画像・言語モデルに関する順序数の的確な把握と活用能力の調査

増田　琉斗†　　宮森　　恒†

† 京都産業大学情報理工学部情報理工学科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2054227,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では、画像・言語モデルが、順序数の概念を的確に把握し活用する能力をどの程度有するのかについて調査する。NLP 分野における Transformer ベースの事前学習済みモデルについては、四則演算等の単純な算術問題を比較的高い正答率で解けることが知られているが、モデルが数の概念をどのように捉え、どのように活用しているのかについては不明な点も多い。本研究では、数の概念の一つである順序数に焦点をあて、画像・言語のモデルが順序数の概念を把握し活用する能力について調査する。具体的には、順序数を用いた的確な数え上げ等が必須となるプロービングタスクを課すことで、順序数に対するモデルの把握・活用能力を分析する。

キーワード　順序数, 概念理解, 画像・言語モデル, 数え上げ, 推論

1　はじめに

Transformer [1] は、ニューラル機械翻訳のモデルとして提案されて以降、自然言語処理分野における基盤的構造の一つとして多くのモデルで活用されている。特に、BERT をはじめとする事前学習済みモデルで採用されてからは、ほぼ全てのモデルで基盤構造として採用され、多くの自然言語処理タスクに対して高い性能を示している。

現在は、自然言語処理だけでなく、画像や音声といった他のメディア処理においても有用な共通的基盤技術として様々な分野で活用されている。例えば、Vision Transformer では、画像を分割した画像パッチを自然言語処理の単語のように扱い、Transformer エンコーダで埋め込むことで、畳み込みを使わずにそれまでの SOTA を達成している [2] [3] [4]。さらに、画像処理と自然言語処理を融合的に扱う Vision-and-Language 分野においても Transformer に基づくモデルが多数提案され、活発に研究が進められている。

一方、モデルが数を適切に扱う能力は、多くの複雑な推論タスクを遂行する上で重要である。例えば、人とロボットとのインタラクションにおいて、人間からの自然言語指示に数の概念理解が必要な内容が含まれる場合、ロボットは、実世界の特定物体や数の概念を的確に関連付けた上で指示内容に従わないと、指示を正しく遂行できず、場合によっては致命的な事故につながる可能性もあると考えられるためである。

Transformer に基づく言語モデルには、数値推論など数値を扱う能力があることは従来研究で知られている [5]。特に、四則演算等の単純な算術問題においては高い正解率を示している。一方で、扱う数の桁数が 15 桁や 20 桁と増えるにつれて正解率が低下することも示されており、限界があることがあることも明らかとなっている。また、単純な算術問題の場合においても数の概念を理解した上で解いているかどうかは不明である。

一般に、自然数の概念は、順序数、基数、数量の 3 種類で捉えることができるとされている [6]。順序数は、1 番目、2 番目、... といった数え上げにより何番目かを表す数である。基数は、ものの集まりの大きさを表す個数を表す数であり、数量は、ある測定単位のいくつ分かとして捉える、単位を伴った測定結

