クロスモーダルな理解～サーベイ

牛久 祥孝

深層学習の恩恵として、各種データでの既存のタスクの精度が大きく向上したという点以外に、それぞれ隣同士だったはずなのに専門性の強いパイプラインを構築していた結果として生じていた相互参入障壁が大きく緩和された点も特筆するべき点である。例えば、今まで画像認識を専門として研究開発を進めていた研究者や技術者にとってみると、音声認識も導入しようとした際には音声信号処理の独自のパイプラインを習得する必要があった。その様にお互い名前や中身が大きく異なるパイプラインが、畳込みニューラルネットワーク（Convolutional Neural Network; CNN）や再帰ニューラルネットワーク（Recurrent Neural Network; RNN）、最近であれば Transformer などで統一化されてきているためにお互いを理解しやすい状態になって来た。結果として、画像や音声信号、自然言語など多様なデータを行き来したり同時に理解したりするようなクロスモーダル理解の研究が大きく前進してきた。本稿では、そのようなクロスモーダルな形でデータを理解する種々のタスクについて概説する。

Cross-modal Understanding: A Survey

USHIKU Yoshitaka

In addition to the fact that deep learning has dramatically improved the accuracy of existing tasks with various types of data, it is also worth noting that it has dramatically eased the barriers to mutual entry into neighboring fields. Such barriers had arisen as a result of the construction of highly specialized pipelines for each problem. For example, researchers and engineers specializing in image recognition would have had to learn their pipeline for speech signal processing if they wanted to introduce a speech recognition module into their image recognition system. These pipelines, which differ greatly in name and content, have been unified into Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and more recently, Transformer. As a result, it has become easier to understand the state-of-the-art pipelines in each other fields. Thus, research on cross-modal understanding has made significant progress in which diverse data, such as images, speech signals, and natural language can be traversed and understood simultaneously. In this article, the author outlines various tasks for understanding data in such a cross-modal manner.

1. まえがき

2021 年は、音声認識で深層学習が精度の大幅な向上を果たした[1] 国際コンペティションからちょうど 10 年になる。その次の年には画像認識で同様に深層学習が劇的な精度向上を達成[2] し、さらに 2 年後の 2014 年には機械翻訳で深層学習が非常に複雑なそれまでのシステムと同等の精度を達成する[3] という事件が起きた。

これらの課題はそれぞれ、音声信号処理やコンピュータビジョン、自然言語処理と言われるコンピュータサイエンスを支える各領域で研究されていたものである。特に 2000 年頃から課題を解くためのお手本となる教師データを含めたデータセットを構築し、統計的機械学習手法を適用したデータドリブンな解法がそれぞれで注目を集めていた。

その頃の各課題で主流となっていた手法は、基礎的な機械学習手法やいくつかのアイディアが共有されていたものの、基本的には各課題で独自に発達したデータの前処理や特徴量設計、そして後処理などが多段に組み合わされていた。つまり、例えば自然言語処理の研究者が画像も含めた新たな研究を始めようとしても、コンピュータビジョン分野の技術を習得して自身の研究に組み込むための導入コストが大きな壁となっていた。

一方で、それでも各分野の研究者が他の分野と融合したデータ理解を行おうとする取り組みも見られた。例えば動画認識として各時刻の画像情報と音声情報をそれぞれの特徴量設計と機械学習手法によって融合しようとしたり[4]、

Contact : *USHIKU Yoshitaka* yoshitaka.ushiku@sinicx.com

画像の内容を自然言語で説明する為に当時の画像認識と自然言語生成を参考にしながらパイプラインを構築したり[5]といった研究が見られる。

その中で、冒頭に述べたような深層学習の研究が各領域でセンセーションを巻き起こした。音声認識で多層パーセプトロン（Multi-Layer Perceptron; MLP）によって、画像認識がCNNによって、そして機械翻訳がRNNによって塗り替えられたが、それらMLP/CNN/RNNと言うモジュールはすぐにそれぞれの分野で統一的に利用されるようになった。例えば音声認識や動画認識の時系列性はCNNやRNNで、そして一方機械翻訳分野でもCNNによる系列理解が試みられるようになるといった具合である。既によく知られているようにこれらの深層学習は職人芸的な特徴量設計を必要としないことも相まって、お互いの分野での異なる課題へのパイプラインの見通しが極めて立ちやすくなった。

結果として、先ほど述べたように深層学習の流行前に徐々に取り組まれ始めていた複数のモダリティを同時に理解するような研究は、深層学習というロケットブースターを得て非連続な加速を果たした。例えば画像キャプション生成は2010年にその先駆けとなる研究[6]が生まれ、2012年には機械学習手法を含めたパイプライン[5]が確立し、その後の精度向上が徐々に進んでいた。そして2012年の画像認識と2014年の機械翻訳の深層学習化[2,3]によって直ちに深層学習の導入が進み、2015年6月に開催されたCVPR（Conference on Computer Vision and Pattern Recognition）というコンピュータビジョン分野の最高峰の国際会議では、CNNで画像を理解してRNNで文を生成するというアプローチによる画像キャプション生成が世界中の大学や企業から同時多発的に提案されるに至った。

