

JR-EAST Innovation 2017 基調講演

「人工知能は人間を超えるか」

東京大学大学院工学系研究科技術経営戦略学専攻 特任准教授 松尾 豊氏

松尾 豊氏 / Project Assoc. Prof. Yutaka Matsuo

1997年 東京大学工学部電子情報工学科卒業。2002年 同大学院博士課程修了。

博士(工学)。産業技術総合研究所研究員、スタンフォード大学客員研究員を経て、2007年より、東京大学大学院工学系研究科准教授、2014年より特任准教授。

専門分野は、人工知能、ウェブマイニング、深層学習。2010年から人工知能学会 副編集委員長、2012年から編集委員長・理事。2014年より倫理委員長。



基調講演ということでお時間いただきましたので、人工知能の話していきたいと思います。今、ご紹介いただきましたとおりで、私自身はずっと人工知能の研究をやっております。先日、日本ディープラーニング協会というのを立ち上げまして、このディープラーニングを使って産業界に貢献していこうとしております。

皆さんご存じだと思いますけれども、Google傘下のディープマインドが作った「アルファ碁」が、囲碁でプロ棋士イ・セドル9段を4勝1敗で破ったというのが昨年ありました。今年5月には、柯潔9段と対局して3勝0敗で勝ったということで、もう人間ではこの囲碁の世界に敵う人がいなくなり、アルファ碁は引退しました。

そのときの対局に備えて積み重ねた、アルファ碁同士の自己対局の棋譜が公開されました。この棋譜を見ますと、人が打つ棋譜とかなり違うということで、プロ棋士の方もびっくりしました。人間ですと長年常識とされていた打ち方があるわけですが、例えば端から3番目・3番目の所、ここは囲める面積が小さいので置いてはいけないと長らく言われてきたわけですが、このアルファ碁は3・3を多用します。一見すると素人のようですが、それでもすさまじく強いということです。

つまり、我々人類が何千年もかけて囲碁の世界を探索してきたわけですが、それでも実際に分かっていたのは囲碁全体の宇宙の中のごく一部にすぎなかったということが、このアルファ碁を通じて見えてきたのかなと思っています。

人工知能をめぐる動向

第1次AIブーム(1956~1960年代):探索・推論の時代

- ダートマスワークショップ(1956)
 - ・ 人工知能(Artificial Intelligence)という言葉が決まる
 - ・ 世界最初のコンピュータENIAC(1946)のわずか10年後
- 数学の定理証明、チェスを指す人工知能等

考えるのが早い人工知能

...冬の時代

第2次AIブーム(1980年代):知識の時代

- エキスパートシステム
- 医療診断、有機化合物の特定、...
- 第5世代コンピュータプロジェクト:通産省が570億円

ものしりな人工知能

...冬の時代

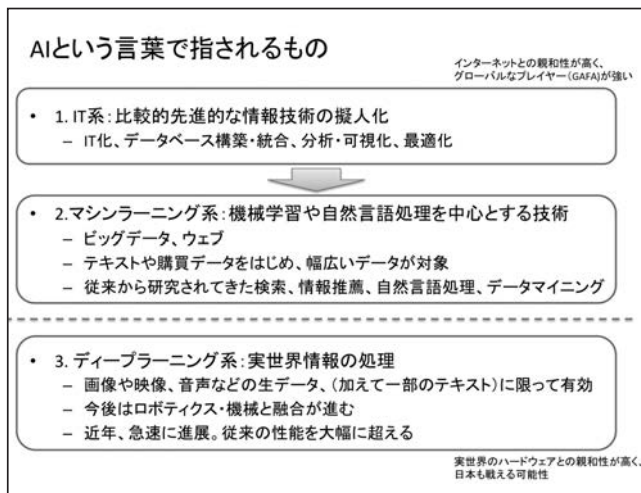
第3次AIブーム(2013年~):機械学習・ディープラーニングの時代

- ウェブとビッグデータの発展
- 計算機の能力の向上

データから学習する人工知能

今、人工知能がすごく世の中で話題になっております。連日のようにニュース等でも見るわけですが、歴史的に見ますとこれは3回目のブームでして、人工知能という分野は1956年から始まっておりますので、ちょうど60年がたったところです。この60年の間に、ブームになっては冬の時代が来て、というのを繰り返しています。この分野、本当にブームになりやすい、過剰な期待が抱かれやすい分野だと思っています。第1次AIブームの頃からいつも、人工知能があんなことこんなことできるのではないか、人間を超えてしまうのではないか、人間を襲うのではないかなど、こういったことが昔から言われています。今

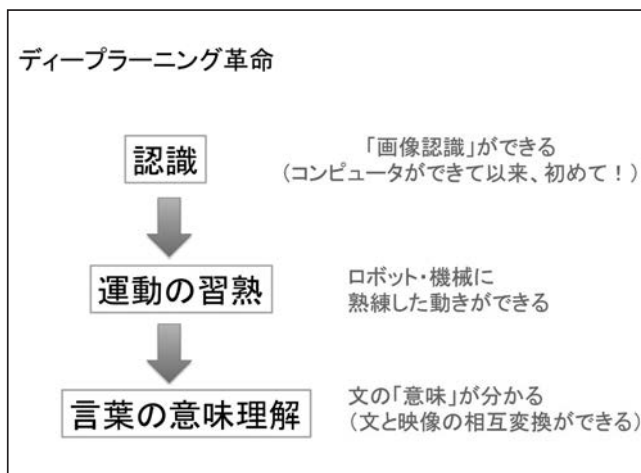
回もそれと似たような傾向があります。やはり何ができて何ができないのか、というのをきちんと見抜く必要があると思っています。いろいろなキーワード出てきますけれども、基本的には60年前からやってきた技術が少しずつ良くなって今に至るということです。ブームになったからといって急に何かできるようになったりということは基本的にはないと思います。ただ、このディープラーニングだけはちょっと別格だと思っており、何十年もできなかったことが次々にできるようになっているということです。



今、人工知能という言葉で指されるものというのは、大雑把に分類すると図のようであると思っています、一つは従来ITと呼ばれていたものをAIと呼び変えるということが非常に多いです。これはなぜかという、そもそも人工知能という言葉の定義が定まっていないため、一番広くとるとサーモスタットですら人工知能と言えてしまいます。環境の温度に合わせてスイッチを切り入れたりというのを賢く振る舞うわけですので。サーモスタットですら人工知能ということは、ありとあらゆるITが言おうと思えば人工知能と言えてしまうわけです。ブームになるとそう言った方が受けがいいですから、そういう言い方をする場合が多いです。

