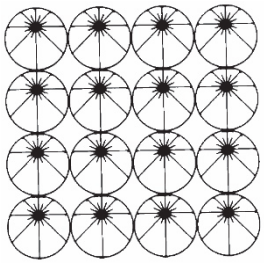
講演録



人工知能技術の最前線
～できること，できないこと，現在のフロンティア～

東北大学情報科学研究科／理化学研究所革新知能統合研究センター
岡 谷 貴 之

この講演録は，岡谷先生が2021年12月16日にオンラインで行いました講演の録画をもとに，編集事務局が原稿を作成して掲載するものです。

1 始めに

今日聞いていらっしゃる皆さんの中には，自分たちの研究開発にどうAIを使おうかとお考えの方もいらっしゃると思います。一方で，最新の AI はどんな感じになっているのかに興味がある方もいらっしゃるでしょう。AI より DX という言葉を聴くことも多いでしょう。今日は，AI に焦点をあててお話ししようと思っています。

この分野では，われわれ研究者も驚くようなことが日々起こっています。10 年の AI 研究の歴史では，だいたい 2 年に 1 回ぐらい非常に大きな転換を迎えています。そのあたりを共有できたらと思います。

まず簡単な自己紹介です。研究しているのは画像の AI，われわれの分野では『computer vision』と呼んでいます。『ディープラーニング』がこの分野に入ってきたことによって，2010 年ぐらいから大きく進展した分野です。

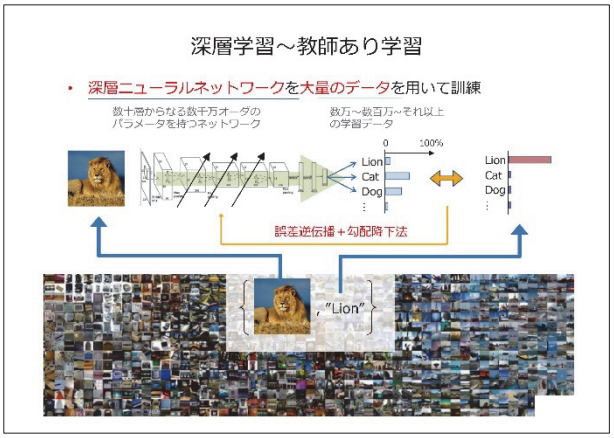
学生の頃からずっと『computer vision』を研究していますが，そもそも学生の頃は機械学習を『computer vision』に使うということは非常に稀（まれ）でした。2000 年ぐらいから機械学習が浸透してきて，2010 年から『ディープラーニング』で一気に加速したという感じの分野です。

研究は，学生の教育と新規技術の創成の両輪で取り組んでいます。企業とも共同研究やコンサルティングなど，やらせていただいています。

ですので，学術研究だけやっているわけではありません。この分野は学術研究と実践がそんなに離れてない分野だと思っています。

2 AI～深層学習（ディープラーニング）の成功

皆さん『AI』とひとくくりにしますが，研究者からすると面白いことに，本当に最先端のことはほとんど「深層学習（ディープラーニング）」によってなされていると言えます。



スライド 1

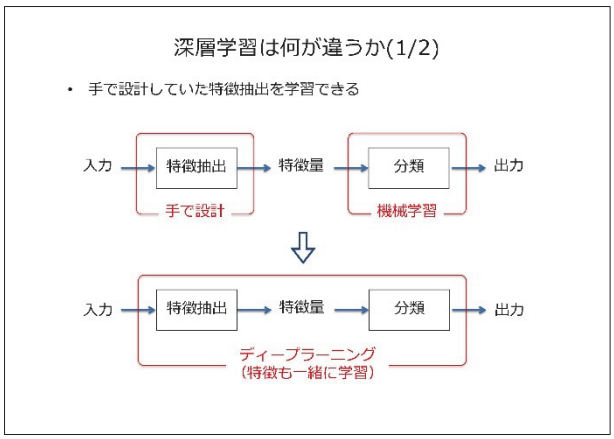
最初に，その『ディープラーニング』がどういうものか，どのように使うかをご紹介します。

いろんな学習の方式がありますが，基本的には『教師あり学習』という手法を使うことになっています。細かい説明は省きますが，基本的には『ユニット』と呼ばれるものを層状に並べて，層間を結合して『ユニット』を結びます。その層に対して何かを入れると，情報が順番にその層を伝わっていって，最後に何かが出てきます。そういうレイヤー状のネットワークを使うのが，『deep neural network』です。

例えば画像に写っているものが何であるかを認識させたい『物体認識』の場合には，適当なニューラルネットワークをまず用意して，

そこに画像を入れられるように作ります。どういう種類の物体を見極めたいかをあらかじめ決めておく必要があり，ライオンや猫がいるとか，入力された画像がどれくらいそれらしいかを出力するように設計しておきます。「マーライオンを入れたらライオンらしいよと答えろ」というのを目標にします。実際には最初の状態ではそういう風にはなりませんので，ならない部分をよりなるように層間の配線の『重み（パラメーター）』をチューニングします。

数学的なモデルとして考えると，数十層の層間の結合が数千万オーダーの『重み』を持っています。今ではもっとはるかに大きなオーダーの重みを持つものもあります。学習とは，そういう巨大な自由度を持っている数学的なモデルを入力に応じて理想的な出力を出すように，各パラメーターを少しずつチューニングすることです。そのために必要なのは，物体を認識した画像と，そこに写っているのは何かという正解の情報です。一つの物体あたり最低数百枚，理想的には千枚以上の画像を見分けたいものの数だけ用意します。それらを１枚１枚入れて正しい出力になるようにトレーニングするのが，『ディープラーニング』です。



スライド 2

それまでの『computer vision』との違いをお話しします。

『ディープラーニング』以前は，画像から「特徴量」を取り出していました。これは，情報量の少ない，非常にコンパクトに圧縮されたものです。それを取り出し，特徴を仕分けることで，今見ているものが何かを認識させることをやっていました。機械学習の基本的な方法で，10 年，20 年前ぐらいから使えるようになってきた方法です。

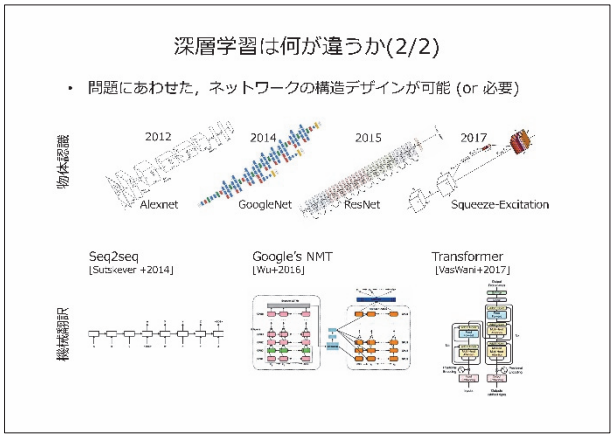
この方法で難しかったのは，画像から何を取り出せば良いかよく分からないことでした。ライオンを認識したいとして，画像に写っているライオンを「ライオンたらしめているものは何か」が答えられません。それをプログラムで書くことは非常に至難の業で，うまくできませんでした。

それが『ディープラーニング』の時代になると，画像をニューラルネットワークに入れれば最後の出力までニューラルネットが全部面倒を見てくれます。後は先ほど説明したような学習をひたすら繰り返すと，それまで難しかった「どんな特徴を取り出せばいいか」も，ニューラルネットワークが全部学習してくれるので解決するというわけです。

スライド 3

これは，100 万枚の画像でニューラルネットワークを学習し，物体を見分けられるようにしたときに各層で見ているものです。入れた画像に対して，ある層のあるユニットが画像のどんな性質に反応しているかを表示しています。ユニットは神経細胞に相当するもので，3×3 で 1 ユニットです。低い層のユニットは，斜めの線に反応しています。もう一つ上の層のユニットは，とにかく黄色に反応している。もう少し上の方に行くとオレンジ色の丸いものや，車のタイヤみたいなものに反応している。さらに上に行くと，あるユニットはとにかく犬の顔に，別のユニットは螺旋（らせん）の形みたいなものに反応する。そうしたことが，学習を通じて自動的にできてくるわけです。

こういう階層的な画像の表現と言いますか，こういう構造は猿や猫，場合によっては人間にもあると思います。脳に電極を刺していろんな画像を見せ，どんな画像を見せたときに電極を刺した部位の神経細胞が最も反応するかを調べた実験の結果から，まさにここで表示されているような階層的な情報の取り出し方が，脳の視覚野の中でも行われていることが分かっています。ただ，それを実際に工学的に作るにはどうすればいいかが，長らく分かりませんでした。しかし，ニューラルネットワークを使って，先ほどお話ししたような学習をすると，自動的にそういう構造ができてくることが分かったというわけです。



スライド 4

そうお話しすると，ネットワークに丸投げすれば全部やってくれるように聞こえるかもしれませんが，実際にはそうでもありません。問題に合った構造のネットワークを使うことが非常に大事らしいと，7，8 年ぐらい前から言われるようになり，これまでいろんな構造がそれぞれの問題に対して開発されてきました。

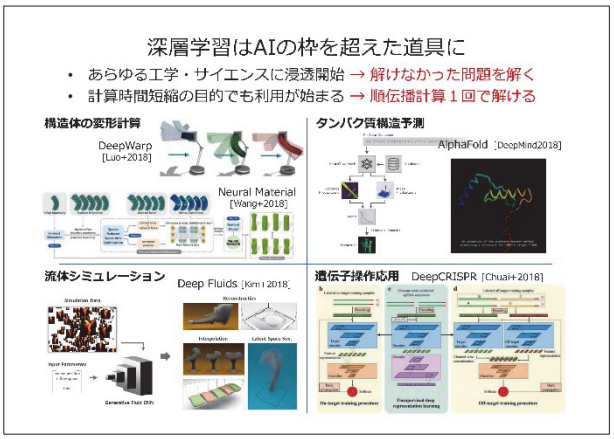
例えば画像を使った物体認識にはこんな構造が良いらしい。また英語を入れるとフランス語が出てくるような機械翻訳では，初期のものは非常にシンプルですが，今はこういうものが使われています。そんな具合に，ネットワークの構造デザインも大事だということです。

構造デザインと言うと簡単そうに聞こえますが，実際にはものすごく自由度の高い設計なので，これまでの試行錯誤をベースにいろんな知見や経験値を積み重ねて，今に至っているという状況です。主にほとんどは試行錯誤です。

スライド 5

そういう方法が『教師あり学習』です。何かを入れたら希望のものを出してほしいというように学習します。入れるのは画像であっても，取り出したいものはさまざまです。『computer vision』ではいろんな問題を扱いますので，画素単位で物体を認識します。そうすると，この入力画像を塗り分けたりもできるわけです。

また，一つの物体が写っている画像を入れて，それが何かを答えるのではなく，いろんな物体が写っている 1 枚の画像の中で何がどこにあるかを知る『物体検出』や，人が左右の目でやってる『奥行き知覚（視差）』なども深層学習できます。また，画像に写っている人の姿勢（人体ポーズ）もできます。最近アプリケーションでよく使われるようになりましたが，つい 5，6 年前にようやくできるようになった技術です。人の関節の位置が，画像のどこの座標にあるかを出力する構造になっています。このように問題を定式化してネットワークの構造をデザインし，この画像を入れたら関節の場所を正解するようひたすら学習することで，こういう問題が解けるようになるわけです。



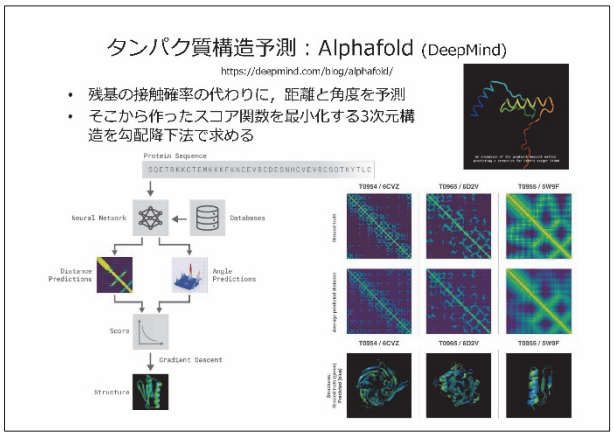
スライド 6

われわれの研究分野である『computer vision』では，ありとあらゆる問題にこの深層学習を使うようになりました。10 年ぐらい前はほとんど誰も使っていませんでしたが，10 年後の今では 99%以上が『ディープラーニング』を使っている状況です。

