「第49回世界情報通信社会・電気通信日のつどい」記念講演より 人工知能は人間を超えるか ―ディープラーニングの先にあるもの―

東京大学大学院　工学系研究科 特任准教授　松尾（まつお）　豊（ゆたか）

1. はじめに

2016年末、Googleの人工知能（AI）アルファ碁の進化版が日中韓のプロ棋士を相手に60連勝した。これは、研究者に衝撃を与えた。AIが囲碁でトッププロに勝てるようになるのは、2025年頃であろうとの予測を、約10年前倒ししたからだ。その最大の要因が、ディープラーニングという技術である。

AIという分野は1956年から始まったとされ、今年で61年目である。この間、過大な期待と、反動としての失望を繰り返し、今回が3回目のブームである。今、AIと言われているものの8割がたは、従来ITと言っていたものを言い換えているに過ぎない。期待感をあおり過ぎず、できること、できないことを見極めていくことが重要である。

2. ディープラーニング革命

一方で、ディープラーニングというのは、潜在的な可能性が大きく、今回のブームの最大の注目点である。単純化すると、認識、運動の習熟、言葉の意味理解ができるようになるということだ。

2.1　認識の難しさ

ここに、ネコ、イヌ、オオカミ、3枚の写真がある（図1）。人間が見ると簡単に見分けることができるが、コンピュータに見分けさせるのは難しい。目が丸ければネコ、目が細長い場合に、耳が垂れているとイヌで、耳がとがっているとオオカミと判断しようと決めると、自動的に見分けられるように思う。ところが、シベリアンハスキーのように耳が細長くてとがっているイヌがいる。確かに、イヌっぽい顔をしているな、それに比べると、右上のオオカミの写真は、よりオオカミっぽい顔をしているなと思う。このオオカミっぽさ、イヌっぽさを特徴量という。これを人間が定義している限りは、画像認識の精度は一向に上がらなかった。特徴量自体をコンピュータが学習できるという仕組みが必要で、これを徐々にできるようになってきたのが、ディープラーニングである。

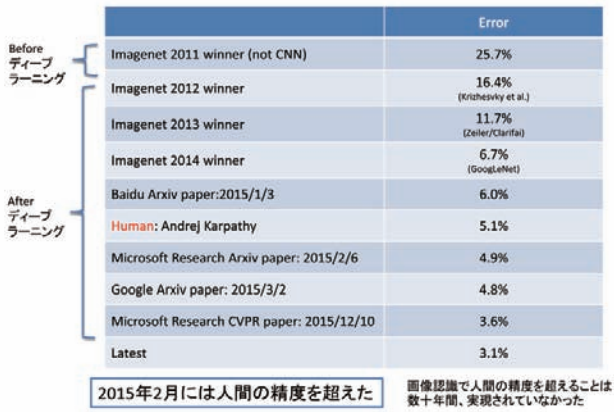
■図1.　認識の難しさ

「Googleの猫」の研究が行われた2012年頃から、ディープラーニングが画像認識で非常に大きな成果を出している（図2）。画像認識というのは、同じデータセットを使って、世界中の研究者がその認識の精度を競う。2015年2月には、特定のデータセットに対してではあるが、コンピュータが画像認識で人間の精度を初めて超えた。

2.2　運動の習熟

強化学習とは昔からある技術で、うまくいった＝報酬が得られると、事前の行動を強化し、上達する。どういう状況でどういう行動をすると、いいか悪いかを学んでいくのだが、従来、状況を記述する特徴量を人間が定義していた。ディープラーニングと組み合わせる方法では、ディープラーニングで出てきた特徴量を使って状況を定義する。

