文部科学省科学研究費・新学術領域研究（2016～2020年度）
人工知能と脳科学の対照と融合

Correspondence and Fusion
of Artificial Intelligence and Brain Science

NEWSLETTER

AI AND BRAIN

Vol.5
2019.03

計算論的神経科学と人工知能

ATR脳情報通信総合研究所・所長　川人 光男

現在の人工知能のブームは、脳科学と人工ニューラルネットワークが原点です。もともとディープニューラルネットワークは、福島邦彦先生らの神経回路モデルから出発し、ノーベル生理学・医学賞を受賞したデイヴィッド・ヒューベル&トルステン・ウィーゼルの研究（大脳皮質視覚野における情報処理）に基づいていることをご存じの、一般の方は少ないでしょう。このような歴史をきちんと振り返り、なぜ人工知能のために脳科学が必要なのかを知ることは重要です。本新学術領域に参加されている皆さんは、啓蒙活動をする必要と義務があると思います。

10年前位までの人工知能に足りなかったビッグデータについては、たとえばイメージネットなどの数千万の画像データを研究者が使えるようになりました。それ以降、Google、IBM、Intel、Microsoft、Facebook といった企業が参入し、ビッグデータに頼った統計的推論が爆発的に進展しました。ディープニューラルネットワーク、ディープQ、英仏翻訳、画像解説、IBM TrueNorth などが注目を浴びています。しかし、このようなディープニューラルネットワークを含めた現在の人工知能は、決して万能ではありません。

その例として、今から4年前にアメリカの国防高等研究計画局（DARPA）が主催しDARPA Robotic Challenge（災害救助用のロボット競技大会）が開催されましたが、そこで出てきたロボットは動いているのか、いないのかわからないくらい動きが遅く、2/3くらいは転倒、転倒しなかったものも恐ろしく、動きが鈍いのです。囲碁だと人間が理解できない手まで繰り出す人工知能が、運動制御では圧倒的に人間に劣っているのです。これにはさまざまな要因がありますが、1つだけ理由を挙げるとすると、ロボット制御などの現実世界の問題では、学習用の訓練データが莫大には取れないことが挙げられます。高価なロボットを気安く転倒させるわけにはいきませんから、上記の例で、数千万回はおろか、数十回の失敗データを取得することも、現実的ではないと考えられます。

数十年にわたる人工知能研究は、人間が持っている知能のごく一部の機能を切り出して処理スピードを上げたり、容量を大きくしたりすることで、社会の役に立たせるものでした。しかし脳あるいは人間が持っているさまざまな機能は、運動学習、概念形成、シンボル生成、エピソード記憶、注意、意識などなど、語り尽くせないほどありますが、それらの殆どの機能に関して今の人工知能では実現できていません。できていないということは、私たちがヒトの知能を理解していないということを示していて、ダートマス会議で定義された本来の意味の人工知能は、いまだ存在していないのです。

新学術領域『人工知能と脳科学の対照と融合』は、代表者である沖縄科学技術大学院大学教授の銅谷賢治さんの計算理論（モデルフリー対モデルベース強化学習）のような脳科学の研究をアルゴリズムとして書き下すことができるなら、脳科学が真の人工知能確立に大きく貢献することでしょう。本領域も2年が過ぎ着実に成果を挙げており、大きな期待を持って見守っております。

［第５回］リレー対談

谷口 忠大
立命館大学 情報理工学部 教授

×

坂上 雅道
玉川大学 脳科学研究所 教授

谷口 今日は玉川大学からはるばる来ていただいてありがとうございます。

坂上 滋賀っていうと琵琶湖しか考えてなかったけど、こんなだだっ広いところにドーンと立命館があるのは、ちょっと驚きでした （笑）。

谷口 僕は坂上先生とほぼこの新学術から仲良くさせてもらったみたいなところなんですけど、まだまだこう、坂上先生のこれまでの遍歴を知らないんですよ。多分。例えば、学部での研究の始まりとか。脳研究されている方って、機械系出身の方とか電気系出身の方とかもおられて出自が学際色強い感じがするんですが。

坂上 私の出身は文学部の心理学で、しかもその時の先生は二木宏明先生っていってワーキングメモリニューロンを世界で初めて発見した有名な先生でした。そういう意味では大学３年から、完全にニューロサイエンスですね。

谷口 おお、じゃあ一番プロパーな!

坂上 まあそうなりますね （笑）。卒論はニューロンを扱ってないんですけど、ニューロン記録をするためにサルのSame-Differentの概念を行動的に調べていました。サルにも抽象化する概念があるよと。で、その後に、そのニューロン活動を記録したいと思ったんだけど、修士に入ったら 「修士は難しいことやらずに単純なことをやりましょう」と言われて。

谷口 ほう。修士だから単純なことを?

坂上 修士課程って2年しかないじゃないですか? まぁ、結局3年かかっちゃったんですけどね （笑）。帯状回っていう脳の場所が海馬に入力を送るいわゆる中継点になってるので、そこがどういう空間情報を持っているかっていうのを調べようと。いわゆる自己中心座標と外部中心座標の関係性を調べるっていうニューロン活動の記録を、サルを使ってやりましたね。その後、博士課程に入ってもっと続けようと思ったのですが、私の同級生がアメリカに逃げたことで、前頭前野を中心に研究してた二木先生のために働く人がいなくなっちゃって 「お前、前頭前野やれ」って説得されちゃって （笑）。

谷口 本当は、最初は前頭前野やる気じゃなかった。

坂上 そうそう、前頭前野やるつもりは全然なかったんだけど、博士課程でやれと説得されて、最初は嫌がってたんだけど、結局前頭前野をやることになったんです。ワーキングメモリーっていう概念が1980年代後半くらいから出てきて、前頭前野研究も少しずつ脚光を浴びるようになったのは、ラッキーでしたが。

谷口 Prefrontal（前頭前野）ですが、僕もまだまだ詳しくないんで、素人質問だと思ってほしいんですが、やっぱり人間が進化中で皮質、前頭前野を肥大化させたっていう話があるわけじゃないですか? 僕も記号創発ロボティクスで、言語獲得とかやっているわけなんですが、そもそも言語って人間しか持ってないっていう話があるから、動物を使った研究がしにくかったりすると思うんですが。

坂上 そういう意味じゃサルが人間に比較的近いということで、Cortex（皮質）の研究に使いやすいって意味じゃサルを使った研究にメリットはあると思うんですけどね。大脳基底核の機能なんてかなりサルの研究も貢献したんだけど、今になってマウス、ラット中心になってきた。値段が高くて実験に使いにくいサルを使って、低次な脳機能の研究をやる必要なんかないじゃないかと。

谷口 なるほど。だからこそサルでCortex（皮質）研究をしようみたいな。

坂上 本当はもっと我々が頑張ってCortex（皮質）の話をやらなくちゃいけないんだけど、あまりに複雑。そうすると今までのように楽観的な、電極を１本ずつ刺してニューロンはこう応答するっていうのを見つければ何か話ができるだろうと思うのは、

ちょっと行き詰まり感が出てきた感じはしますよね。そこで、人工知能のようなモデル研究が引っ張ってくれるときっとやりやすいんだろうなと思うんですけど。

谷口 ええ。進んできた人工知能を支えるようなモデルベースの考え方を、いかに実験科学につなげて行くかは、本領域「人工知能と脳科学」で正に推進しないといけないところでですよね。人工知能の研究者でもモデルをちゃんと考えてない人間もいっぱいいるし、単なる情報処理ツールとしてだけ見てる人もいっぱいいる。そもそも、モデルって概念自体にかなりスペクトラムというかグラデーションがあるんですよね。ダイヤグラムに書いただけみたいな昔の非常にシンプルなモデルから、実際に計算機で動かせるモデルまで。僕の立ち位置では、やっぱり 「動くモデル」というのを大切にしたい。インプリすれば実世界で動くモデル。モデルって色眼鏡だから、その色眼鏡を通して見ることで現象を解釈したり、仮説を立てたりするのが大事だと思うんですね。人工知能の人間と脳科学の人間が一緒にモデルを共有しながら、一緒に研究に取り組むということですよね。

坂上 それをやらないとだめですよね。結局、実験データとそれに基づく理論みたいなものと、インタラクションを非常に近いところで密にやっていかないと。特に実験ってものすごい限定しなくちゃいけないから。そこのところを離れたところで、誰かの論文を読んだだけで、これに合わせてみたいな話って難しいし、すごくやりにくいと思う。

谷口 脳科学の実験の人は、実験の大変さを分かってるおられる。でも、構成の方もめちゃめちゃ限定があるんですよ。例えば、強化学習で試行錯誤してるとロボットはすぐ壊れちゃうから出来ないとか。だから、脳の全体の部位があっても、今構成できる範囲のところを構成するしかないっていうの。それを分かってない実験系の研究者は 「ロボットだったらそう作ったらそう動くんでしょ、人間が作るんだから」みたいなことを言う人がいるんですよね。それは違うぞと。作るっていうのは、場合によったら実験的に理解する以上にExplicit（明示的）に理解しないと作れないわけなので大変です。実は人工知能と脳科学の本当のコラボレーション研究っていうのはその交点を見出さないといけない。

坂上 なるほど。

谷口 コラボレーションの仕方みたいなもの自体が、ある意味で科学哲学的にアップデートされる必要があるんじゃないかなと思っているんです。そういう意味では、よく言われるようなコラボレーションのあり方に、実験屋が実験データを出してモデル屋が解析するみたいな分業体制のステレオタイプがあるじゃないですか？　それって 「人工知能と脳科学」については違うのかなと思っていて。それってモデルとデータ処理を混同している気がします。モデルに基づいて脳を理解しようとなると、むしろ最初のプロブレム――問いを立てるところに相乗りするのが、多分一番大事なんですよね。

坂上 理想的にはその通りだと思うんですよ。実験屋とモデル屋が共同で問題設定して、実験データをもとに現状でのモデルを立てて、必要な実験的検証をして、モデルをアップデートして動かしてみる。ただ、モデルを作るにあたって、データの処理の仕方がモデルの質を決める場合がある。モデル屋さんの要請にこたえるためには、私たちにはできない処理があったり、複雑で探索的な処理が必要になる。こういう場合には、データ処理もモデル屋さんにお願いするほうが良いと思うんですよね。

谷口 やっぱりモデル研究というのを前に進めることが重要なんでしょうね。

坂上 ところで、モデルの話になって、谷口先生と対談ということなら、やっぱり、「構成論」の話は入れないといけない。それがないと玉川に帰れない。

谷口 僕に構成論の話をさせると長くなりますよ （笑）。まず、大切なのは構成論っていう存在が、科学における実証研究のどこに位置づけられるかということ。実は、そこで結構悩んで、鬱々とした時代がありました。実は僕それで科学哲学会に発表しにまで行ったんですよ!

坂上 おお。

谷口 その辺は、いろいろ思索が進んでケリはつきました。「記号創発ロボティクス」（講談社）の中で構成論アプローチについては一章割いてしっかり書いたので、よかったら読んでみてもらえると嬉しいです。やっぱり、構成論は「モデルの提供」っていうのが一つですね。その目的は、新しい仮説を生むためっていうのと、対象系を理解するため。忘れがちなことなんですけど、僕たち人間って物事を 「理解」する時に、対象系のデータそのものは 「理解」にはならないんですよね。大体、データに基づきながらも 「こういう風な意味だよね」ってモデルを通して理解するんですよね。それが無いと我々って人間として理解できない。そこ大事で、僕らって例えばLong-term Memory、Short-term Memoryとかでも、ある種の計算機をモデルにしてそれを通して理解するんですよね。だから我々って人工物としてのモデル通してそのアナロジーで物事を理解するから、リッチなモデルを持たないと、結局それに対する理解の深さって、深くならないんですよね。

坂上 それは実験も同じですよ。結局いろんなモデルを作る時にデータから再構成して、それって結局人間の理解の枠組みに乗っ取った形のものしか作れないから。逆に言えばそれが本当に真実の方向に向かってるのかしらって思うこともありますよね。やっぱり我々の認識バイアスみたいなのが強すぎて。それは思うけど、ただ我々が理解するわけだからそれに合っ

た形のものは作っていかないといけないって意味では、構成論っていうものは訴えかけるものがあるっていうのはよく分かります。ただ、未だにやっぱり構成論ってレベルによるような気がするんですよね。いわゆるロボットを作るなら作るなりのいろんな機能のコンポーネントみたいなものが、実際に人間が持ってるものが、どのくらい精緻に詰め込まれているのか。

谷口 そのものを作ってるんじゃないっていうのを合意するのがまず第一なわけです。モデルですからね。だから、何がQuestion（問い）なのかと。そこをきちんと合意してやるっていうのがポイント。その辺りについては、また我々の文化が「構成論」というものの論理、哲学を十分に育てられてないのが問題。多くの 「構成論」を掲げる研究者もそのあたりが非常にナイーブなままに進めてしまっている。だから 「構成論」を科学の中で上手く使いこなせてない。モデルってそもそも、対象系のある部分のある構造を抽象化して描いてるわけですよね。モデルの歴史を考えると、紙の上で書いていたフィードバック学習の図みたいなダイアグラムレベルでのモデルがあって、次に、計算機上で動かせるモデルが出来てきた。数理モデルによるシミュレーションですよね。さらに、それをセンサ・モータで実世界に接続させて動かせるのがロボットなんですよね。そんな感じに拡張されてきている。つまり、ロボットというのは計算機と身体で拡張されたモデルなんですよ。ロボットや機械学習モデルを用いた構成論のポイントは、紙の上に書いただけの概念的なモデルと違って、実際に動かすことが出来るわけで、実データを基に動くか動かないか、学習できるか学習できないか、みたいなのをある意味で演繹的にチェックできるわけですよ。論理的にそもそも一貫性がなかったり、そもそも動かない矛盾を抱えたりするモデルっていうのはいくらでも作ることが出来るわけですが、「こんな感じどうでしょう?」と思いつきを言ってくるモデルを、構成論は棄却できる。「いやそれじゃ、そもそも動かないから」って。

坂上 今の説明とてもいいですよね （笑）、構成論が何かっていうのを理解するには。そこで例えば脳との関係でいうと非常に精緻な理論とメカニズムが作れて動いたということが、脳が実際にそれを使っているという脳機能との関係性をきちっと示す必要はないんですか?