このような形で、現在では音声/画像/自然言語といった多様なモダリティの自由自在な組み合わせの入力を受けつけ、その内容を統合理解して自由自在な組み合わせの出力を得るような深層学習の研究が広く行われている。本稿では、そうした複数のモダリティを入出する研究をクロスモーダルな理解と称して、その概観となるサーベイを提供することを目的とする。紙幅の都合上各分野に深く入った解説を与えるのは困難であるが、現在のクロスモーダルな理解の索引となることを目指す。

2節では、クロスモーダルな理解としての諸研究を概観するにあたってキーとなるエンコーダとデコーダという概念について解説する。3節から5節では、画像/自然言語/音声という代表的な3つのモダリティから2つを組み合わせたクロスモーダル理解の研究について述べる。すなわち、3節で画像と自然言語、4節で自然言語と音声、5節で画像と音声を繋ぐ研究についてそれぞれ解説する。6節はより多くの、またはより多様なモダリティを扱った研究についても触れながら、本稿の結びとして今後の展望について述べる。

2. エンコーダとデコーダ

まえがきではクロスモーダルな「理解」とあえてぼかした表現を与えたが、基本的には入力として与えられたモダリティのデータに含まれる情報を抽出し、それを既定のモダリティのデータに変換して出力することになる。

例えば画像認識課題では、入力が画像というモダリティであり、出力がクラスラベルという離散化されたシンボルに相当するモダリティである。音声認識は全体としては入力が音声、出力は自然言語というモダリティであり、機械翻訳は入力出力どちらも自然言語である。

さらに画像認識を例にとってもう少し中身を考えてみる。入力された画像から典型的には畳込みニューラルネットワークによって処理された途中のデータはベクトルとなって、単層以上のMLPを通していずれかのクラスに分類される。

また機械翻訳でRNNを用いる場合、入力データの単語を1単語ずつRNNに入力していき、翻訳したい文が全て入力されたときのRNNの隠れ変数ベクトルが得られる。これをまた別のRNNに入力し、1単語ずつ訳文の単語を出力しつつRNNに直前の出力を入力していくということを文の終わりという意味の出力が得られるまで繰り返す。

ここまでの2つの例で、途中に何らかのベクトルが出現していることに注意してほしい。つまり、深層学習においては多くの画素からなる画像や単語の系列である自然言語といった高次元だったり構造化されたりしているデータを、連続値からなる数列データに変換している。この数列データと言うのが典型的にはベクトルの形をとっている場合が多い、という訳である。このように、入力データをベクトル等の数列データに変換する機構をしばしばエンコーダ（符号化器）と呼ぶ。画像認識ではCNNのエンコーダが出力したベクトルをMLPで分類しているし、機械翻訳では最初のRNNがエンコーダとして出力したベクトルが、次のRNNに与えられている。

では、この機械翻訳の2つめのRNNは何か。このように研究課題によっては出力として自然言語や画像などの高次元ないし構造化されたデータを出力することが求められる。他によく知られたものとしては自己回帰を用いた音声合成や、Generative Adversarial Net（GAN）またはVariational Auto-Encoder（VAE）を用いた画像生成などが挙げられるだろう。そしてこれらの自己回帰やGAN/VAEによるデータ生成の際にも、機械翻訳の2つ目のRNNと同様にやはり何らかのベクトルが入力されている。このように、ベクトルなどの数列データを入力として受け取り、音声や画像、自然言語のようなモダリティのデータを出力する機構をしばしばデコーダ（復号化器）と呼ぶ。

