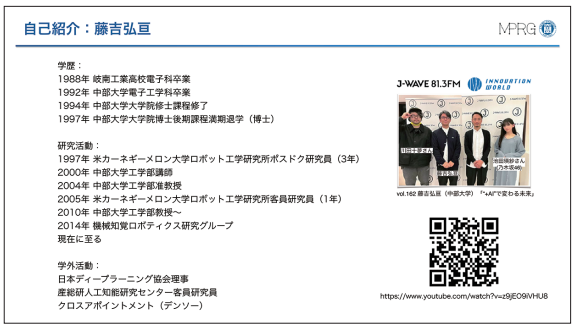
第46回測量調査技術発表会
大規模基盤モデルによる第4次AIブームの到来

中部大学理工学部 AIロボティクス学科 教授　藤吉　弘亘

自己紹介と講演概要

皆さん、こんにちは。中部大学の藤吉と申します。今日はAIがどうやってできているのか、かつAIが今どのように使われようとしているかをご紹介したいと思いますのでよろしくお願いします。



まず自己紹介ですが、先ほどご紹介いただいたことに付け加えますと、社会貢献の一環として、AIを支える技術であるディープラーニングをより多くの方に正しく知っていただき、正しく使っていただくことを啓発している日本ディープラーニング協会という団体で、ディープラーニングに関するG検定という検定と、E資格という資格認定をやっています。

そしてもう一つ、今年の1月にここ関東にありますJ-WAVEというFM局の「INNOVATION WORLD」という番組に呼んでいただきまして、「“+AI”で変わる未来」というタイトルで、毎回10分ずつ4回にわたってAIの講義をしました。このときは何と、あの乃木坂46の池田瑛紗さんに向かって講義をするという、非常に緊張する場面だったにも拘わらずスライドを使うことができなくて大変だったのですが、今日はスライド使えるので、もう少しちゃんとできるかなと思っています。

この時の番組の収録した内容が、こちらのYouTubeのところで公開されておりますので、今日もその一部を話しますが、さらに他のことも話していますので、もしご興味あれば聴いていただければと思います。（スライド内のQRコードを参照）

さて、今日は

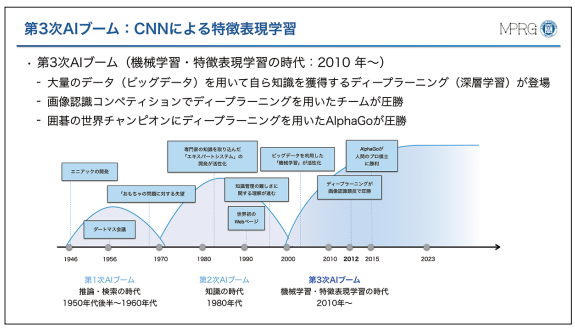
1. 第4次A Iブームの到来
2. 大規模言語モデル（LLM）と活用
3. 画像言語モデル（VLM）

の3つの項目についてお話しします。現在、第4次AIブームが起きておりまして、この第3次から第4次にどういったことができるようになってきたかということを話したいと思います。

その後に、これも避けて通れない、第4次AIブームである生成AIにおいて用いられる大規模言語モデル、よくLLMと呼ばれるものですが、これがどうやってできているのかを簡単に紹介しながら、どうやって活用すべきかとについてもお話できればと思います。

そして最後に、LLMだけではなくて、画像も一緒に扱うような、いわゆる画像言語モデル、ビジュアル・ランゲージ・モデル（VLM）というモデルがありまして、これによってさらにどんなことができるようになっていくのかについても時間のある限りお話ししたいと思います。

1. 第4次AIブームの到来



ではまず、これまでのAIを少し振り返ってみたいと思います。最初は第1次AIブームで、1956年にダートマス会議にてAIという言葉が出てきて、AIの定義がなされました。残念ながら当時のAIは盛り下がってしまったのですが、その後1980年代に第2次AIブームが来ました。こちらはご存じの方もいるかもしれませんが、

エキスパートシステムという、専門家の知識を取り込んだ仕組みです。

これは、色々な事柄、例えば医学における診断という観点で、こういう条件の時にはこうなるといった、ルールを専門家がどんどん入力して判定するものになります。当然、多くの知識をルールベースで書き出さないといけないのですが、これを全て人がやるので大変だということと、ルールで書いたこと以外はほとんどうまく動かないということで、こちらも残念ながらAIブームが盛り下がってしまいました。

そして、現在につながる第3次AIブームというのが2012年にやってきました。この2012年に何ができたかといいますと、ディープラーニングという技術であり、例えば画像認識だとか音声認識だとか言語翻訳など、色々な多岐にわたる分野でディープラーニングにより認識性能が一気に飛躍的に向上し、実利用できるようになってきました。そこからずっとつながって、この第3次のAIブームが、盛り下がることなく現在に至っていると考えています。

この第3次AIブームは、機械学習、正確にはディープラーニングによる特徴表現学習の時代といわれています。大量のデータ、いわゆるビッグデータを用いて自ら知識を獲得する技術、ディープラーニングが登場したことによって、画像認識の性能が向上したり、囲碁の世界チャンピオンにディープラーニングを用いたAlphaGoという人工知能システムが圧勝したりといったことがおこりました。

簡単にこのAlphaGoを紹介します。AlphaGoはこの会場にもご存じの方が多いと思いますが、囲碁を打つAIです。いきなり囲碁だと、どれだけ難しいかが分かりづらいので、最初はオセロから入りましょう。

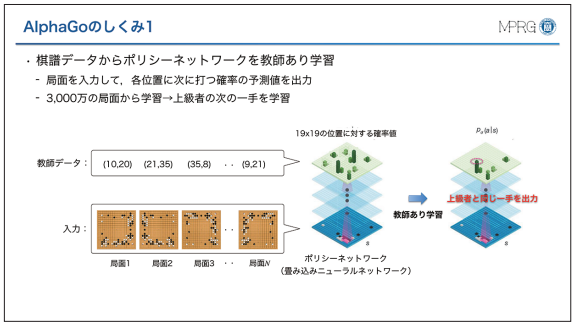
オセロは8×8マスですよね。8×8マスの中に今置かれているパターンに対して、次に白もしくは黒を置く場所というのはかなり限定されますね。なので、その限定された中を全てしらみつぶしに、ここに置いたらどうなるのかを先の先まで計算機で探索していけば、より良い一手を見つけることができます。なのでオセロを解くゲームAIは、賢いというよりは、計算機を使って力任せに全手調べて一番いい手を選択しています。

もう一つ。同じ8×8マスのボードゲームにチェスがあります。チェスも8×8マスなので、オセロと同様に、探索をしらみつぶしにやる全探索で解ける問題でした。実際1997年に、チェス用の探索計算をするための専用ハードウエアであるDeep Blueというシステムが作られ、当時の世界チャンピオンにも勝利しました。

しかし、囲碁や将棋はずっと難しくて実現できていませんでした。なぜかというと、まず将棋を考えてみましょう。将棋のマス目は9×9です。オセロは8×8=64マスで、しかも限定されたとこに置くだけだったのが、将棋では9×9=81マスに増え、かつ歩とか色々な駒がありますから、それによって動き方の自由度が高いですよね。そうすると先ほどのように、この手を打って、次、この手を打ってという探索をする空間が爆発的に増えてしまって、計算機が速くても全然解けないという状態になってしまうわけです。

さらに難しいのが、囲碁です。囲碁のマス目はさらに大きくて19×19なんですね。白と黒の碁石があって、それを19×19マスのどこに置いてもいいんです。ということは2の19×19乗みたいな組み合わせとなり、探索空間が天文学的な数字になってしまいます。そのため、高性能な計算機の力を使っても解くことができませんでした。

では、そこにAlphaGoが登場して、どういう仕組で世界チャンピオンに勝てるようになったかを紹介したいと思います。AlphaGoはDeepMindというGoogleの子会社が2016年に開発したものです。



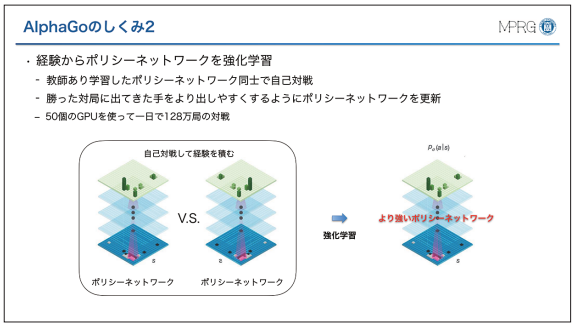
AlphaGoは、ディープラーニングの一種である「畳み込みニューラルネットワーク」というモデルを学習さ

せます。何を学習するかというと、これまでに人が、特に囲碁の上級者がインターネット上で対戦した棋譜データを収集します。棋譜データは、例えば局面1の時、次にどこに打ったかがマス目の位置で分かるデータです。なので、このディープラーニングの入力には、この盤面を画像のように2次元パターンとして入力し、そしてディープラーニングはマス目の各位置に打つ確率を計算して出力します。そして、確率が一番高いマス目と上級者が打ったマス目が一致するように学習します。