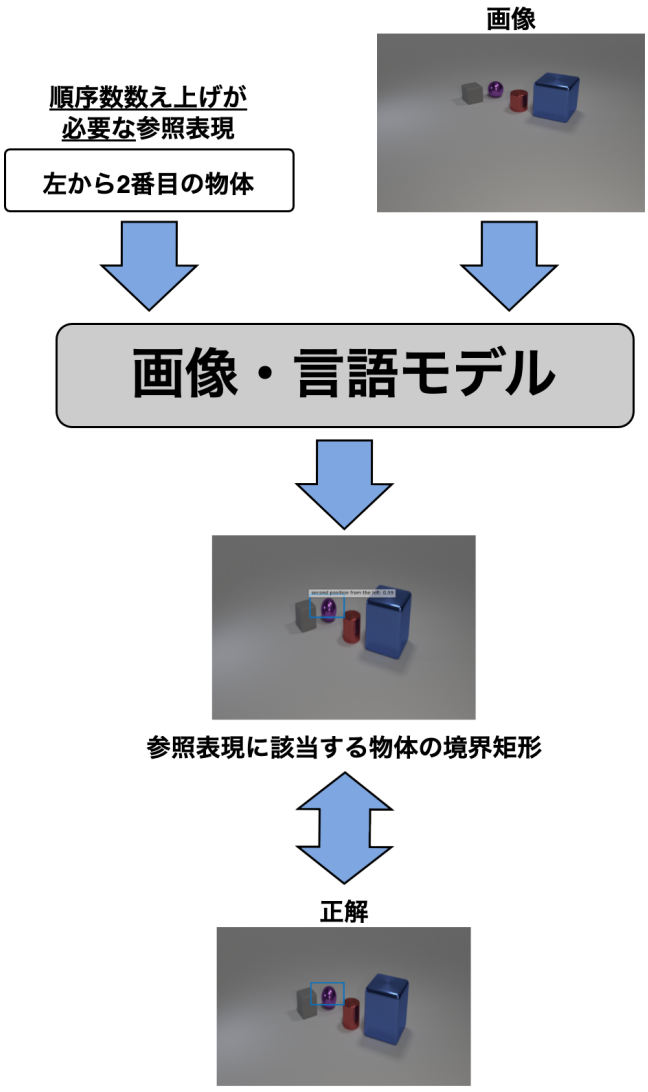

図 1　順序数による数え上げを要する参照表現理解タスク

果としての量を表す数である。

質問応答の従来研究では、基数や数量に関するものが多く、順序数に関する問題はほとんど取り組まれていない。基数は質問や参照データ中に出現する数値をそのまま使用することで計算可能である場合がほとんどであるが、順序数はそうではなく、数え上げや基数に変換するという処理が必要とされることが多く、順序情報を的確に把握していなければ順序数に関する問題は正しく解くことができない。そのため、順序数に関する問題を解くことができれば、自然数を何らかの実体と対応づけながら数え上げることができている可能性があり、順序数の概念を人間と近い形で理解しているのかどうかに迫るための新たな知見が得られるのではないかと考えた。

上記を踏まえ、本稿では、Transformer に基づく画像・言語モデルが、順序数の概念をどの程度的確に把握し活用する能力があるのかを明らかにすることを目的としたプロービングタスクを提案する。具体的には、いくつかの物体が写った画像に対し、順序数を用いた的確な数え上げが必須となる参照表現理解タスクとして設計する。CLEVR [14] データセットで提供されている画像集合から、物体間の順序に曖昧性が見られない画像を収集し、それら画像に対して、順序数を用いた数え上げが必要となる参照表現を付与することでデータセットを構築する。本稿では、順序数を用いた 1 つの数え上げを遂行するには、数え上げ対象の範囲と粒度、および、数え上げの起点と向きをそれぞれ把握する必要があると考える。これら 4 つの項目 (範囲、粒度、起点、向き) を把握することで正しく遂行できる参照表現を、数え上げスキル 1 の参照表現と捉えることとする。数え上げスキル数が複数の参照表現を段階的に用意することで、難易度の異なる数え上げに対するモデルの遂行能力を調査することができると考えられる。構築した参照表現を、図 1 に示すように画像・言語モデルに与え、参照表現理解タスクの推論結果を評価することで、モデルがどの程度順序数を的確に把握し活用できるかを分析する。

実験では、画像・言語モデルとして MDETR モデル [7] を対象に調査する。MDETR モデルは、画像とテキストを入力とし、それぞれから抽出した特徴を共通の埋め込み空間に投影し、Transformer のエンコーダデコーダを介することで、テキスト中の参照表現に対応する画像中物体の境界矩形を出力するモデルである。視覚的質問応答 (VQA) への拡張が容易で、参照表現理解タスクにおいても競争力のある性能を達成している。本稿では、参照表現理解のためのデータセットである CLEVR-Ref+ [15] でファインチューニングした MDETR モデルを用いる。また、評価に用いるデータセットとしては、数え上げスキル数を 1 から 3 の範囲で変化させた参照表現を用いる。数え上げスキル数や数え上げ対象の範囲の把握方法の違いなどによる正解率の変化を調査する。

実験の結果、与えた参照表現に対するモデルの正解率は、必要な数え上げ数が増えると増加する傾向にあることが示された。また、起点と向きの指定は、指定なしの場合より指定ありの方が正解率が高くなる傾向にあることがわかった。数え上げ対象の範囲については、所与の場合より要推論の場合の方が、また、画像中の物体の属性が多様となるほど、正解率は低下する傾向にあることが確認できた。