二つ目に、インターネットの世界ですごく進んできた自然言語処理だとか機械学習、こういった技術を使うことを人工知能という場合もあります。これもインターネットで進んできた技術を他のインダストリーに幅広く展開しようというようなものです。1から2というのは比較的連続的な流れになっています。

一方で、3は画像とか映像、生データを中心に、その解析技術が非常に良くなったということで、1から2の流れとは若干毛色が違うと思っています。例えば、2でやるのは顧客の購買データだとか移動履歴だとかログデータだとか、こういったものを扱う場合が多いのですが、こういったデータというのは実はディープラーニングに向いていません。ディープラーニングで向いているのは、生データ、画像、映像、こういったものであり、しかもこのディープラーニングの技術は物づくりの技術と非常に相性がいい。イメージセンサーだったり機械、ロボティクスと非常に相性がいい。それから、現場を持っている企業にとって非常に活用しがいがあるということであり、3はかなり日本向きの技術で、これをやるのがいいのではないかとずっと言っているわけです。



ディープラーニングで一体何ができるかをお話したいと思います。ざっと言うと、「認識」「運動の習熟」「言葉の意味理解」このようなことだと思っています。画像認識というのは、もともとコンピューターにとって非常に難しいタスクでした。それがここ数年、急激な勢いでできるようになってきていて、2012年にディープラーニングに非常に大きな性能の飛躍がありました。画像認識で、従来画像を見てこれが何かを当てるというタスクで、エラー率、間違い率が28パーセントか29パーセントぐらいであったものが、急激に10パーセント良くなって16パーセントというところまで来たということです。

これが2012年にあったのですけれども、それ以降すごい勢いで画像認識のエラー率が低下してきて、2013年には11.7パーセント、2014年には6.7パーセントと、みるみるこのエラー率が低下しています。人間がやると同じタスクでも5.1パーセント間違えるというのに対して、2015年の2月にはマイクロソフトが4.9パーセント、Googleが4.8パーセントということで、人間を超えるところまで来ています。これが2015年の話ですから、今からわずか2年前に、コンピューターが画像認識で人間の精度を超えたということがもう現実になってしまったわけです。

認識: 2012年以降のエラー率の変化

	Error
Before ディープ ラーニング	
Imagenet 2011 winner (not CNN)	25.7%
Imagenet 2012 winner	16.4% (Krizhevsky et al.)
Imagenet 2013 winner	11.7% (Zeiler/Clarifai)
Imagenet 2014 winner	6.7% (GoogLeNet)
After ディープ ラーニング	
Baidu Arxiv paper: 2015/1/3	6.0%
Human: Andrej Karpathy	5.1%
Microsoft Research Arxiv paper: 2015/2/6	4.9%
Google Arxiv paper: 2015/3/2	4.8%
Microsoft Research CVPR paper: 2015/12/10	3.6%
Latest	3.1%

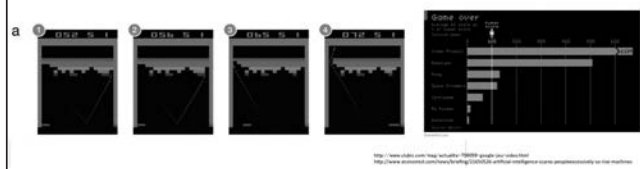
2015年2月には人間の精度を超えた
画像認識で人間の精度を超えることは
数十年間、実現されていなかった

これが意味するところは相当大きいと思っています。人間の仕事の中で、人間が眼で見て認識・判断している仕事がたくさんあります。これが全て自動化・機械化できる可能性が出てきたという意味では、この画像認識の精度が人間を超えたという、この一点だけでも相当大きなインパクトがあると思っています。「YOLO v2」というモデルですが、ここ1年ぐらいで出てきた改良版で、オブジェクトデテクション(物体検知)をするものです。例えば、人だったらパーソンと出ますし、それからチェアだとかセルフォンだとか、こういったものがうまく検知できています。これがすごいのは、この技術がOpenCVに入っているということです。OpenCVというのは画像処理を行うようなプログラム

ですけれども、普通のプログラマーが画像処理をやりたい時に必ず使うもので、極々スタンダードなものです。その中にこういう物体検知をするような、すごい精度の高いモデルがもう含まれてしまっている。このぐらいディープラーニングの技術というのは、急速にコモディティ化しているということです。もう、ありとあらゆる所でこのディープラーニングの技術が使えるようになってきています。

運動の習熟: ディープラーニング+強化学習(2013-)

- 強化学習とは、行動を学習する仕組み。
 - 「報酬」が得られると、事前の行動を強化する。
 - 「状態」「行動」→「望ましき(報酬ありなし)」
 - 古くからある技術だが、これまでは、「状態」を人間が定義してきた。
- 運動の習熟が可能に
 - 状態の認識に、ディープラーニングを使う。
 - DeepMindの研究者(D. Hassabisら)。その後、Googleが買収。
- 試行錯誤することによって、運動が習熟する
 - 最初は下手。繰り返すうちに、うまくなっていく。
 - 最終的には、ブロック崩しでの道路を作ったり、インベーダーゲームでの「名古屋撃ち」も。
 - 「全く同じプログラム」で、異なるゲームを学習。半数のゲームで人間のハイスコアを上回る



これまでが「認識」の話でしたけれども、次に起こっているのがこの認識の技術と機械系ロボット系の技術を組み合わせるというもので、深層強化学習と言われますけれども「強化学習」という技術を組み合わせます。強化学習というのは、やっているうちにだんだん動作が上達してくるというものです。人間の場合も、例えばゴルフボールを何度も打っていると、だんだん上手に打てるようになります。なぜ打てるようになるのかというと、たまたまうまくいったらそのやり方を繰り返すということをやからです。これは、報酬が与えられるとその前にやった行動を強化するというので、強化学習という言い方がされます。従来もあったのですが、従来のやり方では、こういう状況でこういう行動を

すると良い悪いというような、状況を記述する変数を人間が定義していました。ところがディープラーニングと組み合わせる方法は、そこをディープラーニングで置き換えて、変数、特徴量を取り出すというところ自体をディープラーニングにやらせるわけです。