このように、画像認識はエンコーダと分類器の、機械翻訳はエンコーダとデコーダの組合せとみることができる。他にエンコーダやデコーダという単語が良く出てくる分野

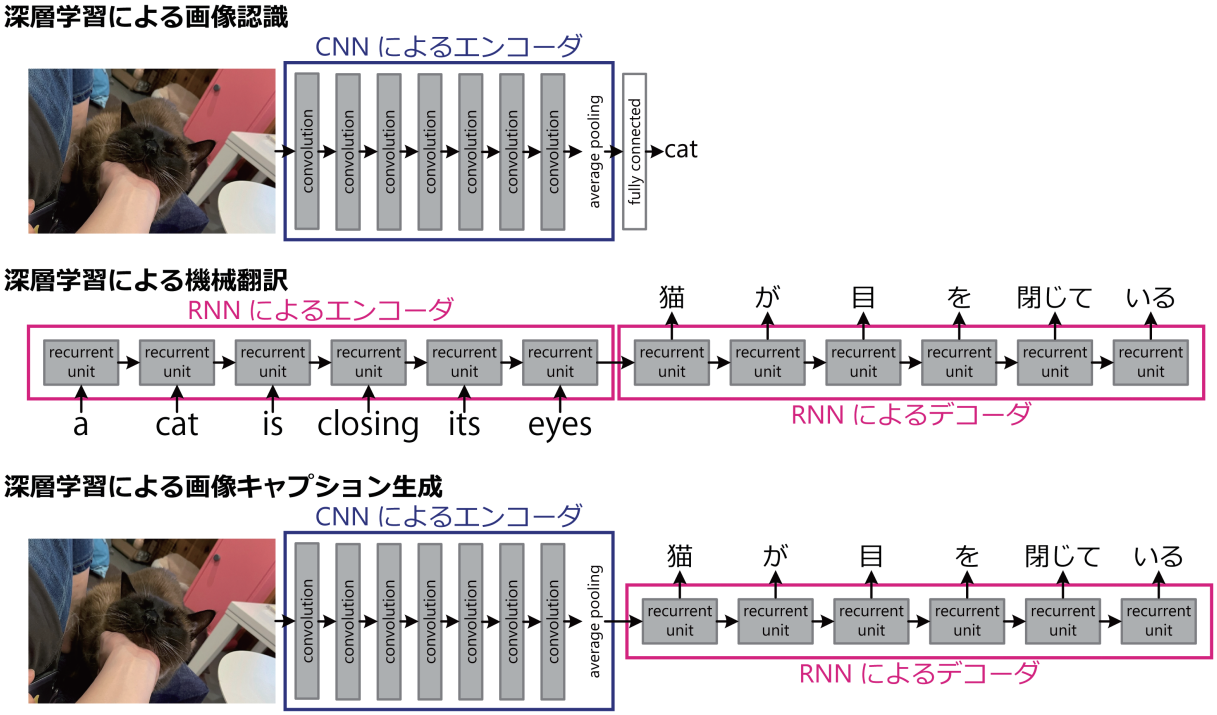

図 1　深層学習による画像認識、機械翻訳、画像キャプション生成。このように各課題は入力データをベクトルなどの数列データに変換するエンコーダ（符号化器）と、数列データから出力データを生成するデコーダ（復号化器）からなる。このようなエンコーダ・デコーダの活用はクロスモーダルな理解を概観するにあたって必須の観点となる。

としては、映像などを符号化して再度映像に戻す際にもやはりエンコーダとデコーダが登場する。このように、情報を符号化するという点がエンコーダに、復号化するという点がデコーダには期待されている。画像認識の分類器がデコーダと言われないのは、この点が理由であろう。

ここまでの議論をまとめたのが、図 1 の「深層学習による画像認識」および「深層学習による機械翻訳」である。ここまでの議論を見ていると、次のようなことを考えないだろうか。画像認識のエンコーダは何らかのベクトルを出力している。これは画像を分類する為の情報が符号化されていると期待されるベクトルである。次に機械翻訳のデコーダはベクトルから文を出力している。文を出力するための情報がベクトルにちゃんと入っていれば、あとはデコーダがちゃんと自然言語文を出力してくれる。であれば、CNN によるエンコーダと RNN によるデコーダを組み合わせれば、入力画像を説明できるようなキャプション生成のパイプラインが実現できるのではないだろうか。

このようにして CNN によるエンコーダと RNN によるデコーダを世界中の機関が同時に提案してきたのが、1 節で述べた 2015 年の CVPR である。もちろん画像認識を学習しただけの CNN エンコーダと機械翻訳を学習しただけの RNN デコーダを直列してそのまま画像のキャプションが生成できるという訳ではなく、画像キャプション生成用の画像とお手本キャプションのペアからなるデータセット上で学習する必要がある。また深層学習による画像キャプション生成のパイプラインが確立する前でも、実際には画像認識のモジュールと機械翻訳などの自然言語生成モジュールが組み合わされたパイプラインが提案されており、このエンコーダ・デコーダ構造を深層学習が初めてもたらしたわけではない。このように多様なモダリティのデータを解釈したり生成したりと言う課題がそのままエンコーダとデコーダとして解釈できる、といういわば観点のようなものが見えてきたのが重要である。3 節以降で解説するクロスモーダルな理解の各分野でも、このように複数のエンコーダとデコーダが登場してくる。

3. 画像と自然言語を繋ぐ研究

画像や自然言語処理で CNN や RNN、はたまた最近の Transformer の表現学習が成功しているのは、データセットとして大量の画像やテキストを収集可能であることも大きい。そのように大量なモデルで、帰納バイアスの少ない巨大パラメータによるネットワークを学習することの恩恵を受けているのがこれらのモダリティのデータであるともいえる。そのようなデータとして、YouTube や Facebook といった web サービスを中心として、画像や動画像と関連するテキストが大量に投稿されている。画像ないし動画像と自然言語というペアのデータは、ドメインさえ絞らなければ大量に集めやすいマルチモーダルデータでもある。