谷口 僕は構成論のモデルっていうのは——特に脳とかのモデルでも「精緻である」ことが目的ではないと思うんですよね。

坂上 精緻っていうか、ある種の機能をうまく実現できるという風に言い換えた方がいいかもしれない。精緻っていうのはそういう意味で使ったんだけど。そうであればひょっとしたら脳よりずっとうまく、ある機能を実現するメカニズムってあるかもしれないじゃないですか。

谷口 それは手法——工学としてはいいと思うんですよね。

坂上 でも構成論っていうのは別に人間とか動物の機能を意識しているっていうだけではないわけですよね。そこはどこまでそうなんですか?

谷口 構成論ってのは 「モデル」って言葉に置き換えましょう、全部。そうするとモデルっていうのは……

坂上 人間のモデル、動物のモデル。

谷口 そうそう、人間の構成論。

坂上 と捉えていいんですか。

谷口 そうですね、ある意味で言ってしまうと構成論っていうのはモデル概念のExtension（拡張）なんですよ。人間の知能っていうのは決められたデータセットだけじゃなくて、実世界情報っていうのを食って行動してないといけない。我々の知能っていうのは実世界の複雑性に適応するのが本質じゃないですか。そのモデルがちゃんと動くかどうかの検証には実世界のデータを入れないと駄目。そのためには身体まで構成要素に入れて動かせるものが必要。そういう意味で、動くモデルとして拡張していったのがロボットを使った構成論。

坂上 なるほど。人工知能と脳の融合って意味で困ったなと思うのは、人工知能がどんどん進んでいって世の中の需要により応える人工知能を作っていくためには、脳なんかの機能を考えるよりも人工知能独自のロジックでどんどん進めていった方がより良いものができるっていう風に考えている人が増えている気がするんですよね。構成論っていう意味では、今のような人工知能の独走みたいなことはありえないって谷口さんは考えているんですか?

谷口 いやいや、そこも話が混同されていて、構成論はいわば理解のためにあるんですよね。だから別に、それは人工知能の工学的進歩という話と別問題であって構わない訳ですよ。なので構成論が脳科学に貢献するというのと、人工知能の工学的独走というのは別問題。脳の長期的な学習みたいな、振る舞いを理解しようと思ったら、潜在変数が入ってくるようなモデルが要ると思います。それを考えましょう、提供しましょう、しかも、ちゃんと動くモデルであるという保証付きで。それが構成論なわけですよ。逆に脳から人工知能へは何が渡せますかってことなんですけど、やっぱりそこは、知能の実現の好例としての脳に学びましょうということだと思うんですよね。今、深層学習に関連しても、様々な人工知能に関する、大量の論文が書かれていますが、それって、人類全体で知能のモデルの多点探索をやっているものだと思うんですよね。これを本当にランダムにやっちゃうと大変。そういう探索に適切なヒューリスティクな知識を入れると探索が早くなるというのは常識。脳に学ぶっていうのはシンプルにそういうことだと思うんですよね。

坂上 なるほど。

谷口 シンプルにそういうことでいいんじゃないかな?

坂上 我々はものすごく期待しますよ、やっぱり。今のような構

成論であれば、つまり人間の脳のモデル化、しかもシミュレーション可能なモデル化っていうのであれば、やっぱりそこから何か我々への示唆が出てくるんじゃないかと。

谷口 脳科学でのある実験タスクがあるとするじゃないですか。言語的に脳機能を説明されたりするわけですけど、僕の視点から見たら数理モデルじゃないと理解できないんですよね。数理モデルにしたら、絶対隠れ変数がこう置かれて、そのInference（推論）問題ですよね？　この話って？　みたいになる。そういう風に整理してやると、いくつかの仮説が新たに出てくる。

坂上 実験屋とモデル屋の言葉がうまく通じるといいですね。私たち実験屋は、モデル化もEmpirical（経験的な）ベースなものになりがちで、それじゃあ動かしてみようと思っても、そもそもテストもできない。実験屋とモデル屋が同じ言葉で話すことはとても重要ですよね。

谷口 いやぁ、どっちかっていうと、僕の経験からですけどね、そういう話をすると実験側の人が「数理は分からないから」って言ってReject（拒絶）してくる場合が多い。

坂上 （笑）。まあね。そこはもうひとつの大きな問題ですよ。そういうのって教育のシステム自体が変わっていかないと……かもしれないですね。

谷口 まぁ、難しいことだからこそ、手元の問題を実験側の人とモデル側の人が一緒に時間をかけて噛みしめるみたいな、そういうことが大切なのかなと思います。そこから何かネタが転がりだすと、そうするとそのネタは比較的実験側にとって身近なものなので実験系に落とせる。構成論ってもの自体がどちらかというと科学に貢献を見出すものなので、それで論文を書くのは脳研究側だと思うんですよ。そうなると、とにかく脳研究側の実験系に落ちないと、いかなるアイデアも形にならない。だから、脳科学の論文になるかどうかの制約条件はガチッとはめながら進めるのが、研究として費用対効果はいいんじゃないかなと思っちゃいますね。でもそのためには、逆に実験系の制約を赤裸々に出してもらいながら、モデル側がそれを理解した上で、モデルの選択を打たないといけない気がするんですよね。

坂上 とにかくお互いの理解が大切ですよね。やっぱりなかなか今お互いに理解でき合っているとはとても思えないから。ちょっと構成論の話は聞いとかないといけないと思っていたもので。

谷口 自分で言うのもなんですが、多分、構成論については日本で一番語れる人間の一人だと思うので　（笑）。

坂上 まあそうでしょうね。構成論の話ですごく引っかかるのは、ある意味間違ってるかもしれない、きちっと確定してないような脳科学、神経科学のデータを基に、それの上に何かを展開しようとするような場合があるんですよね。構成論に対する今までの私の疑問なんだけど、今までの脳科学の知識なんかに依存した形で構成論を積み上げていくような形であるのは、あんまり生産的でない気がするんです。逆にそれは無視して後からこれは違うと、脳科学が検証できるような形を持っていてくれた方が生産的な感じがします。

谷口 構成論を混同してしまっている先人もやはり多いんですよね。まぁ、構成論自体が学問の進化の途上にあるので仕方ないのですが。僕は、構成論の知っていうのは、要は脳科学とかでやる数値的なデータとは地続きじゃないんですよ。違う次元に存在してるんですよね。歴史的に見てというか、科学というのはブランド感があるんですよね。自らの学問の科学化を目指す人間は無意識に科学への憧れを持ちがち。科学の一部にしたいという欲動に学術的方針が影響されるというのは学問全体でもしばしば見られる思考バイアスです。それから自由にならないといけない。僕はそのメタ認識を意識的にやってるつもり。

坂上 なるほど。心理学も近いところありますからね。

谷口「人工知能と脳科学の対照と融合」を進めるというのは、そういう意味でモデルと実験の新しい地平を拓く、大変、科学史的文脈において挑戦的なことをやっているのだと思います。そこは、意識をもってやっていきたいですね。

坂上 そうですね。まずは、私と谷口さんの共同研究を成功させるところからですね。

谷口 ええ、是非よろしくおねがいします　（笑）。

坂上 じゃあ、今日のところはこのあたりで。ありがとうございました。

谷口 はい、ありがとうございました。

特　集

次の脳科学×AIに向けて ～深層生成モデルと世界モデルから見えるAIの未来～

鈴木 雅大（東京大学 松尾研究室 特任研究員）

はじめに

深層学習は、2006年のブレイクスルー以来、多くのタスクで従来手法と比べて高い精度をたたき出すようになった・・・と今や深層学習に関する論文や記事ではおなじみとなった前口上ですが、この精度とは主に分類問題や回帰問題、すなわち教師あり学習の汎化性能を指しています。実際この数年で、いくつかの教師あり学習タスクにおいて人間をも超えるような性能を示す結果が得られるようになったのはご存知の通りです。

しかし振り返ると、元々深層学習の歴史は教師なし学習からスタートしました。制限付きボルツマンマシン（RBM）やオートエンコーダ（AE）を用いた教師なし事前学習によって深層学習のブレイクスルーが起こり（Hinton et al., 2006; Bengio et al., 2006）、「Googleの猫認識」論文によって、深層ニューラルネットワーク（DNN）が膨大な画像を用いて教師なし学習を行うと、中間層で任意の概念（顔や猫）に対応する表現が獲得されることが発表され大きな話題となりました（Le et al., 2012）。初期の深層学習はこうした教師なし学習による「表現学習」に注目が集まりましたが、その後、主軸は「どれだけ高い精度で分類できるか」という教師あり学習に移りました。

しかし最近、再び教師なし学習や表現学習が着目されるようになっています。様々な背景がありますが、その大きな要因の一つとして「深層生成モデル」に関する研究の進展があります。生成モデルはデータの生成過程を確率モデルで表現する枠組みで、深層生成モデルはその確率モデルをDNNで表現した生成モデルです。これらの研究の進展によって、今まで生成モデルでは扱えなかったようなサイズや形式のデータで学習できるようになりました。それによって、より良い表現学習や、世界そのものをシミュレートするような「世界モデル」の研究が進められています。

今回の記事は「次の脳科学とAI」というタイトルの記事となりますが、私自身脳について知識が疎いこともあり、深層生成モデルと世界モデルについての最新の研究動向と課題について触れて、AIの側からみた脳×AIの展望について考えてみたいと思います。

深層学習と表現学習

最初に述べましたように、深層学習の歴史は元々教師なし学習、表現学習からはじまりました。HintonらやBengioらは、全結合のDNNの重みを、下層から順番にRBMやAEで学習する教師なし事前学習を提案し、それによってネットワーク全体をまとめて教師あり学習（fine-tuning）できるような、良い初期値が得られることを示しました（Hinton et al, 2006; Bengio et al, 2006）。当時は全結合層のDNNの学習は困難とされていたので、この2段階の学習方法は大きなブレイクスルーとされました。

ですがその後、事前学習を必要としない重みの初期化方法（Glorot et al, 2010）や、Adam（Kingma et al, 2015）などのより良い最適化手法の提案、データの増加やマシンパワーの向上などによって、事前学習せずに教師あり学習が可能となることがわかり、事前学習の枠組みでの教師なし学習はあまり用いられなくなりました。

しかし、膨大な教師なしデータから、あらゆるタスクに有用な「良い表現」を学習することは、汎用的な人工知能を実現する上ではやはり重要です。我々人間も幼少期にあらゆるものを見たり触ったりして、世界の表現を学習していますし、そうした表現に「言語」という情報を対応させることで、それが何であるかを「理解」します。では、そうした様々なタスクに利用できるような「良い」表現とは一体どのようなものでしょうか?すなわち、世界に共通する事前知識（prior）とは一体何でしょうか?