いくつか動画をお見せします。アルファ碁を作ったディープマインドが2013年ぐらいにやっていた研究で、非常に有名な動画ですけど、ブロック崩しを学習させるAIというのを作っていました。最初は下手ですけど、やっているうちにだんだん上達してきます。スコアを報酬にしまして、スコアが上がるとその前にやった行動を強化するという仕組みによって、どんどん上手になってきます。

しばらくするとすごくなるのですが、このぐらいですと従来の第2次AIブームの頃のAIでもできました。従来はどうやっていたかというと、ボールのxy座標とか自分の動かしているパドルのx座標とか、こういうのをアルゴリズムに当てていました。ところが、この動画のやり方がすごいのは画像を入れているだけだからです。画像を入れるとそこに何があるのかというのを検知して、その関係がどうである時に点が入りやすいというのを学んでいくということをやります。そうすると、左端を狙うというのをやり始めます。このゲーム、左端・右端に通路を作ってボールを上には放り込むとすごく点入るのですけれども、それを見つけて出しているということです。

今と全く同じプログラムで全然違うゲームを学習させることができて、例えばインベーダーゲームなどもブロック崩しと全く同じプログラムで上達させることができます。これができたのが2013年です。直後の2014年の初頭にGoogleがディープマインドを日本円で約500億円で買収しました。

運動の習熟: ディープラーニング+強化学習が実世界へ(2015-)

- ・ 実世界への適用
 - 2015年5月 試行錯誤で部品の取付を習熟するロボットの開発 (UC Berkeley)
 - 2015年5月 試行錯誤で運転を習熟するミニカーの開発 (PFN社, 日本)
 - 2016年3月 試行錯誤でピッキングが上達するロボットの開発 (Google)
 - その他、メリーランド大、EUのプロジェクト等も進展
- ・ 考えてみれば当たり前
 - 犬や猫でもできる。高次元言語能力は必要ない。認識が問題だった。
 - 歴史的には、多数の人工知能研究者がこのことを主張してきた。



いろいろな動作ができるようになります。

今と全く同じような技術を使ってロボットアームに適用したものが、昨年Googleが出したものです。ロボットアームがいろいろな物を持ち上げては落とし、持ち上げては落としということで、つかみ方の練習をしています。右上にカメラがありまして、カメラで画像を撮りながら見えています。持っては下ろし、持っては下ろし、うまく持ち方を学習しています。速く学習させるために、たくさんロボットアームを並べて経験を共有させるということをやります。この場合ですと、14台を並列に並べて14倍高速に学習させるということをやっています。そうすると、いろいろな物を上手に持てるようになります。

何がすごいのかと思われるかもしれませんが、これは相当すごいことでして、従来のロボットや機械にはこのような簡単なことすらできませんでした。このようにたくさんの物が入った箱の中から、目的の物を取り出すというのは相当難しいことでして、従来ですと画像認識の精度がそもそも悪かったので、持ち上げたい物がどこにあるのか見えませんでした。見えたとしても物によって持つべき場所は違っており、長い物を長い方向で持とうとしても持てませんので、長い物は短い方向で持たないといけません。柔らかい物、丸い物、それぞれ持つべき場所が違うわけです。

従来ですと、人間がこういう物はここを持ってと完全に定義すると思っていたのですけれども、ごちゃ混ぜにたくさん入っている物は無理でした。ところがこのやり方は、ディープラーニングで特徴量をつかんで、こういう特徴のときはこういう持ち方をすると良い悪いというのを試行錯誤、強化学習で学んでいきますから、いろいろな物を上手に持つことができるということです。考えてみれば人間もいろいろな物を上手に持てますけれども、なぜ持てるのかというと、赤ちゃんのとき、子どものときに練習しているからで、それと同じようなプロセスを経てこういったロボットも上手に物が持てるということです。

ディープラーニングの人工知能における意味

- ・ モラベックのパラドックス: 「子供のできることほど難しい。」
 - 高度な推論よりも、認識や運動スキルの方が難しい。
 - 比較的易しい: 医療の診断、チェスを打つ、数学の定理の証明
 - 極端に難しい: 画像認識、積み木を上手に積む
 - それがここ3年くらいのあいだに一気にできるようになった
- ・ 現在のコンピュータのパワーでようやく可能に
 - GPUを数十台並列に並べて、数日~数ヶ月計算させてようやく精度が上がる
- ・ アイディアは昔からあった。もともとは日本発
 - 1980年当時、NHK放送技術研究所にいた福島邦彦先生によるネオコグニトロン
 - その後も多くの研究者が試みている
- ・ 初期仮説への回帰
 - 初期仮説「なぜ知能をコンピュータで実現することはできないのか？」
 - できると思っていた→できない理由があった→それが解消された→だとしたら、もう一度できるという仮説を取るべきでは。
 - 産業として非常に大きい可能性を秘めている。

この技術をロボットに適用するとどうなるか、ロボットが上手に動けるようになるのではというので、UC Berkeleyが2015年に、ロボットがおもちゃの飛行機の部品を本体に組み付けるということをやっています。お話したような深層強化学習、ディープラーニングと強化学習を組み合わせたような技術で、行っているうちにだんだん上達していきます。ロボットの目の所にカメラがあり、目で見て上手に動作しているということになります。精密に動くというよりは、見ながら上手に合わせているというような動き方をしています。従来ですと、位置と角度、両方合わないといけなないので難しかったのですが、それもこのロボットは画像を眼で見ながらうまく合わせているということです。このようにいろ

人工知能の世界では、モラベックのパラドックスというのが知られていまして、子どものできることほど難しいと言われていました。これは、大人や専門家でないときになさそうに思える一見して難しい「医療の診断」、「数学の定理の証明」、「チェスを打つ」といったことは、実は人工知能にとってはそれほど難しくなく、人工知能研究の初期の1960年代70年代からある程度はできるようになっています。ところが3歳児でもできるような簡単そうに見えること、画像認識であるとか積み木を上手に積むといったことが、一向にできるようにならないということで、逆説的なのでパラドックスと言われてきたわけです。これが1980年代からずっと言われてきたわけですが、それが今覆りつつあります。