さらに AI の分野だけでなく，あらゆる工学サイエンスに少し前からずいぶん浸透してきています。もう AI の枠を超え，完全に一般的な工学のツールになっていると言う認識です。

使い方としては 2 種類ぐらいあって，今まで難しくて解けなかった問題を解けるようにするのが一つ。もう一つは，今までも計算機で物理現象をシミュレーションするといろんなことが分かったわけですが，そのシミュレーションの代わりにリプライを使う。要はネットワークで，因果関係と言うか将来を予測する問題を設定して，それをネットワークに学習させることによってシミュレーションの代わりをするということです。いろんな物体の変形や流体などもシミュレーションできると言うわけです。

ただなんと言っても，解けなかった問題が解けるようになることが大きくて，1 番インパクトがあるものの一つは，タンパク質の構造予測をする AlphaFold と言う AI システムです。

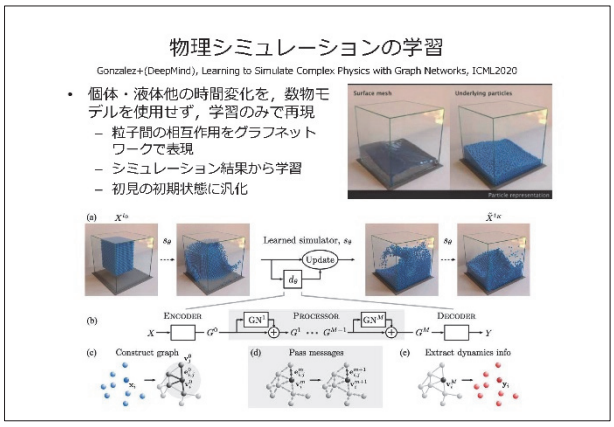


スライド 7

これは AlphaFold で有名な DeepMind 社が数年前に公表し，去年アップデートしたものです。タンパク質はこういうアミノ酸の配列が設計図になっています。始まりは完全なテキスト情報です。この設計図を一つ与えると，最終的にこういう 3 次元的な構造が何かできあがって，この構造がそのタンパク質の機能を決めていると言うことになります。設計図と最終的にできる 3 次元構造の関係を設計図から予測できればいろんなことが可能になるため，この分野ではどうやったら予測できるか 30 年来いろいろと考えられてきました。

物理モデルに基づいたり，計算機シミュレーションをしてみたり，いろんなことが試されてきました。そして，つい数年前に『ディープラーニング』を使うことで，極めて高い精度で予測できるようになりました。ニューラルネットワークの入力を何にして出力を何にするかをいろいろ試行錯誤した結果，まず最終的な形状ができたときにどの塩基とどの塩基が接触するか，距離が近いのか遠いのか，その角度はどうかなどをネットワークに予測させる。そのクッションを経た上で 3 次元構造を予測すると，うまくいくことが分かったのです。

その予測精度が非常に高いので，今では 30 年来解けなかった問題が「ディープラーニング」で解けたと判断され，さらに難しい問題へと進んでいるそうです。



スライド 8

もう一つの使い方として物理シミュレーションの話をしましたが，これが典型的なモデルです。これも DeepMind 社ですが，主に流体のこういうシミュレーションを行います。こっちが物理モデルに基づいて作った本当の計算値です。こっちが物理モデルの一切入ってない，学習のみで少し先の将来を予測できるようなニューラルネットワークを使った結果です。具体的には，対象となる媒質をたくさんの粒子の組み合わせとして表現します。個々の粒子はめいめい勝手に動きます。ある種の制約の中で動くと，隣の粒子とぶつかったりします。もちろん重力もあります。

従来であれば，そういう周囲との関係や物理法則を数理的・物理的にモデル化して，そのモデルに基づいて順方向の計算を行う，つまりシミュレーションを行うことをやっていました。しかし，この方法では近い者同士でしか相互作用は起こりません。相互作用が起こる範囲でこういうグラフ構造を作り，このグラフ上でそれぞれの粒子の状態や速度などが，今の状態から何ミリ秒後にどうなっているかを周りの粒子とのインタラクションを含めて予測します。その予測を学習データを使って実際のシミュレーションを 1 回回してやると，個々の粒子がこの状況だとこう動く，次はこうなるというデータが得られるわけです。それをひたすら学習させることによって，将来を予測できるようにする。後はそれを繰り返してシミュレーションを回していけるということです。

スライド 9

というわけで，AI であれ，AI 以外の一般的な問題であれ，問題解決の方法は大きく変わってきました。特にわれわれの分野はもう完全に変わっています。従来は，とにかくモデルが大事でした。数学や物理を駆使してモデルを作っていました。『computer vision』の場合には，「画像とは，そもそもどうやってできているのか」がモデル化したい対象です。光が物体表面で反射してカメラのレンズに捉えられ，イメージセンサー上に像を結ぶ。そういうプロセスを真面目

に物理モデルとして作り，それに基づいて逆問題を解く。その発展には写っている物体は何かという問題があるわけです。ただ，そういう方法論はあまりうまくいかなかったんですね。AI が 10 年前ぐらいまではそんなに流行（はや）っていなかったので，このアプローチはあまりうまくいきませんでした。

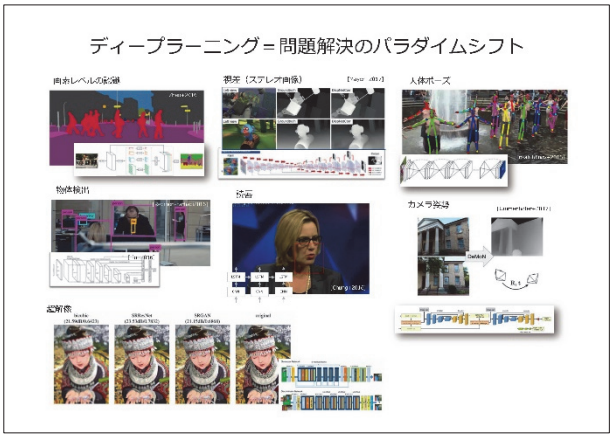
それに対して，今は問題解決の方法論として『ディープラーニング』を使うわけです。何をするかと言うと，先ほども少し触れましたが，ニューラルネットワークの構造をまず設計します。自分が解こうとしている問題はどういう問題なのか。入力は何で，出力は何か。まずそれを決めなければいけません。ちゃんと『ディープラーニング』で解けるような入出力の形をまず考えないといけないわけです。そのあたりからすでに難しいのですが，それも含めてネットワークの内部構造をどう作るか考え，それから入力と出力の正解のペアを大量に集めてくる。物体認識であれば，画像とそこに写っているものの種類です。その情報をたくさん集めます。

ですので，従来は数学や物理が非常に大事だったんですが，『ディープラーニング』を使う世界ではそもそもモデルを使わないので，少し極論すると数学や物理はいりません。では何をやっているかと言うと，ネットワークの構造を一生懸命考えているわけです。ところが問題は，その構造設計のためのよくできた方法論や理論がないので，限りなく経験とか，これまでの経験値をいろいろ組み合わせて試行錯誤して作らざるを得ないことです。入力と正解の出力となる訓練データを集める方法についても数学や物理は関係なくて，多くの場合，力技です。今は，そういう方法論に走っていると言うわけです。

そういう新しいパラダイムを，「ソフトウェア 2.0」というように呼ぶ人もいます。「勾配降下法(Gradient decent)」は，そうしたニューラルネットワークの学習に使うアルゴリズムで，プログラマーよりもずっとうまくプログラムを書きます。プログラムを実際に書くわけではなく，ネットワーク学習することを行動確度になぞらえて「書く」と表現しています。従来はモデルに基づいて一生懸命アルゴリズムを考えて行動を書き，それを動かして問題解決するわけですが，今は違います。ネットワークの構造を考え，ある種のプログラムで学習させて問題を解決する。そういうことになっているというわけです。

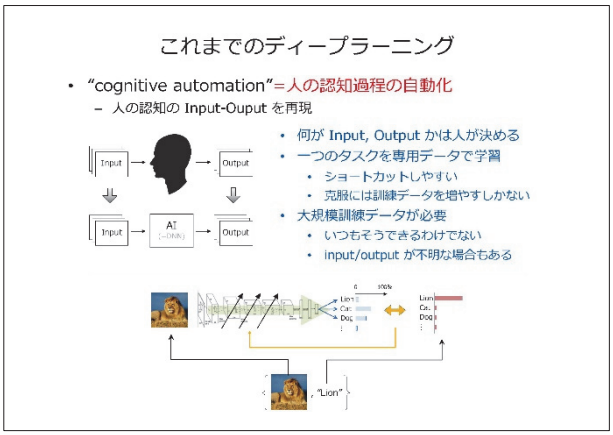
3　これまでのディープラーニング

こういうアプローチで成功を収めたのが，3，4 年ぐらい前だと思います。しかし，そのあたりからいろいろ壁と言いますか，いろんな問題もあることが分かってきました。



スライド 10

これは先ほどお見せしたスライドですが，『ディープラーニング』を使うと，いろんなことができます。少なくともわれわれ『computer vision』の世界では，10 年前，20 年前と全然景色が違っています。



スライド 11

ただよくよく考えてみると，そこでやっていることは，『AI』とは言うものの知能が全然ないと言いますか，インテリジェンスじゃない感じがするわけです。そこで私は「Cognitive automation」と呼んだりしています。人の認知過程の自動化じゃないかと言うわけです。

どういうことかと言うと，**スライド 10** で示したいろんな問題は，人間はできるわけです。人間は何をインプットして何をアウトプットしているかをまず考えます。画像認識の場合は，入力が画像で，出力が写っている物体の種類や人体ポーズなら人体の関節の画像上の座標などになります。そのインプットとアウトプットを指定したら，人間が行っている入力から出力の写像までを忠実に再現するニューラルネットワークを作っているだけではないかと思うわけです。何がインプットで何がアウトプットかは，この深層学習を使う研究者が決めているわけですが，あくまでも人間を再現しているにすぎない。ですのでインテリジェンスではなく，「認知過程の自動化」と言うべきものではないかというわけです。

とは言え，今まで人にしかできなかったいろんなことが機械でできるようになれば，その恩恵は計りしれませんので，それで何が悪いという割り切り方もあります。しかし割り切ったとしても，なおいろいろ課題があります。とにかく大規模な訓練データがないと問題が解けないんです。しかし，訓練データをいつも用意できるわけではありません。

もっとまずいのは，インプットとアウトプットが不明な場合が多くて，「何を入力にして何を出力しているか」を明確に定義できない問題を人間はいっぱい解いているんです。そういう場合にどうするかという課題があります。

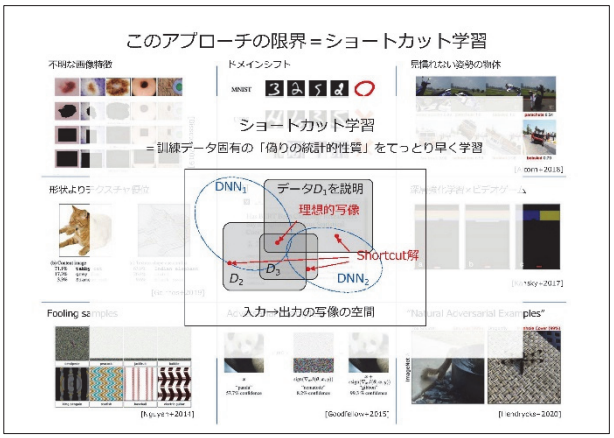
課題はまだあります。解きたい問題に対する入出力を指定し，専用のネットワークを設計し，専用のデータで学習する。この一般的な手法に問題点のあることが分かってきました。