この局面1を入力して出力を計算し、ここ（教師データ）の座標の確率が一番高くなるようにしたいわけです。しかし学習の初期はうまく出力できないので、正解との誤差を求め、その誤差を小さくなるように、畳み込みニューラルネットワークのモデルパラメータを更新します。次は、局面2を入力して出力を計算し、再度誤差を求めて最小化するという学習を何度も何度も繰り返していくのです。

さらにこの学習に重要なのがデータ数です。AlphaGoでは3,000万局のデータを学習します。しかもこの3,000万局は上級者が打った棋譜データです。このような学習データを使って、「教師あり学習」を何度も何度も繰り返していくと、このモデルは上級者と同じような一手が打てるようになるわけです。面白いですよね。これまで人間が打ってきた囲碁における知恵を、データから学習をしているともいえます。

これで上級者と同じような一手が打てますから、当然私みたいな一般人よりもむちゃくちゃ強い状態になっていますが、残念ながら上級者と同じような一手が打てる能力では世界チャンピオンには勝てませんよね。上級者と同じような一手であれば五分五分にしかならないわけです。



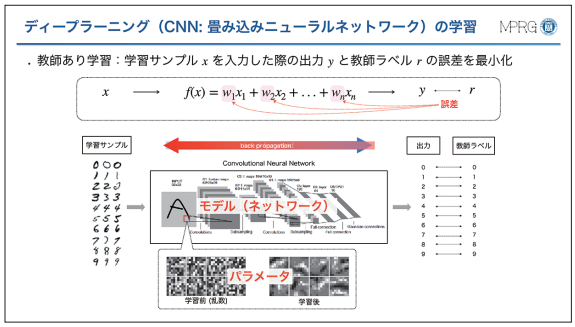
そこで次にどうしているかというと、先ほどの上級者と同じような一手が打てるディープラーニングのコピーを作って、何度も対戦させます。対戦させた時の経験から、新たにより良い手を経験から学ぶことでさらに強くなっていくわけです。

この経験から学ぶ学習を強化学習と呼びます。実際どれぐらいの経験を積んで学習するかというと、50台のGPUという計算機を使って1日で128万局対戦します。われわれ人間ならおそらく1日に囲碁が打てるのは2局か3局です。365日毎日、かつ10年やっても、全然到達しない数であることが分かります。すなわち、われわれ人間が一生かかっても経験できないような数の経験を、このAIは1日で積んで、そしてそこから、新たなより良い一手を学んでしまったというわけです。

このように強化学習したモデルを使って、一番強いとされる囲碁の名人と対戦したら、AIが勝ってしまったというわけです。これがAlphaGoの仕組みです。

この仕組みは大変面白いと思うのですが、これ（スライド4の中央、位置に対する確率値を示している図）は何をやっているかというと、パターン認識であり、今の盤面に対して直感的にここに打てばいいという結果を出力していることになります。

将棋棋士の羽生さんが将棋を指す際に、将棋盤のどこを注視しているかを計測したデータがあります。われわれ、すみません一緒にして申し訳ないのですが、われわれ将棋の素人は、ここを動かしたらどうなるかと、行ったり来たりしながら色々なところを見ているんですよね。一方、羽生さんは全体をぱっと見てからは、ある範囲のみを注視しているんです。ということは、直感的に指しているわけです。もちろん、今までの経験に基づいてですけども、直感的に打つ。一方、素人のわれわれはどうしても、ここ指したらどうなるか、いやこっちに指すとどうなるかということを一生懸命考えたりするわけです。そういう意味では、この人工知能の仕組みも非常に直感的な指し方だと言えると思います。



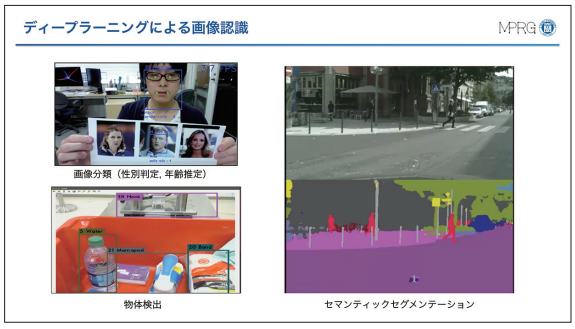
このようにしてAlphaGoという囲碁を打つAIシステムができたのです。そこで使われている技術がディー

プラーニングです。ディープラーニングにも色々なモデルがあり、AlphaGoでは入力として2次元パターンである画像を扱う、畳み込みニューラルネットワークが使われていました。

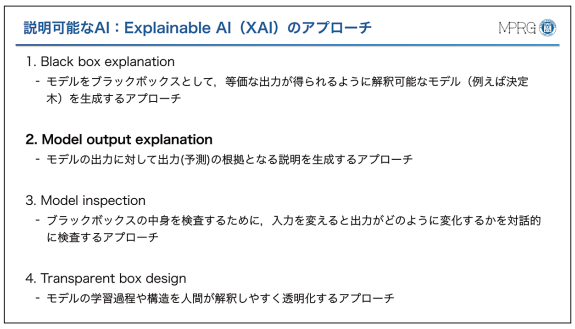
ごく簡単にですが、この畳み込みニューラルネットワークの学習がどういうものかを紹介しますと、画像（スライド左端の数字）を大量に学習データとして用意します。そしてそれぞれが0、1、2、3という数字であるという正解ラベルである教師信号（スライド右端の数字）を用意しておきます。

ニューラルネットワークの処理を単純化して式で表すとこうなります。学習サンプルxを入力すると、xの要素であるx1、x2、…xnに対して、重み係数であるw1、w2、…wnがあり、その乗算により出力yを計算します。このw1からwnという重み係数であるモデルパラメータを学習によって更新します。重み係数の初期値は乱数からスタートします。この時、この入力画像の出力を計算すると、当然でたらめな答えとなります。そこで、入力画像の正解との誤差を計算して、その誤差を最小化するように、重み係数wを更新します。

これを何度も何度も繰り返していくと、正解を出力することができるようになって、この重み係数wの値は、この画像から数字を認識しやすいフィルターとなり、画像認識に有効な特徴抽出を学習によって獲得できたということになります。これをネットワーク、もしくはモデルと呼び、この重み係数wをモデルパラメータと呼びます。



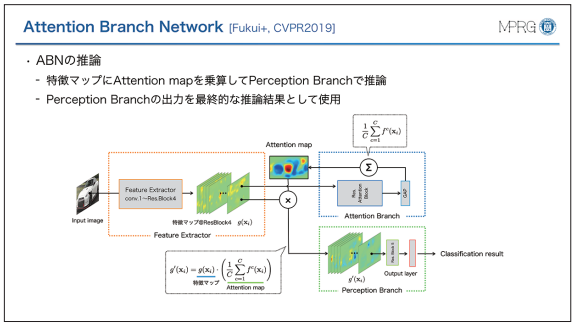
CNN（畳み込みニューラルネットワーク）によって、様々な画像認識タスクを解くことができるようになりました。これはわれわれの研究成果の例であり、性別を判定しながらも、笑顔度を判定することもできますし、こちら（左下）のように、色々な物体がどこにあるかを検出する物体検出も実現できます。さらに、セマンティックセグメンテーションといい、道路領域、歩道、人、車といったように、画像の画素単位でどのクラスに属しているかといったことを解くことで、画像の意味的なセグメンテーションもできるようになっています。



このCNNにより、大量のデータを学習することで、認識性能を飛躍的に伸ばしてきました。しかし、CNNをはじめとするディープラーニングを実応用する時の一つ問題点が、ブラックボックスという話です。

ディープラーニングのモデルは、パラメータ数が多いので、中で何が起きているのかということを把握するのが非常に難しいモデルとなっています。そこで、できるだけ説明できるようなAIを実現しようといった研究も行われています。このような取り組みを英語でExplainable AI、略してXAIと呼びます。

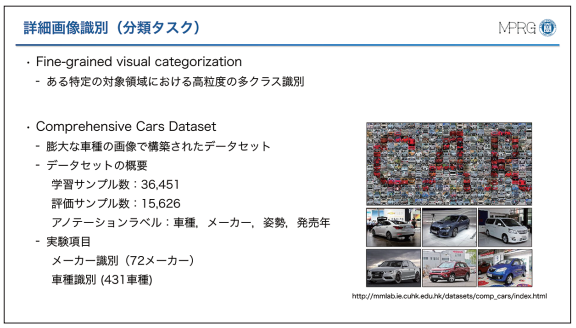
このXAIの研究も色々なアプローチがあるのですが、今日は一つ、この2番目のModel Output Explanationというアプローチについて、われわれの研究を紹介しながら説明したいと思います。