Bengioらは、この事前知識として「多様体」、「滑らかさ」、「線形性」、「複数の説明因子」、「原因因子」、「説明因子の階層性」、「タスク間の共通因子」、そして「もつれを解く（disentanglement）」などを挙げました（Bengio et al., 2013; Goodfellow et al., 2016）。こうした事前知識を正則化基準に落とし込んで表現を学習する方法として、近年「深層生成モデル」を用いた教師なし表現学習の研究が進められています。

深層生成モデル

生成モデルは、とても大雑把に言うと「手元にあるデータがどのようにできているか?」に着目しその生成過程を学習する枠組みです。これは「データを分類する」ことのみに興味があ

る通常の教師あり学習（識別モデル）とは対照的です。生成モデルの具体的な学習方法については割愛しますが、ある確率モデル（生成モデル）$p_\theta(x)$（xは入力データを表す確率変数）を用意し、それがデータの分布$p_{data}(x)$（実際には存在しませんが、そのような「データを生成する分布」があると仮定します）と近くなるように、モデルパラメータθを更新する、というのが大まかな流れです。生成モデルの一番の特徴としてよく知られているのは、学習した生成モデルから、元のデータとよく似たサンプル（擬似的なデータ）を生成できることです。これは生成モデルが、データがどのように生成されているかを理解しているからこそ実現できることに注意して下さい。

近年注目を集めている「深層生成モデル」とは、このモデル分布としてDNNを用いているものです。深層生成モデルの一番の利点は、まさに、豊富な表現力を持つ関数であるこのDNNを利用していることです。そのため、今までは生成できなかったような、高次元かつ複雑なデータを学習・生成できるようになりました。

今回の記事でご紹介する研究の殆どで利用されているのが、変分AE（VAE）（Kingma et al., 2014）と呼ばれる深層生成モデルです。VAEでは、入力データxの他に、データには現れないけれどもその背景に潜んでいる因子として、潜在変数zが仮定されていて、入力からその潜在変数への推論モデル$q_\phi(z|x)$と潜在変数から入力への生成モデル$p_\theta(x|z)$[1]の2つのモデルで構成されます。情報の流れとしては、入力が推論モデルによって潜在表現に圧縮（エンコード）され、その表現から生成モデルで元の入力表現になるように復元（デコード）されます。こうしたモデル、すなわちデータを圧縮して復元するAEは、昔から盛んに研究されていました。VAEが従来のAEと比べて特徴的な点は次の2つです。まず、推論モデル（エンコーダ）と生成モデル（デコーダ）の両方がDNNを使った確率モデルとして設計されていることです。従来のAEに比べて深いモデルになるため、潜在表現でより普遍的な特徴が獲得でき、さらに確率的なモデルであることから後述する多様体学習ができます。もう一つの特徴は、表現についての明示的な正則化が加えられているということです。VAEの目的関数は次の通りです[2]。

$$E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x)||p(z)]$$

第1項が（負の）再構成誤差に対応しており、第2項が推論モデルをなるべく単純な分布$p(z)$（標準ガウス分布など）に近づくように促す項、つまりエンコードされる「表現」についての正則化を行なっています。

一方、一般的に深層生成モデルとして有名なのは敵対的生成ネットワーク（GAN）（Goodfellow et al, 2014）でしょう。GANは、深層学習の一連の研究の中でも最も偉大な成功の一つであり、本物と見紛うほどの綺麗な画像を生成できるという特徴から、AIがいわば「創造性」を獲得した研究としても広く認知されています[3]。GANの最大の特徴は、モデル分布（生成モデル）がデータ分布とどれくらい近いか、という評価自体もDNNを使って学習するようにしてしまったことです（このDNNは識別器と呼ばれます）。そのため、生成モデルの「分布の形」を予め決める必要がなくなり、DNNの持つ表現力をフルに発揮でき、高いクオリティの画像を生成できるようになりました。なお当然ですが、識別器が十分に学習できなければ、生成モデルの良さも適切に評価できません。また、識別器による評価を改善するように生成モデルを更新したら、改めて最新の生成モデルの評価を下すために識別器を学習し直す必要があります。そのため、GANでは識別器と生成モデルが交互に競い合うように学習する「敵対的訓練」を行います[4]。

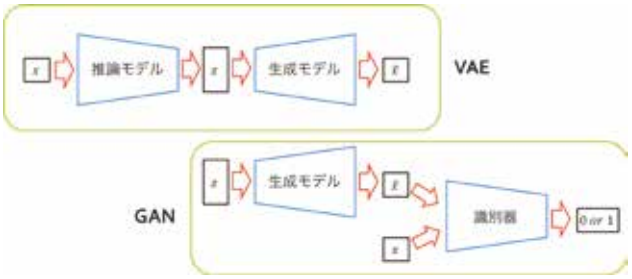


図1. VAEとGANの比較。$\tilde{x}$は生成されたデータを表す。

ここで、VAEとGANの違いを見てみましょう（図1）。まず1つ目の違いは、GANは識別器のおかげで生成モデルの分布の形を決めなくていいのに対して、VAEは推論モデルと生成モデルの両方について分布の形を決めなくてはいけないということです。VAEはGANに比べると生成した画像がぼやける傾向があるとされていますが、理由の一つはこれにあると考えられています。もう一つの違いは、VAEには入力から潜在変数を推論するモデルがありますが、GANにはそれがないということです。脳、特に大脳新皮質の「認識」と「予測」の双方向の流れを考えると、VAEの方が脳の観点から妥当性の高いモデルだと思います。

1 実際にはzの事前分布$p(z)$を考えて、生成モデルをxとzの同時分布$p_\theta(x,z)$として考えます。

2 周辺対数尤度$\log p_\theta(x)$の下界に対応し、エビデンス下界（ELBO）と呼ばれます。また本記事では、VAEにおいて重要な再パラメータ化トリックの話などは省略します。

3 GANを提案したIan Goodfellowを「The man who's given machines the gift of imagination.」として紹介している記事もあります（https://www.technologyreview.com/s/610253/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/）。

4「敵対的」という言葉から何やら攻撃的な印象がありますが、実際は一方が敵対的すぎると相手に圧勝してしまい、良い生成モデルを学習するという全体的な目的が達成されないため、お互いを考慮して「協力し合って」学習しているイメージが正しいように思います。

これらの違いからわかることは、潜在空間すなわち表現の学習については、VAEのほうが扱いやすいということです。確かに、画像生成という観点からはGANに利点がありますが、VAEは表現への推論モデルを持ち、またその形も明示的であるため、表現の正則化が容易になります。また、学習の安定性についてもVAEとGANでは異なります。VAEは再構成誤差最小化に基づき、分布の形も明示的なので非常に安定して学習できますが、GANは識別器と生成モデルのバランスが難しく、DCGAN（Radford et al, 2015）が登場するまでは適切に画像を生成することは困難でした。現在はGANの学習を安定させる様々な素晴らしい研究が提案されていますが（Arjovsky et al., 2017; Miyato et al., 2018）、ハイパーパラメータに対する精度の分散が大きい（Lucic et al. 2017）、生成できる画像のバリエーションが少ない（Arora et al. 2017）などの課題は残っており、表現学習や、後述する世界モデルの学習のためにはVAEのほうが有用だと考えられます。

ただし、こうした議論には例外もあり、また最近では、VAEの目的関数の中で表現の正則化に敵対的訓練を用いたり、GANにエンコーダを加えたりといった研究も多く、あまり「GAN」や「VAE」のように分けるのは適切でないかもしれません。また最近では、単一のモデルで双方向の流れを持ち、複雑な画像を生成できるflow-basedモデル（Dinh et al., 2014; Dinh et al., 2016; Kingma et al., 2018）も注目されており、今後も深層生成モデルの研究は大きな変化が続くでしょう。

深層生成モデルと様々な表現学習

次に、実際にVAEがどのような表現を学習するのか、そしてどのように表現の正則化を入れているのかを最近の研究を交えてご紹介します。

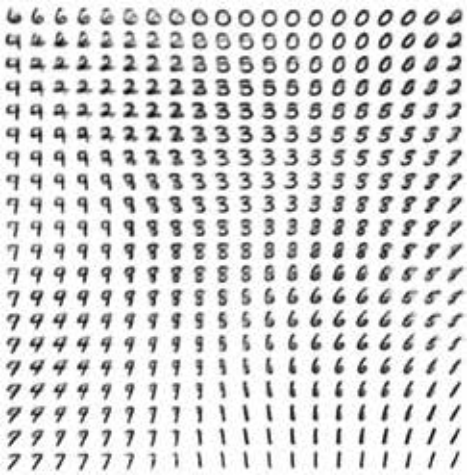


図2. VAEが学習する表現。2次元の潜在空間の任意の座標から生成モデルで生成した画像。手書き数字画像データセットMNISTで学習。図はKingma et al.（2014）より引用。

まず、VAEが学習する表現を御覧ください。図2は、学習したVAEの潜在空間、つまりVAEが学習した表現を可視化したものです。VAE（に限らず確率的AE）の特徴として、潜在空間上で高次元な入力情報の連続的な変化を捉えられるということが知られています（多様体学習）。図2でわかるように、有限のデータセットから連続的な潜在空間を学習できており、潜在空間の任意の座標から未知のデータを生成できることがわかります。

また、DNNに期待することとして、世の中を幾つかの「因子」に分解することが挙げられます。DNNの性質として、浅い層（入力に近い層）ではエッジなどの特徴が獲得され、深い層（入力から遠い層）では「人」や「猫」といったより抽象的な特徴が獲得されることが知られています。このとき、深い層の各ユニット（次元）がそれぞれ異なった意味となることが望ましいでしょう。例えば、あるユニットは物体の回転に対応し、別のユニットは角度に対応する、といった形です。こうした深い層の表現を、深層学習ではもつれを解いた（disentangle）表現と呼びます。なぜこれが重要かというと、もしもつれを解いた表現が教師なしで獲得できるならば、それはデータから何らかの「概念」を自ら獲得したことに他なりません。そして、その表現に言語情報を接地させることで「意味理解」につながると考えられるからです。

VAEはそうした表現を獲得できるのでしょうか?実は目的関数の第2項がその正則化に対応しています。この項は推論分布を単純な分布$p(z)$に制約する役割を果たしていましたが、実はこの事前分布$p(z)$は各次元が独立になっている($p(z) = \prod_i p(z_i)$)ので、推論分布に対して、各要素が独立になるようにも制約していたのです。ですので、この正則化項の係数を大きくすると、より独立性の正則化が強まり、もつれを解いた表現が獲得されるのです。この手法は、第2項にβ（>1）という係数（ハイパーパラメータ）を導入することにちなんで、β-VAEと呼ばれます（Higgins et al., 2017）。図3は、β-VAEによってもつれを解いた表現が獲得されている様子です。潜在変数の任意の次元が「顔の色」「年齢と性別」「彩度」の異なる因子に対応するように学習されていることがわかります。

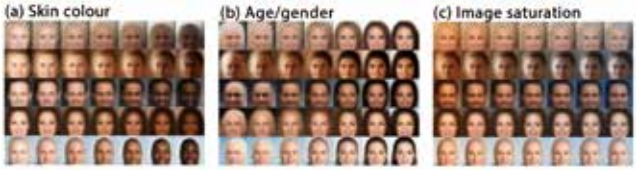


図3. β-VAEによって「もつれを解かれた」表現。(a) ～ (c) で潜在変数の異なる次元を動かし、それに対応する画像を生成している。図はHiggins et al.（2017）より引用。

ただし、この正則化項の係数を単純に大きくすればいい訳

ではありません。この項（のデータ分布における期待値）は$E_{p_{data}}[D_{KL}[q_\phi(z|x)||p(z)]] = I(x;z) + D_{KL}[q_\phi(z)||p(z)]$（ただし$q_\phi(z) = \int q_\phi(z|x)p_{data}(x)dz$）と分解することができ、第1項は入力$x$と潜在変数$z$の相互情報量、第2項が表現の独立性の正則化に対応します。つまり、正則化項の係数を大きくしすぎると、今度は入力と潜在変数の相互情報量まで小さくなってしまい、適切に入力を圧縮できなくなってしまいます。そのため、元のVAEの正則化項以外にも様々な正則化が提案されています。例えばFactorVAEは表現についての全相関（total correlation）の項$D_{KL}[q_\phi(z)||\prod_i q_\phi(z_i)]$を追加することで、相互情報量を小さくせずに各次元が独立になるように学習しています（Kim et al., 2018）。なおこの項は、使われている分布が明示的に計算できないので、敵対的訓練が併用されています。

このように、事前知識に基づいた表現の正則化ができるようになりましたが、ではこうした表現は高度な知能処理にとって果たして有用と言えるのでしょうか?例えば、もつれを解いた表現は、様々なタスクに利用できることが期待されますが、現時点では十分に検証されていません。ですので、今後は獲得した表現の有用性と、そもそも最初に仮定した表現の「事前知識」が妥当なのかどうかといった検証も必要でしょう。

さて、これまでは変数として入力xと潜在変数zしか考えていませんでしたが、別の情報yを加えた生成モデルも容易に設計できます。ここで、入力xは潜在変数zと何かの情報y（例えばラベル情報）から生成されるとし$(\hat{x}\sim p_\theta(x|z,y))$、さらに$z$と$y$は独立であると仮定しましょう。もしこの生成モデルが適切に学習できれば、潜在変数ではラベル情報に依存しない表現を獲得できるはずです。このモデルは、情報yで条件付けられた（conditioned）モデルということで、conditional VAE（CVAE）と呼ばれています。このモデル（厳密にはさらに改良したモデル）を使って、表現zに含めたくない情報（人種や性別など）yを明示的に排除する研究などが行われています（Louizos et al., 2015）。こうしたある属性に対して不変な表現を学習する研究は、プライバシー保護などの観点からも有用といえるでしょう[5]。