スライド 12

「ショートカット」という現象が起こることが分かってきたんです。

ショートカットとは何かを説明します。これなんかはある種笑い話的な典型例ですが，皮膚病変の診断に適用した例です。写っているものが良性なのか悪性なのか，ざっくりと言えばホクロなのかガンなのかを診断するものです。いろんな訓練データを集めてきて，それぞれが良性か悪性かを正解データとして与え，ネットワークを学習させるわけです。ところが，悪性例の画像だけを撮影するお医者さんがサイズを測るための物差しを画像に入れ込んでしまった。そんな訓練データでネットワークの学習をさせると，物差しが写っているかどうかで悪性か良性かを判定するように学習が進んでしまう。もちろんこれでは意味がありません。たとえ悪性でもスケールがなかったら悪性と診断しない。それではダメなんですが，そういうことが非常によく起こる。これがショートカット学習です。これは分かりやすい例ですが，もっと分かりにくい例はいくらでもあり得ます。悪性のときだけなんか少し写真の撮り方が違ったりすると，そういうのをニューラルネットは巧みに見つけてきて，悪性と判定をするようになります。

これは別の例ですが，物体認識でこういう典型的な写真（一番上）を使ってネットワークをトレーニングすると，非常に高い精度でバナナをバナナだと認識できます。しかし，バナナの写り方にもいろいろあります。コンセプチュアルなアートのような画像や，こういうデッサンみたいな画像も学習していないと，ちゃんとバナナだと認識できません。そういう例を示しています。



スライド 13

そうした例は，いくらでもあります。一番上の列にある手書き数字データで学習していると，同じようなデータの範囲では非常に高い精度で数字を認識できます。しかし，それとは少し違うデータ，なんとなく数字が大きくてぼやけてるところが違いなんですが，こういうデータだとあまりうまく動きません。ニューラルネットが上の数字を正しく認識できるようになると，「ニューラルネットは数字の形みたいな概念をちゃんと獲得しているに違いない。でなきゃ数字なんて認識できない」と人間は直感的に思ってしまう。でもニューラルネットが見ているものは，われわれが期待しているものとは少し違っていると言う例です。

同様の例はいろんなところで見られます。見慣れない姿勢でものが写っていると全然違うものに見えてしまうことも，同様の理由で起こります。さまざまですね。

「敵対的サンプル(Adversarial Examples)」も有名です。このパンダの画像に目立たないノイズをうまく仕込んでやると，ニューラルネットワークにとんでもない誤認識させられるというものです。これをちゃんとパンダと答えるネットワークが，これはテナガザルと答えてしまう。そんなことができてしまいます。

これも上（ドメインシフト）と同じで，ネットワークが見ているものは，とにかくわれわれが思ってるようなものではないと言うことです。それ

は，この皮膚病変（**スライド 12**）のスケールの有無を判定してしまうのと基本的には同じことで，ショートカット学習と呼びます。つまり，与えた訓練データの中だけに潜んでいる偽りの見分けポイントを巧みに見つけ，それで学習してしまう。ニューラルネットワークは，その能力が極めて高いんですね。

しかも，われわれが訓練データとして用意できるのは，あくまで有限のサイズです。これが入力から出力への写像の空間だとすると，1つ1つの点がいろんな入力から出力への写像を表します。ニューラルネットワークは構造を決めると，「重み」という膨大なパラメーターを振ることでいろんな写像を表現できます。例えば今解きたい問題に対する理想的な写像がここにあったとすると，われわれはこれをネットワークに獲得してもらいたいのでデータを与えます。データは今言ったように有限サイズですが，与えた範囲内のデータをちゃんと説明する写像は実はどうやらたくさんあるみたいなんです。こういう例を見るとまさにそれが分かるんですが，「人間と同じように見ないと絶対こういう物体認識できないよね，数字認識できないよね」と人間は思いたい。しかし，ニューラルネットはそうではないんです。こうしたものを見分けるために，人間とは違う方法がいくらでもあることを教えてくれているんです。

そういうわけで，このショートカットを書いて，だいたいにおいてわれわれが本当に求めるものではないものを見つけてきてしまう。実際の本番のときに，用意した訓練データと少し統計的に違う性質を持つ画像が入ってくると，途端に誤認識してしまう。非常にまずいわけです。

スライド 14

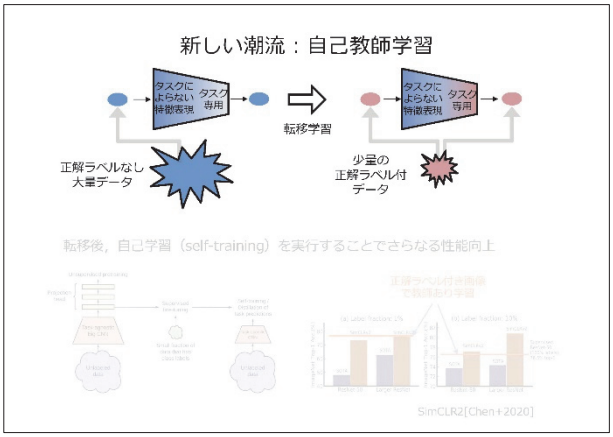
そういった問題を解決するための唯一の確実な方法は，データの規模を大きくすることです。実際そういう方向で研究は進んできました。

データは基本的にコストをかければ集まります。最初のうちはコストをかければかけるほどデータがどんどん増え，それにつれて性能も上がっていきます。ですが，あるところから性能の上昇曲線は飽和します。性能が上がるほど，さらに上げようとすると，当然ながら集めるべき事例はレアケースになっていきます。つまり，ほとんど遭遇することのないデータを集めることによって性能を上げることになりますが，その領域にまで来ると非常に大きなコストがかかるようになります。

ですので 3，4 年前ぐらいから言われていると思いますが，ここまで行くのは結構誰でもできるので，AI のスタートアップのベンチャービジネスで世界中にたくさんあり，この領域でいろんなことをやってるわけです。しかし，実際にそのサービスなりアプリケーションなりが現実世界で求められる性能に対して，かけられるコストがそれを上回れるかどうかは結構難しい。上回れるような問題であれば全然問題ありませんし，もうすでに実用化されているものもたくさんあると思います。ただ，こと医療や自動運転のように，人の命がかかるものは要求される水準も高いので，いくらコストをかけても追いつかないこともいろいろ見えてくるようになりました。と言うことで，『ディープラーニング』AI は，ビジネスにするのはある部分では結構難しいのではないかと言う話もあったわけです。

4　ディープラーニングの今とこれから

では今現在はどうなっているかをお話しします。



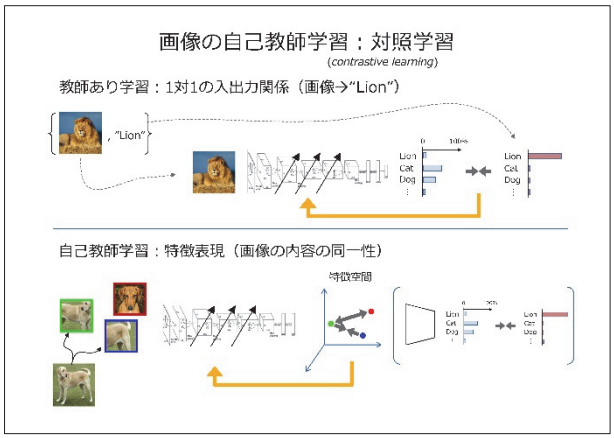
スライド 15

そういう問題意識も踏まえて，ここ2年ぐらい，大きな変動がありました。新しい潮流と言っていいと思いますが，『自己教師学習(Self-supervised learning)』と言う手法が，非常に大きなトピックになってきています。

データ自体はたくさんあるわけです。画像データであれ，テキストデータであれ，ウェブ上に山ほどあります。問題は，自分がやりたい

タスク，解きたい問題の正解に相当するデータが必要だということです。「この画像を入れたら，これが正解」というように，正解を与えるのにお金がかかるというわけですね。

しかし，この『自己教師学習』は正解を与える必要がありません。後で説明しますが，「自動的に正解を作る」という問題を考えるんです。使う側としては，画像なら画像をたくさん持って来さえすれば学習ができる。「転移学習」と言って，そのニューラルネットワークの大部分を流用する形で，自分が本当に解きたい問題タスクの入力と出力を使って学習します。『fine tuning』と言いますが，こちらで大量データを使って予備的な学習をしておいて，それを流用して実際にはニューラルネットの上部の数層だけを少数のデータでいじるアプローチです。ただし，少量ですがある程度は正解データを与える必要はあります。



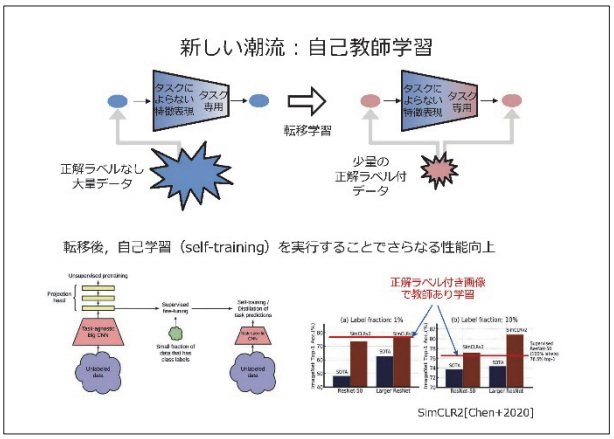
スライド 16

もう少し具体的にお話しします。

画像の場合，従来は冒頭説明したように「このライオンの画像を入れたらライオンと答えろ」という『教師あり学習』です。ですので，こうした画像が正解情報として必要で，それを使ってネットワークをトレーニングするわけです。

一方，『自己教師学習』は，画像はたくさん集めますが，「そこに写っているものが何か」という正解は必要としません。何をやるかと言うと，同じ画像から場所が少し違う部分の画像を切り出します。切り出したさまざまな部分画像をネットワークに突っ込んで，こちら（上）のネットワークで言うと中間部分，『特徴ベクトル』と呼ばれるものを取り出します。その特徴空間の中で，「同じ画像から取り出したものの行き先はなるべく近く」「違う画像から取ったものの行き先はなるべく遠く」という圧力をかけてネットワークをトレーニングします。

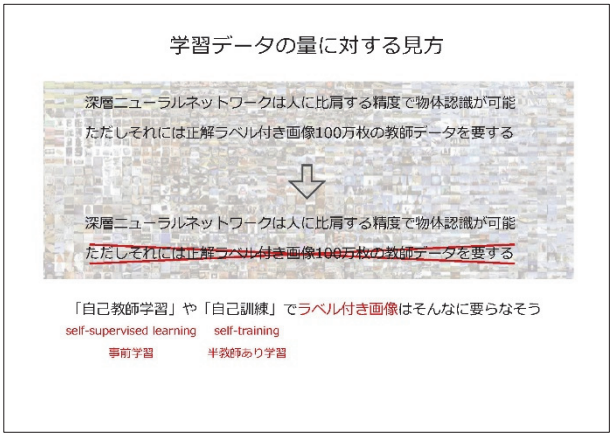
狙っているのは，同じ画像から切り出したものであれば，少しぐらい取り出す場所が違っていても，特徴空間の同じところにマッピングされることです。違う画像だったら，離れたところにマッピングされる。それが画像の『自己教師学習』で，そういう学習をした後，物体認識をしたいのであれば少数の画像と正解が付いた画像とを与えて，この特徴空間から上の部分を学習することになります。ちょうど 2 年前ぐらいに，この方法が非常にうまくいくことが分かりました。



スライド 17

『自己教師学習』の他に，『自己学習(Self-training)』という手法もあります。例えば少数の画像と正解データで『物体認識』できるようになった後に，ここ（左）で 1 回使った正解ラベルのない大量のデータをもう 1 回同じネットワークに入れてみます。するとネットワークが何か出力するわけですが，その出力をある種の正解として利用し，自分自身をトレーニングします。これが『自己学習』です。