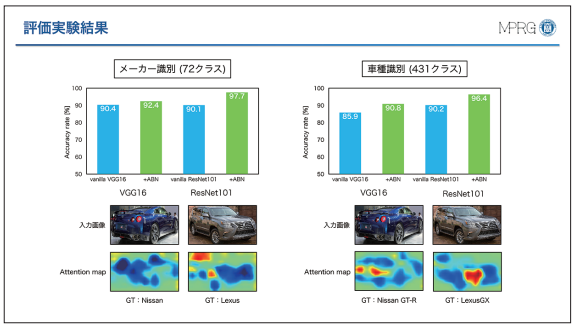
われわれは、畳み込みニューラルネットワークを一部改良して、アテンション・ブランチ・ネットワークという手法を考えました。画像入力すると、色々な畳み込み処理をして特徴（特徴マップ；スライド左側のオレンジ色の矩形）を抽出します。今までは、この特徴から認識処理をしていたのですが、われわれはそこにアテンションブランチ（スライド右上の青い矩形）を導入しました。これは、入力画像に対して、空間的にどこに注目すればいいかというアテンションマップ（スライド

中央の赤いマークのある画像）を出力します。通常、このアテンションマップは、判断結果に対してどこに注目したかという説明のために使われるのですが、われわれはこれを説明性だけではなく識別処理にも利用しました。

このアテンションマップは、空間的にどこに注目するといいよということを教えてくれているわけですから、特徴マップに重み付けします。そして重み付けした特徴マップを用いて識別処理を行い、最終的な結果を出力するような仕組みを考えました。

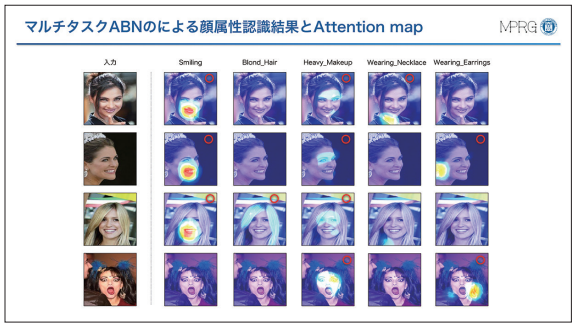


これは詳細画像識別というタスクです。車画像の大量のデータセットの中に、メーカー数は72あるので、72メーカーに分類するというタスクとか、車種が431車種あるので、その車種を判定するというようなタスクを、詳細画像識別といいます。

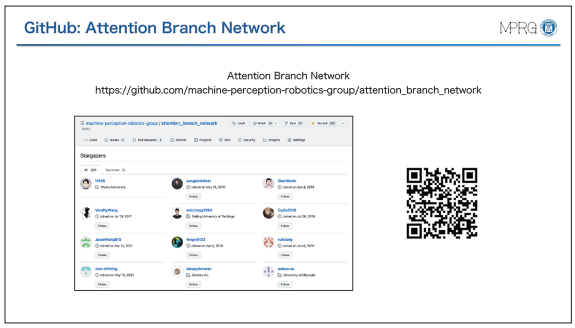


このタスクにおいて、われわれのアテンションブランチを、通常のVGG（Visual Geometry Group）というモデルに追加すると、このように精度が向上し、ResNetというモデルにも、アテンションブランチを追加すると精度が向上しました。メーカー分類だけでなく、車種の判定においてもかなり効果がありました。

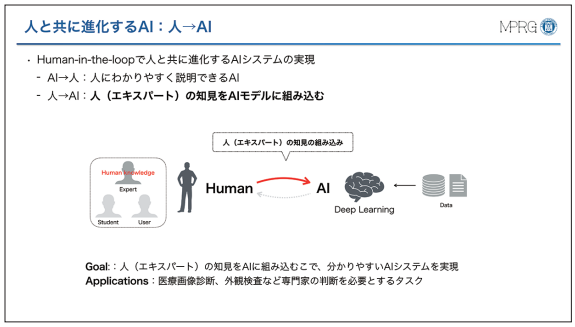
車種を判定した際のアテンションマップをちょっと見てみましょう。この画像は日産のGT-Rという有名な車です。GT-Rと判定するための特徴といえば、この丸いテールランプがかなり特徴的ですよね。アテンションマップにより、このAIモデルはテールランプのところに注目してGT-Rと判断したということが、（右側の車種識別の例の下段左側の図、テールランプに相当する部分が赤くなっていることにより）判断根拠の説明として読み取れるわけです。さらに、このアテンションマップを使って重み付けして識別処理することで、精度向上もできました。



こちらは、これらの画像（一番左の列）に対して笑顔と判定した際にどこに注目しているかを示すアテンションマップです。笑顔の判断基準がどこにあるかというと、やはり口だということが、このように（左から2列目、Smilingの上から3人までの女性の口の部分が赤くなっていることにより）分かります。アテンションマップによる説明性を高めながら、それを識別の精度向上にも繋げるアプローチに取り組んできました。



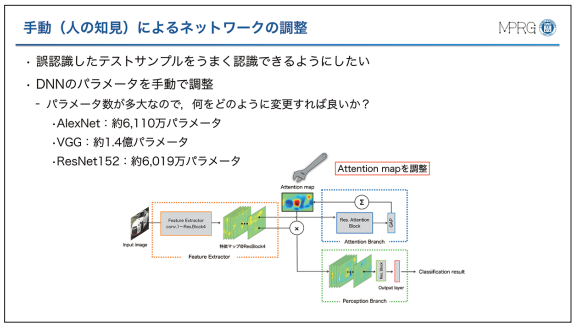
このアテンション・ブランチ・ネットワークは、GitHubというサイトでソースコードを公開していますので、ご興味ある方は、こちら（スライド内のQRコード）からアクセスしていただければ幸いです。



AIを使用するユーザである人に対して、ディープラーニングを分かりやすく説明できるようにしようという研究とともに、逆も考えられるのではないかと思っています。より良いAIを作るために、われわれ、人も何らかAIに介入して良くすることができるのではという考え方です。

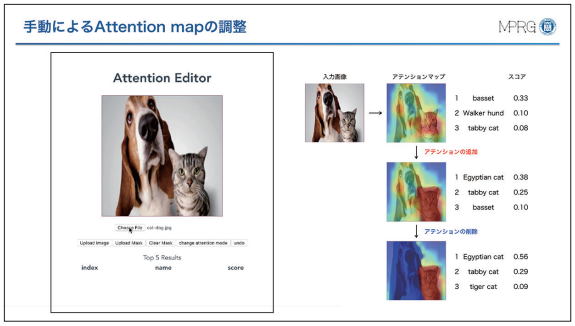
データドリブンで学習したディープラーニングはより良い性能を発揮しますが、学習データが不完全なデータであれば、当然、学習したAIも良いものができません。ではそういった時、そのAIモデルをより良くするにはどうすれば良いでしょうか？おそらく一番の解決策は大量にデータを集めて完全なデータを作るということになります。しかし、データを集めるといっても簡単ではありませんし、教師あり学習なので正解をアノテーションしないといけませんから大変です。

そこで、そのタスクをこれまで人がやっていたのであれば、その専門知識を持った「人」の知見をこのAIモデルに組み込むことができれば、データと人のハイブリッドで、解釈性の高い、より良いAIモデルができるのではという考え方です。

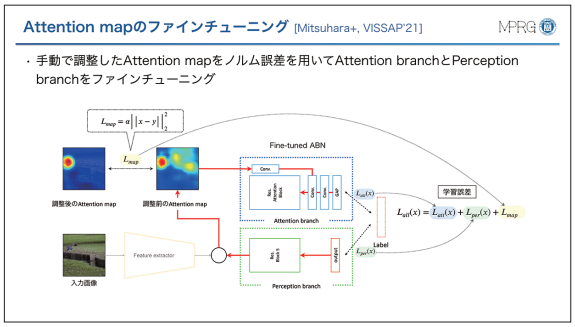


最近のディープラーニングのモデルは、パラメータ数がどんどん多くなっています。例えばAlexNetというモデルは6,110万のパラメータ数がありますし、VGGというモデルに至っては1.4億のパラメータ数です。このようなパラメータ数が多いモデルに、人の知見を導入することは、非常に難しい問題です。この膨大なパラメータに手作業で反映させることは、かなり難しいということがお分かりになると思います。

しかし、われわれのモデルには、ちょうどここ（図中央の中央が赤くなっている矩形）にアテンションマップがありました。このアテンションマップは、われわれ人間にとって視覚的に分かりやすいものです。そこで、このアテンションマップを人の知見に従って調整し、調整したアテンションマップを用いて推論すればいいのではないかということです。

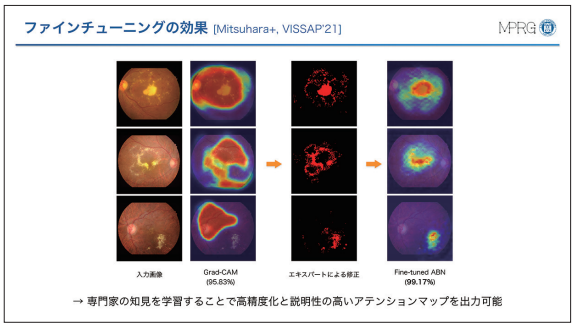


イメージはこんな感じです。この画像を入力して推論をすると、1位がバセット（犬の種類）という結果が出てきました。でもここでは犬を扱うタスクではなく、猫を扱うタスクだったとします。その場合どうしたらいいかというと、このように（右側真ん中の画像）赤い色で、ここに注目するようにアテンションマップを修正して推論すると、ちゃんと1位がエジプシャンキャット(猫の種類)に結果が変化します。次に、青色で塗ります(右側下の画像)。青色は消しゴムのイメージで、ここは注目しないようにアテンションマップを修正して推論をすると、結果は1位から3位まで全て猫の種類の結果となります。