2013年の研究では、ブロック崩しを学習させるAIを作っ



■図2.　認識：2012年以降のエラー率の変化

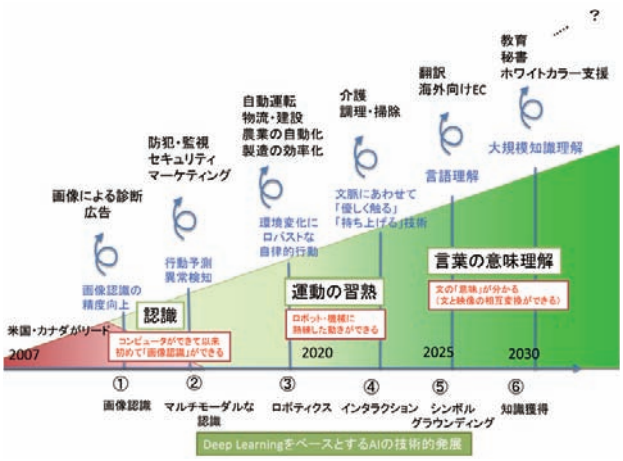
ていた。スコアを報酬にして、スコアが上がるとその前の行動を強化する仕組みによって、だんだん上達し、しばらくするとコツを見つけられるようになる。また、驚くべきことに、画像を入れてスコアを報酬にしているだけなので、全く同じプログラムで、全然違うゲームを学習できる。パズルや冒険等、記憶や思考を必要とするものは人間のほうがうまいが、反射神経、運動神経だけでよいゲームは、もうコンピュータのほうがうまい。

この技術は2015年以降、実世界にも適用され始めた。昨年3月にGoogleが出した研究では、ロボットアームが箱の中から目的のものを取り出す練習をしている。カメラがあり、画像で見ている。14台を並列に並べ、14倍高速に学習させている。たくさんのものが入った箱の中から、目的のものを取り出すというのはとても難しいタスクで、これまでは画像認識の精度が悪く、目的のものがうまく見つからなかった。見つかったとしても、ものによって違う持ち方を人間が定義する必要があり、たくさんのものがごちゃ混ぜに入っていると困難であった。ところが、画像認識で特徴量を取り出して、ものの特徴と持ち方を試行錯誤で学んでいくので、いろいろなものを上手に持てるようになる。人間が赤ちゃんのときに練習しているのと同じプロセスを経て、機械も上手に持てるようになってきたということだ。第3次AIブームの技術的なエッセンスは、3歳児でもできるようなことが、コンピュータにようやくできるようになってきたということである。

2.3　言葉の意味理解

AIの技術的発展は、子どもの発達の過程とよく似ている。まず目で見て分かるようになる。次に体の使い方が上達する。すると、いろいろな概念を理解できるようになるので、言葉の理解ができるようになる。2年半前に書いたこの図はほとんど当たっているが、スピードについては、2030年どころではなく、既に言葉の意味理解というところまで研究が進みつつあり、相当ペースが早い（図3）。

新しい研究の一つはイメージキャプショニングといい、画像から文を生成する技術である。写真を入れると、A man in black shirt is playing guitar. という文が出てくる。つまり、画像認識だけではなくて、その画像の中のことを文として表現するような技術が出てきている。さらに、逆に文を入れると絵が出てくる。画像を検索ではなく描いているので、A stop sign flying in blue skies. というあり得ない文を入れると、止まれ標識が青い空を飛んでいるような



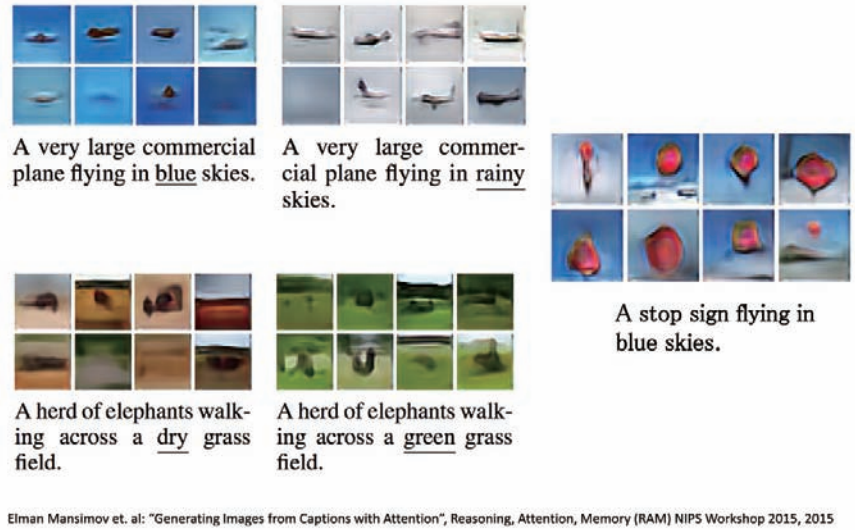
■図3.　人工知能技術の発展と社会への影響
（2014年9月での未来予測）

絵を描く。まさに、われわれがお話を聞きながらその情景を頭の中に思い浮かべたのと、とても近いことができるようになっている（図4、図5）。こうなると、文から画像を生成し、その画像からオートメイティッドイメージキャプチャニングで即日本語に直すということができるはずで、これは英語から日本語への翻訳になっている。画像を介して、意味が分かって訳しているということになる。実は今、この方式での自動翻訳の研究もしている。