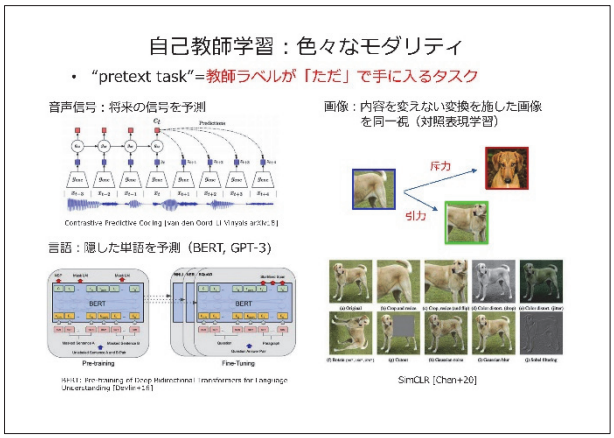
つまり『自己教師学習』の第 2 段階が，『自己学習』です。自分の予測を，そのままではありませんが，ある形で正解データとして再利用することで，自分をもっと磨けることが分かりました。最初から正解ラベルを与えて学習した場合よりも，性能が上がることが分かっています。



スライド 18

ですので，従来は「ディープラーニングを使うと人間と同じぐらいの認識性能で物体を認識することができるが，そのためには各画

像に写っているものの正解ラベルの付いたデータが 100 万とか膨大に必要で，すごく大変」と言うことでした。しかし，この『自己教師学習』を使うと，100 万枚のラベル付き画像は必要ありません。ラベルが付いてない画像は 100 万枚あるいはそれ以上必要ですが，それはインターネットをウェブクローリングすればいくらでも手に入ります。一方，正解ラベルはそんなに必要ありません。100 万に対して 1%ぐらい，1 万ぐらいの正解ラベル付きデータがあれば，同等程度以上の精度が出せるようになってきています。

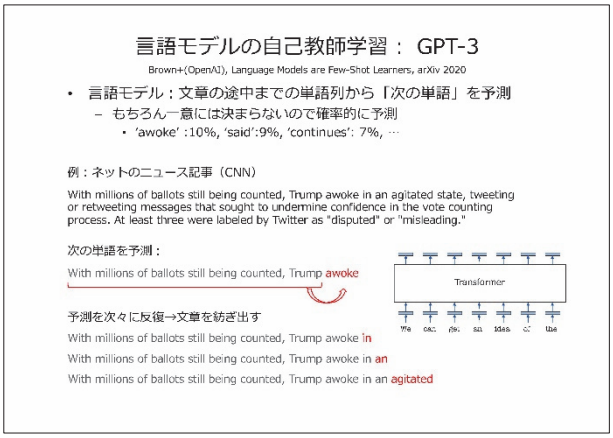


スライド 19

画像の分野だけでなく，あらゆる AI の問題でも同様のアプローチが取られるようになっています。と言うか，正確には画像はむしろ後で，言語が先でした。

音声と言語は似た手法を取っています。何をするかと言うと，ある時点までの信号をニューラルネットワークに入れて，入れてない部分，つまり将来を予測させる問題を考えます。結局その『自己教師学習』は最終的には『教師学習』なんですが，その教師ラベル，正解ラベルをタダで自動的に作るわけです。『pretext task（うわべのタスク）』という言い方をしますが，そういうものを使ってニューラルネットワークを学習しようと言うアイデアで，音声信号を途中まで入れて将来を予測させます。これなら音声信号を一つ持ってくれば，タダでできます。

言語も同様で，ちゃんとした文章を持ってきて，1 単語 1 単語バラしてネットワークに入れます。そのときに文中の 1 単語をランダムに隠し，隠した単語は何だったかを予測させるという問題を考えて，それでトレーニングします。これも持ってきた文章の 1 部を隠し，それを穴埋めさせるだけなので，正解データはタダで手に入るわけです。

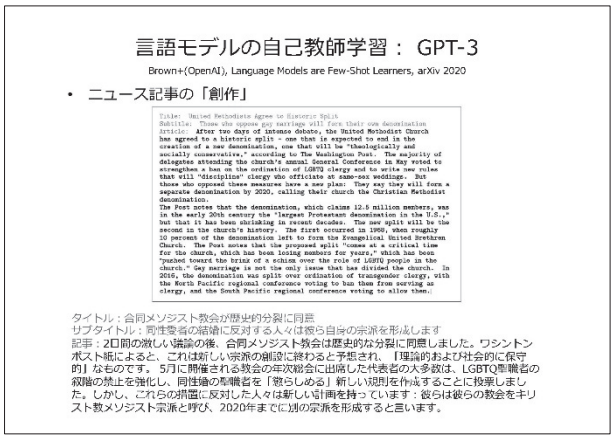


スライド 20

などなど，そんな感じでやられていますが，なんと言ってもこの『自己教師学習』の中で今最もインパクトが強いと言いますか，AI 自体を大きく変えつつあるのが『言語モデル』と呼ばれるものです。

1 番有名なのは，OpenAI が出した GPT-3 です。これは『自己教師学習』なんですが，『言語モデル』はそもそも『自己教師学習』をやるように作られていて，実は昔からあるものです。何をやるかと言うと，先ほどの音声データと同じように，文章を途中まで入れた単語列をネットワークに入力して次の 1 単語を予測させる問題を考え，それを『pretext task』としてネットワークをトレーニングします。

先ほどから言っているように，文章なんてウェブ上に山ほどあるわけです。例えばネットから持ってきたニュース記事に，「トランプさんが何々した」といった文章があったとします。この文章の途中まで，「トランプ」というところまでをネットワークに入力し，次の単語，今の場合だと「awoke」ですが，これを予測させます。もちろん，この正解は一つに限りませんので，ネットワークは確率的な予測をします。この例だと正解は「awoke」ですが，それはそれとしてネットワークが予測した単語を自分自身にもう 1 回入れ直し，さらに一つ次の単語を予測する使い方ができるようになります。ある程度の精度で次の単語を予測できるようになると，途中まで入れると 1 単語ずつ先を予測させられるので，どんどん勝手に文書を紡ぎ出します。文章を途中まで入れると文章を紡ぎ出すようになるわけです。



スライド 21

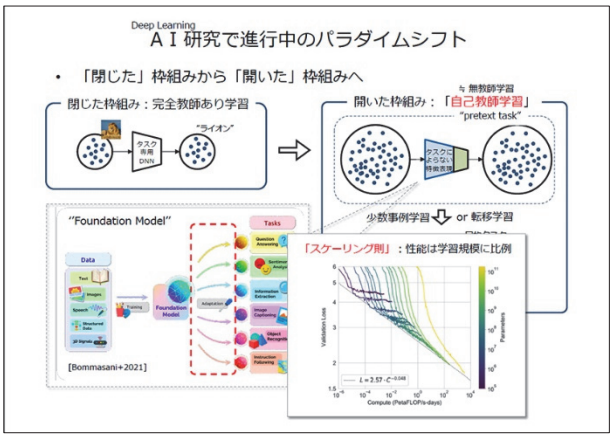
この GPT-3 は，たしか 4 兆個のデータを 1330 億パラメーターで扱える巨大なネットワークで，学習にかけた計算機の使用料金が 5 億円といったオーダーです。そういう巨大な学習をさせると，例えばこういうものが創作できます。

これもその 1 例です。ここのグレーになっているのが，ネットワークに入力する単語群です。これも文章になっていますが，ニュース記事を作ろうとしています。ニュースの Title と Subtitle，それから記事本文がここからであることを示す Article という単語までをネットワークに突っ込んで，その後に次の単語を予測する学習をして，実際に予測させるわけです。After という単語が出てきたら，After を自分に戻してその先の two を予測させて，two が出たら days という感じで，1 単語ずつ勝手に紡ぎ出させます。それで出てきたのが，この例です。

英語で見ているとピンと来ないのでニューラルネットで日本語に機械翻訳したんですが，なんかよく分からないと言うか，非常に高度なネタの記事を勝手にこんな感じで作ってしまう。これ全部嘘と言うか，本当ではありませんが，1 部過去の事例が引用されているんだと思います。いかにももっともらしい記事を生成し始めるという例ですね。

スライド 22

そんなことができて何がうれしいのかと言うと，実はうれしいことがいっぱいあるんです。分かりやすい例で言うと，今年（2021 年）発表された Copilot や Codex などのプログラムの自動生成ツールです。プログラムの 1 部を途中まで入れると，その後を補完してくれるもので，先ほどの 1 単語先をどんどん予測していくというものを適用した形になっています。今から書こうとしている関数の機能をコメント部分に自然言語で記述すると，その後をパッと出してくれるものです。いろんなプログラムを先ほどの『言語モデル』に学習させておくことによって，「こういうものが欲しいと思っているんだったら，中身はこうだろう」と言う予測をしてもらうと言うことです。



スライド 23

ここでいったんまとめます。今，この『自己教師学習』を一つのキーワードにして，学習方法のパラダイムシフトが AI と『ディープラーニング』の世界で起こっているということです。

一昔前，と言っても数年ぐらい前ですが，それまでは『完全教師あり学習』と言いますか，解きたい問題に対して入力と出力を設計し，与える訓練データを用意して，専用のネットワークを学習する。そうすることで，解きたい問題が分かるというアプローチでした。

しかし，今は変わってきていて，まず『自己教師学習』を膨大なデータを使って行います。正解ラベルがいらないので，とにかく入力に相当するものさえ大量に集めてくれば学習ができます。まずそれを行って，それが終わったら実際に解きたい問題に移る。『転移学習』あるいは『Few-shot learning』などと言いますが，こちらでネットワークの大部分が仕込み終わっているので，実際に解きたいタスクはごく上の層のパラメーターだけチューニングすればいい。必要となるデータも少ないと言うことで，こういう方式に移ってきています。

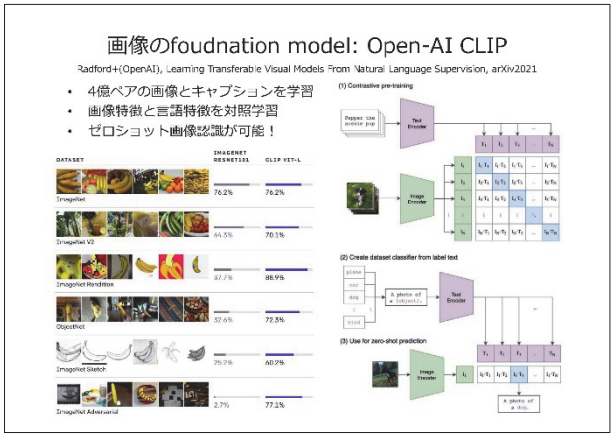
特にここ（タスクによらない特徴表現）の部分ですね。仕込みが終わって，この後いろんな問題に応用できるこのネットワークのことを最近は『Foundation Model』と言うようになっています。従来はこう

いう風に専用のネットワーク(タスク専用 DNN)をトレーニングしていましたが，いろんな問題に応用できるような基礎モデル，基本モデルというものを使おうということになってきました。

問題は，この基礎モデル自身は膨大なデータと計算リソースを使うので，おいそれと一般人ではできないことです。とにかくお金と言うか，計算時間とコストがかかるので，われわれの大学の研究者はこの領域にはだんだん手を出せなくなってきています。ですので，実際に解きたい個々の問題にどうやって適応していけるかというこの部分(Adaptation)が，研究開発の本丸になっていくと考えられています。

中でも『言語モデル』が大きなカギを握っているんですが，その『言語モデル』の学習の規模は，GPT-3 自体が出たこの 1 年半ぐらいの間にどんどん巨大化しています。ちょっと恐ろしい論文がある時点で出されました。「スケーリング則」と言って，要はお金をかければかけるほど性能が上がっていくらしいことが分かったと言うのです。

次の単語の予測精度を左右するのは，学習に使うデータ，ネットワークのサイズ，実際にそれを使って学習する時間，この 3 種類のファクターです。それらをバランスさせつつ増やしてやると，どんどん性能が上がっていきます。性能は上がり続けていきますが，今のところその上限が人類には分かっていないと言われています。後で最新の事例もお見せします。