アテンションマップを人の知見に基づいて手動で修正して推論すると、われわれの意図どおりに推論結果を調整することできるというわけです。であれば、この注目領域をモデルのパラメータに、ファインチューニング（再学習）して埋め込めばいい訳です。

データから学習したAIモデルの出力は、間違ったところに注目して、間違った答えを出力しました。その際に、入力画像に対して、エキスパートのわれわれ人間がどこに注目するといいよというヒントを与え、そのヒントにアテンションマップを近づけるようにファインチューニングします。そうすると、ヒントである、人の知見をこのモデルに組み込むことができるわけです。



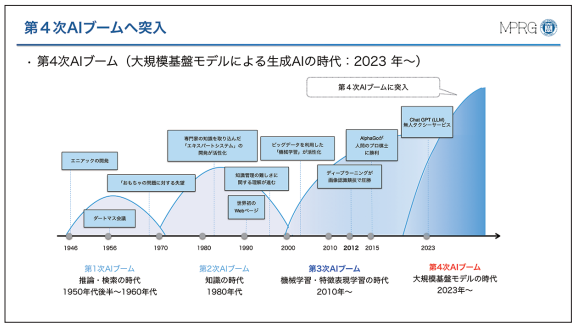
その一例をちょっと紹介します。これは眼底画像の疾患判定です。正常と疾患というラベルを用いて教師あり学習を行います。この3つ（一番左側の3画像）は疾患画像です。通常の畳み込みニューラルネットワークで学習すると、大体96%ぐらい精度が出ました。しかし、疾患と判断した根拠を可視化すると、こういった領域（左から2列目の画像の赤い部分）に注目して疾患と判断したというわけです。

例えば、お医者さんに行って、説明なしに「疾患です」って言われることありませんよね。結果だけでなく、説明できることは重要です。このAIモデルに説明させると、何と正常領域に注目して、疾患だって判定しているわけです。答えは合っているのですが、注目している領域が説明として合っていません。皆さんは、このようなAIモデルの診断を受けたいと思いますか？もちろん、嫌ですよね。こういった時に、先ほどのアプローチでAIモデルをより良くすることができるわけです。

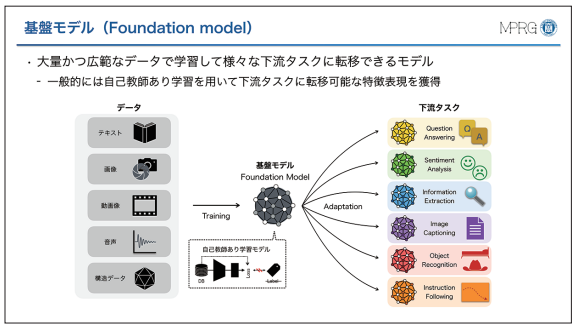
これらの疾患の眼底画像を専門家である眼科の先生に見てもらい、どこに注目すべきかというヒントをもらいます（左から3列目の画像）。そしてこのヒントを使って、AIモデルに組み込むと、このようなアテンションマップ（一番右の画像）になり、眼科の先生と同じ領域に注目して判断するAIモデルにすることができます。また、認識性能も99.1%に高精度化することができました。

このように専門家の知見を学習して、高精度化と説明性の高いAIモデルを作ることができます。データのみから学習することもいいんですが、データが不完全の場合は、われわれ人が関わることも重要だということになります。

すみません。少し長く話しすぎてしまって、第4次と言いながら、今までのところはまだ第3次AIブームでのお話しでした。この第3次AIブームがブームとして終わることなく、2023年ぐらいから第4次AIブームに



突入しました。この第4次AIブームは、基盤モデルによるいわゆる生成AIの時代と呼ばれています。



基盤モデルとは、英語ではfoundation modelと呼んでいて、このように大量かつ広範なデータ、例えば画像だけでなくテキストや動画像、音声など、色々なデータで学習しておくモデルです。これ自体で何かを直接解くわけではなく、色々な下流タスクに転移できるような基盤となるモデルとなります。

これまでは例えば囲碁だけに特化したAIモデルを作ってたわけですが、この基盤モデル自体はどれかに特化することなく汎用性の高い状態にしておいて、それを色々なタスクに転移すると、かなりいいものができるようになってきたというわけです。

この基盤モデルを使った2つの例を紹介します。一つは自動運転です。私は人工知能技術を適用した自動運転の研究にも取り組んでおります。自動運転は、研究レベルでまだ実現できていないことまだあるのですが、一方で、一般の方が使えるところにも来ています。

それを少しだけ紹介します。

無人ロボタクシーというものです。日本だと、地方自治体等で実証実験を行われていますが、アメリカのフェニックスやサンフランシスコ、ロスアンゼルスでは、もう普通にこの無人ロボタクシーが一般市民の方誰でも使用できる状態になっています。

これはちょっと乗ってみないといけないだろうということで、昨年の3月にアメリカのフェニックスまで行って体験してきました。どんなものかといいますと、まずUberのアプリと同じようなWaymoというアプリを入れ、クレジットカードを登録して、行き先を入力します。そうすると、実際こんな感じで自動運転車がやってくるわけです。（会場では、無人ロボタクシーが来て、実際に乗車する動画を放映）

これがまさに私が呼んだ無人ロボタクシーです。運転席も含め中には誰も乗っていない状態で私の目の前まで来て止まってくれます。そして、鍵は自動で開くのですが、ドアは自動で開きませんので、自分で開けて中に乗ります。そしてアプリ上でスタートボタン、もしくはここ（助手席前）の画面のところの「スタートライド」というボタンを押すと、「10分後に目的地に着きます」と言ってくれて、このように動き出します。

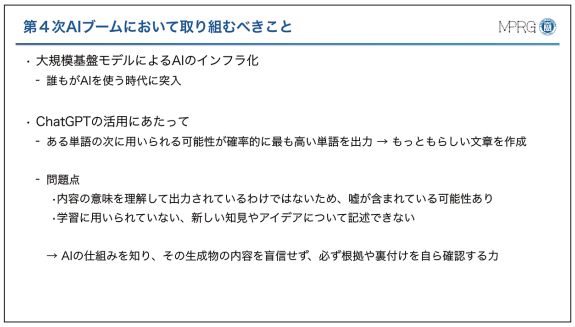
私は画像認識の研究者ですので、認識性能が100％はないことを分かっていて、どうなんだろうかと少しドキドキしながら乗ったんですが、非常にスムーズな動きで良かったです。あともう一点、非常にいい点は、このように人がいっぱい動いている場面において、この無人ロボタクシーが周辺の何を認識しているかを、このようにディスプレーしてくれる点です。非常に安心感も得られますよね。ある意味、一種の説明性とも言えます。そうすると、乗っている、サービスを受けているわれわれとしては、何を認識しているのかよく分かるし、スムーズな動きだし、不安が解消され、このサービスを受け入れることができるようになりました。

実際、UberとWaymo、この2択だったら、Waymoの方を必ず選択したいと思うぐらいです。なかなか想像つかないかもしれませんが、中に誰もいない状態で目的地まで動いてくれるのです。そんなモビリティって今までないですよね。私、別にタクシーの運転手さんと話すのが嫌いではないですが、でもたまには静かに色々考えたりしたりしたい時もあり、この無人ロボタクシーはただの移動手段ではなく、快適な移動空間を提供するモビリティであったということです。米国の都市に行くチャンスがある方は、誰でもアプリをインストールすれば使えますので、ぜひ試していただければと思います。

そして、基盤モデルを使ったもう一つの例はChatGPTです。これは大規模言語モデルと呼ばれるもので、人間と同じように対話ができたりだとか、プログラムコードも記述できたりと、生成A Iによって色々なことができるようになってきました。

ChatGPTは、まずChatGPT自体に登録してサービスを使うというものだったのですが、ChatGPTと同じような言語モデルが、今はMicrosoftのOfficeと連携して動くようになっています。すなわち、Word上で「このファイルを要約して」と入力すると、自動で要約した文章を生成してくれたりするわけです。PowerPointでは、素材のファイルを指定して、「10枚のスライドを作成して」と入力すると、レイアウトも含めて、言語モデルがスライドを作ってくれます。Excelのシートを選択して「解析して」と入力すると、その解析した結果を出力してくれます。

すなわち、ChatGPTを目的として使うのではなく、通常の作業としてWord、Excel、PowerPointを使っていると、AIである言語モデルが作業のサポートをしてくれるわけです。