3歳児でもできるような画像認識であるとか、あるいは積み木を上手に積むということが、ようやくコンピューターにできるようになってきたということです。

第3次AIブームの技術的なエッセンスはもうここにしかないと思っており、人間にとって難しそうなことがある程度できるのは第1次、第2次AIブームも同じです。3歳児でもできるような簡単そうに見えること、このほうがコンピューターにとってはよっぽど難しく、それが今ようやくできるようになってきたということです。それによって世の中がどう変わるのか、産業社会がどう変わるのかということを考えていかないといけないということです。

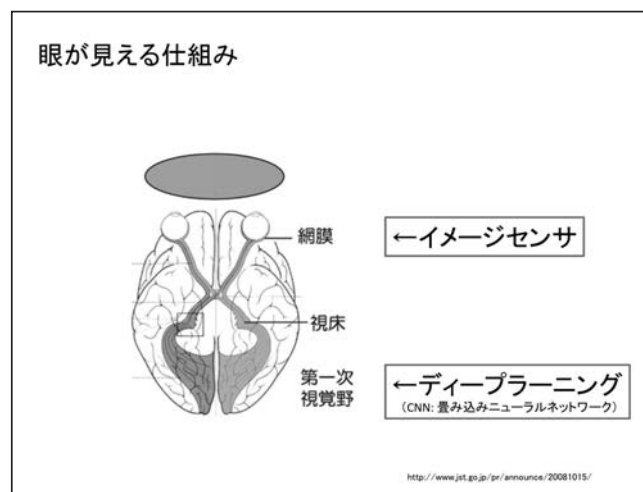


ディープラーニングの技術にはいろいろあるのですが、全部まとめて一言で言うと眼の誕生ということだと思っています。『眼の誕生』という本があります。これはアンドリュー・パーカーが10年以上前にカンブリア爆発について書いた本です。地球ができて46億年たちますが、46億の中の5億4200万年前から1000万年という非常に短い期間の間に、現存する全ての生物の種、「門」が出そろったという期間がありまして、これをカンブリア爆発といいます。

なぜカンブリア爆発が起こったかというのは諸説あり、いまだに決着がついていないのですが、このアンドリュー・パーカーが言った説が光スイッチ説といいまして、眼ができたからだ

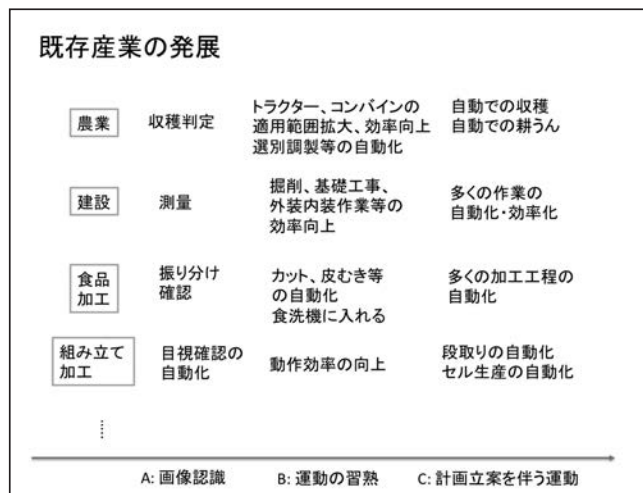
と。それまでの生物は眼を持っていなかったで、匂いを頼りにのろのろと近づいていってぶつかると食べるとか、ぶつかられると逃げると非常に緩慢な動きしかできませんでした。ところが高度な眼を持った三葉虫が現れ、遠くから見えますから、生存上圧倒的に有利だったわけです。大繁殖しました。そのうち逃げるほうも眼を持ち始めて、見つかったと思い速く逃げようということで速く泳ぐ能力が重要になったり、見つかったと思って隠れようとか、見つからないように擬態しようとか、いろいろな作戦が出てきました。生物が眼を持つことによって生物の生存戦略が多様化し、それによって種が多様化したというのがこの光スイッチ説なわけですが、これも、これと全く同じことがこれから機械・ロボットの世界で起こると思っています。

つまり、今までの機械とかロボットは眼がなかった、眼が見えてなかったわけですね。それが見えるようになる。それによってできるタスクの幅が一気に広がります。だから、従来ですとほとんど考えられないぐらいの規模、いろいろな多様性を持った機械とかロボットが今後社会で使われていくはずだと思っています。



眼といっても、カメラ、イメージセンサーは昔からあると言われるかもしれませんが。イメージセンサーは、人間でいうと網膜に当たります。人間も網膜だけで見ているわけではなくて、網膜で映った信号を脳の後ろの方にある視覚野という部分で処理をすることによって見えます。この視覚野の処理が、ディープラーニングに当たるということです。ですから、イメージセンサーとディープラーニングを組み合わせ、初めて機械に眼が見えるようになるということです。

それでは、機械に眼が見えるようになると何が起るのか。環境を整えることのできない所での自動化、機械化というのは、今まで非常に難しかったわけです。それは眼がないので、環境に合わせて動作するというのが不可能だったからです。その典型例が農業とか建設とか食品加工だと思っていまして、このような分野ではほとんど自動化、機械化というのができていません。



例えば農業でいうと、トマトはすごくマーケットも大きいですし、それから収穫に工数もかかっているのですけれども、いまだにトマト収穫ロボットはありません。なぜかという、トマトがどこに成っているのか見えなかったからです。農業機械にはいろいろな機械があるように思いますが、収穫用だとコンバインとかトラクターでジャガイモを収穫するとか、結局こういうものは全部根こそぎ取るものです。根こそぎ取るにおいては認識能力は必要ないので、眼のない機械でも自動化できたのですけれども、トマトのような園芸作物ですとか、ミカン、リンゴ、桃のような果樹は根こそぎ取るわけにいけないので、そうすると人が眼で見ながら取る必要がありました。この部分が代替できないので

機械化できずに人手がかかっていました。農業って考えてみれば人手がかかってしょうがないわけですから、これはほぼ全ての作業が眼を必要としています。だから、そこがどうしても機械化、自動化できませんでした。だから人がやるしかなくて人手がかかって困っています。

それと同じことが建設にも言えまして、建設現場行くと人がたくさんいて、鉄筋を組んだり溶接したりコンクリートを入れたり固めたり内装のボードを取り付けたりということをやっています。なぜ人がやる必要があるかというと、建設において全ての行程にやはり眼が必要でした。したがって自動化、機械化できませんでした。だから人が必要で、人手不足で困っていたということになります。