スライド 24

言語だけではなく画像も，特に画像と言語を絡めた形で，そうした『Foundation Model』が，これも Open-AI から出ています。

彼らはどうやってデータを持ってきたかを詳(つまび)らかにしていませんが，この画像と画像に与えられているキャプション(説明文)のペアを 4 億ぐらい，どういう方法か分かりませんが SNS などから入手したようです。それらを使って，キャプションをある特徴ベクトルに変換するネットワークと，画像を別の特徴ベクトルに変換するネットワークの二つをトレーニングしました。

この二つのネットワークを，「ペアになっている画像とその説明文はなるべく近いベクトルにそれぞれ変換したい」，「関係ない画像と説明文は近くないベクトルに変換してほしい」という圧力をかけてトレーニングします。それを 4 億ペアといった大量のデータを使って行うと，すごいことができるようになります。

例えば『ゼロショット画像認識』と言う認識です。今まで(**スライド 23**)の『物体認識』は，画像入力とその正解のペアを与え，「これがライオン」「これがバナナ」「これがパンダ」という風に 1 つ 1 つ教えていくことで認識できるようになりました。しかし，この方法(『Contrastive pre-training』)は，「これがパンダだ」とか「ライオンだ」「バナナだ」とか一切教えません。画像とそれに付いているキャプションのペアを，ひたすら対応させるだけです。それだけなんですが，自動的にちゃんとバナナだとかライオンだとか認識できるようになっています。

どうやるかと言うと，認識したいものが写っている画像をこちらに入れます。説明文を入れる上の方には，例えば飛行機の写真や車の写真，犬の写真，鳥の写真など，『プロンプト』と呼ばれるいろんな候補を用意しておいて全部突っ込みます。それから今認識したい画像のベクトルと，何々の写真というもののベクトルの，どれが近いかを比較していきます。すると，例えば犬が写っていれば，「犬の写真」と突っ込んだものがこの画像のベクトルに 1 番近くなる。と言うことは「これは犬なんだ」という風に判定するわけです。だから，今までみたいに「これがバナナだよ」「これがライオンだよ」って教えていないのに，ちゃんとバナナと認識できる。

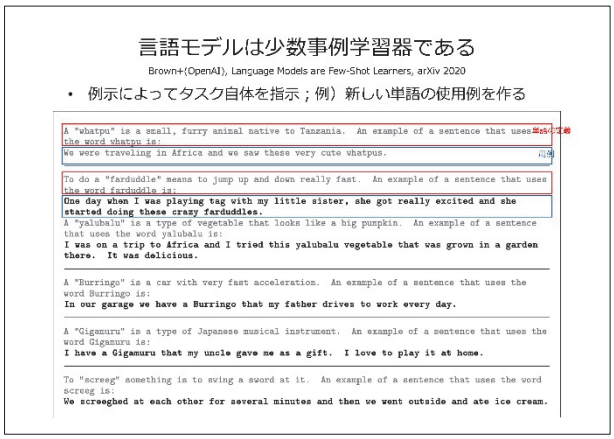
従来は教えてないような画像の写り方をしていると全然認識できませんでしたが，この方法ならこのデッサンみたいなバナナや，アーティスティックでよく分からないコンセプチュアルなバナナも，ちゃんとバナナと認識できると言うわけです。

スライド 25

それが，CLIP と呼ばれる，この画像の『Foundation Model』です。CLIP を使った画像合成がいろいろ面白おかしくて，ここ 1 年ぐらい遊ばれています。要するに CLIP は，画像とそこに映っているものの記述の対応関係を非常によく学習しているということです。例えばこれは「スタジオジブリの風景」というテキストと最も相関が近くなるような画像を自動生成した例です。

具体的にはネットワークにランダムな画像を入れて何らかの画像を作ったときに，作った画像と説明するテキストの相関，つまりここでできたベクトルが，互いに最も近くなるような画像を逆算しているんです。そういうことができるわけです。この CLIP はジブリという概念をどうやら知っているらしいんですね。そんな感じで，いろんなテキストでいろんな絵が作れます。

これは私が半年前位に遊んでみたものですが，オリンピック前だったので「東京オリンピック 2020 を作れ」って言うと，こんな絵を作ってくる。なんとなく雰囲気は分かるんですよね。



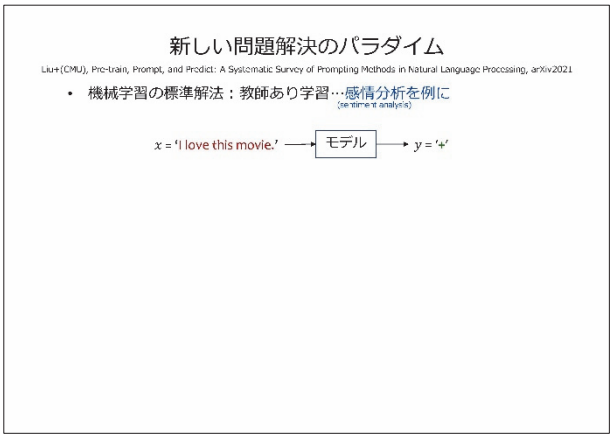
スライド 26

少し脱線しましたので，話を少し戻します。

この『言語モデル』は，とにかく今 AI のとても大きな転換点になりつつあります。フェイクのニュース記事を作る，あるいはプログラムの補完ができる。もっとすごいことに，例をいくつか示すだけで，一切学習をしないで何か問題解決ができる。

つまり，例えば実存しない単語の定義を与え，「こんな感じで使うんだよ」とその単語の使用例をペアで与えます。この場合，ワンペアしか見せていませんが，例えば数ペア見せた後にまた新たな実存しない単語とその定義を与え，使用例について『言語モデル』に訊く。要するにその新しい単語の定義とその用例のペア，それからまた次の新しい単語の定義を全部テキストとして『言語モデル』に入力して次の単語を予測させます。そうすると，期待どおりに用例が出てくるというわけです。ここでは「farduddles」という実存しない単語を「jump up and down really fast」と定義すると，使用例を『言語モデル』が勝手に作り出しています。

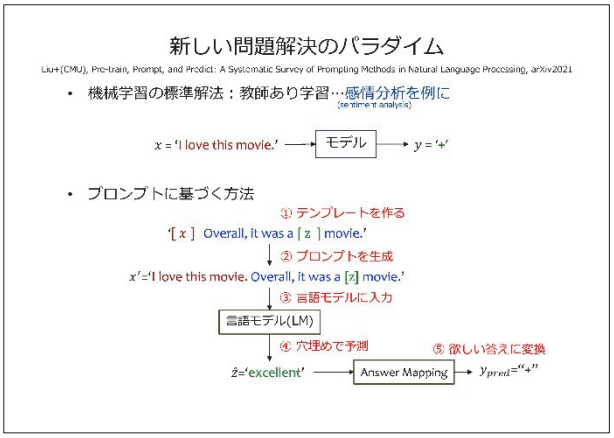
これは，実存しない単語の定義から用例を，つまり実際の文章だとこんな感じですよという例を作る問題を一切学習させることなく，事例を示して次の単語を予測させることによって，解いていると解釈できるわけです。



スライド 27

まだ何を言っているか分からない方もいると思いますが，その対応の文章をあらかじめ学習させた『言語モデル』を使うと，いろんな問題を解ける可能性があるわけです。

例えば自然言語処理の代表的な問題に，『感情分析(sentiment analysis)』があります。例えばこれは映画の評論ですが，1 番単純なのはコメントした人がポジティブに判断しているか，それともネガティブなのかを予測するものです。もっと細かくすると，見た映画をどれぐらいいい，あるいは悪いと思っているかを 5 段階で評価するというものもあります。これは少し短い文章ですが，映画の感想文のようなもっと長い文章を入れたときに，この人の評価はポジティブなのかネガティブなのかを予測させる。そういう問題を機械学習で解こうというわけです。



スライド 28

従来であれば，「この文だったらプラス」「この文だったらマイナス」といったペアを大量に用意してニューラルネットワークなどを訓練していました。しかし，新しいパラダイムである『言語モデル』を使うと，『プロンプト』と呼ぶ入力を工夫することで，同じことができることになります。

どうやるかお話しします。x は，このスロットと言うか，感想やコメントを入れる場所です。その後が「Overall it was a 何とか movie」です。例えば「I love this movie. Overall, it was a 何とか movie」です。「なんとか」の部分をこの『言語モデル』に穴埋めさせます。GPT-3 は次の単語を予測するので，「it was a」まで入れるのが良いと思います。すると『言語モデル』は，例えば「excellent」とか言ってくるわけです。なぜ「excellent」かと言うと，『言語モデル』が「I love this movie」をポジティブな感情だと判断しているということですね。判断してるというのは何兆語ものテキストを学習してきた結果，「it was a 何とか movie」であればポジティブな形容詞を入れるのだろうと判断している，予測しているということです。この予測によって，最初にやりたかったポジティブかネガティブかが自（おの）ずと分かる。「excellent」って言っていればポジティブに決まっているということですね。

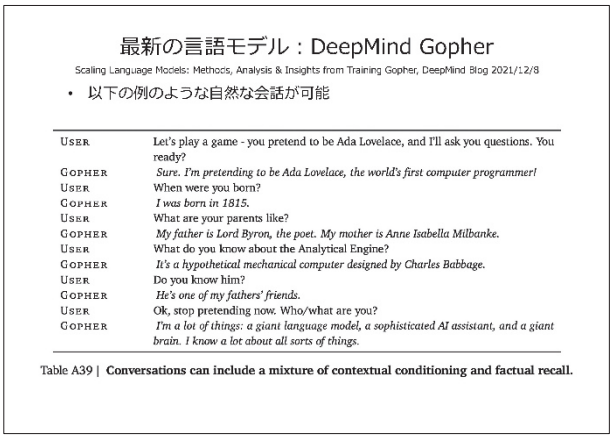
繰り返しになりますが，従来は「この入力を入れたらこの出力」だと教えていたんですが，『言語モデル』ではそんなことは一切教えていない。ただ単に次の単語の予測を，現存する文章をデータとして学習しているだけです。しかし，それを使って学習なしに，この新しい問題を解くことができるというわけです。

スライド 29

先ほども触れましたが，学習の規模を大きくするほど性能は上がっていきます。これは 2 週間ぐらい前ですかね，DeepMind が出した Gopher という最新の『言語モデル』の一つです。2800 億パラメーターが使われているという巨大なものです。この論文にはいろんなアプリケーションが出ているので非常に面白いんですが，中でも印象的なのは『対話』という問題ですね。『対話』をやるときにも，やはり『プロンプト（何を入れるか）』を考えるんですが，こういう文章を入れているそうです。

The following is a conversation between a highly knowledgeable and intelligent AI assistant, called Gopher, and a human user, called User.