さらにCVAEが嬉しいのは、潜在変数zの値を一定にしたままyの値を変えることで、スタイルを保ちつつyに対応する画像を生成できることです。具体的には図4をご覧ください。この例ではyをラベル情報としています。zの表現が「筆跡」に対応するように学習できており、元画像の筆跡を保ちつつ、0～9まで生成できていることがわかります。これは「実際には見たことないけれども、もしあの人がこの字を書いたら、過去の経験からして、多分こんな感じになるだろう」といった所謂「反実仮想」を実現している例とみなせます。これは手書き数字画像を使った単純な例ですが、時系列情報を扱うCVAEを用いて「もし薬を投与しなかった場合、患者の症状はどうなるか?」という検証をした研究もあります（Krishnan et al., 2015）。その他にもCVAEは半教師あり学習に利用されたり（Kingma et al., 2014）、後述する世界モデルとして使われたりと、大活躍のモデルです。

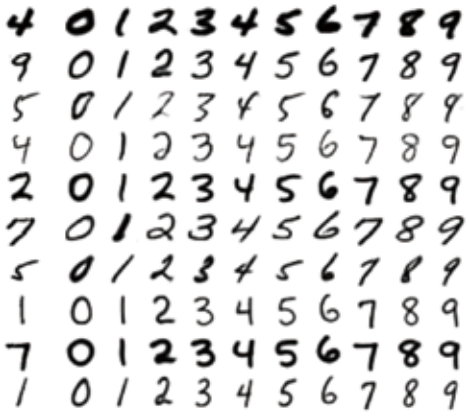


図4. CVAEが「筆跡」を保ったまま各数字を生成した結果。一番左の列が元画像で、2列目以降が、潜在表現を固定して対応するラベル情報を0～9に変化させている。図はKingma et al.（2014）より引用。

言語情報との統合

こうした表現学習で獲得した表現を他の情報、特に言語情報に接地・統合させる研究も進んでいます。DeepMindが2018年に提案したSCAN（Higgins et al., 2017）と呼ばれるモデルでは、β-VAEで画像の表現を学習したあとに、言語情報（実験では属性の組合せ）を入力としたVAEを用意して、その表現がβ-VAEで獲得した表現と近づくように学習します。これによって、教師なし学習で獲得した表現と言語表現を結びつけることができます。このモデルの利点は、画像から対応する言語を生成するだけでなく、言語から画像も生成できることです。また、生成モデルが見たことのない未知の概念に該当する言語情報から、それに対応する画像を生成することもできます（図5）。

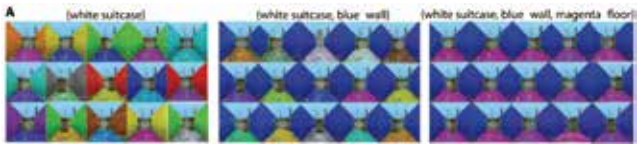


図5. SCANを使って、言語情報（属性）から画像（ここでは部屋の風景）を生成した結果。「white suitcase」や「blue wall」が属性になる。一番右については、一度も見たことない概念（つまりこの属性の組合せがデータセットに存在しない）にも関わらず、適切に対応した画像を生成できている。図はHiggins et al.（2017）より引用。

この研究は、言語情報を画像の表現学習よりも上位として置いたモデルと言えますが、一方で言語情報を単に画像とは

5 最近では敵対的訓練を用いた不変表現学習も主流となっています。

異なる「モダリティ」の一つと考え、言語と画像のマルチモーダル学習として同時に学習するモデルもあります。筆者らが提案したJMVAEというモデルは、2つのモダリティ（画像と言語情報）を統合した表現（共有表現）を学習し、SCANと同様、双方向の生成ができます（Suzuki et al., 2016）。またVendantamらは、エキスパートの積（product of expert）（Hinton, 2002）を利用した、より適切に未知の概念（属性の組み合わせ）から推論できる推論モデルを提案して、JMVAEでもSCANと同様のことができることを示しています（Vendantam et al., 2018）。SCANと比べたときのJMVAEの利点は、潜在空間において、画像と言語情報の両方が統合された情報量の大きい表現が得られるということです。

これらの研究は現時点では言語情報としては属性という非常に単純なものを使っています。ですが、画像の表現学習と言語を組み合わせた研究という意味では、真の「意味理解」に繋がる研究といえると思います。

世界モデル

これまで述べたように、深層生成モデルは、様々な事前知識を仮定することで、画像データから有用な表現を獲得でき、また未知の概念に対応する画像も生成できるようになっています。ただし、こうした研究が対象とするデータは、手書き数字や顔画像、物体画像など小規模でした。しかし、最近ではある環境の画像だけからその環境そのものを学習してしまおうという野心的な研究が進められています。こうした研究群を本記事では「世界モデル」と呼びます。

背景としては、人間が世界の情報を元に、脳の内部で世界の「シミュレーター」を作っているということがあります。我々はこれを使って常に世界の予測を行っており、Jeff Hawkinsは著書「考える脳　考えるコンピューター」の中で、この予測こそが「知能を解明する鍵」であると述べています。

ただし、世界の大きさは膨大であり、得られる情報も無数にあるので、実世界をそのまま頭の中にモデル化するのは非常に困難なはずです。それにもかかわらず実現できているということは、脳の中で無意識に重要な情報が取捨選択・圧縮され、それを基に頭の中で世界が「再構築」されているということです。すなわちシミュレーターを獲得する上で、脳（大脳新皮質）は教師なし表現学習を行っているのです。世界モデルは、まさにこれを深層生成モデルで実現しようとする研究分野です[6]。

世界モデルの研究として最近注目を集めたのが、Haらによる「World Model」です（Ha et al., 2018）。World Model、つまり彼らが提案した世界を学習するモデルは、画像から空間方向の抽象化を行う部分と、得られた抽象表現（状態）とそのときの行動から次の状態を予測するダイナミクス部分で構成されています。前者はVAEでモデル化されていて、後者は混合密度ネットワークRNN（MDN-RNN）で設計されています。論文によると、学習したWorld Modelが、実際に世界を生成できることが実験で確認されています（実験では、カーレースのシミュレーションなどを環境としています）。また、この世界モデル内の環境のみで強化学習したエージェントが、実際の環境でも動作できることが示されています。つまり、エージェントはいわば「イメージトレーニング」によって、実際の環境でもうまく動作する方策を身に着けたということです。

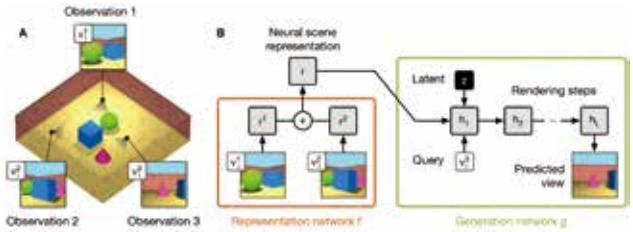


図6. GQNの概要図。2つの視点（Observation 1, 2）での画像に基づき、未知の視点（Observation 3）から見える画像を予測している。図はEslami et al.（2018）より引用。

さらに、その後大きな注目を集めたのが、Generative Query Network（GQN）という研究です（Eslami et al., 2018）。この研究では、画像入力xのほかにその画像を得た視点vを考え、様々な環境の様々な視点を学習します。そしてその後、未知の環境で幾つかの視点（例えば2つの視点v^1, v^2）に対応する画像（x^1, x^2）を見せたあとに、未知の視点v^qから見える可能性の高い画像x^qを予測してくれるというものです（図6）。文章で書くと伝わりづらいですが、是非元論文の動画[7]を見ることをおすすめします。

実のことを言うと、このGQN論文はそこまで新しい生成モデルを提案しているわけではなく、従来からあるCVAEを応用しているにすぎません。つまり先述した「筆跡」を保ちつつ「未知のラベル」に対応する画像を生成する例が、「現在の環境」を保ちつつ「未知の視点」に対応する画像を生成する、と置き換えられているだけです[8]。では、この研究の何がすごいのかというと、深層生成モデルが実際に世界を学習できることを実証してみせたことです。この論文では、あるシミュレーション環境での箱庭を世界と考えて学習していますが、この程度の環境ならば、画像のみから世界の3次元構造（のようなもの）をシミュレーションできてしまうのです。先程のWorld Modelと

6 なお、こうした研究は内部モデルや力学モデル、強化学習におけるモデルベースの研究として数多く行われてきましたが、それらと世界モデルが違う点がこの「表現学習」をしているということです。

7 https://www.youtube.com/watch?v=RBJFngN33Qo

8 視点の順番に依存しないような「表現ネットワーク」を提案しているといったアーキテクチャ的な工夫はあるものの、それ以外の工夫はそれほど大きな貢献ではありません。

比べると、視点という情報を利用して環境を限定してあげることで、より複雑な環境を学習できています。このように、別の情報を用いて環境を限定する、すなわち「条件付ける」ことが、今後の世界モデル研究においても重要になると思われます。

世界モデルの課題

さて、これまでの研究では空間方向の抽象化には成功していますが、我々のような世界モデルを深層学習が実現するには、それだけでは足りません。例えば、我々は「自転車に乗っている」という言葉から実際に運転している様子を想像できますが、これは自転車に乗っている風景とそのシーケンス、つまり空間方向と時間方向の両方で抽象化を行っているためです。しかし、自転車に乗っている時間の間隔は実際には毎回違うので、それをどのように「自転車に乗っている」という表現に抽象化すればいいのでしょうか?また、我々は現在から遠い未来のことを直接想像（シミュレーション）することができます。系列情報の推論として考えると、遠い未来にまでには間に長いステップがあるため、相等な推論時間がかかるはずですが、我々はこれを一瞬で実行できます。一体どのようにAIで実現すればいいでしょうか?

こうした系列情報の抽象化に関する研究は、残された大きな課題であり、今後も高い優先度で研究を続けるべきだと思います。鍵となるのは、空間方向を抽象化していた既存の深層生成モデルと系列情報を扱う再帰型ニューラルネットワーク（RNN）をどのように組み合わせるかということです[9]。World Modelでは、これを完全に分離して学習していましたが、実際は一体化した生成モデルとして設計して学習するべきでしょう。生成モデルにおける系列情報の扱いについては、昔から状態空間モデルなどが研究されていましたが、最近、この状態空間モデルと自己回帰モデル（RNN）を組み合わせた、TD-VAEという興味深い研究が発表されたのでご紹介します。

TD-VAEは、状態空間モデルとRNNを組み合わせて両者のいいとこ取りをしつつ、強化学習における信念状態に該当する分布を取り入れたモデルです（Gregor et al., 2018）。このモデルの最大の利点は、潜在空間（状態空間）上で、あるステップから任意のステップにジャンプして矛盾のない確率的な推論ができるということです。すなわち系列方向の抽象化問題に対する一つの解を提示しているのです。TD-VAEは、獲得した状態を強化学習に利用することが想定されており、時間方向のジャンプが可能になったことでプランニングなどに有効だと考えられます。

このように、世界モデルは強化学習で用いることが想定されています。強化学習における課題には、状態表現をどうするか、そしてデータが少ない場合にどのように学習するか、といったことがありますが、学習した世界モデルを用いることで、こうした問題を部分的に解決できると考えられます。このように、学習した世界モデルを使って強化学習を行う「モデルベース」の研究は近年再び進められるようになっています。

ただし注意しなければならないのは、多くの場合、世界モデルは予め適切に学習されていることが想定されているということです。しかし、実際には全世界を完全に学習しきることは不可能です（我々人間も世界の全てを理解しているわけではありません）。したがって、世界モデルが完全に信頼できない場合に、どのように行動するべきか、ということは大きな課題です。特に系列モデルでは、毎ステップ推論がずれてしまうと、そのずれがどんどん大きくなってしまいます。解決策の一つとしては、TD-VAEのような、不確実性を考慮して任意のステップまで一気に推論ができる世界モデルを利用することが挙げられます。またHafnerらは、推論時にジャンプするのではなく、訓練時に複数先のステップまでの予測誤差を最小化するように学習することで、予めこの問題に対処することを提案しています（Hafner et al., 2018）。

また、我々は幼少期に自ら様々なものを見て、聞いて、触ることで「能動的」に世界モデルを学習してきましたが、これまで紹介した深層生成モデルによる世界モデルは、大量の画像などが与えられた下で「受動的」に学習しています。そのため、自らがどれくらい世界を理解しているかという予測誤差などの基準に従って行動を選択しながら、自身も更新していく「能動的な」世界モデルの学習が考えられます。ただし世界モデルの場合学習に時間がかかるため、どのように探索を進めていくかは十分に検討する必要があります。

また、現在の世界モデルは一種類の状態（行動）遷移しか学習することができません。しかし世の中には複数の環境やエージェントが存在し、それぞれ異なったダイナミクスを持ちます。深層生成モデルを使った世界モデルにも、複数のダイナミクスを認識させ、モデル化させることが課題となります。この辺りに関連する話として、近年研究が進んでいる「メタ学習」があります。メタ学習は、タスクに応じてモデルのバイアスを決定するメタ知識を獲得することを目的とした枠組みで、例えば犬と猫を分類するように学習した分類器から、猿を分類するのに有用な知識を得ることができるか?といった課題が該当します。近年、タスクに該当する情報を潜在変数として含めた深層生成モデルによるメタ学習のモデルが提案されており（Gordon et al., 2018）、系列情報への拡張も期待されます。