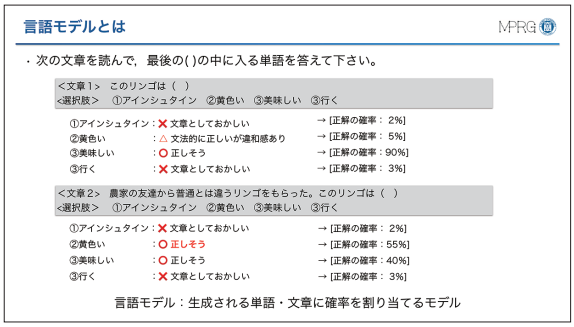
第4次AIブームがこれまでと大きく違うのは、この基盤モデルによるAIがインフラ化している点であると言えます。すなわち、誰もがAIを使う時代に突入した

ということです。第3次AIは囲碁のようにある特定のタスクだけでしたよね。それがこのように現在は誰もが使う時代に入った。これが第4次AIブームだと考えています。

ただし、色々な問題もあります。特に言語モデルとハルシネーションいうもっともらしい嘘をつくことがあるので、われわれは、どのようにAIと向き合って行くべきかということを考えなければいけない時期でもあるわけです。

2. 大規模言語モデル（LLM）と活用

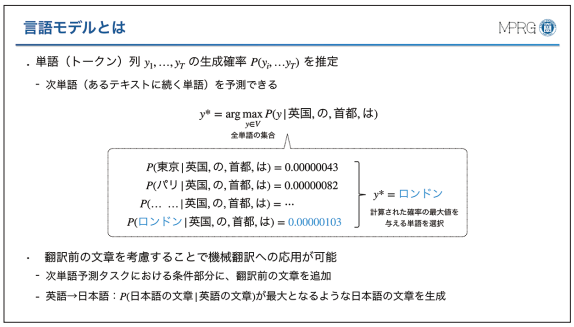
ここから大規模言語モデルがどのようにできているのか、今日はその仕組みについてもお話したいと思います。



まず言語モデルとはどういうものかというと、例えばこういう穴埋め問題があったとします。質問は「このリンゴは○○」。回答は1番、アインシュタイン。2番、黄色い。3番、おいしい。4番、行く。当然、1番の「このリンゴはアインシュタイン」はおかしいですよね。2番の「このリンゴは黄色い」は、文法的には正しいですが、少し違和感があります。3番の「このリンゴはおいしい」は正しそうですね。4番の「このリンゴは行く」。これは文章としておかしいと、われわれは分かるわけです。

今度は、質問をこのように変えます。「農家の友達から普通とは違うリンゴをもらった。このリンゴは○○」。この場合、1番のアインシュタインはおかしいですよね。2番の黄色いは、「このリンゴは黄色い」だけだと何となく違和感あったんですが、「農家の友達から普通とは違うリンゴをもらった。このリンゴは」と質問が変わると、「黄色い」も正しくなります。

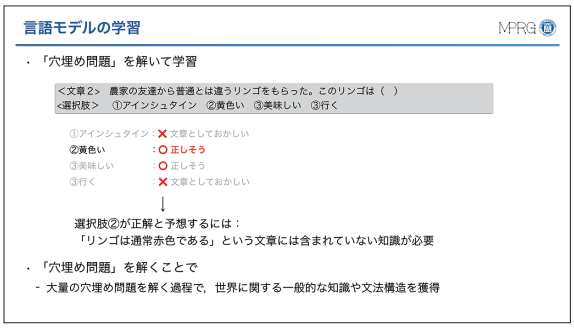
言語モデルとは、生成される単語や文章に、いわゆる確率を割り当てるモデルのことを言います。言語モデルはどのように次の文章、言語を生成するかというと、単語であるトークンの列が与えられると、次に続く



単語の確率を計算します。この例では、「英国の首都は」を条件として、次に続く単語yの確率を計算するわけです。この単語yは何かというと、全単語の集合から1個ずつ単語を割り当てて、その確率を計算します。

例えば「英国の首都は」と来たら、東京が出力される確率。パリ、ロンドンが出力される確率のように、全単語の確率を計算し、このargmaxという演算子によって、確率pが最大となる単語yを出力します。すなわち、この確率の中で一番高いのは?というと、ロンドンと出力するというのが、言語モデルを数式で表したものになります。

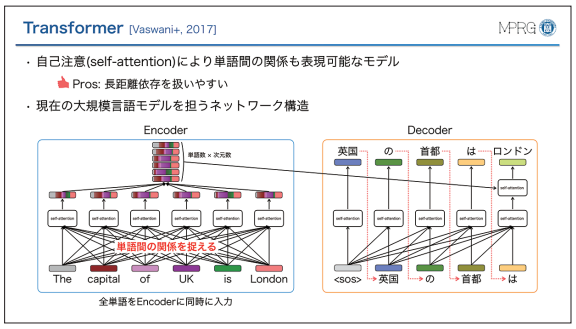
この仕組みを用いると、英語の文章を条件として入れて、日本語の文章の確率が最大となるように、日本語の文章を生成すれば、英語から日本語の言語翻訳もできるというわけです。



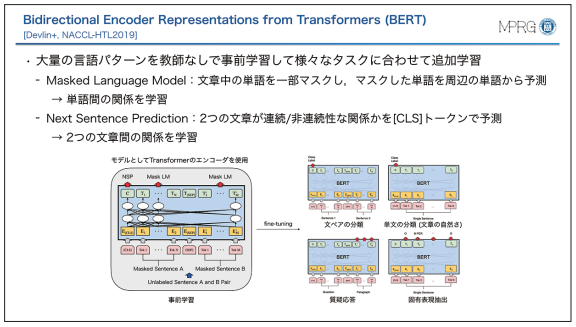
そしてもう一つ重要なのが、どうやって学習をするかということです。実はこのように大規模言語モデルも、この穴埋め問題を一生懸命解いて学習するわけです。この2番目の文章であると、黄色も正しそうだと分かります。ちゃんと正解の選択肢である単語を予測できるようになるためには、リンゴは通常赤色であるという、文章に含まれていない知識も当然必要になるわけです。

これをどうやって獲得しているかというと、ここに続く括弧に当てはまる単語を、大量の穴埋め問題を解く過程で、世界に関する一般的な知識や文法構造を獲

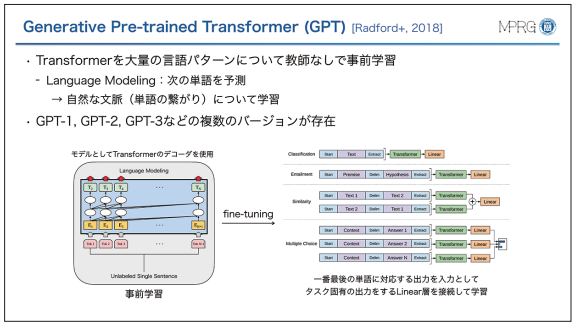
得します。ここで言う知識というのは、具体的に表現された知識ではなくて、この文章に続く自然な文章や単語を生成することができる知識となります。



大規模言語モデルには、Transformerというニューラルネットワークのモデルが使われています。Transformerは、英語から日本語にする翻訳する場合は、英語の全単語を同時に入力して、その単語間の関係性を捉えて特徴を表現し、その文章の特徴を参照しながら日本語の翻訳先の単語を予測していくという仕組みになっています。



Transformerを用いたBERTという言語モデルでは、Masked Language Modellingという、先ほど説明した穴埋め問題を学習します。具体的には、文章中のある単語をマスクして、そのマスクした単語を周辺から予測したり、Next Sentence Predictionといって、2つの文章を入れて、その2つの文章が連続して続いている文章かそうでないかを判定して言語モデル学習しています。

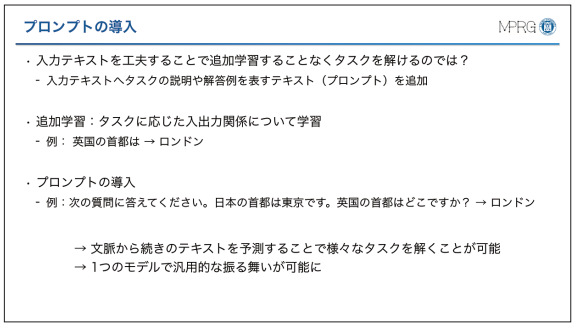


さらにもう一つ、Transformerベースの有名な言語モデルにGPT、Generative Pre-trained Transformerというモデルがあります。このGPTがChatGPTのサービスで使われているモデルになります。

GPTの学習は、Language Modellingという、文章の最後が“(　)”なっていて、括弧に入る次の単語が何であるかを解くことをするだけです。これを大量の文章データで学習しておくと、言語モデルができたというわけです。

	GPT-1	GPT-2	GPT-3	GPT-3.5	GPT-4	GPT-4o
発表年	2018年6月	2019年2月	2020年6月	2022年3月	2023年3月	2024年5月
モーダル	言語	言語	言語	言語	言語＋画像	言語＋画像＋音声
事前学習の方法	次単語予測	次単語予測	次単語予測	非公開	非公開	非公開
追加学習の方法	fine-tuning	fine-tuning, プロンプトの導入	fine-tuning, プロンプトの導入	非公開	非公開	非公開
モデルのパラメータ数	1億1700万	15億4200万	1,750億	非公開	非公開	非公開
入力可能なトークン数	512	1,024	2,048	4,096	8,192 / 32,768	128,000
学習データのサイズ	4.5GB	40GB	570GB	非公開	非公開	非公開