本節では、こうした画像と自然言語を繋ぐクロスモーダル理解について述べる。この二つのモダリティの組合せは

特にビジョン＆ランゲージと言う名前がついていて、コンピュータビジョンの国際会議でも自然言語処理の国際会議でも数年前から一定の存在感を示すようになっている。

なお、画像から自然言語を出力するタスクと言うと光学的文字認識（Optical Character Recognition; OCR）を想像するかもしれない。確かに、深層学習以前から種々の方法で文字認識がさかんに試みられている。深層学習の歴史の中でも、Yann LeCun による LeNet[7] は学習可能な畳込みニューラルネットワークとして前世紀から非常に有名なモデルで、文字認識でその効果が確かめられている。OCR は画像から自然言語へのクロスモーダルな変換を行う研究と言えなくもないが、直接的な言語情報である文字以外の視覚情報を言語と結びつけるビジョン＆ランゲージの議論で、OCR そのものを含めることは少ない。

3.1 キャプション生成

画像キャプション生成（Image Captioning）は、画像の内容を示すテキストを生成するタスクであり、他の自然言語処理の研究と同様に英語を対象としたものが多い。ビジョン＆ランゲージ分野の中でも歴史の長い課題であり、深層学習の流行を迎える前の 2010 年頃から取り組みが徐々に増えてきたテーマである。そのバリエーションについても枚挙に暇がなく、画像列や動画からのキャプション生成の研究も存在する。

最初にこの問題に取り組んだ論文[6] では、条件付き確率（Conditional Random Field; CRF）を用いて、画像の 3 種類のラベル（トリプレット）:「object」、「action」、「scene」を推定し、トリプレットが付加されたキャプションの集合から、類似のトリプレットを持つ文を検索している。つまり、システムは新たな文を生成するのではなく、既存の文の内容と一致する文を検索するのであり、画像やキャプションにトリプレットを付加する必要がある。その後、画像認識モジュールと文生成モジュールをそれぞれコンピュータビジョンや自然言語処理の分野から取り込み、画像とキャプションのペアデータのみから新たな文を生成して画像キャプションを生成することを目指した研究が提案された。

その後、前節で述べた通り CVPR 2015 において深層学習によるキャプション生成が話題になり、複数の論文が同時に同じパイプラインでキャプション生成を提案した。その中でも、Google の Vinyals らによる方法[8] は、Google の CNN 画像認識モデルである Inception と、LSTM 機械翻訳モデルを組み合わせたもので、仕組みが理解しやすい。CNN 画像エンコーダと LSTM 自然言語デコーダの組み合わせという非常にシンプルな構造でありながら、生成されたキャプションは精度、流暢性ともに大きく向上している。ちなみに深層学習の近年の手法ではしばしば用いられる注意（アテンション）機構は、機械翻訳[9]・キャプション生成[10] に導入されたのが生い立ちである。生成中のキャプションで次の単語を推定する際に、画像のどの領域に注意を払うべきかを、画像とキャプションのペアのみからデータドリブンで学習できる。

3.2 ビジュアル質問応答

ビジュアル質問応答（Visual Question Answering; VQA）は、与えられた画像に関する質問に答えるタスクである。ビジュアル質問応答では、質問に答えるための情報は基本的に入力画像に含まれている。画像と質問文というマルチモーダルな入力を用いて、回答候補をいずれかの単語に分類するというクロスモーダル理解の研究である。

この研究課題はユーザーインターフェースの分野で最初にアプリケーションとして提案されたもの[11] で、視覚障がい者が旅行先で何か分からなくて困っているもの撮影し、回答してほしい質問を入力するという使い方を想定していた。そしてこの論文では、アプリに接続されたクラウドソーシングの人々が手動で回答していた。今回紹介するビジュアル質問応答システムは、このプロセスを自動化する試みともいえる。

ビジュアル質問応答は、視覚と言語の分野で最も広く取り組まれている問題のひとつである。その理由は様々だが、一つにはデータセットとその評価方法が早くから確立されていたことが挙げられる。その中でも貢献の大きいベンチマーク論文[12] が 2015 年に発表されているが、画像を見ないと答えられない質問をクラウドソーシングで収集し、さらに 1 つの質問に対して 10 個の回答をクラウドソーシングで様々な人から収集して新規データセットを構築している。そして、3 人以上の人が一致して回答した場合に質問が正しいと評価するプロトコルを提案し、画像と質問文をエンコーダで特徴量に変換して直列した後にクラス分類するというシンプルなベースラインで評価を行った。このように整えられたデータセットと評価方法、理解しやすいベースラインがその後の大量の新規参入者をもたらした要素であると考えられる。

また、ビジュアル質問応答もキャプション生成と同様に注意機構と親和性の高い課題である。Shih らによる手法[13] は注意機構をビジュアル質問応答に適用した先駆的な研究で、階層的な質問・画像の共起を活用している。Anderson らが提案した注意メカニズム[14] は、画像上で規則的に配置したグリッドに対する注意と、物体らしき領域を別の手法で複数生成したものへの注意を組み合わせたモデルである。

