受験番号	※記入不要
氏　名	[redacted]
現在の専門	機械学習
希望研究室	自然言語処理学研究室

小　論　文
（情報科学区分）

取り組みたい研究テーマ： 機械学習によるマルチモーダル環境における自然言語，および概念の獲得

1.　はじめに

私が奈良先端科学技術大学院大学で取り組みたい研究テーマは「機械学習によるマルチモーダル環境における自然言語，および概念の獲得」である．本稿では，私のこれまでの修学内容と，研究テーマについて，最後に NAIST を志望する理由について述べる．

2.　これまでの修学内容

私は，現在の研究室で「マルチモーダルなタスクにおけるグラフニューラルネットワークの適用可能性」について研究している．今は特に画像データについて，文章や音声といった他のドメインでも解釈，および応用が可能な潜在表現についての研究に力を入れ，実験と考察を重ねている．

また，論文読解や既存手法の実装を積極的に行っている．

2.1.　背景

マルチモーダルとは，文章と画像に限らず，音声や動画を含む複数の情報（モダリティ）を同時に扱うことが可能であるものを意味する．マルチモーダルなモデルを用いることで，従来の 1 種類の情報しか扱えない機械学習モデルよりも頑強で多様なタスクを同一モデルで扱えるようになる利点がある．また，入力が異なる情報を統合的に扱うことができる仕組みは，人間の動作を模倣する人工知能を作る上でも重要だと考えられる．

私は，このようなマルチモーダル推論において用いられる潜在表現について，グラフ表現などを用いることでより柔軟でデータ形式に依存しない表現が可能になるのではないかと考え，研究を行っている．

2.2.　研究目的

現在の私の研究目的は，マルチモーダルな機械学習に向けたより良い潜在表現の構造を提案することである．

2.3.　先行研究

従来のマルチモーダルな AI には様々なものがあるが，そのほとんどは既存の 1 種類のモダリティを扱うモデルを組み合わせたものとなっている．

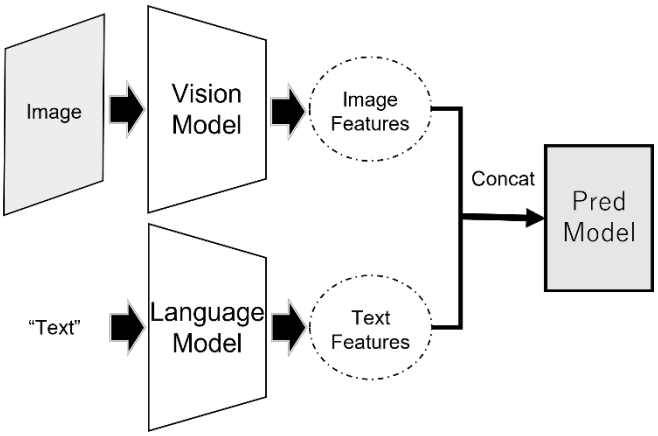

図 1　MDETR の構造

2021 年に公開された MDETR [1] は，画像中の物体検出モデルである DETR と，自然言語処理モデルである RoBERTa という 2 つのモデルを組み合わせたマルチモーダル推論モデルである．

MDETR は，物体検出モデルをベースとして，入力画像からバウンディングボックスを学習および推論する過程で，画像の潜在表現に，事前学習済みの言語モデルから得られた説明文の潜在表現を結合させている．このときに用いる潜在表現は画像，文章ともにベクトルのシーケンスであり，結合によりベクトルの単一シーケンスとする．これを，バウンディングボックスを推論するモデルへの入力として扱っている．

2.4.　課題と解決方法

従来の深層学習では，入力された情報について，画像や文章などモダリティごとに異なる手法でベクトル化し，シーケンス構造などに変換した上で，それぞれのモダリティに特化したモデルで処理をしていた．しかしこの手法では，モダリティごとに得られた情報の相互関係を十分に学習することは難しい．また，構造が異なるモデルを統合することはモデルの複雑さを高め，計算量が増大する．