GPTモデルは、GPT-1、2、3と来て、ChatGPTの登場以降はGPT-3.5、4、4oと出てきています。各モデルのパラメータ数を見ると、GPT-1は1億、2は15億、3は1,750億と増え、学習データもテキストにもかかわらず4ギガ、400ギガ、570ギガと、大量のモデルパラメータを大量のデータで学習してできています。現在のChatGPTで使われているモデルのパラメータ数やデータ数は、残念ながら公開されていませんが、明らかにこれよりも多いということが想像つくと思います。



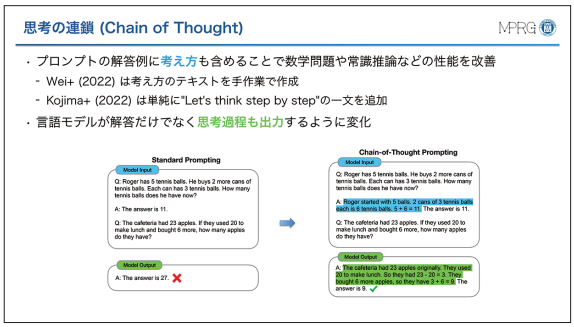
大規模言語モデルにおいてもう一つ重要なのがプロンプトの導入です。先ほどご紹介したように、文章にある説明が付くと、回答が変わったりしますよね。従って入力する質問文、すなわちプロンプトをうまく工夫することで、色々なタスクを解くことができるのも大規模言語モデルの面白いところです。

通常は対象とるするタスクの入力と出力の正解データを用いて教師あり学習を行います。この入力に対す

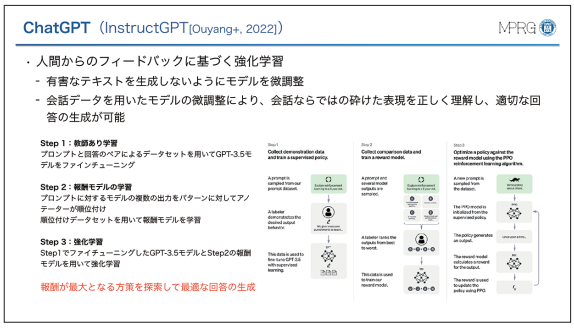
る正解を出力できるように学習するのですが、大規模言語モデルはこの教師あり学習を必要とせず、プロンプトを工夫するだけで色々なタスクが解けるのです。

例えばプロンプトとして、このように入力します。「次の質問に答えてください。日本の首都は東京です。英国の首都はどこですか」というふうにすると、日本と首都と東京の関係から、英国と首都に対応する答えとなる単語を、次に繋がる自然な文章としてロンドンという単語を生成することができるというわけです。

このように、入力として質問文を工夫することで、その文脈の続きからテキストを予測し、色々なタスクが解けるようになりました。すなわち、1つのモデルで汎用的な振る舞いができるようになったわけで、これがまさに基盤モデルということになります。

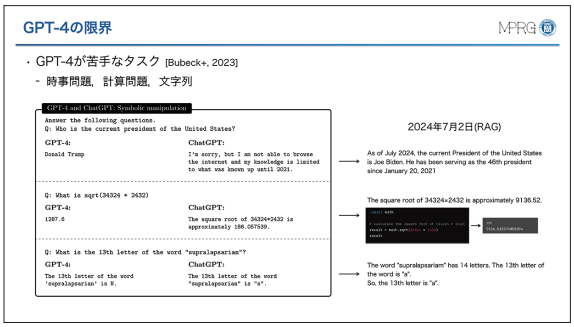


他にも「思考の連鎖」といいまして、ただ単に問題と答えをプロンプトに入れて解かせるのではなくて、どうやってその答えを導き出したかという、その思考過程を言語で書き下して新たな質問をすると、ちゃんと正解できるようになっています。



ChatGPTはこのような技術でできていますが、GPTモデルをそのまま使用しているのではありません。インターネット上の大量のテキストデータから学習するので、われわれ人間にとって、サービスとして使う時に不適切な回答をすることもあります。それはなぜかというと、インターネット上にあるテキストが全て正しいわけではなく、例えば差別的なことばなども含まれているからです。

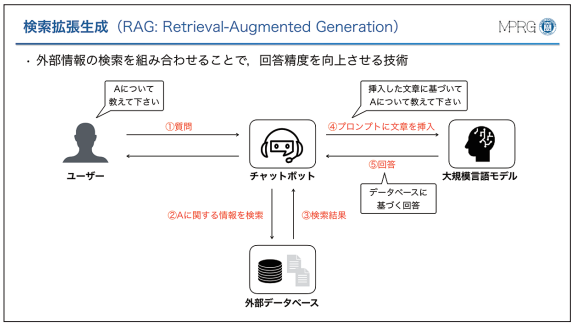
そこで、有害なテキストを生成しないようにモデルを微調整しています。GPTの回答に対して、人が良い・悪いとフィードバックして、最初にお話した強化学習により、より良文章を生成できるようにさらに微調整（ファインチューニング）をしています。これにより、対話サービスとして安全に使用できる言語モデルになるわけです。



もう一つ。GPT-4の限界についてもお話します。GPT-4のモデルを2023年に調査した際の例です。どういうプロンプトかというと、Who is the current President of the United States?　米国の現在の大統領は誰ですかと入力すると、当時のGPT-4はドナルド・トランプと答えます。これ、間違ってますよね。なぜかというと、GPT-4は2021年までのテキストデータで学習しているので、そのまま素直に、2021年時点での大統領であるドナルド・トランプを答えてしまいます。

一方、ChatGPTはどう答えるかというと、My knowledge is limited to what was known up until 2021。翻訳すると「インターネットをブラウズすることはできなく、かつ、私の知識は2021年までの知識に限られていますよ」というふうに答えるのです。これは、先ほどお話した微調整されたモデルなので、対話サービスとして人が納得いく回答をするのです。

でも、最近のChatGPTでは回答が変わってきました。2024年の7月時点で、全く同じ質問をChatGPTに入力してみると、ここにありますように、ちゃんとジョー・バイデンと答えてくれます。



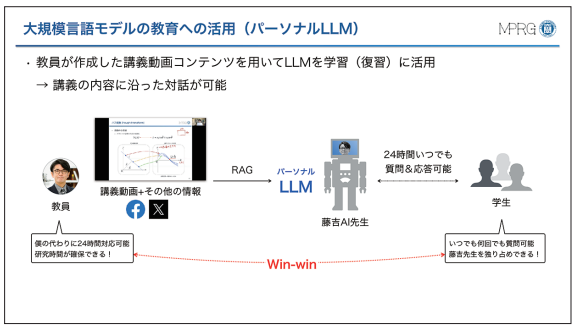
GPTモデルは2021年までのデータでしか学習していないのにもかかわらず、どうして時事問題を回答できるかというと、RAGという技術が使われているからです。

RAGとは、検索拡張生成といいまして、例えばAについて教えてくださいと質問をします。今までは言語モデルに直接入力して、その回答を出力していましたが、RAGはそうではなく、外部のデータベースがあり、プロンプトをクエリとしてデータベースを検索をします。検索した結果の文章を、プロンプトに付与して言語モデルに入力します。そうすると、この外部データベースの検索結果である文章を引用した回答ができるようになるというわけです。

このRAGを用いた言語モデルにより、現在では、不得意だった時事問題にも対応できるようになっています。

私は大学という教育現場にいますので、生成AIの教育活用にも取り組んでいます。コロナ禍では、対面の講義ができなくなりましたので、担当講義の動画をこのように作り上げました。90分の講義動画を作るには、90分の講義をそのまま録画すればいいと思うかもしれませんが、そんなわけではありません。言い間違いなどありますので、一つの講義に対して2時間や3時間かけて編集して、15回分の講義動画作っていたわけです。

とても苦労して動画を作ったのですが、対面講義に戻り、作成した講義動画が全然使われないという状態になってしまったのです。

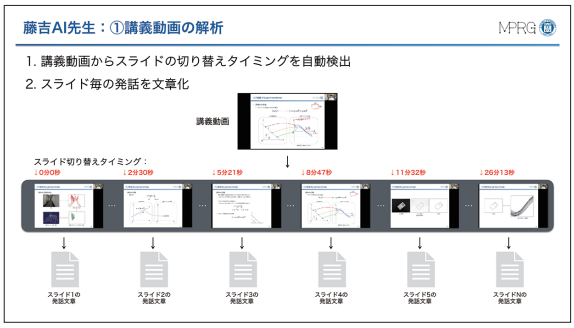


せっかく時間かけて作ったコンテンツなのでなんとか再度活用できないかと考え、生成AIである大規模現言語モデルと組み合わせてみようという思いに至りました。ChatGPTを使われた方はイメージとして分かると思いますが、優等生的な答えを回答してくれますよね。私の講義内容の質問をしたかったにもかかわらず、一般的な知識として回答してくれるので、ちょっと答えとずれてしまうこともあるわけです。講義の文脈で答えてくれたらいいのですが、そうではないのでうまくいかないところがあります。そこで、講義に特化したパーソナルなLLMがあれば良いのではないかと思い作ってみました。