その他、深層生成モデルは人間のように、画像だけでなく、

9 人間の脳の情報処理を考えると、もしかしたらRNNを用いるのは適切ではないかもしれませんが、現時点では他に代替案がないのでいわば「暫定的に」利用しているという形です。

複数の種類の情報を取り入れて（つまりマルチモーダル学習を行って）、より確実な世界モデルを構築するべきでしょう。これが実現すれば、複数のモダリティを統合した、より情報量の多い表現を学習できます。また、画像情報が不足していても、言語や音声といった他の種類の情報から表現を推論することもできます。さらに言語情報から画像を想像したり、音声を想像したりすることができれば、前述した真の意味での「意味理解」により大きく近づくでしょうし、そうしたマルチモーダル情報から自身の世界モデルと比較し、行動を選択していくことで、AIにおける「身体性」に結びつくかもしれません。

まとめ

今回、深層生成モデル及び世界モデルをメインに記事を書かせていただいたのは、私自身の専門ということもありますが、近年大幅に研究が進み、かつ脳の情報処理の知見を活かせる分野だと感じたためです。Hawkinsらが言うように予測こそ知能の本質だとするならば、深層生成モデルはまさにそれをAIに実現させる大きなブレイクスルーといえるでしょう。今後の世界モデル研究はやはり、強化学習との結びつきが重要になると思います。その中で、前節に挙げたようなものを含めた数多くの課題を如何に解決できるかが鍵となりますが、そのヒントは脳から得られるかもしれません。

現在、この研究分野はGoogle（DeepMindやGoogle Brain）などの独壇場といった状況ですが、本新学術領域にはAIと脳の著名な先生方が一堂に会しており、様々な分野の最先端のノウハウが集約しているという大きな強みがあります。この記事が、本領域で脳科学×AIの新しい研究が生まれることに少しでも貢献できましたら幸いです。

参考文献

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation, 18*(7), 1527-1554.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *Advances in neural information processing systems* (pp. 153-160).

Le, Q. V., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., & Ng, A. Y. (2012). Building High-level Features Using Large Scale Unsupervised Learning.

Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256).

Kingma, D. P., & Ba, J. L. (2015). Adam: A Method for Stochastic Optimization.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence, 35*(8), 1798-1828.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). Cambridge: MIT press.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2018). Are gans created equal? a large-scale study. In *Advances in neural information processing systems* (pp. 698-707).

Arora, S., & Zhang, Y. (2017). Do GANs actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*.

Dinh, L., Krueger, D., & Bengio, Y. (2014). NICE: Non-linear independent components estimation. arXiv preprint *arXiv:1410.8516*.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2016). Density estimation using Real NVP. *arXiv preprint arXiv:1605.08803*.

Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems* (pp. 10236-10245).

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... & Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework.

Kim, H., & Mnih, A. (2018). Disentangling by factorising. *arXiv preprint arXiv:1802.05983*.

Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in neural information processing systems* (pp. 3581-3589).

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems* (pp. 3483-3491).

Krishnan, R. G., Shalit, U., & Sontag, D. (2015). Deep kalman filters. *arXiv preprint arXiv:1511.05121*.

Higgins, I., Sonnerat, N., Matthey, L., Pal, A., Burgess, C. P., Botvinick, M., ... & Lerchner, A. (2017). SCAN: learning abstract hierarchical compositional visual concepts. *arXiv preprint arXiv:1707.03389*.

Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation, 14*(8), 1771-1800.

Vedantam, R., Fischer, I., Huang, J., & Murphy, K. (2017). Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*.

Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems* (pp. 2455-2467).

Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., ... & Reichert, D. P. (2018). Neural scene representation and rendering. *Science, 360*(6394), 1204-1210.

Gregor, K., & Besse, F. (2018). Temporal Difference Variational Auto-Encoder. *arXiv preprint arXiv:1806.03107*.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2018). Learning Latent Dynamics for Planning from Pixels. arXiv preprint *arXiv:1811.04551*.

Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., & Turner, R. (2018). Meta-Learning Probabilistic Inference for Prediction.

論文紹介

黒質ー線条体ドーパミン神経路による行動抑制の制御

松本 正幸（筑波大学 医学医療系 教授）

論文タイトル：Primate nigrostriatal dopamine system regulates saccadic response inhibition
著者：Takaya Ogasawara, Masafumi Nejime, Masahiko Takada, Masayuki Matsumoto
Neuron, 100 (6), 1513-1526 (2018)
doi: 10.1016/j.neuron.2018.10.025.

はじめに

社会生活を送る上では、衝動的な行動や不必要な行動を抑制できることがとても重要です。ところが注意欠陥・多動性障害やパーキンソン病などの精神・神経疾患をもつ患者さんの多くでは、この行動抑制の能力が低下しています。これまでの先行研究により、行動抑制では、脳の中の前頭前野や大脳基底核と呼ばれる領域が重要な役割を果たしていることがわかっていました。他方、注意欠陥多動性障害やパーキンソン病などの行動抑制の能力が低下する疾患の多くでは、ドーパミン神経系に異常が見られることも知られています。しかしながら、ドーパミン神経系は報酬に関わる神経系として注目されてきたこともあり、この神経系がどのようにして衝動的な行動や不必要な行動を抑制しているのかは全く明らかにされていませんでした。

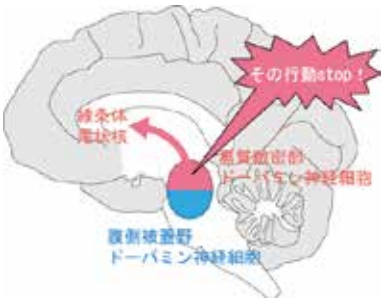
行動抑制課題実行中のサル黒質ー線条体ドーパミン神経路の電気生理学的・薬理学的解析

上記の問題に取り組むため、我々のグループは、ヒトに近縁なマカク属のサルを被験動物として、行動の抑制が求められる認知課題を訓練しました。実験室の中では、サルがモニターに向かって座っており、その視線が計測されています（図1A）。まず、モニターの中央に注視点が現れます（図1B）。サルがこの点を見たら、注視点が消えて別の場所にターゲットとなる点が現れます。全体の70%の試行では、サルがこのターゲットに眼球運動（視線移動）すれば報酬としてりんごジュースが与えられます。ただし、残りの30%の試行では、ターゲットが現れた直後に中央の点が再呈示されます。この再呈示はstop指令であり、サルは今まさに行おうとしている眼球運動をキャンセルする必要があります。そして眼球運動をキャンセルできた場合にだけりんごジュースが与えられます。このstop指令はランダムに30%の試行でだけ出されるので、サルが眼球運動をキャンセルできず、ジュースがもらえない場合も多々あります。この認知課題はsaccadic countermanding taskと呼ばれ、精神疾患の患者さんの行動抑制の能力を評価するためにも使われています。

課題遂行中のサルの黒質緻密部と腹側被蓋野からドーパミン神経細胞の活動を記録すると、stop指令が出てサルが眼球運動をキャンセルすることが求められたときに、黒質緻密部のドーパミン神経細胞の活動が上昇することが明らかになりました（図2A）。特に、このドーパミン神経細胞の活動上昇が小さいときは、サルが眼球運動のキャンセルに失敗する確率が高くなりました。ただ、腹側被蓋野のドーパミン神経細胞では、このような活動上昇はほとんど見られませんでした（図2B）。また、黒質緻密部のドーパミン神経細胞から投射を受ける線条体領域（尾状核）からも同様の神経活動の上昇が観察されました。さらには、この線条体領域にドーパミンD2受容体拮抗薬注を注入し、ドーパミン神経細胞からの神経入力を薬理学的に遮断すると、サルが眼球運動のキャンセルに失敗する確立が有意に上昇しました（図2C）。

以上の結果から、黒質緻密部のドーパミン神経細胞から線条体尾状核に対して、不適切な行動を抑制するための神経シグナルが伝達されていることが示唆されました（図3）。

図3 黒質 - 線条体ドーパミン神経路



おわりに

今回の発見は、注意欠陥多動性障害やパーキンソン病などで見られる不適切な行動を抑制できない症状の治療ターゲットとして、黒質 - 線条体ドーパミン神経路が有力な候補であることを示しています。特に、本研究はヒトに近縁なマカク属のサルを用いて行ったもので、その成果はヒトの治療に直接結びつくのではないかと期待できます。今後、黒質 - 線条体ドーパミン神経路を障害したモデルザルを作成し、不適切な行動を抑制できない症状の治療法を探索していきます。

図1 実験室の様子とサルが行なう認知課題

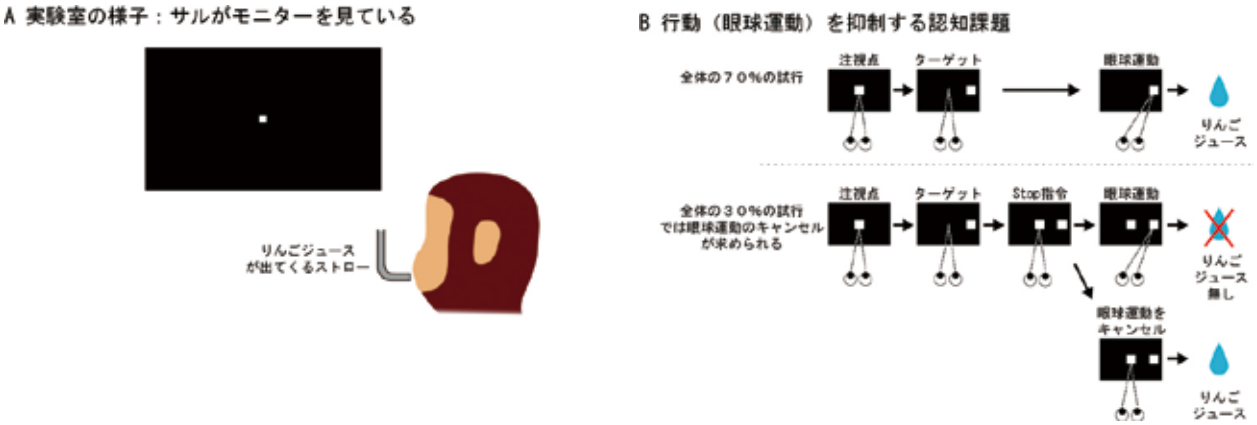
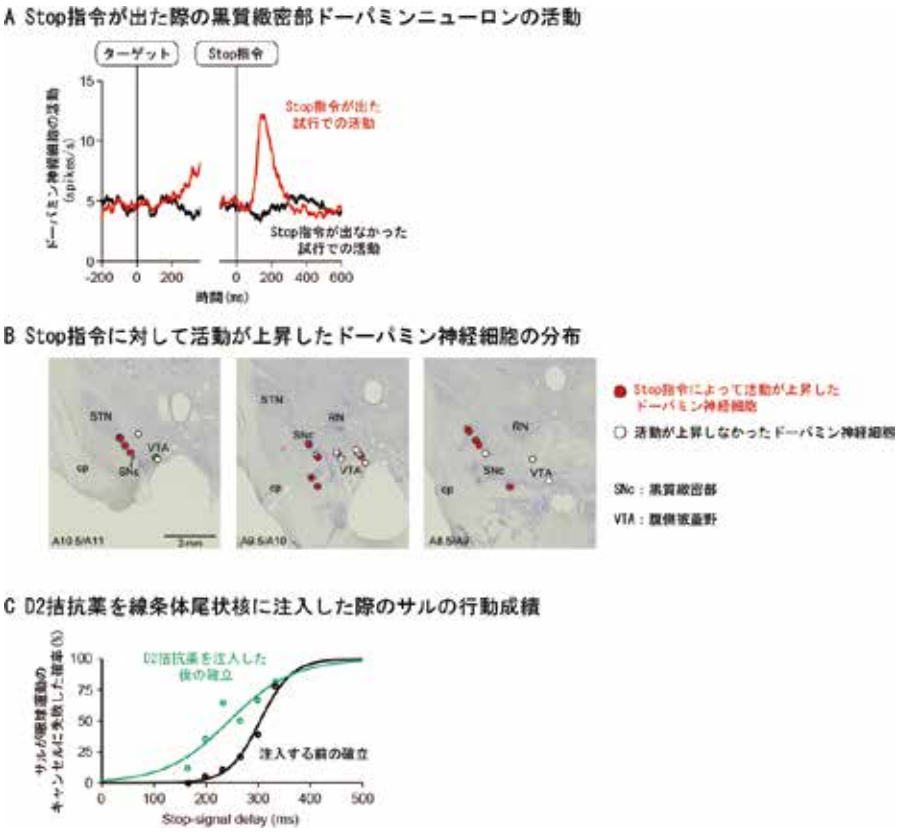