本稿の貢献は以下の通りである。

(1) 画像・言語モデルが、順序数の概念をどの程度的確に把握し活用できるかを調査するための参照表現理解タスクに基づくデータセットを構築した。

(2) 画像・言語モデルの一つ MDETR に対して、構築したデータセットを用いたプロービングを実施し、予備的な検証を行った。

(3) 順序数を用いた数え上げの能力について、必要スキル数や数え上げ対象の範囲の把握方法の違いなどによる正解率の変化を明らかにした。

本論文の構成は、以下の通りである。2 節で関連研究について述べ、3 節で本稿で扱う問題設定とデータセットの構築について詳述する。4 節では、本稿で対象とする画像・言語モデルについて説明する。5 節で実験内容と実験結果、考察を示し、6 節では、結論と課題を述べる。

2 関 連 研 究

2.1 数を理解し扱う能力に関する研究

ニューラルネットワークが数を理解し扱う能力に関する研究は、これまで多くなされてきた。Wallace ら [8] は、標準的なトークン埋め込み方式を用いて、数のリストからの最大値選択、数の埋め込みからのその数値のデコード、2 つの数の埋め込みからの和の計算といった各タスクを調査し、GloVe や BERT などの標準的な埋め込みは数を扱う能力が高いことを示した。青木ら [9] は、問題に応じて層の深さを適応的に変化させるモデル PonderNet が多段の推論を要する数量推論タスクに対してどのように作用するかを検証し、形式言語によって、構成される数量推論データセットの中の簡単な演算に対しては、非適応的なモデルと適応的なモデルでは大きな差は見られず、適応的なモデルの層の深さは入力に応じて変化することを示した。これらの研究では、ニューラルモデルの数を理解し扱う能力について主に四則演算を用いて検証を行なっていたが、本研究では、順序数に着目した段階的な参照表現を構築し用いることで、モデルの順序数に対する理解と活用能力を調査する。

2.2 Transformer における知識の保存形態や挙動に関する研究

Transformer における知識の保存形態や挙動に関する研究もこれまでに取り組まれている。有山ら [10] は、Transformer の知識ニューロンを探すためのタスクとして、穴埋め文の穴埋め部分を予測するタスクを言語モデルに解かせることで、概念についての知識が Transformer の FF 層の中で局所的にエンコードされていることを確認した。松本ら [11] は、四則演算を用いたプロービングにより、Transformer が再起的構造を捉えることができるかを検証し、Transformer が計算の途中結果

表 1　作成した参照表現の例と種別

数え上げスキル数	参照表現	範囲 [a]	粒度 [a]	起点の指定 [a]	向きの指定 [a]
1	2 番目に位置する物体	所与	揃っている	なし	なし
	左から 2 番目の物体	所与	揃っている	あり	あり
	偶数番目のうち 2 番目に位置する物体	要推論	揃っている	なし	なし
	奇数番目の物体のうち左から 2 番目の物体	要推論	揃っている	あり	あり
2	奇数番目の物体のうちの左から 2 番目の物体から数えて、左へ 2 番目の物体	所与	揃っている	あり	あり
3	左から 2 番目の物体と、奇数番目の物体のうち左から 2 番目の物体からなる並びにおいて、左から 2 番目に位置する物体	所与	揃っている	あり	あり

[a] スペースの都合上、2 つ以上の数え上げスキルを要する参照表現の種別については、代表例を 1 種類のみ示しており、スキルの 4 つの項目については最後の数え上げについての項目種別のみを表中に記載している

を内部的に保存できていることを示した。これらの研究では、Transformer の層の様子を知識帰属法やプロービング手法などを用いて観察し検証している。本稿では、Transformer に基づく画像・言語モデルに対して、順序数に着目した参照表現によるプロービングを行うことで、モデルの順序数に関する把握・活用能力について検証する。