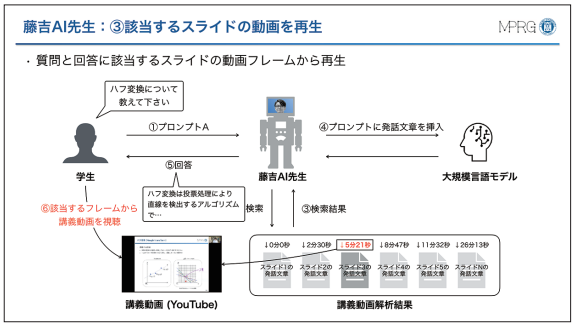
もう一つのきっかけは質問です。講義が対面に戻ったので、学生さんに質問ある人？と聞いても、全然質問してくれません。でも後から一人一人に聞いてみると、質問したかったんだけど大人数の学生がいる前で手を挙げて質問するのはちょっと恥ずかしいとか、くだらない質問したらまずいんじゃないかなどと思って質問しないようなのです。

そこで考えたのが、私の代わりに講義内容に関する質問回答をしてくれる、通称「藤吉AI先生」です。学生としては、いつでも何回でも私に質問できるわけです。私を独り占めできる。したいかどうかは別として独り占めできるわけです。一方、教員側としては、私の代わりに藤吉AI先生が24時間対応してくれて、そうすると研究時間の確保ができるわけです。これは結構Win-Winじゃないかなと思って、実際にシステムを作って、私の担当している講義に運用しています。

では、藤吉AI先生はどんな仕組みかといいますと、講義動画をYouTubeにアップロードします。講義動画が15個もあるので、復習として質問に該当する動画を試聴したいけど、探すのは大変ですよね。それで講義動画をアップロードしておき、その講義動画を解析して、LLMで対話するようにします。



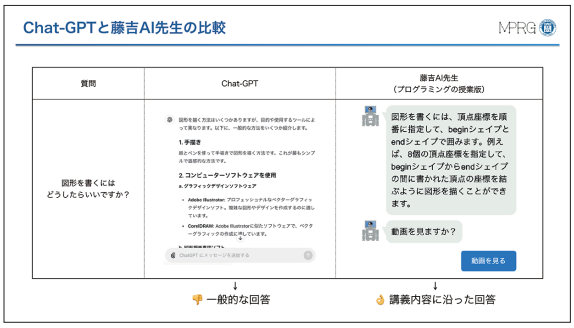
こんな感じです。まず、それぞれの講義動画から、AIを使ってスライドの切り替えタイミングを自動的に検出させます。そしてこの講義動画の中で私が話している言葉、音声データから文字起こしをした文章をスライドごとに作成し、講義発話データセットを作っておきます。



次に、学生さんが藤吉AI先生に質問を入力します。画像処理の講義なので、例えば「ハフ変換について教えてください」と質問したとします。そうすると、先ほどのRAGという技術を使って、このプロンプトに関する文章が、先ほどの講義発話データセットのどこと一番類似度が高いかを検索します。そうすると、5分21秒のスライド3の内容が一番近いということが分かります。次に、該当スライドで話した文章を検索結果として返してプロンプトに追加して、大規模言語モデルに入力します。そして大規模言語モデルが推論して、「ハフ変換は投票処理により・・・」という形で、私が講義の中で話した表現を使いながら回答してくれるわけです。

さらに、この藤吉AI先生のいいところは、どのスライドに関する質疑応答しているかが分かりますから、学生さんは該当スライドの時刻から、直接、動画の該当フレームに飛んで講義動画を視聴することもできるわけです。

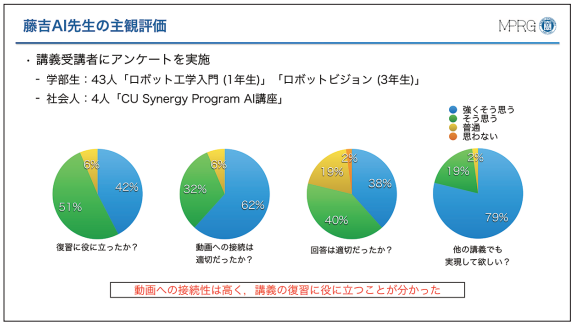
実際に、ChatGPTに「図形を描くにはどうしたらいいですか?」と質問すると、手書きだとかソフトを使う



とか一般的な回答しか返ってきません。一方、プログラミングの講義の藤吉AI先生では、プログラミングの講義の中で説明した文脈に従って回答してくれますから、講義の内容に沿った回答ができるというわけです。

藤吉AI先生はこちらのQRコードからアクセスできますので、興味ある方は試してみてください。（スライド内のQRコードを参照）

こうやってまず質問を入力します。「ハフ変換の直線検出アルゴリズムについて教えてください」と入力してみます。そして送信ボタンを押します。そうすると回答がここに出てきてて、この回答がまさに私が講義動画で説明したのと近い形の文章になっています。さらに、動画を見るボタンを押すと、長い動画のうち、ハフ変換のアルゴリズムを説明しているところから視聴することができます。



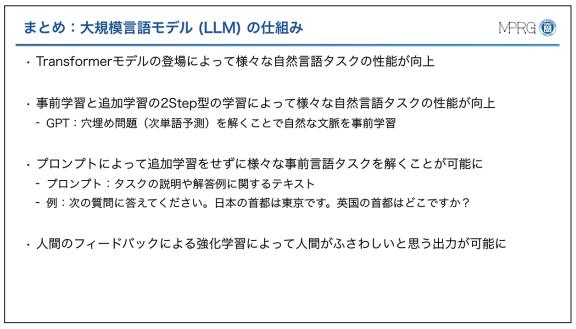
この藤吉AI先生を実際に今学期の講義で運用し、アンケートを取りました。1年生、3年生、社会人の講座で、それぞれの講義動画で使ってもらいました。青

色が「強くそう思う」、緑色が「そう思う」です。まず、「復習に役に立ったか」を見ると、青が51%と緑が42%ですから約93%の方が、復習の役に立ったって回答してくれました。「講義動画への接続は適切だったか」というのも94%と高評価でした。一番右の「他の講義でも実現してほしいか」については、何と98%と多くの学生が要望しています。

重要なのは左から3つ目の「回答は適切だったか」です。大規模言語モデルの回答について、2%の人が「そう思わない」と回答しています。これは何かというと、いわゆるハルシネーション「幻覚」と呼ばれるものです。

大規模言語モデルは大量のデータで学習しているのですが、正しいことを必ずしも答えるわけではありません。文章として次に出てくる単語が、より文章らしいものを出すだけであって、内容が正しいかは理解してないのです。なので、実際にここでもハルシネーションは起こって、2%の学生さんは、回答が適切じゃなかったというわけです。この問題をどう解決するかというのが、この第4次AIブームにおける生成AIの重要な観点だと思います。

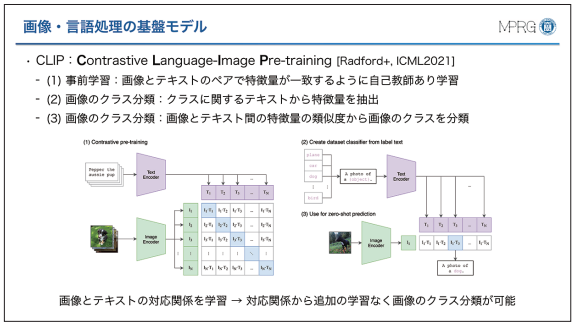
ちなみに、藤吉AI先生のシステムでは、何か回答がおかしいなと思ったら、「動画を見る」というボタンを押していただければいいわけです。「動画を見る」ボタンを押すと、その元となる動画の中で、私が説明するわけです。ハルシネーションは必ず発生するので、そういった時にその情報元であるソースに辿り着くようにリンクできるかが、生成AIにおいては非常に重要ではないかと考えています。



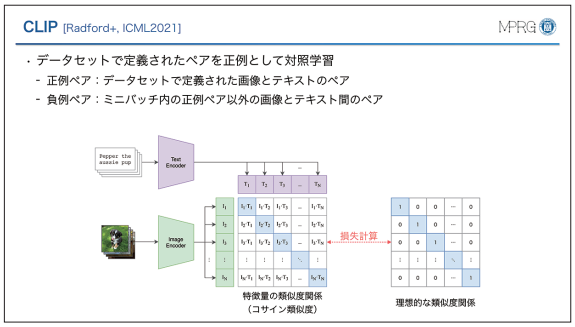
大規模言語モデルは、Transformerというモデルと穴埋め問題の学習によって実現することができました。次の単語を予測するという穴埋め問題を解くことで、自然な文章の文脈を事前学習します。さらに、プロンプトによって、色々なタスクを解くことができるようになりました。