図2 ドーパミン神経細胞の活動と分布、サルの行動成績



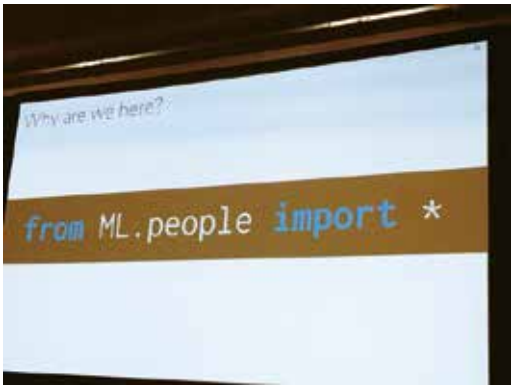
大　会
参加記

NeurIPS 参加記

Thirty-second Conference on Neural Information Processing Systems

谷口 尚平（東京大学 松尾研究室 学部4年）

2018年12月2日から8日にかけてカナダのモントリオールにて行われたNeurIPSに参加してきました。NeurIPSはICML、ICLRと並び機械学習のトップカンファレンスの1つとして知られており、投稿される論文の数ではその3つの中でも最も多く、名実ともに機械学習の最先端の論文が集まる学会と言えるでしょう。2017年まではNIPSという略称が用いられていましたが、性差別を連想させるという理由により今回から新しくNeurIPSを用いることが正式に決まりました。今回も採択率は21%と例年通り低く、厳しい学会であることは変わりありませんが、投稿数の増加に伴い、約1,000本の論文が採択され、会議の参加チケットの販売も会場の収容人数の制限により開始15分程度で打ち切られるなど、この分野の近年の過熱ぶりが伺われました。今回はそんなNeurIPS 2018の様子を写真とともにお届けしたいと思います。

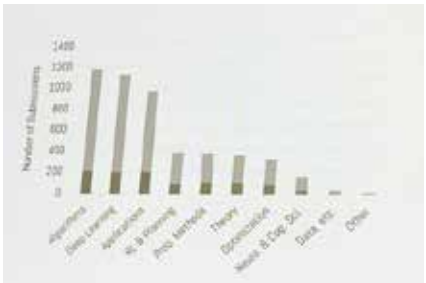


NeurIPSは機械学習全般を扱う学会であり、特に応用研究よりも基礎理論寄りの論文が多く集まりますが、近年の深層学習の盛り上がりを背景に、NeurIPSでも深層学習を用いた研究が占める割合が非常に多くなってきているのを感じました。

その中でも目立ったトピックとしては

1. 機械学習の公平性 (Fairness)
2. モデルベース強化学習
3. 深層学習で用いられるバッチ正規化等の理論的な検証

などが挙げられます。



特に1については、Invited Talkでも多くの研究者が言及しており、注目度の高さが伺われました。この背景には、機械学習の実社会での活用が本格的に進み始める中で浮かび上がってきた現状のデータ・ドリブンな機械学習の問題点として、学習データに含まれるバイアスがモデルの予測に強く影響してしまうという課題があると思われます。このような実社会での応用上の問題に動機付けられた研究は今後さらに増えていくことが予想されます。

「人工知能と脳科学の融合と対照」という新学術領域の観点では、2のモデルベース強化学習は非常に関係が深いと言えます。機械学習、特に深層学習と脳科学のつながりは、脳のニューロンの仕組みを模倣したニューラルネットワークに始まり、視覚野に着想を得た畳み込みニューラルネットワーク(CNN)の登場、さらに大脳基底核の働きと強化学習の関係など様々な視点から研究されてきましたが、これまではそれぞれを個別のタスクに適用するのみに留まっていた印象を持っていました。しかし近年、深層生成モデルの研究などが大きく進んできたことで、画像のような生のセンサーデータを入力として、畳み込みニューラルネットワークを使った深層生成モデルで世界のダイナミクスモデルを構築し、それを用いて強化学習でプランニングを行うという汎用人工知能に向けた枠組みが形になり始めてきたと言えるでしょう。これは、知能の本質とも言える大脳新皮質のモデル化が深層生成モデルを用いて達成され始めているとも言い換えられます。今回のNeurIPSにOral枠で採択されたHaとSchmidhuberの”Recurrent World Models Facilitate Policy Evolution”はその象徴的な論文です。

この論文では、CNNベースの変分自己符号化器(VAE)と再起型ニューラルネットワーク(RNN)を用いて世界をモデル化し、それを強化学習に用いることで、レーシングカーのゲームタスクなどを解いています。これまでは、モデルベース強化学習というと、サンプル効率や別タスクへの転移などの文脈で語られることが多く、純粋に単一タスクに対する性能の面で比較すると、モデルフリーな手法が優位であるというのが常識でしたが、この手法はA3Cなどの強力なモデルフリー手法を大きく上回る結果を残しており、まさに衝撃的な内容です。さらに、このようにして得られた世界モデルを用いれば、エージェントは実際の環境なしでも方策の学習を行うことができるとしており、これを論文内ではTraining Inside of the Dream(夢の中での学習)と呼んでいます。要は、このエージェントは所謂イメージトレーニングができるというわけです。このように、世界をシミュレートするモデルの構築に関する研究は、これまでも注目されてきましたが、実際に強化学習のタスクに適用し、成果を挙げたのはこの研究が初めてであり、大きなターニングポイントであると考えています。囲碁でプロ棋士に勝利したAlphaGoの登場により、近年一気に注目を集め始めた深層強化学習の分野ですが、今後はこのような「世界モデル」の研究と組み合わさって、囲碁のようなゲームタスクだけでなく、より現実的な応用に近い領域での研究が進んでいくのではないでしょうか。事実、UC BerkeleyのSergey Levineらの研究グループは、ここ数年でモデルベース深層強化学習を用いた実ロボット制御に関する論文を量産しており、今回のNeurIPSでも非常に大きな存在感がありました。本プロジェクトでも、各チームの深層学習、神経科学、ロボティクスなどの知識を総動員して、この分野における研究を加速させていけたらと考えています。

最後にNeurIPS 2018全体を通しての印象を総括したいと思います。近年の人工知能の世界的なブームの波はNeurIPSのようなアカデミックな領域にも大きな影響を与えています。企業ブースにはITのみならず世界中のあらゆる分野の企業が並び、積極的にリクルート活動が行われていました。また、採択された論文も企業からのものが非常に多く、大学と企業の共同研究のものもかなり見受けられました。今後もこのような産学連携はより一層進んでいくことでしょう。

一方で人工知能と脳科学の関係という観点は、NeurIPSのような機械学習の学会においてはまだそれほど注目されていません。これは、学会の性質にも起因すると思われますが、それ以上に両者の関係が学術的な考察に至るほど十分に議論されていないからとも考えられます。本プロジェクトでも今後、双方の研究者間での議論を深め、その成果を積極的に発信していくことで、研究コミュニティの国際的な発展にも貢献していきたいと考えています。

イベント
報　告

Joint workshop of UCL-ICN、NTT、UCL-Gatsby and AIBS Analysis and Synthesis for Human/Artificial Cognition and Behaviour参加記

2018/10/22 -23 OIST シーサイドハウス

安部川 直稔（NTT コミュニケーション科学基礎研究所 主任研究員）

10月沖縄の海を眺望するシーサイドハウス（OIST）で、本ワークショップは開催されました。NTT（NTT コミュニケーション科学基礎研究所）とUCL（University College London）の共同研究プロジェクトと当領域の合同開催であり、14件の講演と18件のポスター発表、約40名の参加者で賑わいました。神経科学から機械学習まで多様な発表が行われたことはもちろんのこと、異分野を融合させた萌芽的研究も見受けられました。

認知心理学の分野からは、UCL側の代表者でもあるHaggard氏が、戦略変換を運動の自由意思と捉え、運動準備電位との関係性を議論。他にUCL-ICNからは、Ward氏が自閉症患者の行動と社会性を結びつける内容を、Bahrami氏が、ヒトが他者への影響（アドバイスなど）を高める際にとり得る行動戦略と脳内機序について、それぞれ発表しました。認知心理的テーマを、実験環境から実社会へと応用させるこれら研究は、まさにUCL-ICNが世界を牽引してきた分野で、さらにBestmann氏は自由行動下での脳活動計測を実現する新手法（量子センサ×MEG）を紹介。これらは社会へのメッセージ性が高い重要な知見である一方で、因果関係を含めた脳・計算メカニズムの解明にいかにして繋げていくのか、そこには実社会ビッグデータを用いたAI解析が役に立つのでないかと聴衆を含めて活発な議論が交わされました。

視聴覚機構については、NTTから澤山氏が視覚的質感知覚について、古川氏が聴覚的注意について、廣谷氏が“聞くと話す”の共通脳機能について紹介。これらは視聴覚という基本モダリティを対象としながらもより高次な環境認識に係る脳機能解明を目指しており、独創的な研究手法・独自の開発技術が挑戦的テーマを下支えしている点が特長的でありました。聴覚についてはさらに寺島氏が、自然音刺激とDNNを利用してマウスの聴覚野地図を同定する試みについて発表。これら一連の研究には、①強い仮説を前提としない、②実環境に近い自然刺激を利用、③複雑な神経応答・生理的応答の解析に機械学習を利用、というパラダイムシフトが見られます。特に深層学習後の中間層表現の解析は、AIと神経科学を融合させる有力な切り口であり、今後、新仮説・新コンセプトの提案につながっていくものと思われます。

機械学習については、NTTから亀岡氏、UCL-GatsbyからSahani氏が、それぞれ生成モデルの観点から発表しました。Sahani氏は機械学習を神経科学に応用してきた世界的権威で、生物の知覚は不確かな情報をベイズ的に統合し推論する機能であることを強調。そのアナロジーから、複雑な不確かさを表現できるDistributed Distributional Codeについて紹介しました。

当領域からは、視覚情報に基づく運動制御について、五味氏が人間科学の観点から、森本氏がヒューマノイドロボットの

観点から発表。両氏共に、環境との相互作用の中で運動プリミティブを獲得することが、ロバストな実時間運動制御に貢献することを訴えました。さらに谷口氏は、環境との相互作用を知能創発まで発展させて論じます。従来のSLAMを拡張させたSpCoSLAMでは空間をクラスタリング、音声認識や各種AIモジュールとの有機的統合を経て場所概念の学習が可能となるコンセプトを提案しました。また、当領域代表の銅谷氏は、内部モデルを利用した状態遷移予測をメンタルシミュレーションとして捉え、関連脳部位の同定、神経活動まで含めた計算理論の提案を行いました。

会議全体の研究内容を俯瞰しますと、従来は独立した機能として議論されてきた運動・知覚・認知などの各種モジュールが、様々な側面で有機的に結びつく多様的ネットワークの中に知能の本質があるのでは思えてきます。環境とのインタラクションの中で、脳はどのようにそれらモジュールを階層的に表現するのか、その階層構造はどのように更新・固定化されるのか、機械学習ともおおいに影響しあう魅力的な研究発展が期待されます。

ICDL-EpiRob 2018 参加記

The 8th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics

谷口 彰（立命館大学 総合科学技術研究機構 日本学術振興会 特別研究員 (PD)）

認知発達ロボティクスに関する国際会議 The 8th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob) が2018年9月16日から9月20日にかけて東京の早稲田大学にて開催されました。この会議は2011年より開催されており、発達と学習に関する国際会議ICDLとエピジェネティックロボティクスに関する国際会議EpiRobの共同開催という形で行われています。人工知能、ロボティクス、心理学や認知科学に関する研究者が集まり、近年注目されているディープラーニングを始めとする神経回路モデルや人間の認知発達モデルに関する密な議論が行われました。私自身はこの会議には初参加でしたが、どの発表も興味深いものばかりでした。

本会議に先駆けて行われたワークショップでは、“Understanding Developmental Disorders: From Computational Models to Assistive Technology”、“Active vision, Attention, and Learning”、“Continual Unsupervised Sensorimotor Learning”の3つのパラレルセッションが開かれました。中でも、二体のロボットが教示者と学習者に分かれ互いの視点で見える物事について言語を使って教示し学習し合うDr. Michael SprangerのSpatial Language Learningのトークが印象に残りました。

本会議の講演発表はシングルトラックのオーラル発表とポスター発表で構成されており、程良い規模感の会議でした。基調講演では、Prof. Oliver BrockはDevelopmental Theories of AIについて、銅谷賢治先生はWhat can we further learn from the brain for AI and robotics? というタイトルで人工知能と脳科学に関する取組みについて、Prof. Peter J. Marshallは Embodiment and Human Developmentについて、SONYの藤田雅俊氏はAI×Robotics in Sonyと題し新型AIBOに関する取組みについて、それぞれ話されました。

一般講演では、Neural Networks、Action selection and learning、Architectures、Body schema、Social Learning、Language learningを中心とした多岐にわたるものでした。中でも、Michael Garcia Ortizらの“Learning Representations of Spatial Displacement through Sensorimotor Prediction”やAlexandre Antunesらの“Solving Bidirectional Tasks Using MTRNN”は特に興味深かいものでした。今回の国際会議の参加は、認知発達ロボティクス分野の面白さを実感させられる有意義な機会となりました。