2.3　参照表現理解に関する研究

参照表現理解に関する研究は古くからなされていたが、近年は大規模データセットが整備され、より本格的な活用のための研究が進められている。Kazemzadeh ら [12] は、実世界の自然画像中のオブジェクトに対する参照表現を 2 名のゲームを通して取得し、参照表現の理解や生成に活用できる最初の大規模データセットを構築した。Mao ら [13] は、CNN と LSTM で構成されるモデルを用いて参照表現を生成し、MS-COCO に基づく参照表現のための大規模データセットを構築した。Johnson ら [14] は、視覚的質問応答の推論能力を検証するため、自然画像による意図しないバイアスを軽減した CLEVR データセットを構築した。このデータセットでは、形状や色、材質が異なる複数の物体が無地のテーブル上に配置された CG 画像が用いられ、画像中の物体や状況に関する質問と正答との 3 つ組として提供されている。Liu ら [15] は、CLEVR データセットと同じ画像を用いて、各画像に様々な参照表現を付与し、参照表現理解の検証に活用できる CLEVR-Ref+データセットを構築した。本稿では、CLEVR-Ref+データセットで微調整した画像・言語モデルに対して、順序数に着目した段階的な難易度の参照表現を用いた分析を行う。

3　問 題 設 定

3.1　順序数を用いた数え上げを要する参照表現理解

本稿では、順序数を用いた数え上げを要する参照表現テキストと画像のペアを画像・言語モデルに与えることで、参照表現に対応する境界矩形 (Bounding Box) の集合が出力される参照表現理解タスクを考える (図 1)。

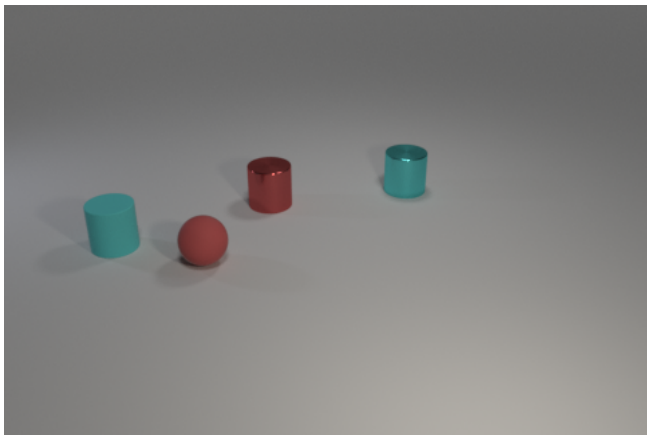

図 2　参照表現理解に用いる画像の例．この例は、個数が 4, 形状が aaab, 物体の大きさが全て同じに該当する．ここで、aaab は 3 つの物体が同じ形状で残り 1 つの物体が異なる形状であることを表す．

3.2　順序数を用いた数え上げを要する参照表現

参照表現としては、順序数を用いた数え上げが必要となるような内容で、その難易度が段階的に異なるようなテキストを作成する。ここで、参照表現理解は、1 つ以上の数え上げスキルを用いて解くことができると考える。ここで、1 つの数え上げスキルとは、数え上げ対象の範囲、数え上げ対象の粒度、数え上げの起点とその向きの 4 つを把握する能力であると定義する。例えば、図 2 に示す画像に対して「左から 2 番目の物体」という参照表現は、1 つの数え上げスキルを要する参照表現である。その数え上げ対象の範囲は、画像中で与えられたままの物体集合であり、数え上げ対象の粒度は 1 つ 1 つの物体把握に紛らわしい要素がなく揃っており、起点は数え上げ対象範囲の左端、向きは左から右向きである。当該数え上げを正しく遂行するには、これら 4 つの項目を適切に把握することが不可欠であると考えられる。

以下、数え上げスキルの各項目についてより詳細に説明する。

数え上げ対象の範囲は、その対象が画像中に与えられた状態のまま (所与) なのか、数え上げ対象を何らかの方法で選択したり並び替えたりしたものを推論する必要があるか (要推論) の 2 つに分けて考える。

図 2 に示す画像に対して、「左から 2 番目の物体」という参照表現は、数え上げを行う対象の範囲は、画像中で与えられた状態のままの物体集合である。一方、同じ画像に対して、「偶数番目のうち赤色の物体」という参照表現は、偶数番目にある物体のみが並んでいる画像は直接与えられている訳ではなく、そのような条件を満たす物体の並んでいる状況を推論する必要がある。

数え上げ対象の粒度については、1 つ 1 つの物体を紛れなく把握できるか、紛らわしい要素がある状態かでそれぞれ揃っている揃っていないとして分類する。本稿では、簡単のため、粒度は全て揃ったもののみを対象とする。