食品加工とか外食、これも同じでして、我々仕事が終わってどこか食べに行こうとレストラン行った際、後ろで人が調理しているのは当たり前だと思っていますけれども、これも当たり前ではなくて、調理というのは人がやる以外のソリューションが今のところありません。調理は完全に眼を必要としますから、人がやるしかありません。だから食品のバリューチェーンというのは、基本的には取れたものを何とか飲食店か家庭に運んで、あとはそこでやってくださいねという構造をするしかないということになっています。この食のバリューチェーンというのも大きく変わると思いますし、外食産業の後ろは10年か20年か分かりませんが、いずれは全て自動化すると思っています。

もちろん全て自動化するというのと、どこから先に自動化するかというのは全然別の話だと思っていまして、一番コストに合うところ、技術的な難易度が低いところから入ってくるはずで、そこがどこなのかというのを見抜く必要があります。例えば、調理は非常に複合的な作業なので、調理がいきなりできるとは全く思わないのですけれども、例えば食洗機にお皿を入れるというような作業であれば、今でも頑張ればできるはずで、ディープラーニングの技術とハンドの技術があれば頑張ればできるはずで、そこが業務用で入ってくればその周辺領域がどんどん広がってくると思っています。

このような形で機械をベースにしてそこに眼を付けていく、それでプラットフォーム化していくということは、日本なりの闘い方ではないかと思っています。機械系・ロボット系というのは日本は非常に強いですから、ここに対して眼を付けて作業を自動化していく、そこでデータを集めてプラットフォーム化していくことだと思っています。

機械・ロボットのカンブリア爆発

- ・ 介護施設や病院等での見守り・介護ロボット
- ・ 医療（X線、CT、皮膚、心電図、手術ロボット）
- ・ 警備、防犯技術
- ・ 顔による認証・ログイン・広告技術、表情読み取り技術（サービス業全般に重要）
- ・ 国家の安全保障、入国管理、警察業務、輸出入管理業務における活用
- ・ 防災系（河川、火山、土砂崩れを見張る）
- ・ 重機系（掘削、揚重）、建設現場系（セメント固め、溶接、運搬、取り付け）
- ・ 農業系（収穫、選果、防除、摘花・摘果）
- ・ 自動操縦系（ドローン、小型運搬車、農機、建機）
- ・ 自動運転系、物流
- ・ 産業用ロボット系（特に組み立て加工等）
- ・ 調理系（牛丼、炊飯、ファミリーレストラン、外食全般）
- ・ ペットロボット系
- ・ 片付けロボット（家庭、オフィス、商業施設）
- ・ 新薬発見や新素材の開発（遺伝子の認識・分析、実験ロボット）
- ・ 廃炉系（深海や鉱山、宇宙も含めた極限環境）

農業・建設・食品加工だけでなく、医療や介護、製造、廃炉なども。

頑張ればできるような段階にきていると思っています。

片付けロボットが家の中に入ると何が起きるかという、朝会社に行って夜家に戻ってくると、全ての物が元の場所に戻っているということが起きます。私はこれを家のホテル化と言っていて、相当居心地良くなるはずですし、このようなものが1回入れば散らかしたい放題、汚したい放題になりますから、人は相当堕落します。1回堕落してしまうと元に戻れませんから、相当高くても売れるのではないのでしょうか。

家の中でもそうですし、それからオフィスでも片付けという業務はどこにでもあります。商業施設でもあります。この会場も終わったら誰かが片付けるはずで、世界中の人、世界の75億人が1日あたり片付けに使っている時間は一体何十億時間なのかと考えるとこれは相当大きいはず。運転している時間よりもだいぶ大きいのではないかと考えると、片付け一つ取ってですら、相当大きな産業がこれから立ち上がるのではないかと考えています。いろいろな点検をしたり安全確認をしたり、もちろん自動操縦したりいろいろな意味で鉄道にも関わってくるようなところはあると思います。

当然海外のベンチャーはどんどん実施しており、一番早いのが医療画像です。医療画像は画像認識だけで、ほとんどハードウェアなど関係なく行えますので、先ほどのディープマインドもそうですし、いろいろな所が実施しています。例えば眼底の疾患の診断、加齢黄斑変性とか糖尿病網膜症、あるいはレントゲン、CT、MRI、それから皮膚病、内視鏡、病理診断、こういった画像だけで判断できるものは、軒並み人間のお医者さんを超えるような精度が達成されています。医療画像に関しては、かなり戦いが中盤戦から終盤戦に差し掛かっているということで、この2年の間に相当な勢いで決着がつき始めているというような感じです。これが機械系とかロボット系に関わる場所では、まだまだこれからというところもありますし、ここで日本企業が何とか勝負していかないとならないのではと思っています。

Deep30

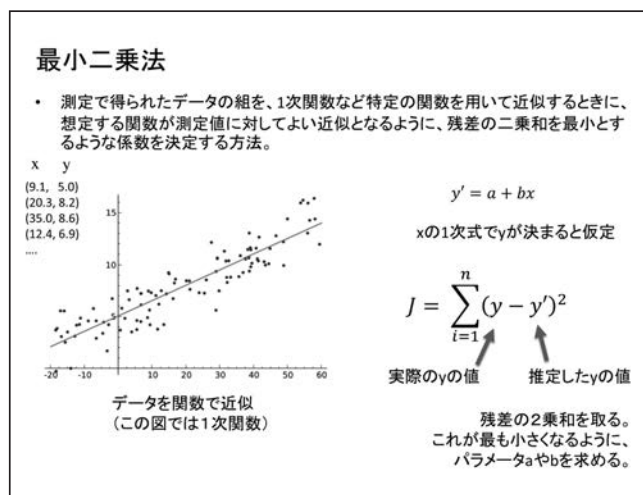
- ・ Deep 牛：牛丼が世界を制す。ゼロオベの牛丼チェーンを運営する会社
- ・ Deep 片付け：片付け業務の自動化。片付けロボットの設計・開発・生産、リース・運用保守ならびに片付け業務のサービス提供業。
- ・ Deep Toilet：トイレをいつでもきれいに。自動のトイレ掃除ロボットの開発／トイレ管理サービス業
- ・ Deep ドア：自動ドアに眼を。DL式の自動ドアの設計・開発・生産。ならびに空調管理のコンサル業務。
- ・ Deep 建設：建設業務におけるDL活用。重機、建機、溶接、コンクリート等の工程の自動化
- ・ Deep 化工：化学プラントにおけるDL活用。
- ・ Deep 顔認証：オフィスの入退室管理、会員へのおもてなし、本人認証。
- ・ Deep Beauty：美をDLで科学する会社。化粧品、美容院等へ展開。
- ・ Deep アグリ：農業の収穫、間引き、防除等を行う。
- ・ Deep 魚市場／畜産市場：魚の仲買人の目利きをDL化し、グローバルな市場での価格設定、取引を行う会社