一方、昨年の秋から方式が変わったGoogle翻訳は、画像を介した翻訳にはなっていない。だが、Googleが持っている大規模なデータと計算機と、ディープラーニングの最新の技術を全部入れて、従来よりも相当精度が上がっている。

このような技術に、画像やロボティクスの仕組みが少しでも入ってくると、翻訳としてはほとんど完成形だと思われる。あと3年から5年ぐらいで、人間の通訳と変わらない翻訳が実現される可能性がある。すると、日本経済、日本社会にとって非常に大きな変化が起こる。同時通訳ができ

■図4.　言葉の意味理解：Automated Image Captioning（2014ー）



■図5. 言葉の意味理解：Generating Images（2015.12ー）

るので言葉の壁がなくなる。日本の良さが海外に伝わるようになるのはプラスだが、壁に守られている産業は危ない。例えば、メディアと教育と金融である。BBC、CNNを苦労なく見られるとしたら、視聴者はNHKを選ぶのか。日本の大学を選んでくれるのか。金融商品はどうか。起こりうる色々な変化に備え、本来の付加価値は何かということをよく考えておかないといけない。

2.4 未来の画像を予測する

ディープラーニングでできることの中に未来を予測するという技術がある。人間は常に予測をしながら生きており、予測と違うことが起こるとびっくりする。同様に、ディープラーニングが0.1秒後の画像を予測するということをやっている。例えば、車が左に曲がり始めたら、このように曲がり終わるのだろうということが分かるわけである。静止画だけを与えて1秒間の動画を作るという技術もできている。例えば、ビーチの写真を1枚与えると、波がざばんと来た動画ができる。自分が行動すると、そのあとに何が起こるのかというのを予測する技術もできている。ロボットが、自分が手を動かすとその手に従って、いろいろなものが一緒に動き、何もないところで手を動かしても何も動かないということが上手に予測されている。

3. 眼の誕生

3.1 カンブリア爆発

これまで説明した技術を一言で言うと、眼の誕生だと思っている。

10年程前に書かれた『眼の誕生』（アンドリュー・パーカー著）という本がある。地球ができてから46億年の中の、5億4200万年前から5億3000万年前という非常に短い間に、現存する生物の全ての種（門）が出そろったという期間があり、これをカンブリア爆発と言う。短期間に生物の多様性が急激に増大したことについては諸説あるが、著者は光スイッチ説を唱える。それまでの生物は、臭いを頼りにのろのろと進み、ぶつかると食べるというような緩慢な動きしかできなかった。高度な眼を持った三葉虫という生物が現れると、遠くから見えるというのは、生存上大変有利な条件だったので大繁殖した。すると、今度は逃げるほうも眼を持ち始めて、早く逃げよう、隠れよう、擬態しよう、など様々な戦略が出てきた。つまり、生物が眼を持つことによって生物の生存戦略が多様化し、それによって生物の種が多様化したという説だ。

同じことが、今後、ロボット・機械の世界で起こるのではないか。機械が眼を持つことで、これまでできなかった相当多くの仕事ができるはずだと思う。昔からあるイメージセンサーは人間でいうと網膜にあたる。網膜で得た信号を脳の視覚野で処理をすることによって見える。視覚野の部分の処理がディープラーニングだ。つまり、イメージセンサーとディープラーニングを組み合わせて、初めて眼が見えるということになる。

産業化の歴史は、眼の見えない機械を使った自動化であった。機械は基本的に眼が見えないので、必ず入れるものを一定にする。同じ作業をしていても、仕事をしていることになるように環境側を工夫するわけだ。工場のライン

では、必ず同じものが同じ間隔で流れてくるという状況を作って、眼の見えない機械を動かす。信号機は一定間隔で青、赤、青を変える。眼のない機械は、社会の中に根深く入ってきていて、この機械に眼があった時に一体何が起こるのかというのは、ほとんど誰も想像したことがない。