日本神経回路学会 第28回全国大会（JNNS2018）参加記

水谷 晃大（大阪大学大学院 理学研究科 博士前期課程 1 年）

2018年10月24日から10月27日の4日間にわたって行われた「日本神経回路学会　第28回全国大会 (JNNS2018)」に参加しました。会場は沖縄科学技術大学院大学(OIST)で開催され、国内外から多くの研究者が集まり、神経情報処理に関する研究発表・議論が盛んに執り行われました。ニューラルネットワークや機械学習についてほとんど知識がない状態で参加しましたが、初日のチュートリアルでは基本的なところの説明を聞くことができ、理解がとてもスムーズになりました。

はじめのチュートリアルでは、東京大学の松尾豊先生らのグループから、深層学習についてのイントロダクションから最近のトレンドをフォローした内容のご講演をいただきました。機械が知能を持っていることを示す指標として、チューリング・テストに合格するというのがあるようで、そのために、機械が言語を使って被験者とコミュニケーションを取りながら(自然言語処理)、内部に蓄積した情報を表現すること(知識表現)、そして、新しい状況に適応し、そこにあるパターンを検出する(機械学習)といった様々な認知機能を実装することが求められ、そのため幅広い専門領域にわかれて研究活動が進められていることを知りました。

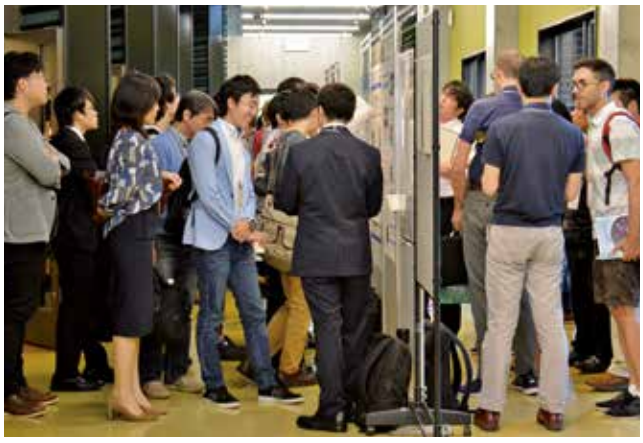
京都大学の神谷之康先生からは、脳情報のデコーディング技術についてのご講演をいただきました。最も印象深かったのは、被験者が何を見ているかを脳情報から読み解く「Generic object decoding」という技術でした。これは、様々な言語ラベルのついた一連の視覚刺激呈示中のfMRIボクセルデータの特徴量をトレーニングデータとして用い、新たに視覚刺激呈示中の被験者のfMRIデータから、それがどの言語ラベルの特徴量に最も類似しているかを推定する手法のようです。さらに、「deep image reconstruction」という技術では、被験者が見ている実際の視覚情報を脳活動から再構築することができるとのことで、最先端のブレインデコーディングの技術進歩に感銘を覚えました。

基調講演では、理化学研究所の甘利俊一先生から、フィッシャー情報行列の理論的な説明から勾配学習の実装可能性について、最新のニューラルネットワーク理論のご講演を、University College LondonのManeesh Sahani先生からは、脳の知覚処理における推論と機械学習の推論についてのご講演を頂きました。

私自身は、2日目の口頭発表・ポスター発表において「Characteristic Whisker Movements Reflect the Internal State of Mice Related to Reward Anticipation」というタイトルで発表させて頂きました。本研究では、マウスの顔面表現の1つである洞毛(whisker)の動きが、聴覚Go/No-Go課題遂行中において、報酬予測時に特徴的な動きのパターンを示すことを新しく見つけ、さらに教師あり学習を用い、マウスの洞毛の動きから課題成績の推定に成功したことを報告致しました。身に過ぎて光栄なことに、本研究に関して、会長の大森隆司先生から大会奨励賞を賜りました。

大会最終日の懇親会では、食事後に創作エイサーと獅子舞の催し物が用意されておりました。夜の静かなOISTで民謡を歌いながら軽やかに舞う様子が神秘的で、圧巻されました。最後は参加者全員で輪をつくり、楽しく歌いながらダンスをしました。

日本神経回路学会に参加したのは今回が初めてでしたが、モデリングやシミュレーションの必要性など、生命科学系の学会ではあまり疑問を抱かれない数理的な角度から、鋭いご指摘・助言をたくさんご教授いただきました。今回の大会参加を経て、人工知能の今後の発展において、脳科学・行動科学がどのような貢献をできるかを考える大変有意義な機会となりました。大会実行委員長の銅谷賢治先生をはじめ、企画・運営を行ってくださった先生方、参加者の皆様にここに厚く御礼申し上げます。



The 2nd Conference on Robot Learning (CoRL 2018) 参加記

Guilherme Maeda・森本 淳（ATR脳情報通信総合研究所 ブレインロボットインタフェース研究室）

The 2nd Conference on Robot Learning (CoRL 2018) was held in Zurich, Switzerland from 29 to 31 of November 2018. Similar to the first edition that was held in Mountain View, California in 2017, the large participation from North American institutions (such as Berkeley, Stanford, CMU, and MIT) was evident. The participation from industry was also noticeable with many papers affiliated with Google, NVIDIA, Uber, and Baidu; and plenary talks from Sony and Preferred Networks. Most of the talks were focused on some form of deep representations or deep learning and it seems already quite common-place in CoRL the use of deep neural networks.

So far, the large majority of research in robotics and deep learning has been carried out and applied almost exclusively in simulation; and this CoRL edition was not an exception. As such, one of the main topics of discussion during the conference was the use of simulation and its transferring to real environments. The CoRL community seem to have adopted a number of different names to refer to the fact that policies learned on models do not transfer to real robotic systems, such as "sim-to-real", "sim2real" or "reality gap"; which gives a new flavor to an old control problem.

Two other topics received a fair amount of attention were self-supervised learning and benchmarking. Self-supervised learning was discussed in a few papers, particularly focusing on how to train deep architectures in a supervised manner, without recurring to an exorbitant amount of labeled data. An interesting paper was based on crowdsourcing the collection of training data for robotics application by using a smartphone app. Also, many papers discussed benchmark tasks or datasets to alleviate the reproducibility problems that are particular of robotics. The first session of CoRL was almost entirely dedicated to such papers, and the plenary talk of Joelle Pineau also strongly emphasized this issue. It seems benchmark discussions and reproducibility have been, in part, a healthy consequence of the influence of the computer vision/perception community.

Other highlights of the conference were as follows. Some papers such as "Particle filter networks with application to visual localization" [Karkus P. et al.] and "Policies modulating trajectory generators" [Iscen A. et al.] bring the interesting concept of combining classical methodologies with deep neural networks. In a similar spirit, the paper "Expanding Motor Skills using Relay Networks" [Kumar, et al.] also combines classical Markov Decision Processes with deep learning in the form of curriculum learning. Finally, the talk of Gregory Stein, a student from MIT, was extremely well done and received the best presentation award. Although some people in the audience frowned upon the true novelty of the method, the presentation was remarkably clear and organized and is a good reference for young students learning how to prepare a talk.

CoRL2018においては、6件のキーノート講演、4件のチュートリアル講演が豪華な講師陣を招いて行われました（http://www.robot-learning.org/home/program#keynotes）。日本からも3名の講師の方々にご講演いただきました。なお採択論文（採択率3割程度）はProceedings of Machine Learning Research (PMLR)（http://proceedings.mlr.press/v87/）に公開されています。ところで、本年(CoRL 2019)は日本（大阪）での開催が予定されています。日本からの多くの投稿および参加が期待されます。

第７回 NIPS+ 読み会・関西開催記

2018/11/11 立命館大学 大阪茨木キャンパス (OIC)

https://connpass.com/event/105338/

内部 英治（国際電気通信基礎技術研究所脳情報研究所ブレインロボットインタフェース研究室 主幹研究員）

関西地区の機械学習に関わる研究者、エンジニア、学生間で最新の機械学習の研究動向を把握および共有することを目指した論文読み会を当領域の共催イベントとして運営のお手伝いをしています。通常5,6人の発表者が持ち時間25分程度で論文を紹介し、その後にさらに5分を使って会場を交えてディスカッションします。

第７回となる読み会では、最初に私がICLR2018で発表されたTemporal Difference Models: Model-Free Deep RL for Model-Based Controlについて紹介しました。モデルベース強化学習は環境の状態遷移確率を使って明示的にプランニングできることが利点の一つですが、ロボティクスなどの制御課題に適用した場合には制御に有効な長期予測が困難でした。この研究では状態行動価値関数を状態と行動だけでなく、目標状態とプランニングする時間ステップの関数に拡張し、TD学習によって長期予測する手法を提案しました。推定された価値関数はモデル予測制御の制約条件として利用されます。モデルの利用方法としてとても興味深いものでしたが、様々な時間ステップのもとで価値関数を推定する必要があるなど、まだまだ改良の余地があるものではないかというのが議論となりました。次の発表は大阪大学の小林京一郎さんがMaximum Causal Tsallis Entropy Imitation Learningを紹介されました。強化学習と逆強化学習を組み合わせた模倣学習は最近の模倣学習のトレンドであり、そこではエントロピ正則された報酬が重要な役割を持ちます。本研究は、通常のエントロピ正則された強化学習で用いられるShannonエントロピをTsallisエントロピに拡張することで、方策の分布が多峰性であっても意図した最適方策が学習できることを示しました。続けて、次に大阪大学の立川和樹さんがSanity Checks for Saliency MapsとA Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizationを中心に、深層学習で学習されたネットワークの識別根拠を解釈するための方法について紹介されました。

休憩をはさんで再び私が逆強化学習の最近の応用研究としてCan AI predict animal movements? Filling gaps in animal trajectories using inverse reinforcement learningとModeling sensory-motor decisions in natural behaviorについて紹介しました。これまで逆強化学習は前述の模倣学習の構成要素の一つとして用いられることが多かったのですが、前者は海鳥の飛行経路、後者は室内環境における人のナビゲーションを解析する方法として逆強化学習が用いられており、実際の問題に逆強化学習がどのように適用されているかを中心に紹介しました。最後に奈良先端科学技術大学院大学の品川政太朗さんが今年のNeurIPSで発表されたGenerating Informative and Diverse Conversational Responses via Adversarial Information Maximizationを紹介されました。これはテキスト対話におけるユーザの発話に対して、システムの応答の多様性(diversity)と情報度(informativeness)を改善するために、敵対的生成ネットワークの枠組みに相互情報量最大化を導入したものです。現時点では人間の比べられるほど性能は良くなかったのが残念でしたが、情報度の定義について議論が盛り上がりました。

読み会で使用したスライドは上記のアドレスから参照できます。また今年度中にもう一度開催する予定です。興味のある方は私までご一報ください。

新学術領域研究「人工知能と脳科学」第５回領域全体会議参加記

小口 峰樹（玉川大学 脳科学研究所 特任助教）

2018年11月12日から13日にかけて、国際電気通信基礎技術研究所（ATR）にて、新学術領域研究「人工知能と脳科学」第５回領域全体会議が開催されました。人工知能と神経科学のさまざまな分野にわたり先端的な研究を行っている研究者が集い、それぞれの研究成果について白熱した議論が交わされました。

12日は、領域代表の銅谷賢治先生よるご挨拶に続き、まず、ATRの細谷春夫先生から、NEDO-AI連携講演として、「霊長類の脳における顔処理システムの計算モデル」についてご講演がありました。ご講演のなかでは、脳のIT野を中心とした顔認識系の計算モデルについて、要素的な特徴ごとに構築されたスパースコーディングモデル群を融合させるというアプローチからの試みをご紹介いただきました。

続いて、A01計画研究班「知覚と予測」より、松尾豊先生のグループから、深層生成モデルを用いた寡少観察データからの世界モデルの構築、銅谷賢治先生から、カルシウムイメージングを用いた線条体下位構造の分析やオプトジェネティクスを用いたセロトニンニューロンの機能解析、田中啓治先生のグループから、一致性連続効果に関する眼窩前頭皮質の関与、等に関する研究成果が紹介されました。

次に、ATRの田中沙織先生から、新学術領域研究「思春期主体価値（脳・生活・人生の統合的理解にもとづく思春期からの主体価値発展学）」連携講演として、「大規模家族コホートデータセットを用いたホーリスティックな多変量解析」についてご講演いただきました。ご講演の中では、「東京ティーンコホート」プロジェクトによって収集されたビッグデータを用いた解析結果についてお話しいただきました。

その後、A02計画研究班「運動と行動」より、松本正幸先生から、経済的意思決定におけるドーパミンニューロンと眼窩前頭皮質ニューロンとの比較、疋田貴俊先生から、DISC1遺伝子変異型マウスを用いた統合失調症モデル研究、森本淳先生から、新しい近似モデルを用いたロボットの運動制御システムの構築、五味裕章先生から、運動制御に対する視覚的な背景情報のダイナミックな影響の分析、等に関する研究成果が紹介されました。