今、農業、建設、食品加工についてお話ししましたが、それ以外にもいろいろな領域でこのような眼の技術、眼を持った機械のカンブリア爆発ということは起こると思っています。例えば医療とか介護、製造などたくさんあります。例えば、家の中に片付けロボットはいまだにありません。お掃除ロボットルンバがありますけれども、ルンバは認識能力がないので、動かす前に片付けしないとけなく面倒です。なぜ片付けができないかというと、片付けは散らかり方が毎回違いますから、認識能力がない限り片付けは絶対できませんでした。ところが今、ディープラーニングで眼の技術ができましたから、あとは物を見て元の場所に戻すということをやればよいということで、技術的には今

いろいろな可能性があるので、研究室からどんどんベンチャーが出たらいいと、勝手に30社ぐらい作りたいと思っています。一番やりたいのはDeep牛というもので、牛丼を作るベンチャーです。調理というのは絶対に大きくなります。私が30年後40年後に孫に向かって、若い頃は調理は人がやっていたんだよ、おじいちゃん本当なの、大変じゃないみたいな会話をしているのが目に浮かびます。やはり今、洗濯などを手でしないですね。掃除も掃除機があるのと同じように、昔はやはり洗濯にすごい時間かかっていたわけです。

それでは、なぜ調理ができないのかというと、やはり眼が必要だったからです。これほど重要なものはないし、これほど人がお金を使っていることもないですよ。1日かなり食費に使っている、しかもおいしいものを食べたいという欲は、かなり際限なくあるということを見ると、この調理というのは自動化されるはずだし、自動化されたら次はレシピの配信というビジネスが起こるし、さらに顧客の顔認証、あるいは表情認識もできますから、どういう人が何を食べておいしそう顔をしたのかしなかったのかというデータが取れます。これは、巨大なグローバルプラットフォームになると思っています。

さらには、日本は食がおいしいということを考えると、このあたりも非常に大きな産業になると思っています、それが実現されれば、単品でマーケットが大きくて、かつ眼がないとギリギリ自動化できてない食品となると牛丼なのです。そこでDeep牛と言っているのですが、それ以外にももちろんたくさんありまして、先ほどの建設業務に関わる所、化学プラントに関わる所等もありますし、それから交通違反等を見張るとか採掘とか労務管理をするとか、例えば安全確認をする、こういった業務なども自動化できるはずですし、設備の点検等の業務も当然できていくと思います。衛星画像でいろいろ見るとか、搬入だとかこういった物流関係も、たくさん可能性があると思います。このようなあたりを、全く新しい産業がこれからどんどん立ち上がり、例えて言うとエンジンができたとか、あるいはトランジスタができたとか、そういったものに匹敵するような大きな変化がこれから起きてくるのではないかと思います。



次に、ディープラーニングとは一体何をやっているのかについて、お話ししていきたいと思います。魔法のように聞こえますけれども、実は全然そのようなことはなくて、やっていることは比較的シンプルです。最小二乗法というところから話が始まりますので、これを知っていただくと大変ありがたいのですが、統計の割と最初のほうに出てきます。

例えば、気温と飲料の売り上げというデータ(何月何日に温度が何度で、飲料がどのくらい売れた)があったとします。データを散布図としてプロットすると、図のような感じになります。温度が上がると飲料が売れますというようなものです。これにエクセルなどにある近似直線を追加すると、図のような直線が引けま

す。この直線をどのように求めているのかというのが最小二乗法です。気温をx、飲料の売り上げをyと置くと、このyをxで近似する1次式を立てます。つまり、これはこのyの予測式になっているわけですが、この予測した値と実際の値がどのくらい差があるのかというのを引き算し、これを二乗してサンプル点ごとに和を取っていく、二乗和を取るというわけです。この二乗和が最小になるように、一番小さくなるようにaとかbのパラメータを決めるのが最小二乗法です。そうするとこのプロットにうまく乗るような点が出てきます。

今、気温という1変数にしましたが、気温と湿度の2変数にします。2変数にすると何が起きるか、3次元になり平面になります。でもやることは一緒で、このyをx1、x2の2変数で近似し、それを実際の値とどのくらい乖離しているのかというのを引き算して二乗和を取ります。二乗和が小さくなるようにaとかbとかcを決める、ということをやればよいわけです。

猫関数

$$f_{\text{猫}}(x) = k_0 + k_1 x_1 + k_2 x_2 + \dots + k_{10000} x_{10000}$$

- 100×100の1万個の変数から、猫か(y=1)、猫でないか(y=0)を出力する関数を求める問題
 - 基本的には、最小二乗法で求められる
 - 各画像が猫か猫でないかの教師データが与えられれば、データの当てはまりが最もよくなるように、 $k_0, k_1, \dots, k_{10000}$ のパラメータを求めればよい。
- 細かい補足
 - 回帰ではなく分類
 - 実際は、ロス関数として、最小二乗和ではなく、クロスエントロピーや対数尤度
 - 実際は、線形ではなく、非線形の階層をもった関数で近似
- 求め方は、繰り返し反復による勾配法
 - 1. 適当に $k_0, k_1, \dots, k_{10000}$ の初期値を決める。
 - 2. そのkにおいてロス関数を微分し、勾配を求める。
 - 3. その方向に少しだけkを動かす。
 - 4. 収束すれば終了。でなければ、1に戻る。

そこで、今度は1万変数にします。やることは一緒で、このyをx1からx1万までの式で近似し、これが実際の値とどのぐらい乖離しているかを二乗和を取って、この二乗和を最小にするようにk0からk1万のパラメータを決めましょうということをやればいわけです。