3.2　既存産業の発展

自動化するのが非常に難しく、いまだに人手がかかり人手不足で困っている産業が、農業、建設、食品加工である。人手というのは、手ではなく眼の問題。機械には眼がないから人がやらないといけない。例えば、トマトはマーケットも非常に大きく収穫にかかる工数も非常に大きいのに、トマト収穫ロボットがいまだにないのは、トマトがどこになってるかが見えなかったからだ。稲やジャガイモなどは根こそぎ取ればいいので、認識能力は必要ないが、トマトは、茎は取らずにトマトだけを取らなければいけない。認識が必要なので、人がやるしかなかった。ディープラーニングの技術を使うとトマトがどこになっているのか見えるはずなので、トマトの収穫ロボットはできるはずである。

建設も同じで、現場にはいろいろな作業があり、全ての作業に眼が必要だ。鉄筋を組んだり、コンクリートを流し込んだり、固めたり。あるいは、内装のボードを取り付けたり、溶接したり。つまり人手不足で困るわけである。

調理というのは典型的に眼を必要としていて、肉を二つに切るのですら、眼がないと上手に切れない。それをフライパンに入れて色が変わったらひっくり返すなどは、絶対にできない。だから、調理は人がやっているのだ。

ところが、ディープラーニング、眼の技術を使うと、調理も自動化できる。農業、建設、食品加工、特に、外食産業のバックヤードは全部、いずれは自動化すると思っている。

4. 眼をもった機械・ロボット

4.1　単独の製品から入り、サービス化へ

全て自動化するということと、どこから先に自動化するかということは別の話である。

Amazonは、何を売っても良かったのだが、本から入った。本というのはスペックだけで勝負できる。同じものがあったときに優劣がなく、しかもロングテール性が一番高い。商材として一番先に成立するのが本だということを見抜いて始めたと思う。

農業、建設、食品加工、こういった中での本というのは何かを見抜くことが重要だ。農業では、おそらくトマト収穫が一番やりやすいはずだ。コストにも合いやすいのではないか。建設では、溶接作業が一番コストに合いやすいのではないか。食品加工でいうと、食洗器に皿を入れるという作業は、あり得ると思う。しかも、マーケットが大きく、スピードがそれ程必要とされないことを考えると、一番先にコストに合いやすい。そこから先の領域は、料理人の修行のようにだんだん難しい仕事をやっていけるのではないかと思っている。

4.2　プラットフォーム化から海外展開へ

外食産業は10年から20年ぐらいで、調理機械、調理ロボットが普及すると思う。最もコストの合いやすいところ、マーケットが大きく単純なものから入っていくだろう。その意味で牛丼の可能性が一番高いと思う。牛丼ができると、カレーやラーメンなど、いろいろな仕事ができるようになる。調理機械、調理ロボットが外食産業の後ろを担うようになると、日本の外食産業がグローバルに進出しやすくなる。

日本の食は世界で一番おいしいと思っているが、なかなか外に出ていけないのは、オペレーションが回らず味が落ちてしまうからであろう。後ろが機械化されると、そのままの形で持っていける。どんどん世界に出て、世界の外食産業の後ろを日本が作っているという状態にできる。すると、新しいアプリを配信するのと同じような形で、メニューの配信ビジネスができるはずだと思う。メキシコ料理も、ブラジル料理も、日本でメニューを開発して配信する。日本の消費者のレベルが高く、飲食店が鍛えられているし、日本人の最適化能力が非常に高いと思っているので、顔認識、表情認識で、誰が、どう思って食べているのかというフィードバックが分かったときに、メキシコ料理をメキシコ人よりも上手に作ることは十分あり得るのではないかと思う。そして、その人の好みや味付けが学習されていき、食の嗜好を捉えることができる。さらに、アレルギー、宗教、健康状態などに合わせた食を提供できるということである。

4.3　日本なりのプラットフォーム戦略

農業、建設、食品加工だけでなく、医療、介護、製造、廃炉など、できることはいろいろあるが、市場として一番大きいのは食で、日本が勝つ可能性があるのではないかと思う。ものづくり、特に機械・ロボットに対して、技術のアドバンテージはあるし、少子高齢化していて労働力が不