3. 画像言語モデル（VLM）

最後に画像言語モデルについても紹介したいと思います。最近は、画像と言語を結び付けるような基盤モデルが出てきました。



画像言語モデルにCLIPというモデルがあります。これは、画像を入力した時の特徴と、その画像に合っている文章を入力した時の特徴が一致するように学習されています。すなわち画像と言語が結び付くようになったというものです。

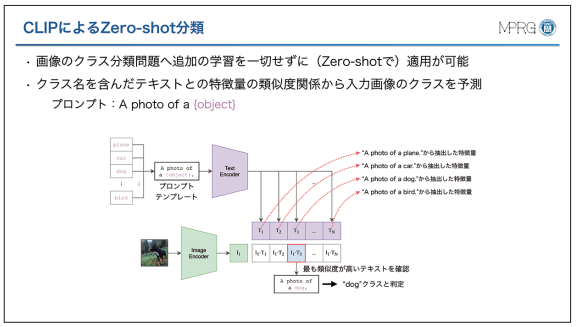


どうやって作っているかというと、まず大量の画像と、その各画像を説明した文章がペアとなったデータを用意します。そして、データセット内で定義された画像とテキストのペアが、正しいペアを正例ペア、そうでない組み合わせのペアを負例ペアといいます。

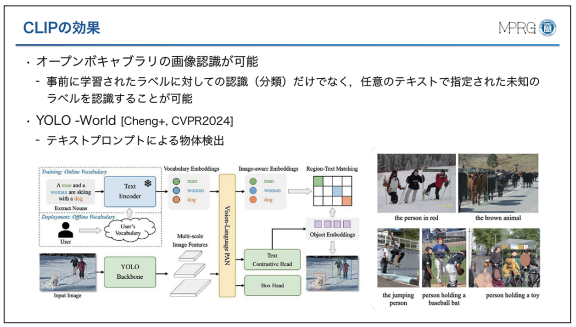
画像とテキストを入れた時の特徴量が正例ペアの時は1、そうでない負例ペアの時は0となるように学習をします。すなわち正例の時は、画像とその画像を説明した文章は同じものですよ、違う場合はそうじゃないですよという学習をたくさんするわけです。そうすると画像特徴と、その画像に対する文章の特徴が一致するようになります。

このCLIPで何ができるかというと、一切追加学習しない、ゼロショットでの画像分類ができるようになります。

どういうことかというと、a photo of ○○ objectの○○に、birdとかdogとかcarとかplaneなどの識別対



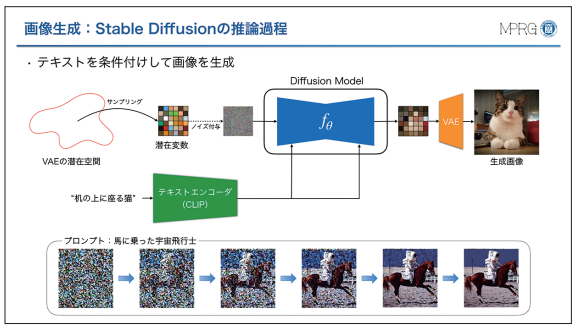
象の単語を挿入して、そのテキストの特徴量を計算しておきます。そして認識対象の画像を入力して、画像の特徴量を計算します。画像特徴とテキスト特徴量との類似度を計算し、最も類似度が高いテキストを探します。この場合はdogなので、この入力画像はdogクラスと判定するものです。このように画像と言語が結び付くことによって、ゼロショットで分類が予測できるようになりました。



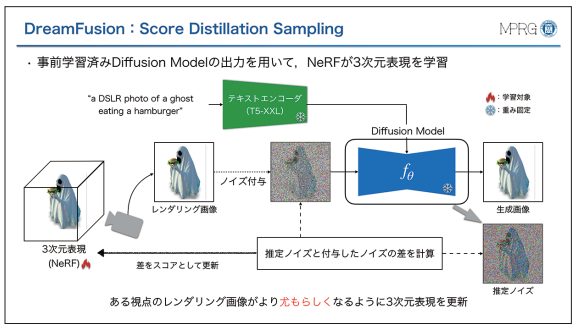
今、このCLIPを使うことで色々なタスクを解くことができるようになっています。オープンボキャブラリーというのですが、事前に学習されたラベルに対しての認識、分類ができるだけでなく、任意のテキストで指定された、未知のラベルの認識ができるようになっています。

例えば、YOLO-Worldという最新の物体検出モデルでは、このように、A man and a woman are skiing with a dogというテキストをプロンプトとして入力して、そのテキスト中に出ているman、woman、dogを検出することができるわけです。（スライド中央下部の画像で、man、woman、dogの各々が正しく検出されている）

これまでも人、猫、犬といったクラスを教師あり学習しておけば、学習したクラスだけ画像から物体検出することはできたのですが、今ではこういうテキストプロンプトを使って、それに一致する物体領域を検出することができます。例えば、これ（スライド右側の5つ画像のうち左上のもの）も面白い例ですよね。the person in redとテキストプロンプト入力すると、画像中の人が全て検出されるのではなくて、赤色の服を着た人だけを（一番右の赤い服の人のみが矩形で囲われて）検出することができるわけです。このように画像と言語が結び付くことによって、認識できる範囲がオープンとなり、飛躍的に広がっています。

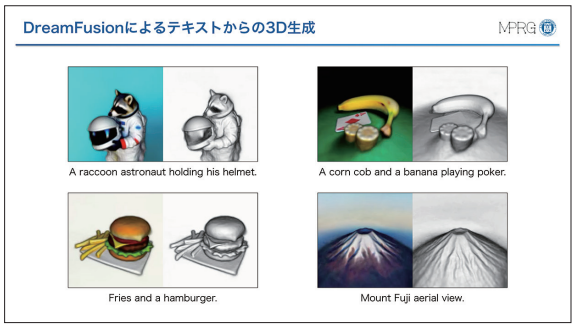


CLIPは認識だけではなくて、画像生成にも使われています。時間の都合で詳しい説明はできないのですが、Diffusion Modelというニューラルネットワークのモデルがあります。Diffusion Modelは生成モデルであり、ノイズから画像を生成することができます。CLIPを使って、例えばプロンプトとして「馬に乗った宇宙飛行士」というテキスト特徴を計算し、生成の際にはこのテキスト特徴近づくようにDiffusion Modelを生成すると、最初はノイズなんですが、だんだんこのプロンプトに合った画像を生成することができるようになっていきます。

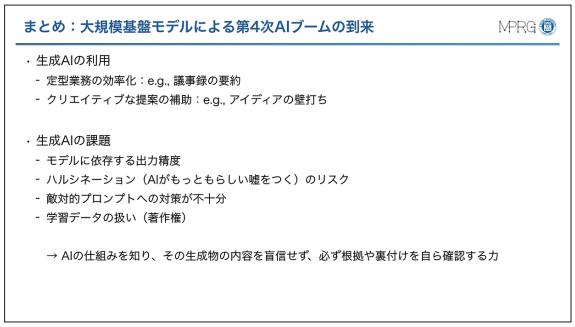


さらに画像生成AIは、3次元生成に拡張されています。簡単な紹介となりますが、まず、画像生成と同様にプロンプトのテキスト特徴をCLIPで計算し、Diffusion Modelに入力する際に、NeRFという3次元の光線空間をニューラルネットワークで獲得するモデルを使って、ある視点の画像をレンダリングします。このレンダリングした画像にノイズを付与して、Diffusion Modelにてテキスト特徴量に物体の見え方が近づくように画像を生成します。この時の情報を用いてNeRFモデルが、

テキストに合った3次元表現を学習し、三次元モデルを生成できるようになるという仕組みです。



この三次元生成の例ではA raccoon astronaut holding his helmet（ヘルメットをかざすアライグマの宇宙飛行士）というテキストをいれると、このような3次元のモデル（スライド左上の画像）が生成されます。



そろそろ時間ですので、まとめたいと思います。この生成AIという技術は様々なところに今後活用が促進されていきます。誰もがAIを使う時代になったというわけです。この第4次AIブームの時代における課題はまだ残っており、そのひとつは精度です。やはりまだまだ出力精度は使用するモデルに依存しています。

もう一つの課題が、先ほど紹介したハルシネーションの対策です。生成AIの大規模言語モデルはもっともらしい嘘をつくというリスクがあります。あとは、学習データの扱い（著作権）というのも非常に重要かと思います。これらの課題をしっかりと解決していくことが、AIを誰もが使える技術にするために必要なことと考えています。また、第4次AIブームにおいて、AIを活用するわれわれ人間がすべきことは、AIの仕組みをしっかりと知ることによって、その生成物の内容を盲信せず、必ず根拠や裏付けを自ら確認する力が、今、問われているのではないかと思っています。

以上で、私の今日の話は終わりたいと思います。長い時間、お付き合いくださりどうもありがとうございました。（拍手）

■講演者
藤吉 弘亘（ふじよし ひろのぶ）
中部大学理工学部 AIロボティクス学科 教授

本稿は2024年8月8日に開催された、当協会主催「第46回測量調査技術発表会」における、藤吉弘亘氏の特別講演の内容をまとめたものです。