それでは1万変数の問題は世の中にあるのかというと、1万変数の問題はありまして、例えば100×100の画像に猫が写っているかどうかを判定する問題は、このような関数を見つける問題です。100×100というのは、要するに1万個の画像の画素の値が入ってきます。そのときに猫だったら1、猫でなかったら0という、このような関数を見つけようということなので、基本的には

最小二乗法でできます。

このように猫関数があったとすると、猫関数はx1からx1万までの式で表されて、k0からk1万までのパラメータをどのように決めるかという最小二乗法でやればいい。つまり、この画像が猫だ、この画像は猫でないというようなデータがたくさんあれば、それに一番うまく乗るような線を引けばいいのです。したがって、ディープラーニングはデータがたくさん必要だと言えます。なぜなら、このようにパラメータが非常に多い式を近似しよう、求めようとしているのですから、データがたくさん必要だということと言えます。

ディープラーニングの場合は、中で最適化を行います。どうやって最適化するかというと、最小二乗法をやりますから、この誤差関数、二乗和というのを最小にするようにパラメータを動かします。そのときに、偏微分というのを取ります。二乗和に対しての偏微分を取って、パラメータをどちらの方向に動かせばこの二乗和が減るのかを計算しながら最適化していくことをやりますので、中が数学、最適化の問題になっているということです。

ところがこの猫関数、今は線形な式で書いていますけれども、このような線形な式ではなくてもっと難しい式になっています。ディープラーニングは、深いわけですね。なぜ深いのかというと、深いほうがいろいろなものを上手に表せるからです。どのように行おうかというと、猫関数をいきなり表していたものを、いったんf1とかf2とかf100のような中間的な関数を用意します。中間的な関数を使って、この最終的な猫関数というのを表します。つまり一段噛ませるということをやります。これは、要するに2層のニューラルネットワークを作っていることに相当します。ところが、線形線形で行いますと結局いくら合わせても線形になりますから、非線形な要素を入れたいということで、シグモイド関数というものを使います。このシグモイド関数というのがニューラルネットワークの語源になったところですけども、ニューロンの働きを模擬しています。ニューロンは、刺激がある一定の値を超えると発火し、そうでないと発火しないということで、それを模擬するように値が正のときは1で、正でないとき負のときは0というような、これをうまくつなぐような、このような関数を使ってこのシグモイド関数を作ります。このシグモイド関数を入れていくことによって、非線形な関数で表されることになります。

ただし、このパラメータをどのように決めているかというと、これは最小二乗法でデータがたくさんあったときに、この二乗和というのを最小にするように求めていけばいいということになります。

なぜ深いことが重要か

- 1. 世界の階層性
 - エッジ-ヒゲ-猫の顔
 - 世界の構造がそもそも階層的になっているので、学習するモデルも、素性がフラットに並んだ線形のモデルのようなものよりも、階層性をもったモデルのほうが良い。
- 2. 多様体
 - 線形なモデルで切り取れる領域は、直線で囲める範囲である。
 - 非線形なモデルで階層を重ねれば重ねるほど、自由に領域が指定できる
 - 通常、我々が扱いたい概念は、入力の高次元空間よりも低次元の多様体を形成する
- 3. 効率
 - 1層の隠れ層を持つニューラルネットワークを用いれば、任意の連続関数を任意の精度で近似できる。(普遍性定理)
 - ところが、1層だと、場合によっては無限に多いニューロンの数が必要になる。
 - 階層を増やすことで、指数的に効率的に記述することができる。
- なので深くない！

猫を表す画像というのは非常に非線形な領域をしています。この非線形な領域をうまく囲むには、非線形性の強い表現力の高い関数を使わないといけないということで、階層性を持たせると非常に表現力が上がります。これらによって階層的なほうが良いとされています。

なぜうまくいくようになったか： 活性化関数の進化

Sigmoid

$$\sigma(x) = 1 / (1 + e^{-x})$$

tanh

$$\tanh(x)$$

ReLU

$$\max(0, x)$$

Leaky ReLU

$$\max(0.1x, x)$$

Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha (\exp(x) - 1) & \text{if } x \leq 0 \end{cases}$$

ReLUが使われるようになったのが大きい。
2000年に提案されたが、2012年ごろから使われるようになった。

それでは深いとなぜよいかというと、これは答えるのがそんなに簡単ではありません。いくつか答え方がありますが、一つ目はそもそも世界が階層的だからということです。例えば、今お話ししていますけれども、ここはホテルの中にあって、ホテルの中の3階にあって、3階の中の部屋にいて、この部屋の中のステージにいる、というようになぜか階層構造をしています。猫の顔というのも、猫の顔は目とか鼻とか口とかヒゲから構成されていて、ヒゲはまた3本の線から構成されていて、というように世界はなぜか階層的なので、階層的な関数を使ったほうが良いという理由です。

二つ目に、そもそも全ての画像、ありうる画像の空間の中で

このような深い階層を持ったニューラルネットワークを学習させたいというのは、長年研究者が思ってきたことですが、これがうまく学習できませんでした。それが今はできるようになってきました。それは、データが増えてきたことと、それから計算機のパワーが上がってきたということが一番大きな要因ですけれども、それ以外にもいろいろなトリックがあります。例えば、先程シグモイド関数を使うと言いましたが、今のディープラーニングではシグモイド関数はほとんど使われません。代わりに使われるのがReLUと言いまして、リクティファイドリニアユニットというものが使われます。これはどのような関数かと言いますと、 $\max(0, x)$ という関数で、負の領域だと0で正の領域に入るとxになるという関数が使

われます。

なぜかという、シグモイド関数はこのxが大きい値になると、この値が1でほとんど動かなくなるからです。したがって、最適化するときに微分を取ってもほとんど微分が効かなくなることが起こります。ところがこのReLUという関数にすると値が上がりに続けますから、微分を取ったときに効きやすいということで、このReLUというのが使われることが増えてきました。

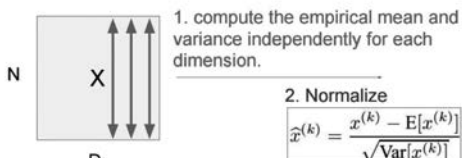
ところがこれが使われるようになったのは2012年です。ディープラーニング、ニューラルネットワークの研究はもう何十年も行われていながら、このようなReLUという関数を使った方がよいと誰も思わなかったわけです。確かに言われてみれば、シグモイド関数は格好いいんですね。1/(1+e^{-x})というのはすごいエレガントで微分を取っても綺麗ですし、何となくこれがいいとみんな思い込んでいたわけです。

一方でこのReLUというのすごい馬鹿みたいな関数で、微分不可能な点があって、すごい素人っぽいんですね。それで嫌だったのですが、ところがやってみると実はこちらの方がうまくいくということが分かってきて、今はもうみんなこちらを使っているということです。

なぜうまくいくようになったか: Batch Normalization (2015)

- 勾配消失、勾配爆発に対して、かなり強力な武器
 - これでほとんど気にしなくてよかった。
- バッチごとに正規化する。そのための層を入れる。 [Ioffe and Szegedy, 2015]

"you want unit gaussian activations?
just make them so."