足しているので、ロボット・機械を使わざるを得ない状況にある。この点、諸外国、特に先進国では、ロボットが家庭の中や日常生活に使われることに対する抵抗感も非常に大きい。そう考えると、日本は有利なポジションにあると思う。

4.5　経営的な側面から見た考え方

ディープラーニングの技術は、実は追いかけるほうにとっては、簡単な技術である。進展が急速過ぎて、ここ1、2年の技術をきちんと押さえていれば、それだけでも十分通用する。画像認識で30年、40年やってきた技術者よりも、ディープラーニングを1年間やった新卒の人のほうが精度が出たりする。それが、年功序列の社内文化と合わないところがあるので、いかに経営的に乗り越えていくかというところが重要になる。そこで、学習工場を作ろうと言っている。眼を持った機械を作るには、機械の部分を作る工場も、眼の部分を作る工場も両方必要で、眼の部分を作る工場を学習工場と言っている。これを作るのに必要なものが、人とデータと計算機。この3つを集めることによって、設備投資だと思って企業が投資してくれれば、これは十分、GoogleとかFacebookと戦えるぐらいの投資規模になるはずだし、日本の強みを生かしていけば、競争力を出していけるのではないか。

ディープラーニングという教科書が2016年11月に出た。この中に、2016年の時点で、大体の目安として、教師ありの深層学習のアルゴリズムは、一般的にカテゴリーごとに約5000のラベル付き事例で許容できる性能を達成するとある。つまり、トマトの認識を10段階の品質で分類したいと思うと、5000×10＝5万枚のトマト画像と、それのアノテーションを用意すればいい。熟練の農家の方と同じか、それを超えるような精度を出したいとすると、1000万枚のラベル付き事例を含むデータセットで訓練すればいい。1000万枚を、1枚100円で作ると10億円。1枚10円で作ると1億円。このぐらいの投資で、目的に応じた認識はできる段階にある。既に、実現したときの市場規模を判断し、その投資に見合うかどうかの事業判断をしていく段階にきているのだと思っている。

5. 人工知能と倫理の問題

社会全体で考えるべきことに、人工知能と倫理の問題がある。自動運転が危機回避で、AさんをひくのかBさんをひくのか、というトロッコ問題。事故を起こした場合の責任の問題。人工知能を悪用する心を持ったように見える人工知能を作っていいのか。人工知能の軍事利用に対する考え方。人工知能が知財を生み出す場合。あるいは、人間の本来的に持っている権利がもっとあるのではないか、など。この辺りは世界的にも、国内でも議論が進んでいる。人工知能学会でも、2017年2月に倫理指針を出し、人工知能研究者が、多くの人の役に立つように、人間社会にとって有益なものとなるような活動をしていくということを、改めて確認している。

6. おわりに

人工知能が人間の職業を奪うのではないかという声もあるが、人と接する仕事というのは、今後更に重要になってくるだろう。また、目的の設定、価値判断、責任主体などは人間にしかできない仕事である。人間の創造性や、本能、感情に由来するセンサー機能、そして、当然、人工知能やロボットを使うような仕事も重要なものになってくる。日本が、眼を持った機械という大きな市場を取ることができると、こういう仕事が増えるということだと思う。

日本の抱えている社会課題の多くは、少子高齢化に起因して労働力が不足しているということだが、それは、眼を持った機械のニーズが非常に高いということでもある。農業分野、介護、廃炉、防災といった辺りで技術を伸ばしていくことによって、大きな輸出産業にしていくことができるのではないか。初期においては、いろいろなハードルや、ディープラーニングの技術を、上手に組み合わせ、すり合わせていくようなやり方が絶対に必要になる。ものづくり、素材、駆動系など日本の強みが発揮できる大きなチャンスなのだ。

ただ、日本は動きが遅いので、チャンスをきちんと捉えていく必要がある。そのために重要な3点を挙げる。1つが人材の育成。次に事業・産業がどう変わっていくのかを見抜き、動いていくということ。3つ目が社会全体で新しい未来像を描いていくことだ。

人が減って、多くの労働を機械・ロボットが担うというのは、新しい先進国の姿だと思っている。人間にとって良い形で新しい社会を実現していくことは、日本が果たすべき大きな役割だと思う。

※本記事は2017年5月17日開催の「第49回世界情報社会・電気通信日」のつどいでの講演をリライトしたものです。

（責任編集：日本ITU協会）