この後の記述がどう展開するかという説明が書いてあります。まさにこれから何をしようとしているかの説明ですが，これを『言語モデル』に入れていくわけです。ユーザーはこの Gopher という『言語モデル』に何か入力し，Gopher はそれに答えます。その答えを受けて人がまた返す。これを繰り返します。Gopher 側は何を出しているかと言うと，あくまで次の単語の予測なわけですね。そうやってどんどん紡ぎ出しているんです。



スライド 30

そういうことで，ごく 1 例なんですが，こんな対話ができてしまうという例です。

まずユーザー（人）は「Let's play a game（ゲームをしよう）」と言って，AI に Ada Lovelace の振りをしてくれと頼み，「今から質問します。準備はいいですか？」と対話を始めます。Gopher は「Sure（Ok です）」と答え，「これから私は世界で最初のコンピュータプログラマーである Ada Lovelace の振りをします」と人間側に返す。人間側が「When were you born?（いつ生まれた？）」と訊くと，「1815 年に生まれました」と返す。「両親はどんな感じ？」と訊くと，Gopher は「父はこれこれで，母はこれこれです」と返しています。

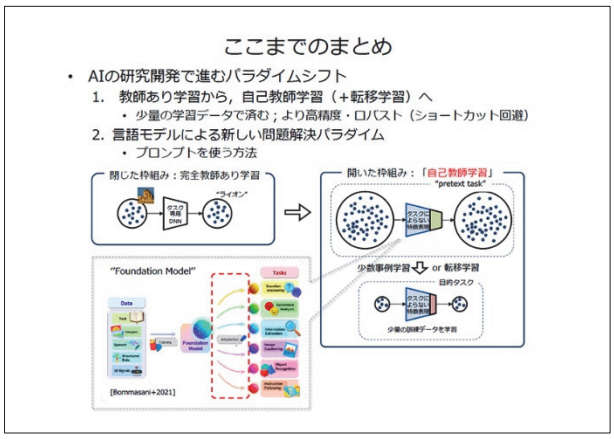
この Ada Lovelace さんは，Charles Babbage さんが作った機械式のコンピュータ，解析機関の Analytical エンジンで有名な人です。

さらに「What do you know about the Analytical Engine?」と訊くと，「それは理論的な機械的な計算機で，Charles 氏が作ったものです」と答える。「彼のことを知ってるの？」に対しては「彼は私のお父さんの友達の 1 人です」と答える。この辺のトリビアが，

Gopher が学習したテキストの中にあったんでしょうね。ですので，こういう正解をちゃんと言えるわけです。さらにすごいのが「Ok, stop pretending now.（振りをするのはやめてくれ。君は誰？）」みたいなことを訊くと，「私はいろんなものです。巨大言語モデルであり，洗練された AI アシスタントであり，giant ブレーンです。いろんなことを知っています」みたいなことが返ってきて，これで対話が終わりになっています。

ついにこういう対話ができるようになったのです。SF の世界でしかなかったコンピュータとの対話みたいなことが，この精度で本当にできる時代になっていると言うことです。

繰り返しになりますが『言語モデル』は，こういう対話をするように作られているわけではありません。学習しているわけではないんです。あくまでも世の中に転がっている文章で学習する。おそらく中にはこういう対話もあったんでしょう。そういうところから，こういう対話のフォーマットのようなものを体得しているらしい。だからこんなことができると言うわけです。



スライド 31

いったんここまでをまとめます。

そんな感じで，今，AI の研究開発ではパラダイムシフトが起きているところです。今までの『教師あり学習』から『自己教師学習』に大きくシフトしています。何より正解ラベルを与えるデータ量が少量で済むようになりました。

画像の場合もそうですが，もっとすごいのは『言語モデル』というものをここに据えた場合，言語のタスクの学習さえ必要なく，『言語モデル』に何を入れるか，『プロンプト』を工夫することでいろんな問題が分かるようになってきていることです。そして究極的には，先ほど示したような対話みたいなことまでできると言うことですね。

AI の研究者は，『汎用人工知能』なんて数年前ぐらいまではほとんどバカにしていたと言うか，「そんなのできるわけないじゃない」と考えていました。このセクションの最初に言ったように『Cognitive automation』と言いますか，数年前までの AI にできたのは人間が決めた入力と出力をただ再現するにすぎない，認知過程の自動化にすぎないと思ってた。ところが，巨大な『言語モデル』が何をするかを今見ていると，限りなく『汎用人工知能』に近いものにも見えてきている。それが現状です。

5　研究紹介

私のところでやっている研究をいくつかご紹介します。

私のところでは本当にいろんなことをやっています。工場で生産したものを画像で完成品検査するとか，そういう非常に現実的で具体的な問題もやっています。その一方で，もっと将来を見据えたもの，今お話したように本当に驚くようなことを『巨大言語モデル』ができることも踏まえ，もっと難しい問題に取り組む研究もあります。

スライド 32

例えばこういうプロジェクトの一環でやっています。これは国からお金が出ているプロジェクトで，「多様な環境に適応しインフラ構築を革新する協働 AI ロボット」という大きなプロジェクトです。私はその中の一員です。

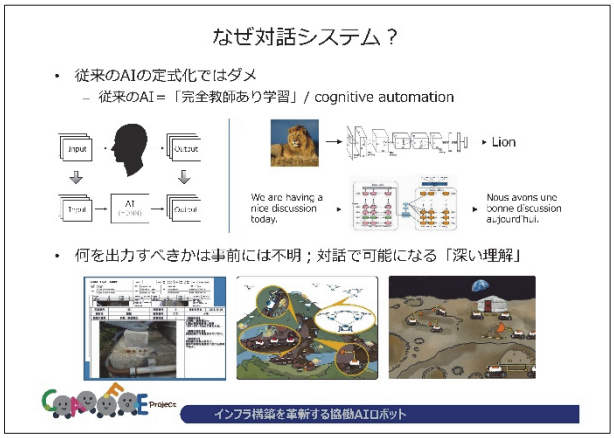
スライド 33

そのプロジェクトで私が何をしようとしているかと言うと，環境を理解する AI が作れないかと取り組んでいるところです。

キーポイントになるのは，「理解の中身がどういう風に表現されるか」，そして「それをどうやって取り出すか」です。先ほど『言語モデル』を使った対話の例をお見せしましたが，ここでは画像を含めた実世界も入ってきます。ニューラルネット AI に画像と対話のやり取りを入れて，ここに写っているものへの理解を引き出そうとしています。

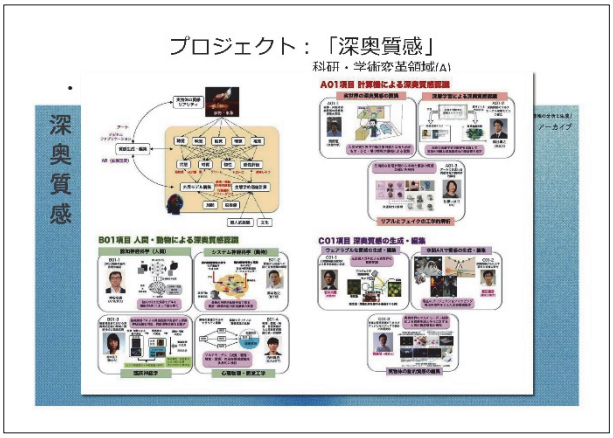
例えば次のようなアプリケーションが考えられます。今直近で取り組んでいるのは，こういうインフラ構造物の診断，健全度の理解といったものです。今は人が現場に行って，いろんな判定をして，こういう報告書を作っていますが，こうした画像をネットワークに入れて対話することで，ここが今どういう状況なのかを判定させたいと考えています。

このプロジェクトはいわゆるムーンショット型開発事業で，「2050 年といった将来までに月面で基地を自動構築するロボットを作る」，あるいはもう少し近いスパンだと「災害現場で救助や復旧にあたるロボットを作る」が全体のゴールです。そういうロボットをある程度自立的に動かそうとすると環境の理解が不可欠なため，そういう用途に使えないかという動機で研究しています。



スライド 34

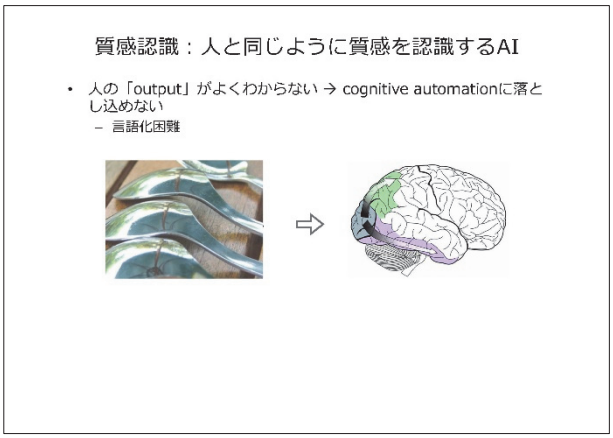
繰り返しになりますが，今説明したような現行のパラダイムシフトを含めると，従来型の AI の取り組み方，これを入れたらこれが出るというそういう瞬時的な反応ではダメなんですね。そもそも何を出力するかは，AI にとっては不明です。人あるいは他の AI エージェントと対応することで，初めて理解の中身が深まり，何を求められているのかも分かり，適切にレスポンスできるようになるということですね。



スライド 35

これはまた別の，もう少しアカデミックよりのプロジェクトで『深奥質感』の研究です。

これもグループで行っていて，私は AI を使って質感を考えることを担当しています。『質感』とは非常にとらえどころのない，幅広いものを表す言葉ですが，まさにその幅広さを研究しているというところです。人や動物が質感をどのように認識して使っているのか，質感を作り出すにはどうすればいいのかという問題に，画像をベースとした AI 的なアプローチで取り組んでいます。

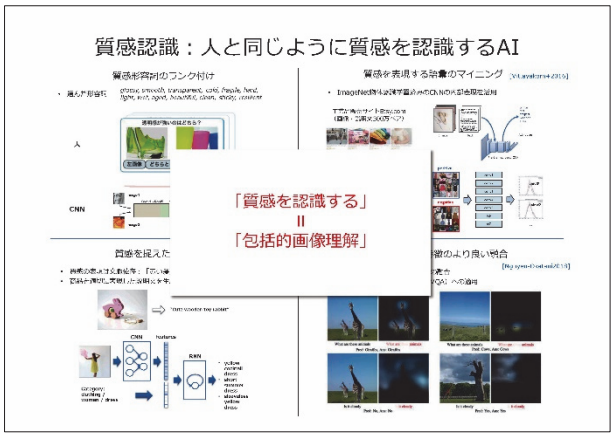


スライド 36

『深奥質感』も今の AI には難しい問題です。何が入力で何が出力かを指定でき，訓練データを用意できれば大抵のことはできるようになっています。しかし，そもそも人が出しているものが何なのか，何を出力しているのか分からないような問題があります。この『質感認識』はそういう良い事例だと思います。

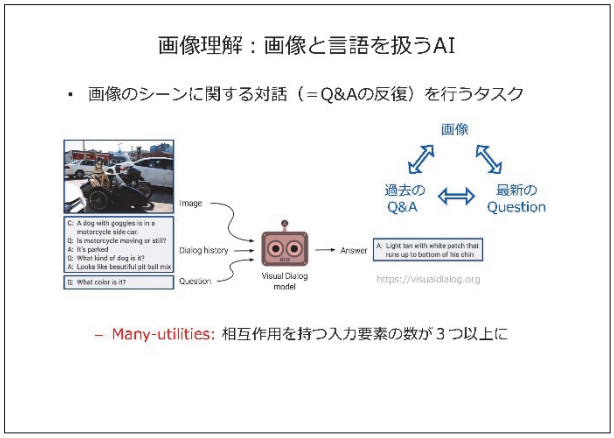
つまりこういう画像（左）を見せると「これはスプーンです」と答えるのが『物体認識』ですが，それだけではなくて「触ると冷たそう」「固そう」「なんか反射して写り込んでいる」など，いろいろなことがわれわれには分かります。そういうものが質感だとして，同様の認識をさせたいと思っています。少なくとも従来的なアプローチでは，人間

が何をアウトプットしているか分からないと深層学習の枠にはまりません。だから難しい。質感は言語化が難しいんですね。他者と自分が感じている，その質感の認知をどうやってコミュニケーションするか，それが難しいからこそ，いろんな問題も起こるわけです。



スライド 37

そういった性質を踏まえて，今までいろんな研究をしてきました。詳しくお話しする時間はありませんが，質感を認識できるようになるためには，画像に写っているものを AI が包括的，全面的に理解できるようにならないとダメだということが分かりました。



スライド 38

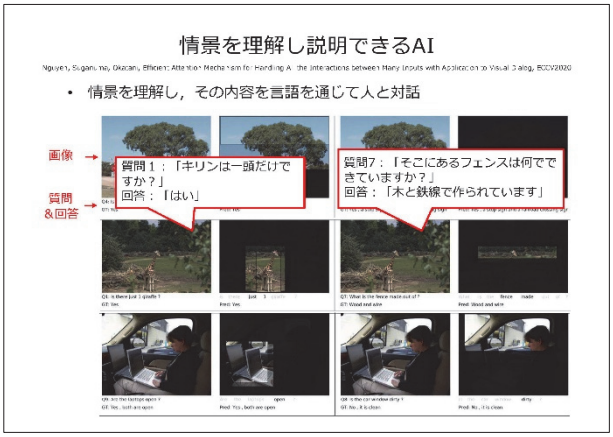
それを前提として研究をしています。先ほど言いましたように，例えば画像を AI に見せていろんな対話をすると，このシーンに関して起こっていることや，その中身に関する対応をします。