そこで，潜在表現をグラフ構造にすることで，より良い学習が可能ではないかと考えた．

グラフ構造はデータの関係を直接表すことが可能であり，シーケンス構造などよりも柔軟な表現が可能であると予想される．

受験番号	※記入不要
氏　名	[redacted]
現在の専門	機械学習
希望研究室	自然言語処理学研究室

小　論　文
（情報科学区分）

例えば VisionGNN [2] は，グラフ構造の柔軟性に着目した画像処理手法である．画像内のデータ点をノードに見立て，ノード間の関係をもとに画像をグラフへ変換し，GNN で学習をしている．

この手法では，画像認識と物体検出のタスクにおいて従来手法と比較して優位な結果が得られたと報告している．

さらに，知識間の関係をグラフにした知識グラフや，画像中の物体間の関係をグラフにしたシーングラフのように，データ間の関係性をグラフで表現する技術が知られている．これらのことからも，グラフ構造は直感的で解釈しやすい表現であると考えている．

以上の考えから，「マルチモーダルなタスクにおけるグラフニューラルネットワークの適用可能性」として現在研究中である．

3.　貴学において取り組みたい研究内容

取り組みたい研究内容は「機械学習によるマルチモーダル環境における自然言語，および概念の獲得」である．以下に，概要を述べる．

3.1.　背景

近年，大規模言語モデルの急速な発展により機械学習への期待が一段と高まっている．ファインチューニングにより，大規模言語モデルを活用したソフトウェアでは文章に限らず画像や音声を扱うようになり，さらにはゲームやパソコンの操作といったタスクもこなせるように進化している．

一方で，このような大規模言語モデルベースの手法は，あくまでも生成モデルであり，人間のように能動的な行動ができるとは言い難い．

この理由の一つは，扱うデータをモダリティごとに異なるアーキテクチャで，異なる表現をしているためだと考えられる．

また，このような大規模言語モデルが性能を発揮している要因として，自然言語そのものにも着目している．事象を伝達するための方法として自然言語が発生したのならば，言語の成り立ちはより良い潜在表現を考える上で重要であると考えられる．

3.2.　課題と解決方法

これまでの修学内容の項でも述べたように，既存のマルチモーダルなモデルは，異なるモダリティで共通の潜在表現を得ることはできない．例えば，画像中の「犬」と説明文中の「犬」，それを読み上げた音声の「いぬ」は，それぞれが同じものを指していたとしても，モデル内では全く異なる潜在表現で扱われてしまう．

これらを解決する手法として，時系列や周辺の状況を含めた情報を入力した際に，モダリティによらず，同じ事象は共通の潜在表現となるような学習モデルを検討する．

先行研究のように，それぞれのデータに特化したモデルによる潜在表現を結合して扱うのではなく，例えば異なるモダリティ間の入力において対照学習による自己教師あり学習を行う，あるいはグラフなどの共通表現をあらかじめ用意し，それを入力により更新していくような学習モデルが考えられる．

異なるモダリティで共通の潜在表現を獲得できれば，それはすなわち知識を獲得したことになり，その表現がグラフなどの解釈しやすい形状であれば，いわゆる説明可能な AI，そして汎用人工知能を開発する上での一歩にもなると考える．

貴学には，理化学研究所をはじめとする外部の強力な機関と連携し，マルチモーダル認識に関する最先端の研究環境が整っている．『ChaSen』をはじめとした有名な自然言語処理ツールを開発した研究室があるなど，自然言語処理に関して全国でも有数のノウハウを持っているといえる．

最先端の研究環境や知見のもと，上述した機械学習モデルを実現するため，貴学でこの研究に取り組むことを志望する．

4.　まとめ

これまでの修学内容と現在の研究内容，NAIST において取り組みたい研究内容について述べた．

挑戦的な研究内容であるため，入学後は，実現可能性や教員，学生の知見も踏まえ，研究を進めていきたい．

参考文献

[1] A.Kamath et al., “MDETR – Modulated Detection for End-to-End Multi-Modal Understanding,” ICCV, 2021, oral.

[2] K. Han et al., “Vision GNN: An Image is Worth Graph of Nodes”, NeurIPS, 2022.