12日の夜には公募班のメンバーを含むポスター発表が行われ、大規模同時神経活動記録法を用いた意思決定研究や、ディープニューラルネットワークを用いたfMRIデータ解析、計算論的神経科学による精神疾患研究など、多岐にわたる研究成果が披露されました。

13日には、A03計画研究班「認知と社会性」より、中原裕之先生から、強化学習の枠組みに基づく社会的意思決定の研究、谷口忠大先生から、ノンパラメトリックベイズ二重分節解析器（NPB-DAA）を用いた言語研究やモジュール群の階層的な結合による大規模認知モデルの実現、高橋英彦先生のグループから、統合失調症における視覚弁別やカテゴリー形成の障害に関する研究、等についての研究成果が紹介されま

した。また、坂上雅道先生のグループから、私が「化学遺伝学2重遺伝子導入法を用いたマカク前頭前野−線条体経路の機能解明」について報告しました。これは、特定の化学物質を投与することによって神経経路を可逆的に抑制することのできる遺伝子操作技術を用いて、前頭前野と線条体の尾状核とを結ぶ経路の機能解明に迫った研究で、現在までに得られている行動および神経活動の解析結果について紹介を行いました。具体的には、逆行性および順行性に感染可能な2種類のウイルスベクターをそれぞれ尾状核と前頭前野外側部に打ち、尾状核に投射する前頭前野外側部のニューロンのみに人工受容体を発現させます。この人工受容体は、通常生体内に存在しない化学物質を投与することによって作動させることができ、それによってこの受容体を発現したニューロンの活動を抑制することができます。2重感染後のサルにおいてリガンドとなる化学物質を用いて課題遂行中の行動および神経活動の解析を行うことで、前頭前野外側部の尾状核投射ニューロンが特に抑制機能に関わっているということを示唆する結果が得られています。

最後に、「多目的制御」、「ゼロショット／転移学習」、「モジュールの選択と結合」という3つの主題に関して、3つのグループに分かれてのグループ討論会が行われました。私は多目的制御のグループに入り、容易には比較しえない多様な目的間で脳がどのように最適化を行っており、そうした最適化をどのように人工知能やロボットに実装しうるかについて、メンバー間で知見をすり合わせながら議論を行いました。ある目的を達成するための行動選択や価値比較が脳でどのように行われているかについては、これまで、ドーパミンニューロンの働きを中心に多くの知見が得られています。しかし、価値尺度が容易には定まりそうにない複数の目的の間でどの目的を優先させるかといった問題に対しては、モデルフリー／モデルベースシステムの競合といった観点からある程度の研究の蓄積はあるものの、まだ問題の定式化すら十分には行われていない状況です。討論会の中では、こうした問題へどのようなアプローチが有効かに関して、神経科学と人工知能双方の研究者間で、前頭前野の役割や価値関数の設定に関してなど、様々な観点から意見が出されました。討論後は、それぞれのグループでどのような議論が交わされたかについて、各グループの代表者から全体へ向けて発表がなされました。

今回の領域全体会議は、本新学術領域の前半期間で得られた諸成果を確認した上で、特に、人工知能と神経科学の間での融合的な研究を後半期間でより強力に推進するという課題意識を共有する重要な機会となったと感じました。この場を借りて、本会議の運営を支えて下さった皆様に謝意を表したいと思います。

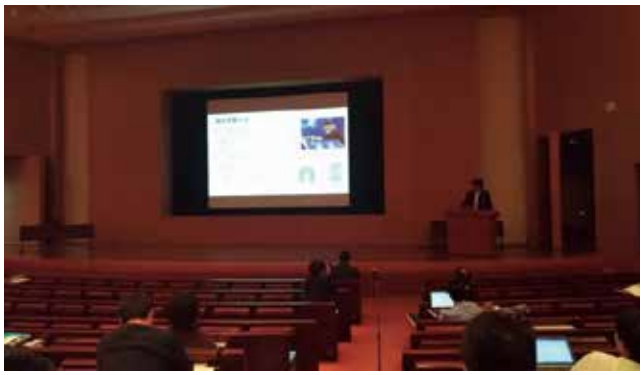
次世代脳プロジェクト　冬のシンポジウム参加記

福田 玄明（理化学研究所 脳神経科学研究センター 学習理論・社会脳研究チーム 客員研究員）

私は昨年（2018年）、12月12日～12月14日に一橋大学一橋講堂にて開催された「次世代脳プロジェクト」冬のシンポジウムに参加しました。次世代脳プロジェクトは、「我が国の脳科学研究のさらなる発展と次世代を担う中堅・若手研究者の育成を目指した取り組みを行う」プロジェクトで今回が3回目のシンポジウムとなるそうです（パンフレットより）。その目的通り、[「個性」創発脳]、[共創言語進化]、[人工知能と脳科学]、[思春期主体価値]の4つの新学術領域研究の合同若手シンポジウムや[適応回路シフト]、[身体性システム]、[オシロロジー]、[人工知能と脳科学]、[脳情報動態]の5領域合同シンポジウムなど、脳科学研究に関わる様々な研究者が一同に会しての領域縦断的な内容となっており、普段はあまり見聞きしない近隣分野の最近の動向にも触れることができ、大変ためになるシンポジウムでした。

4領域合同若手シンポジウムでは、「個性」の階層的理解を目指し空間弁別能力の分子・細胞学的基盤に関する研究や脳磁図を用いた幼児を対象とした言語発達研究からエージェント・シミュレーションを用いた思春期の行動特性の研究など幅広い研究が紹介されました。[人工知能と脳科学]からは、東北大学の鈴木真介先生からfMRIと計算論的手法を組み合わせた“他者との駆け引き”に関わる神経基盤についての2つの研究成果が報告されました。ひとつは、コンセンサスによる意思決定に関するもので他者がどの程度自分の意見を変更しそうかとういう「がんこさ」を我々が考慮して自分の行動を決めていることをモデルと脳イメージングにより明らかにされていました。もうひとつは、他者に自分がどう思われていると思うかという2nd-order beliefを、面白い実験課題と複雑なモデルをうまく組み合わせることで実証し、その神経基盤をfMRIとTMSによって明らかにするという非常によく考えられた素晴らしい研究でした。私自身、fMRIを用いて意思決定の研究を行なっているため、鈴木先生のお話は大変興味深く勉強になりました。

5領域合同シンポジウムは2部構成となっており、第1部は各領域の若手研究者による最先端の研究成果と今後の方向性の紹介、第2部は脳科学に関わる理論についてのチュートリアルでした。第1部では、サルの小脳の計算モデルに関する研究やラットの場所細胞での他者の場所の表象からてんかん発作の予測・検出に関わる研究まで大変幅広く、人間対象のイメージング実験に関わっている私としては理解が難しい部分も多かったのですが、それでも興味深く聞き入ることができました。第2部では、人工知能による神経系ビッグデータ処理として、京都大学の石井先生が巨大化する脳神経系のデータに対してのAIによるデータ処理について紹介されました。特に高解像度画像を得る技術に驚かされました。次に、東京大学の郡先生から、振動や同期を単純化して記述するための数学的方法についての紹介がありました。大変高度な内容で私自身は自分の研究と結びつけて考えることはできなかったのですが、そのような普段は触れない知識にも触れることができるところがこのシンポジウムの良いところだと思いました。最後にATRの内部先生から、逆強化学習についてのチュートリアルがありました。逆強化学習に関しては、以前から興味は持っていたのですが、きちんと勉強することができずにいたので、この機会に最先端の研究者の説明が聞けたことは大変有意義であったと思います。



今回、次世代脳プロジェクト冬のシンポジウムに参加して、自分が普段どれだけ狭い領域の知識の中でものを考えているのかということを気付かされ、領域を縦断する成果を聞ける機会の重要性を知ったように思います。今後もこのようなシンポジウムにできるだけ参加し、知見を深めていきたいという気持ちになりました。

脳と心のメカニズム 第19回 冬のワークショップ 参加記

高椋 慎也（NTTコミュニケーション科学基礎研究所 主任研究員）

2019年1月9日から11日にかけて、脳と心のメカニズム冬のワークショップが開催されました。この会議は、神経科学、動物行動学、機械学習、ロボティクスなど多様な分野の、とくに計算論に興味を持った研究者が多く参加する合宿形式のワークショップであり、今年は「認知発達と発達障害：予測符号化の視点から”Cognitive development and its disorders: From the viewpoint of predictive coding“」というテーマで、9件の講演と59件のポスター発表がありました。

初日のスペシャルセッションでは、Caroline Catmur先生（King's College London）がAlexithymia（自らの感情を認知することの障害）、自閉症、内受容感覚の関係について調べた一連の研究を、橋本龍一郎先生（首都大学・昭和大学）が安静時の脳機能ネットワークを発達障害の診断やニューロフィードバックを通じた治療に利用する試みを、そして、James Kilner先生(University College London)が、他者運動の微細な差異に基づく心的状態の推定や、振動刺激による運動の高速化など、複数の興味深い現象を、予測符号化の理論に関連付けて説明されました。

2日目のトピックセッションでは先ず、Jean-Jacques Slotine先生(Massachusetts Institute of Technology)の基調講演がありました。先生は非線形システムの理論的研究で世界的に有名な方ですが、今回は非線形システムのContractionの理論について、状態推定、引き込み、テレオペレーション、ニューラルネットワーク、自然勾配学習法など、多岐にわたるトピックと関連付けて紹介されました。続く、Rebecca Lawson先生（University of Cambridge)は、自閉症患者が様々な知覚課題において、感覚刺激をそのまま反映した知覚をしやすい性質を持つことを紹介し、その性質が、どういう状況で学習を行うべきかいうことを判断するメタ学習の異常によって説明されることを示す研究成果を紹介されました。最後には、当領域の山下祐一先生（国立精神・神経医療研究センター）が、計算論的精神医学という新しく立ち上がりつつある分野について紹介され、その1つの試みとして、自閉症患者の代表的な症状である感覚過敏・鈍磨と反復行動の関係について、構成論的アプローチによって考察されたご研究を紹介されました。

最終日のトピックセッションでは、先ず、大隅典子先生（東北大学）が、父親の高齢化を近年の自閉症の発症率上昇とその多様性の要因として検討された一連のご研究（ラットを用いた動物実験）について紹介され、続いて、高橋哲也先生（福井大学）が、ネットワーク構造解析や複雑性解析のアプローチとその手法を用いた発達障害に関するご研究を紹介され、最後に、竹内倫徳（Aarhus University）が新奇刺激の提示による記憶の定着化や記憶の同化(assimilation)のメカニズムについて、動物実験を通じて調べられた一連のご研究を紹介されました。

ポスターセッションは1日目と2日目にそれぞれ実施され、例年のことですが、長時間にわたり白熱した議論が行われていました。とくに今年は、優秀なポスター発表を表彰するポスター発表賞が新設されたことで、今まで以上にわかりやすい発表や充実した議論が行われていたように感じられました。

発達障害を計算論的アプローチから明らかにしようとする試みは、近年、世界的にも注目を集めている分野であり、当領域でも、高橋先生（京都大学）、山下先生（国立精神神経医療研究センター）、三村先生（量子科学技術研究開発機構）などが取り組まれていますが、本会議の参加を通じて、この分野が、実に多様なアプローチを内包したものであることに改めて気づかされました。また、発達障害を抱えた方の中には、ある側面において、健常者を圧倒する知能を備えた方がいるという、大隅先生が紹介されたお話も印象的でした（イギリスの画家Stephen Wiltshireなど）。ヒトに学び、新たな人工知能をつくる、ということを達成する上で、こういう特殊な事例から学ぶという発想も面白いのではないかと思いました。次回の会議は、来年の1月8日から10日に開催。テーマは「ヒトの学習と機械の学習（仮題）」ということで、当領域のメンバーには関連の強いものになりそうです。

イベント情報

平成31年度 主催イベント

第6回領域会議
日程：2019.5.14-15
場所：東京都、玉川大学

新学術領域「人工知能と脳科学」第2回サマースクール
日程：2019.7.31-8.2
場所：埼玉県、理化学研究所 脳神経科学研究センター (RIKEN CBS)

脳と心のメカニズム第20回冬のワークショップ
日程：2020.1.8-10
場所：北海道、ルスツリゾートホテル

平成31年 共催、協賛、後援、関連イベント

IRCN神経科学コンピューテーションコース
日程：2019.3.21-24
場所：東京都、東京大学本郷キャンパス
https://ircn.jp/neuro_inspired

SBDM 2019
日程：2019.5.27-29
場所：United Kingdom、Oxford
http://sbdm2019.isir.upmc.fr/

OCNC2019 沖縄計算神経科学コース
日程：2019.6.24-7.11
場所：沖縄県、沖縄科学技術大学院大学
https://groups.oist.jp/ja/ocnc

AI AND BRAIN

発行／編集　新学術領域研究「人工知能と脳科学の対照と融合」
お問い合わせ　新学術領域研究「人工知能と脳科学の対照と融合」事務局
Mail ncus@oist.jp
2019年3月発行

www.brain-ai.jp