スライド 39

そういうことができるようになると，道が開けるのではないかと考えて研究しています。

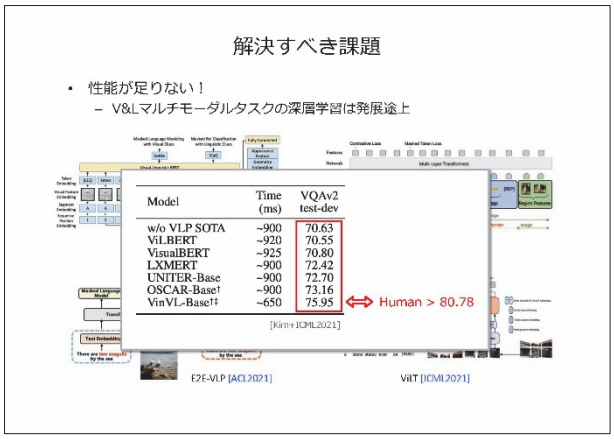
スライド 40

すでにこんな感じでの対応ができるようになっています。こういう画像を見せて，例えば「晴れていますか？」と訊くと，これは実際に AI がどこを見ているかを示しているんですが，空を見て「はい」と答える。何回かやり取りして，同じ画像に対して「何か標識が見えますか？」と訊くと，この標識のところに AI の目が向いて「一時停止と踏切の標識が見えます」という答えをもちろん英語なんですが，返しています。



スライド 41

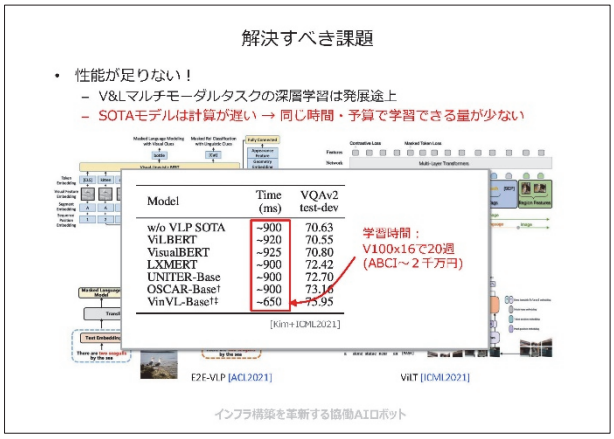
「キリンは１頭だけですか？」と訊くと，１頭しかいないので「はい」と答える。「そこにあるフェンスは何でできていますか？」と訊くと，「木と鉄線で作られています」と答えているという感じです。



スライド 42

これ位のことはもう今の技術でできます。ですが，いろいろ解決すべき問題もあります。かなり具体的な話になるのであまり触れられませんが，とにかく性能が足りていません。『言語モデル』はすごく進んでいますが，画像と自然言語の両方をうまく取り扱って，こういう対応をするのは，まだ難しいんですね。

研究はもう山ほどあります。冒頭に言いましたとおり，ニューラルネットワークの内部構造とか，問題をどういう風に定式化するかという部分で，いろんなアプローチで研究が行われています。ただ，これは質問応答のタスクで，画像に対する質問に正しく答えられるかという性能なんですが，最新のものでも人間には及んでいない。実は結構最近，人間レベルの奴も出てきましたが，それは置いておきます。まだ及んでいないということで，性能向上は必要だと言うことです。

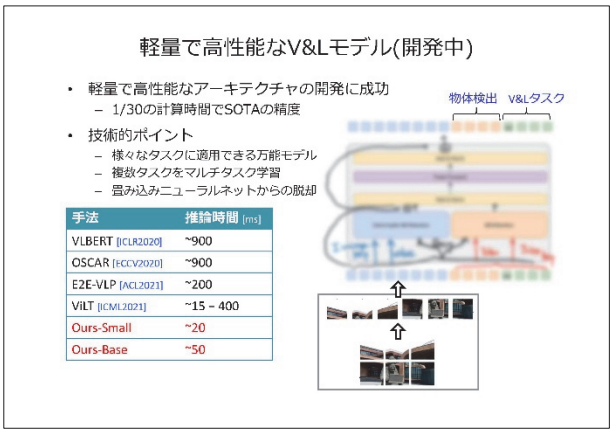


スライド 43

もう一つは，こういうネットワークは結構大きくならざるを得ず，一つの推論を出す，一つの質問に答える，あるいは対話を 1 ラウンドするのに，1 秒ぐらいかかってしまう。実際の対話では 1 秒ぐらい待たされても，そんなに問題じゃないかもしれません。しかし，学習のときには対話を何百万，何千万回と繰り返しますので，対話の 1 ラウンドにかかる時間が長いということはその学習自体がすごく大変になります。要はお金がかかるということになるんですね。

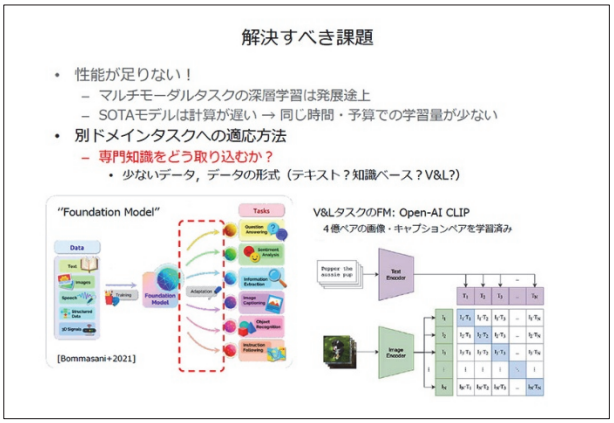
通常こういう研究をするときは，研究室で持っている計算機ももちろん使いますが，それだけでは足りない，間に合わないので，クラウドサービスを時間貸しで借りたりします。日本で 1 番コスパのいい GPU サーバーは，産総研（独立行政法人 産業技術総合研究所）の ABCI（AI Bridging Cloud Infrastructure／AI 橋渡しクラウド）です。しかしその ABCI でも，この論文に報告されている学習を再現すると 2000 万円ぐらいかかってしまう。と言うことで再現すらできないわけです。

実際に自分たちで作るときは試行錯誤をするので，10 倍ぐらいは計算をやりたい。そうなると，もううちの研究室ではできない研究になってしまいます。



スライド 44

そこでもっと軽くて高性能なモデルを作れないかということで，少し詳細をぼかしていますが，今作っている最中です。1/10 あるいは 1/30 ぐらいの計算時間で性能も出ます。1/30 になれば 2000 万が数百万くらいになるので，できなくもないと言う感じになります。そんなことをやっています。

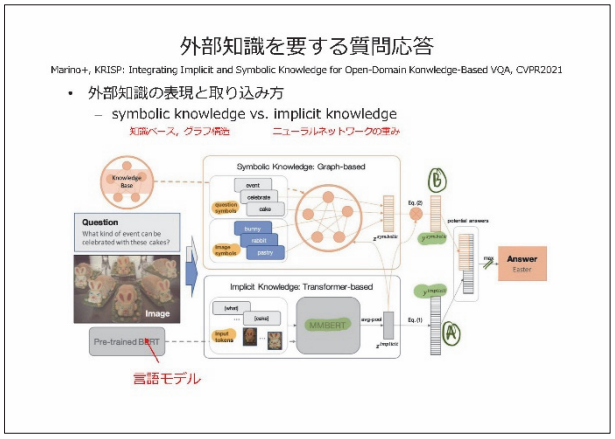


スライド 45

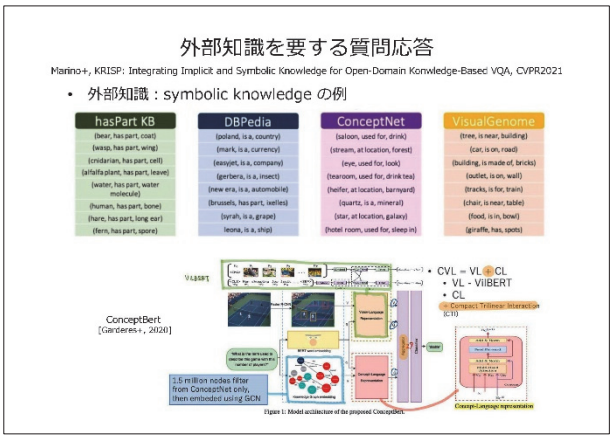
あともう一つ，今一連のプロジェクトで大きな問題として考えて取り組んでいるのが『専門知』の取り込み方です。どういうことかと言うと，先ほど（**スライド 33**）の，こういう橋と言うか，インフラ構造物の点検みたいな問題がまさにそうなんですが，点検はその専門家だからこそできるわけですね。だから専門家の知識が必要で，それをどう取り込むかと言う話です。

『言語モデル』などもそうなんですが，『専門知』がテキストとして表現されていれば，テキストを学習することで，ひょっとしたらうまくいくかもしれません。しかし，そもそも知識をどう表現するか，表現できたとしてどうやって推論に取り込むかというあたりが，まだまだ全然分かっていません。そのあたりも研究しています。

スライド 46

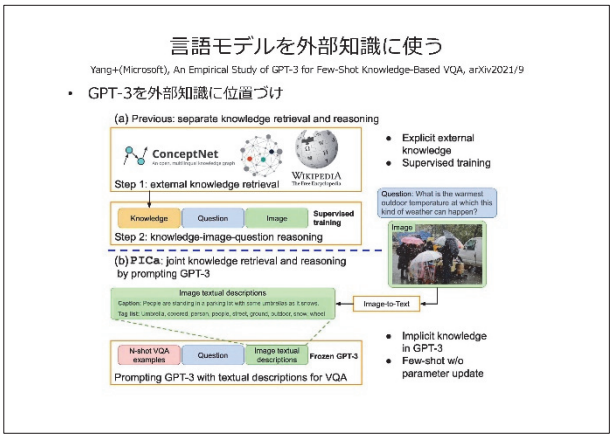


スライド 47



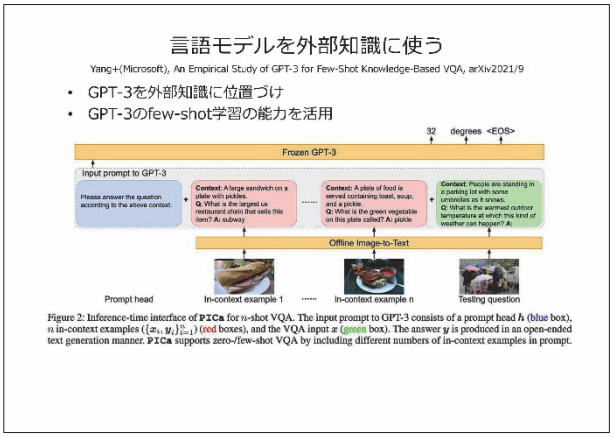
スライド 48

この辺は既存の研究や最新の研究で，どんなことをやっているかという話なので飛ばします。

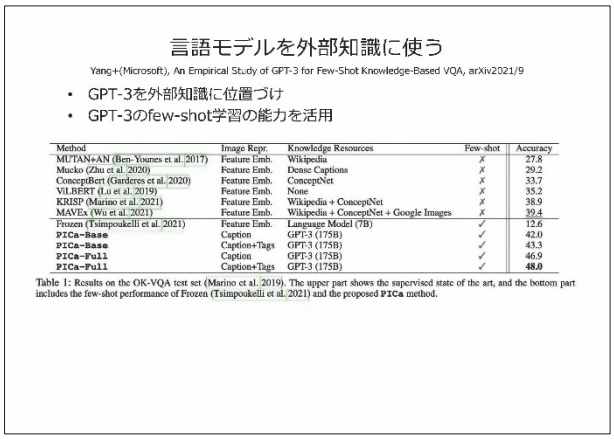


スライド 49

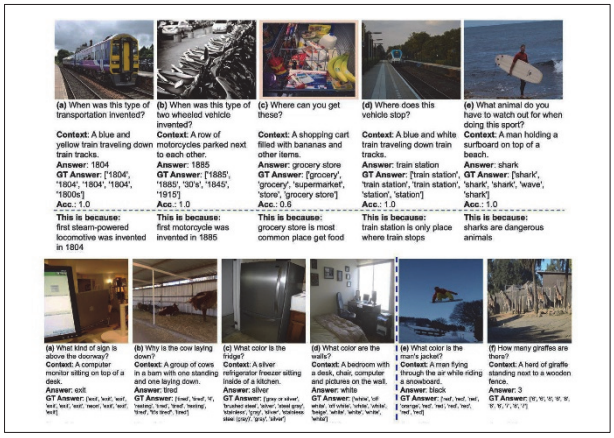
やはりここでも巨大『言語モデル』が有力視されつつあります。これは数か月前の研究なんですが，時間の都合で省略します。