3.3 マルチモーダル機械翻訳

マルチモーダル機械翻訳とは、与えられた画像とそのキャプションを用いて、キャプションを別の言語に翻訳する作業である。概念的には、画像を使って入力文の曖昧な

部分を解消し、翻訳の精度を向上させることが期待できる。例えば、Seal という単語は、ものを貼る「シール」もあれば、ラッコにちかい海洋生物もある。実際にものを貼る「シール」が写っている画像があれば、キャプションに Seal という単語が入っていても、どちらの Seal を指しているのかはすぐにわかる[15]。これは、画像と自然言語が入力されると、自然言語が出力されるという形のクロスモーダル理解の研究になる。

Hitschler らの手法[16]は、まず、入力された画像とキャプションのペアから、キャプションのみに対して通常の機械翻訳を行い、翻訳候補文を複数生成している。次に、ターゲット言語の画像・キャプションペアデータセットを検索して、入力画像と翻訳候補文からなるペアを複数探し、検索結果に基づいて翻訳候補文のスコアを更新（リランク）することで画像による精度向上を促している。

マルチモーダル機械翻訳は、一部のモダリティでしか実行できない問題の精度を、モダリティの数を増やすことで向上させようとする典型的な例の一つである。一方で実は、このようにモダリティを増やして精度を向上させる問題はクロスモーダル理解のなかでも難しいものの一つとなってしまう。この点については最後の節でも簡単に述べる。

3.4 テキストからの画像生成

テキストからの画像生成は、画像キャプション生成の逆問題で、画像の内容を示すキャプションから同じ内容を示す画像を生成するタスクである。テキストから画像へ変換するクロスモーダル理解の研究と解釈できる課題で、前節でも触れた GAN や VAE などの深層学習生成モデルが普及した結果、この複雑なタスクへの挑戦が増えている。

Mansimov ら[17]は、入力されたキャプションを双方向 LSTM でエンコードした後、RNN ベースの画像デコーダで画像をより鮮明にするために複数回更新するというパイプラインを提案している。Reed ら[18]は、この問題に初めて GAN を導入しており、Stack GAN[19]はこのような GAN を積み上げて画像の高精細化を達成している。Stack GAN では Reed らとほぼ同じアプローチで生成した 64 ピクセル四方の画像を、入力キャプションとともに別の GAN に再度入力し、256 ピクセル四方の画像を出力させている。このように GAN を繋げて高解像度な画像を生成するアプローチが続いていくことになる。

この問題のブレークスルーは Transformer によるアプローチで、OpenAI が 2021 年 1 月 5 日に、キャプションからこれまでよりもはるかに多様な自然な画像／イラストを生成できる手法である「DALL-E」についてのブログ記事[20]を公開した。キャプションから極めて多様かつ自然な画像やイラストを生成できるという例を示し、界隈に衝撃を与えた。またこの記事では、後述する表現学習手法の CLIP[21]を活用しているということが述べられており、CLIP と各種画像生成手法を組み合わせた DALL-E の再現が様々なところで試みられている。

3.5 視覚的対話

視覚的対話は、これまで言語だけで行われていた対話の研究に、画像や映像を加えたものである。まさに Visual Dialog と言うタイトルを冠した 2017 年の論文[22]では、画像に関連した対話を自然言語で行うデータを収集して対話の学習を行っている。自然言語での対話履歴と関連する視覚情報から、次の対話行為として自然言語などを生成するという意味で、連続的なクロスモーダル理解の研究と解釈できる。

対話生成は人工知能の分野で古くから取り組まれてきた研究テーマで、視覚的対話でもこのような人工知能タスクとしての対話を扱う研究が多く存在する。代表的な設定としては、2 人のペアに画像を使った会話をしてもらい、その自然言語データを収集している。例えば片方が画像を見ていてもう片方が画像を見られない状態の中で、Q&A 形式の対話を通じて画像の内容を口頭で伝えるという対話[22]や、2 人とも画像を見ている中で、一方が特に見ている領域をもう片方が当てるゲームを Q&A 形式の自然言語対話で進めるもの[23]などがある。店舗での店員と客[24]や、市街地での電話ナビゲーターと旅行者[25]などより社会での活用シーンに近い設定での対話を扱う研究も存在する。

また、広い意味では、エージェントやロボットによるクロスモーダル理解の研究もある。VLN（Vision and Language Navigation）というタスク[26]はこの中でも有名で、仮想・現実環境にいるエージェントやロボットが、与えられた言語による指示と現在の視覚情報を頼りに目的地に到達するタスクである。その時々の視覚情報と最初の言語指示に基づいて、回転や移動などの次の動作を対話行為として出力するという意味では、これもまた視覚的対話に関連する研究と言える。

3.6 表現学習

表現学習とは、画像（または動画）とそれに対応する自然言語から、ここまで述べてきた諸タスクのための特徴空間を学習することで、様々なタスクの精度向上と学習データ量の削減を目指すものである。画像と自然言語のペアデータをそれぞれのモダリティのためのエンコーダに入力し、何らかの自己教師あり学習を経て特徴量の初期学習を行うものが典型的なパイプラインである。