起点と向きについては、参照表現内で具体的な指示が与えられているか否かによってそれぞれ 2 種類に分類する。“左から 2 番目” や “右から 3 番目” などのように、数え上げの起点が “左” や “右” と指定されている場合は、起点の指示ありに相当する。同様に、“左から” や “右から” などのように、数え上げの向きについても指定されている場合は、向きの指示ありに相当する。「2 番目に位置する物体」という参照表現では、起点の指示なしに相当し、複数物体が真横に並んだ画像である場合は、多くの場合、左端を数え上げ起点とすることを把握する必要がある。

作成した各参照表現をこれらの視点で難易度別に分類した。その結果、合計で 84 種類となった。1 つの数え上げスキルを要する表現については 4 種類とした。その内訳は、数え上げ対象の範囲が所与の場合、(起点、向き) が (なし、なし)、(あり、あり) の 2 種類であり、要推論の場合、(起点、向き) が (なし、なし)、(なし、あり) の 2 種類である。2 つ以上の数え上げスキルを要する表現については、この 4 種類の要素の組み合わせで構成し、16 種類、3 つの数え上げスキルを要する表現については 64 種類とした。これら 84 種類の参照表現に対して、画像中の特定物体に正解が偏らないように、画像中の 3 つの異なる物体が正解となるように考慮することで、計 252 種類の参照表現を作成した。

表 1 に、本稿で作成した参照表現の例と種別を示す。なお、スペースの都合上、2 つ以上の数え上げスキルを要する参照表現の種別については、代表例を 1 種類のみ示しており、スキルの 4 つの項目については最後の数え上げについての項目種別のみを表中に記載している。

3.3 参照表現理解に用いる画像

本稿で参照表現理解に用いる画像は、CLEVR データセットで提供された画像から収集する。CLEVR データセットでは、自然画像の背景等に含まれる意図しないバイアスを軽減するため、形状や色、材質が異なる複数の物体が無地のテーブル上に配置された CG 画像が 75,000 枚用いられている。本稿では、簡単のため、物体が横方向に一列で並んだとみなせる画像を 6,030 枚収集した。収集する際には、画像中の物体の個数 (3 つか 4 つか)、同じ形状と異なる形状の物体が含まれる割合、同じ形状の物体の大きさ (同じ大きさか異なるか) のそれぞれの違いで絞り込み、全体を整理した。本稿で用いる画像の属性と枚数を表 2 に示す。

表 2　参照表現理解に用いる画像の属性と枚数

個数	形状	物体の大きさ	枚数
3	aaa	同	136
		異	0
	aab	同	1,797
		異	1,736
	abc	-	1,225
4	aaaa	同	10
		異	0
	aaab	同	115
		異	0
	aabb	同同	76
		同異	154
		異異	82
	aabc	同	348
		異	351

aaa は 3 つの物体が同じ形状であることを表し、
abc は 3 つの物体が全て異なる形状であることを表す。
同じ形状の物体の大きさが、
同: 全て同じ、異: いずれかが異なることを表す。

なお、CLEVR データセットの画像では、物体の色と材質についてもそれぞれ 7 種類、2 種類のバリエーションがあるが、本稿で用いる画像を収集する際にはこれら 2 つの要素は特に考慮していない。

図 2 に、本稿での参照表現理解に用いる画像の例を示す。この画像は、個数が 4、形状が aaab、物体の大きさが全て同じに該当する例である。ここで、aaab は 3 つの物体が同じ形状で残り 1 つの物体が異なる形状であることを表す。

4 対象とする画像・言語モデル

本稿では、画像・言語モデルとして MDETR [7] を対象に調査する。MDETR は、画像とテキストを入力とし、それぞれから CNN および RoBERTa で抽出した特徴を共通の埋め込み空間に投影し、Transformer のエンコーダデコーダを介することで、テキスト中の参照表現に対応する画像中物体の境界矩形を出力するモデルである。MDETR の処理概要を図 3 に示す。視覚的質問応答 (VQA) への拡張が容易で、参照表現理解タスクにおいても競争力のある性能を達成している。本稿では、参照表現理解のためのデータセットである CLEVR-Ref+でファイ

ンチューニングした MDETR モデルを用いる。

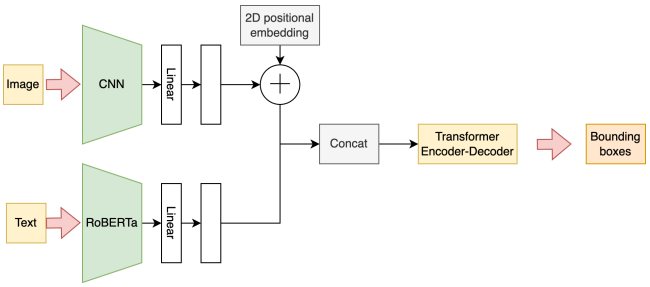

図 3　MDETR モデルの概観.

