、可能な行動のひとつにしてしまうことである（ここにある種の再帰性が発生する）。これによって、抽象的な概念の操作、あるいは、ルールによって定義された概念操作を、自由に行えるようになる。これが人間の記号的な思考の奥深さを生み出したのではないだろうか。

## 4. 結論

本稿で言いたいことを別の言い方でまとめると、意味とは、深層生成モデルによって潜在変数からデータが生成されることである。これは、その深層生成モデルがどのように訓練されたかによって大きく異なるが、画像や映像、マルチモーダルなセンサ、センサとアクチュエータの複合体、そしてそれらを抽象化したものまでさまざまな段階がある。言語と潜在変数が結びつくことにより、複数の潜在変数を指定したデータの生成が可能になる。これが「想像」であり、この想像を記号的な処理自体にも適用することで、高次の潜在変数を得られることになる。

深層学習の急速な進展を考えると、上記のような議論が現実にシミュレートできる段階が近いと思われる。そのために、今後の研究として重要なのは、以下の２つであると考えている。

(1) 深層生成モデル、特に「世界モデル」と呼ばれる、世界の時間発展をモデル化するような技術。ロボットを使った深層強化学習にも大きく寄与すると考えられる。
(2) 記号が、学習されたニューラルネットワークのモデルに統合される仕組み。

このような技術の進展をガイドするために、これまでの意味論の研究、あるいは言語哲学の研究と深層学習の融合が重要である。

[Goodfellow14] I. Goodfellow et al: Generative Adversarial Networks, Proc. NIPS2014, 2014
[Kingma13] D. P. Kingma and M. Welling: Auto-Encoding Variational Bayes, Proc. ICLR2013, 2013
[Zhu17] J. Zhu et al: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, 2017
[Mirza14] M. Mirza, S. Osindero: Conditional Generative Adversarial Nets, 2014
[Kingma14] D. P. Kingma, D. J. Rezende, S. Mohamed, M. Welling: Semi-supervised learning with deep generative models, Proc. NIPS2014, 2014
[Mansimov15] E. Mansimov, E. Parisotto, J. Lei Ba, R. Salakhutdinov: Generating Images from Captions with Attention, 2015
[Reed16] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee: Generative Adversarial Text to Image Synthesis, Proc. ICML2016, 2016

[Hudson19] D. A. Hudson, C. D. Manning: GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering, 2019
[Zhang17] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. Metaxas: StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, Proc. ICCV2017, 2017
[Li19] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, K. Chang: VisualBERT: A Simple and Performant Baseline for Vision and Language, 2019
[Mathieu15] M. Mathieu, C. Couprie, Y. LeCun: Deep Multi Scale Video Prediction Beyond Mean Square Error, 2015
[Owens16] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, W. T. Freeman: Visually Indicated Sounds, Proc. CVPR2016, 2016
[Suzuki16] M. Suzuki, K. Nakayama, Y. Matsuo: Joint Multimodal Learning with Deep Generative Models, 2016
[Sasaki16] K. Sasaki, K. Noda, T. Ogata: Visual Motor Integration of Robot's Drawing Behavior using Recurrent Neural Network, Robotics and Autonomous Systems 86, pp.184-195, 2016
[Zhong19] J. Zhong, M. Peniak, J. Tani, T. Ogata, Angelo Cangelosi: Sensorimotor Input as a Language Generalisation Tool: a Neurorobotics Model for Generation and Generalisation of Noun-Verb Combinations with Sensorimotor inputs, Autonomous Robots 43(5), pp.1271-1290, 2019
[Hinton87] G. Hinton and S. Nolan: How Learning can Guide Evolution, Complex Systems, No.1, pp.495-502, 1987
[Mahadevan18] S. Mahadevan: Imagination Machines: A New Challenge for Artificial Intelligence, Proc. AAAI2018, 2018
[Hamrick19] J. B. Hamrick: Analogues of mental simulation and imagination in deep learning, Behavioral Science, 2019
[Ha18] D. Ha, J. Schmidhuber: World Models, 2018
[Eslami18] S. M. Ali Eslami et al.: Neural scene representation and rendering, Science, No.15, Vol.360, 2018
[Hassabis17] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick: Neuroscience-Inspired Artificial Intelligence, Neuron, 2017
[ハラリ 16] ユヴァル・ノア・ハラリ　(著), 柴田裕之 (翻訳)、サピエンス全史（上）　文明の構造と人類の幸福 サピエンス全史　文明の構造と人類の幸福、河出書房新社、2016