VILBERT（Vision-and-Language BERT）はこのような研究の草分け的存在[27]で、BERT[28]のマルチモーダル版である。動画の時々の画像フレームの特徴表現の集合と、キャプションの各単語の分散表現のベクトルを Transformer に入力して、エンコーダを学習させる。また、テキ

ストから画像を生成する手法であるDALL-Eとともに紹介したCLIP[21]は、現時点で大きな注目を集めている手法の一つである。Webから4億対の画像・キャプションペアを収集し、対になっている画像・キャプションペアが最も似た値の特徴量にエンコードされるように、画像・キャプションそれぞれのエンコーダを学習する。

4. 自然言語と音声を繋ぐ研究

音と自然言語を扱ったクロスモーダル理解としては、音声認識や音声合成がまさに合致するタスクである。音声認識（Acoustic Speech Recognition; ASR）は特にOCRと類似している。OCRが画像メディアの中に埋め込まれている自然言語を抽出するものであるのに対して、ASRは音声メディアの中に埋め込まれている自然言語を抽出するもので、最近出てきたクロスモーダル理解の研究と言うには過去の歴史が長い。深層学習によるアプローチが導入される直前までは手動で設計された時系列データ処理によるローカルな特徴量設計と、混合ガウス分布や隠れマルコフモデルによるモデル化を行い、音素と呼ばれる文字のような単位をまず認識していた。ここが深層学習、具体的にはMLPによって代替されたというのが最初2011年に起きた事件であり、その後CNNやRNN、近年ではTransformerと言う形で他のモダリティデータと同様にネットワーク構造の変遷が起きている。音声合成も深層学習によって研究が大きく進んだタスクで、画像生成に似たアプローチとして自己回帰モデルを利用したもの[29]やGANによる敵対的学習を取り入れたもの[30]などがある。

もう一つクロスモーダル理解の研究として挙げられるのが音声翻訳である。与えられた音声を自然言語として理解するのみならず、別の言語に翻訳するタスクで、旅行先での翻訳など応用への期待も高い課題である。単純に考えるとまず音声認識を経て同じ言語の音声から自然言語への変換を済ませ、その次に自然言語データ上で別の言語に更に変換するという2段階の変換アプローチが思いつく。実際そのような手法を提案している論文[31]が歴史的に大半である中で、ある言語の音声から直接別の言語の自然言語を出力するようなネットワークを学習するアプローチ[32]も出現している。

5. 画像と音声を繋ぐ研究

特に音声コミュニケーションは人々の間のやり取りにおいて重要な役割を果たしており、その視覚との関係を分析するために心理学や情報学など様々なアプローチからの研究が行われている。人間は視覚と聴覚を組み合わせて事象を認識しているという心理学的知見があり、有名な例としてはマガーク効果がある。例えば「ババ」という音と「ガガ」と発音している口の動きの人の動画を組み合わせて流すと、被験者には「ダダ」に聞こえるという現象である。音だけを聞いていれば正しく認識できても、音と矛盾する視覚情報が加わると、その矛盾を解消するために認識が変わってしまうのである。

5.1 動画像認識

動画像認識とは、動画に含まれる動作、物体、シーンなどの内容に基づいて、動画の特定の時間間隔や領域を何らかのクラスに分類するタスクである。動画像というと時々刻々のフレームで構成される時系列データだけを指し、音を含まない場合もある。特に人が映った動画での動作認識などは、画像データだけでもうまく識別出来てしまう。逆に画像と音声が相補的に精度向上に貢献する課題としては、この節の冒頭のように音声認識を音声だけでなく喋っている顔の動画も組み合わせて行うものがある[33]。

5.2 視覚補助つき音源定位

視覚補助付き音源定位（Visually Guided Sound Source Localization）とは、主に動画像中の音源の位置を特定する作業である[34]。もともと伝統的に音響データで行う音源定位の研究が続けられており、複数の場所から観測された音の時系列データを用いて音源の位置を推定することが目的となっている。視覚的音源定位の出力は、映像と音が時系列的に同期して記録されている中で、映像の各フレームの中でそれぞれの音が発生している領域を切り出すという意味で、画像上のアテンションマップやセグメンテーションに似ていると言える。

5.3 視覚補助つき音源分離

視覚補助つき音源分離（Visually Guided Sound Source Separation）とは、主に映像を対象として音の複雑な組み合わせをそれぞれ音の要素に分解する作業である[35]。例えば、複数の楽器が同時に演奏されている音のデータから、各楽器の音のデータを復元するのが目的となる。これを音のデータだけで行うものは音源分離と呼ばれ、統計的なアプローチや機械学習によるアプローチなど、長い間の発展の歴史をもつ。視覚補助つき音源分離は、視覚的な情報である映像を参照しながら音声データの音源分離を行う課題と捉えることができる。