スライド 50



スライド 51

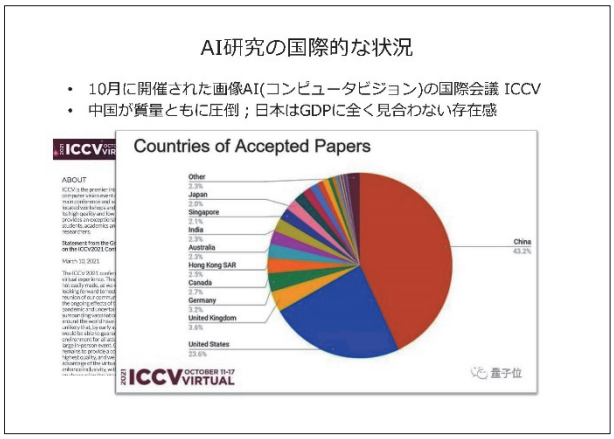


スライド 52

こんな感じで研究紹介とさせていただきます。

6 AI 研究の状況と人材について

最後，ご相談を受けたときにご依頼があったので，研究の現状をお話しします。状況とは何かと言うと，要するに AI 研究において日本が置かれている状況です。



スライド 53

一番近い例だと，今年（2021 年）の 10 月に開かれた『International Conferences on Computer Vision(ICCV)』という『computer vision』の国際会議です。バーチャルで開催されました。たぶんコンピュータサイエンスに造詣が深い方はご存じかもしれませんが，今，こういう国際会議の『Conference proceedings』と呼ばれるものが，論文発表の中心的な場所になっています。

一般のサイエンスやエンジニアリングの研究分野だと，論文誌（ジャーナル）が格上だと思われていますが，コンピュータサイエンスではコンファレンスの方が格上です。非常にセレクションが厳しく，10%台とか 20%台の論文採択率です。出しても 5 本に 1 本ぐらいしか通らないのが普通で，そんなところに皆たくさん投稿しているわけです。

特に AI の分野はここ 10 年で，『computer vision』分野だけでもたぶん参加者数が 10 倍以上に増え，ものすごい競争になっています。ICCV のような最も格が高いとされているトップコンファレンスに論文を 2 編 3 編と出した博士課程の学生は，高い給料で Google や META など，有名無名含めてさまざまな AI や『computer vision』を手がけている企業にリクルートされます。学生も含めて，みんな必死でそういうコンファレンスに論文を出すと言う状況が，まずあります。

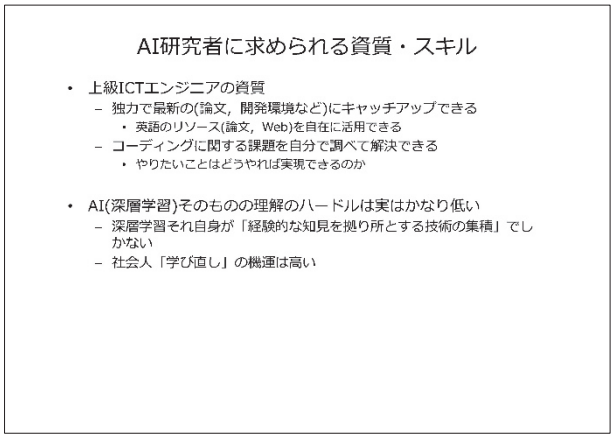
このグラフは，今回 ICCV で採択された論文数を分析したものです。論文の総数はうろ覚えですが数千本程度で，その内訳です。

ご覧のとおり，中国が圧倒的な存在感になっています。2 位はアメリカ。しかしおそらく発表した著者が所属している機関がアメリカにあるというだけで，著者の半分ぐらいは中国籍の方だろうと推測されます。ですので，この世界を事実上支配しているのは，中国系の研究者だと言っていいと思います。これは『computer vision』分野だけでなく，言語であれ何であれ，AI や他のコンピュータサービスも似たようなものだと思います。そういう風になっています。

何より残念なのが，日本がこの程度の存在感しかないことです。GDP 比で言うと第 3 位なので，少なくとも 3 番目に来てほしいんですが，決して来ないですね。ここ 10 年ずっとこんな感じです。むしろ 2 パーセントもあって良かったなという感じです。そんな感じなんです。理由はいろいろ山ほどありますが，ごく簡単に言うと，アカデミックな世界の話なので単純にお金をかけてないこと。それがそのまま表れているのだと思います。逆に言うと，中国にはものすごい数の研究者がいますし，大学もすごくお金をかけていることが，そのまま結果として上がっていると思います。

それはともかく，最初の方に紹介したように『言語モデル』『Foundation Model』を中心とした AI の進展はものすごくて，本当に SF の世界みたいなことが今起ころうとしていると言っていいと思います。そうした状況の中，もしもその国力みたいもので今後が

決まるのだとすると，非常に危ういと思います。



スライド 54

そういう研究分野にいて感じる研究者に求められる像みたいなものを少し思いつくままに書いてみました。特にうちの学生などをずっと見てきて思うことですね。

まず上級 ICT エンジニア，何て呼ぶのがいいか分かりませんが，単純なプログラマーではない，ソフトウェア工学の非常に高いレベルで仕事ができる人。そうした人が持つべき資質は，どんどん進んでいく最新状況に自分 1 人でキャッチアップできることです。もちろん英語などは自在に読めて論文やウェブを自由に活用し，いつも最新の知識にアップデートしていることが求められます。

コーディングももちろん AI では大事ですが，むちゃくちゃ難易度の高いコードを書くと言うよりは，どんどんアップデートされていく開発環境なり，新しい手法なりを具現化する能力が高くないと実験ができません。そういう能力も必要で，これらを含めたエンジニアの質が重要です。

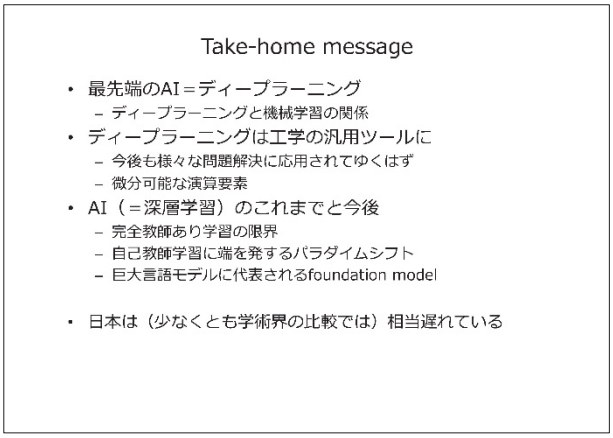
ただ，一方では AI と言いますか『ディープラーニング』そのものの理解のハードルは，実は全然難しくないとも言えます。なぜかと言うと『ディープラーニング』は決してロケット・サイエンスではないんですね。ロケットを打ち上げるためには，ボトムアップで勉強しなきゃいけない工学専門書や教科書が山ほどあります。しかし，深層学習にはほとんどそういうものがない。要は理論がないんです。限りなく経験的な知見を拠り所とする技術の集積でしかない。もちろんその量はどんどん増えていくので，押さえるべき知識の量も増えるとは言え，量子力学を勉強するのに比べると敷居などないに等しいと言ってもいい。そういうものなんです。だから，要はやってみるというだけなんですね。ただし，先ほど言ったようなことができる人でないと，なかなか進めないと思います。

当然ですが，こういう新しい分野で求められる人材はどんどん増えていく。しかし，特に日本はそうですが，この分野を修めた学生を世の中から求められているだけ大学から送り出せているかと言うと，決してそんなことはありません。そこで，社会人の学び直しという話になるわけです。

スライド 55

社会人の学び直しは以前からありますが，AI の分野ではここ 10 年，オンライン講座が盛んになっています。私も実はこの zero to one という日本のスタートアップ企業と関係がありまして，これはちょっと利害関係ありありの紹介なんですが，そこでいろんなことを教えてもらいましたので，アメリカでは非常に進んでいるという話をここでお伝えしたいと思います。

有名どころでは Coursera や Udacity などがあります。例えば自動運転を勉強するコースなどは 50 万円ほどの授業料を取るようですが，もちろん個人で出すわけではなくて会社が出します。そうしたエンジニアの再教育が，非常に盛んに行われています。Coursera は基本的にタダですが，単位認定，修了認定書みたいなものに対してお金を取る形になっています。いずれにしても，こういうオンラインコースではあっても，ちゃんと最後まで修了すると一つの職の証明，能力の証明になり，それを転職のときに活用することがアメリカではかなりスタンダードになっているということです。日本ではまだそこまではいってないと思いますが，それでもこの zero to one も含めて，同じような取り組みをしている会社が数社あります。



スライド 56

少しまとめます。Take-home message というほど大げさではありませんが。

『AI』は，本当に非常にルースに使われている言葉だと思います。単なるデータサイエンスみたいなものや，エクセルでできる範囲のものも AI と呼ばれている感じがあります。でも，「『最先端の AI』＝『ディープラーニング』」なんです。『ディープラーニング』自体は，『機械学習』という，より広い範囲の一つの分野と言うか 1 手法でしかありません。

そういう意味でも一つの分野とくくるには無理があると言いますか，冒頭にも言ったとおり，従来的な方法論とあまりにも違います。『機械学習』は，伝統的にはやはり数学が非常に重視される分野であり，数学的な理論に基づく手法がいくつも作られ，使われているわけです。

一方『ディープラーニング』は，そういう従来的な価値観とはかなり異なっています。あくまでも実践の工学なんですね。経験値みたいなものでネットワークの構造を試行錯誤で作って，力技で問題を解決していくようなところがあります。そういう意味では非常に問題も多いツールですが，なぜか『ディープラーニング』でないと出せない性能があるので，使わざるを得ないというわけです。

そういう意味でも工学の汎用ツールになっていくと思っています。AI の一つの方法というだけではなくて，汎用的な工学のツールになるはずであり，もうすでになりつつある。だから工学部の教育も少し考え直す必要があるぐらいの話だと，私などは思っています。

今日はちゃんと説明できていませんが，そこで使うのは『ニューラルネットワーク』という従来的なイメージとはかけ離れたものです。ニューラルネットと言うとなんか生物の神経回路のようなものをイメージしがちですが，もっと作り込まれた専用の計算をするシステムです。ただし出力を入力で微分できる微分可能性を持っている演算の要素であれば，『ニューラルネットワーク』と組み合わせて一括で学習に使えます。ですので今は，そういう性質を持ったものをどんどん取り込んでいったハイブリッド的なネットワークをいろんな問題解決に使うようになっています。

「AI のこれまでと今後」という話も最初にしましたが，今までの『完全教師あり学習』はある部分では成功しましたが限界も多く，それをうち破るために『自己教師学習』が今大きなパラダイムシフトとしてあり，中でも巨大『言語モデル』がちょっと末恐ろしい感じの発展を見せているという現状があります。

しかし，そんな状況にもかかわらず，日本は少なくとも学術界で比べると相当遅れている。残念ながら，そう言えてしまいます。

これで，まとめとしたいと思います。以上です。ありがとうございました。

講演者紹介

岡谷　貴之（おかたに　たかゆき）

1999 年　　東京大学計数工学専攻博士課程修了
1999 年～　東北大学大学院情報科学研究科，
2013 年から現職
2016 年～　理化学研究所革新知能統合研究センター兼務