他にも、Normalization Propagation(2016), Weight Normalization(2016)など、改良した方法が次々と提案されている。

このような例が他にもありまして、バッチノーマライゼーション、これも今ディープラーニングで非常に標準的に使われる方法です。これは何かというと、そもそも深い階層を持ったネットワークを最適化するときに、この微分を取って微分を層ごとに順番に伝えていくということをやります。誤差逆伝播と言いますが、その時にこの微分が0に収束したり無限大に発散したりということがよく起こります。そのような数値的に良くない挙動がよく起こるので、それがゆえに深い構造のネットワークは、なかなか学習できないと言われてきたわけですが、このバッチノーマライゼーションで何をするかというと、それが問題になるのだったら各層ごとに出る出力を平均0、分散1になるように正規

化したらいいいのではないかというアイデアなのです。平均0、分散1になるように正規化すると、確かに微分が上手に伝わるようになって、最適化がうまくできるようになりました。

今、ディープラーニングの手法だとバッチノーマライゼーションが非常によく使われておりますけれども、これが提案されたのが2015年です。これは驚くべきことで、普通データを使っている人だったら正規化は誰でもやります。ところがこれをやりませんでした。なぜやらなかったかという、これは格好悪いのですね。各ニューロンが出力を正規化することは、一体どのようなことなのか。それは、全体の合計値とか分散の合計値とかを取らないといけなくて、ニューラルネットワーク、脳のニューロンがそんなことをしている訳がないのです。そんな合計を取るニューロンがいるのですか、そんな訳ないので合計を取るとか分散を取るとか、そういうことと自体、非常にナンセンスな感じがするのです。

ところが、やってみるとうまく行きましたということで、もう今はみんな使っています。このように、要するに何十年もそもそもできない、こういうものだと思い込んできたのでみんなそれに捉われていて、ところが一旦できると分かったら、こういうことでもできるんだ、ああいうことでもできるんだというような発見が、今相次いでいるということです。

こういうのがたくさん積み重なることによって、層の数がどんどん増えてきて、それとともにエラー率がすごく下がってきているということです。

あと一つだけお話ししますと、ディープラーニングには、CNN (コンボリューションニューラルネットワーク) というタイプと、RNN (リカレントニューラルネットワーク) というタイプがあります。CNNは画像認識に向いていて、RNNは時系列に向いています。

機械翻訳などもディープラーニングでできます。今、NMT (ニューラルマシントランスレーション) と言いますが、翻訳精度がすごく上がっています。翻訳もどのように行っているかというと、結局最小二乗法です。例えば入力英語だとします。出力が日本語の文です。これで入力と出力をx、yだとみなします。そうした上で最小二乗法を適用します。つまり、入力となる英語の文を入力すると何らかの答えが出てくるようなモデルを作って、出力した英語の翻訳のようなもの、これが本来出すべき英語の翻訳とどのくらいずれているかという誤差の関数を定義します。この誤差の関数を最小化するように、中のパラメータをチューニングしていくということをやります。つまり、英語の文と日本語の文のペアがたくさんあれば、最小二乗法でこの誤差を最小化するようにパラメータを学習していくという仕組みによって、翻訳ですらこのディープラーニングでできるようになるということです。つまり、今いろいろな領域で使われるようになってきていますけれども、結局何を入力したら何が出るのかというのを、パラメータを介したモデルで作り出して、あとはデータがたくさんあれば最適化することによってパラメータを求めればよいという手法がすごく広がってきているということです。これによっていろいろなことができるようになってきているということです。

そう考えると、実はディープラーニングはやっていることはすごく単純で、データさえあればもうできます。ディープラーニングの手法自体は実はそんなに大変ではなくて、これはやればできます。ただし、一番大変なのはどうやってデータを作るかというところで、ここはすごく労働集約的ですし、ここのチューニングというか、どのような品質でどれだけたくさんのデータを作るかによって、出てくるモデルの質が決まります。そのような意味では、これもまた非常に日本向きというか、日本的な頑張って努力すればどんどん精度が上がるような、こういう領域ではないかと思っています。早くこのディープラーニングの技術を各社自社内の技術として取り込んでいただいて、そこにハードウェアとか現場のノウハウとかを入れ込んでいくことによって、競争力につながるようなそのような体制になればと思っています。

最後に、このディープラーニングの手法を今のような最適化、学習することによってシステムを作り上げていくという、このアプローチはソフトウェア2.0という言い方をしている人もいますが、恐らくかなり大きな変革をもたらすと思います。同時にこれは日本にとって非常に大きなチャンスだと思っており、なぜならこの眼の技術、ディープラーニングの技術というのは、物づくり、機械とかロボットと非常に相性がいいと同時に、少子高齢化で生産性が低下しているというときに、この眼を持った機械というのはさまざまな観点で役に立つ可能性があります。それが建設だったり物流であったり介護であったり、それからもちろん鉄道もそうですし、それから自動運転などもそうだと思いますし、いろいろな領域で活用できる可能性があるということです。

ただし、一方ですごく大きな危機感も感じております。先週深圳に行っていましたけれども、中国のAIの技術力は非常に高い、あっという間に先進国になりました。ここ2年ぐらいの間に急激に力を付けて、アメリカが敵わなくなりそうぐらい、質量ともに圧倒しつつあります。一方で、ハードウェアの技術も力を付けていますので、こういったものが融合する、眼を持った機械というのは放っておくと取られてしまうだろうなという危機感も感じています。

ですから、やはりスピード感をもってやっていかないといけないと思いますし、こういった世界の先にどのような社会をつくりたいのか、そういったビジョンを描くということもすごく重要ではないかと思っています。

以上になります。どうもありがとうございました。