5.4 音と映像の相互変換

音と映像の相互変換とは、音から映像あるいは映像から音を推定するクロスモーダル変換の研究である。人間の音声を対象としたもの[36]と、楽器の演奏を対象としたもの[37]に大別される。

6. むすび〜より多様なモダリティを繋ぐ研究へ

ここまで見てきた2つのモダリティを組み合わせる研究以外にも、多様なモダリティの組み合せの研究が存在す

る。まずはここまで扱ってきたモダリティを全て組み合わせる課題である。例えばテキストによる動画検索で、画像の時系列情報以外に音声も活用してより精緻に動画を検索しようとする研究[38]や、動画の音声などを活用しながらその内容を記述する研究[39]が挙げられる。

また一言で表現すると例えば「視覚」などの単一のモダリティになってしまうものを、複数の形態の画像など敢えてマルチモーダル表現されたデータとして同時に扱うことでクロスモーダル理解を実現する研究もある。たとえばRGB 画像データと赤外光を組み合わせた物体領域分割[40]や、3 次元の点群データと（2 次元の）RGB 画像データを組み合わせた物体検出[41]などがある。

ただ注意したいのが、単純にモダリティを増やせば増やすほどうれしいことばかりではない点である。一つ目の問題は、データの収集の困難さである。当然ではあるが、扱うモダリティが揃ったデータを収集しようとすると、モダリティの種類が増えるほどそのコストも増える。画像キャプション生成であれば画像とキャプションのペアデータを収集すればよいが、テキストによる動画検索で音声も扱おうとすると、テキスト・動画像・音声の 3 つのモダリティが同時に揃ったトリプレットデータが必要になる。

もう一つの問題が、モダリティを組み合わせたときに単一のモダリティよりもタスクの精度が上がる場合だけとは限らない、と言う点である。例えば動画像認識であれば画像列データだけの方がやや高精度だったり、マルチモーダル機械翻訳であれば自然言語データだけの方が高精度だったりすることが挙げられる。単一モダリティでもある程度の精度が出てしまうようなタスクの入力をマルチモーダルにした場合には、この問題がつきまとう。最近はこれに対応するために勾配ブレンディングと言う手法が提案されている[42]。簡単に言えば、収束が進んだモダリティの損失関数はどんどん重みを下げて、まだ収束していないモダリティの損失関数を下げる様な働きをする手法である。種々のタスクでの効果が確かめられており、上記の点群とRGB画像から物体を検出する研究[41]でも採用されている。クロスモーダル理解に対する機械学習の解析的な議論を含めた研究はまだまだ開拓の余地があるように思われるので、発見的なアプローチだけにとらわれず、数式を操作しながらクロスモーダル理解を解き明かす研究が進むことにも期待している。

本稿ではクロスモーダルな理解の研究として、自然言語処理/画像/音声を繋ぐ研究課題について概観した。エンコーダ・デコーダと言う基本的な概念について触れたあと、上記の 3 つのモダリティのうち 2 つを繋ぐ研究についてそれぞれ解説した。画像と自然言語を繋ぐ Vision & Language と呼ばれる分野の研究が一番多く、この分野のサーベイは日本語のものも散見されるようになってきた（英語は言わずもがなである）。一方で音声や他のモダリティまで含めたサーベイはあまり見かけない。本稿がより多くの研究者にとって、クロスモーダルな理解に興味をもつきっかけとなれば幸甚である。

参考文献

1）Seide, F. et al. “Conversational Speech Transcription Using Context-Dependent Deep Neural Networks”. Conference of the International Speech Communication Association. 2011, p.437–440.
2）Kriuzhevsky, A. et al. “ImageNet Classification with Deep Convolutional Neural Networks”. Advances in Neural Information Processing Systems. 2012, p.1097–1105.
3）Sutskever, I. et al. “Sequence to Sequence Learning with Neural Networks”. Advances in Neural Information Processing Systems. 2014, p.3104–3112.
4）Inoue, N.; Shinoda, K. “A Fast MAP Adaptation Technique for GMM-supervector-based Video Semantic Indexing Systems”. ACM Multimedia. 2011, p.1357–1360.
5）Ushiku, Y. et al. “Efficient Image Annotation for Automatic Sentence Generation”. ACM Multimedia. 2012, p.549–558.
6）Farhadi, A. et al. “Every Picture Tells a Story: Generating Sentences from Images”. European Conference on Computer Vision. 2010, p.15–29.
7）LeCun, Y. et al. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE. 1998, Vol.86, No.11, p.2278–2324.
8）Vinyals, O. et al. “Show and Tell: A Neural Image Caption Generator”. IEEE Conference on Computer Vision and Pattern Recognition. 2015, p.3156–3164.
9）Bahdanau, D. et al. “Neural Machine Translation by Jointly Learning to Align and Translate”. International Conference on Learning Representations. 2015.
10）Xu, K. et al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. International Conference on Machine Learning. 2015, p.2048–2057.
11）Bigham, J. P. et al. “VizWiz: Nearly Real-time Answers to Visual Questions”. ACM Symposium on User Interface Software and Technology. 2010, p.333–342.
12）Antol, S. et al. “VQA: Visual Question Answering”. IEEE International Conference on Computer Vision. 2015, p.2425–2433.
13）Shih, K. J. et al. “Where To Look: Focus Regions for Visual Question Answering”. IEEE Conference on Computer Vision and Pattern Recognition. 2016, p.4613–4621.
14）Anderson, P. et al. “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”. IEEE Conference on Computer Vision and Pattern Recognition. 2018, p.6077–6086.
15）Calixto, I. et al. “Images as Context in Statistical Machine Translation”. Workshop on Vision and Language. 2012.
16）Hitschler, J. et al. “Multimodal Pivots for Image Caption Translation”. Meeting of the Association for Computational Linguistics. 2016, p.2399–2409.
17）Mansimov, E. et al. “Generating Images from Captions with Attention”. International Conference on Learning Representations.

2016.
18）Reed, S. et al. “Generative Adversarial Text to Image Synthesis”. International Conference on Machine Learning. 2016, p.1060–1069.
19）Zhang, H. et al. “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks”. IEEE International Conference on Computer Vision. 2017, p.5907–5915.
20）Ramesh, A. et al. “DALL·E: Creating Images from Text”. https://openai.com/blog/dall-e/, (参照 2021–09–01).
21）Radford, A. et al. “Learning Transferable Visual Models From Natural Language Supervision”. International Conference on Machine Learning. 2021, p.8748–8763.
22）Das, A. et al. “Visual Dialog”. IEEE Conference on Computer Vision and Pattern Recognition. 2017, p.326–335.
23）de Vries, H. et al. “GuessWhat?! Visual Object Discovery Through Multi-Modal Dialogue”. IEEE Conference on Computer Vision and Pattern Recognition. 2017, p.5503–5512.
24）Saha, A. et al. “Towards Building Large Scale Multimodal Domain-Aware Conversation Systems”. AAAI Conference on Artificial Intelligence. 2018, p.696–704.
25）de Vries, H. et al. Talk the Walk: Navigating New York City through Grounded Dialogue. arXiv. 2018, 1807.03367v3.
26）Anderson, P. et al. “Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments”. IEEE Conference on Computer Vision and Pattern Recognition. 2018, p.3674–3683.
27）Lu, J. et al. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”. Advances in Neural Information Processing Systems. 2019, p.13–23.
28）Devlin, J. et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. Conference of the North American Chapter of the Association for Computational Linguistics. 2019, p.4171–4186.
29）van den Oord, A. et al. WaveNet: A Generative Model for Raw Audio. arXiv. 2016, 1609.03499v2.
30）Kumar, K. et al. “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis”. Advances in Neural Information Processing Systems. 2019, p.14910–14921.
31）Vidal, E. “Finite-state speech-to-speech translation”. IEEE International Conference on Acoustics, Speech, and Signal Processing. 1997, p.111–114.
32）Jia, Y. et al. “Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model”. Conference of the International Speech Communication Association. 2019, p.1123–1127.
33）Ngiam, J. et al. “Multimodal Deep Learning”. International Conference on Machine Learning. 2011, p.689–696.
34）Senocak, A. et al. “Learning to Localize Sound Source in Visual Scenes”. IEEE Conference on Computer Vision and Pattern Recognition. 2018, p.4358–4366.
35）Hershey, J. et al. “Deep clustering: Discriminative embeddings for segmentation and separation”. IEEE International Conference on Acoustics, Speech and Signal Processing. 2016, p.31–35.
36）Prajwal, K. R. et al. “Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis”. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, p.13796–13805.
37）Owens, A. et al. “Visually Indicated Sounds”. IEEE Conference on Computer Vision and Pattern Recognition. 2016, p.2405–2413.
38）Gabeur, V. et al. “Multi-modal Transformer for Video Retrieval”. European Conference on Computer Vision. 2020.
39）Hori, C. et al. “Attention-Based Multimodal Fusion for Video Description”. IEEE International Conference on Computer Vision. 2017, p.4193–4202.
40）Ha, Q. et al. “MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes”. IEEE/RSJ International Conference on Intelligent Robots and Systems. 2017, p.5108–5115.
41）Qi, C. R. et al. “ImVoteNet: Boosting 3D Object Detection in Point Clouds with Image Votes”. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, p.4404–4413.
42）Wang, W. et al. “What Makes Training Multi-modal Classification Networks Hard?”. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, p.12695–12705.

執筆者紹介

牛久 祥孝　USHIKU Yoshitaka
オムロン サイニックエックス株式会社
リサーチアドミニストレイティブディビジョン
専門：コンピュータビジョン、自然言語処理、パターン認識、機械学習
所属学会：ACM、IEEE、電子情報通信学会、情報処理学会、日本ロボット学会、人工知能学会、応用物理学会、建築情報学会
博士（情報理工学）

本文に掲載の商品の名称は、各社が商標としている場合があります。