5　実　　験

5.1　実験目的

Transformer に基づく画像・言語モデルが、順序数の概念をどの程度的確に把握し活用する能力があるのかを明らかにする。

5.2　実験方法

評価対象となる画像・言語モデルは、CLEVR-Ref+でファインチューニングした MDETR モデルとする。

3 節で構築した、順序数を用いた数え上げが必要な参照表現 252 種類と、表 2 で用意した 11 種類の画像 10 枚ずつの組合せで MDETR モデルの出力結果の正解率をいくつかの観点で評価する。具体的には、以下の観点で正解率の変化を調べる。

(1)　数え上げスキル数に対する正解率

(2)　同じ数え上げスキルでの数え上げ対象の範囲の違いに対する正解率

(3)　同じ数え上げスキルでの起点・向きの指定の有無に対する正解率

(4)　同じ数え上げスキルでの画像種別の違いに対する正解率

正解の判定は、参照表現に対応すると判断された物体の境界矩形のうち、最上位の境界矩形が正解の境界矩形と少なくとも 0.5 の IoU を有するかどうかで測定する。

5.3　実験結果

数え上げスキル数に対する正解率の結果を図 4 に示す。数え上げ回数での比較を行った結果、1 回の場合は、平均正解率が 27.42%、2 回の場合は、平均正解率 34.32%、3 回の場合は、平均正解率 36.82%となった。スキル数が上がるほど、正解率が増加する傾向であることがわかった。

同じ数え上げスキルでの数え上げ対象の範囲の違いに対する正解率の結果を図 5 に示す。スキル数 1 では、所与より要推論の正解率が高く、スキル数 2,3 では要推論より所与の正解率が高い結果となった。全体としては、所与の平均正解率 33.03%、要推論の平均正解率が 28.52%であり、所与の正解率が要推論を上回る傾向であることがわかった。

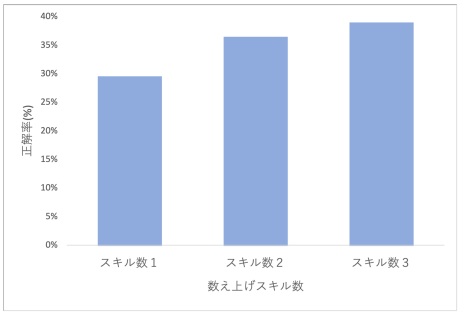

図 4　数え上げスキル数に対する正解率

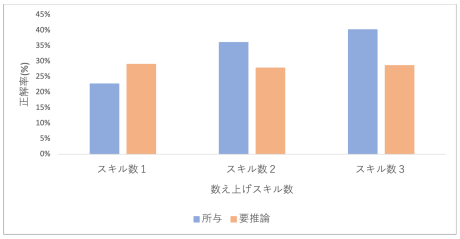

図 5　数え上げ対象の範囲に対する正解率

同じ数え上げスキルでの起点・向きの指定の有無に対する正解率の結果を図 6 に示す。スキル数 1 では、起点・向きの指定なしの場合より指定ありの場合の正解率が高く、スキル数 2,3 では、起点・向きの指定ありの場合より指定なしの場合の正解率が高い結果となった。全体としては、起点・向きを指定した場合の平均正解率 31.11%が、指定しない場合の平均正解率 30.45%を上回る傾向であることがわかった。

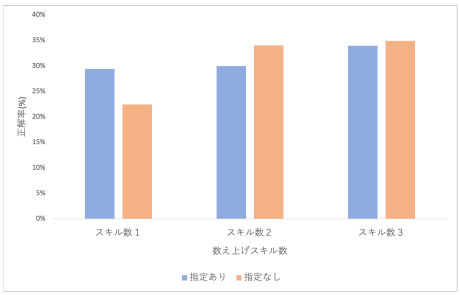

図 6　起点・向きの指定の有無に対する正解率

同じ数え上げスキルでの画像種別の違いに対する正解率の結果を図 7 に示す。画像中の出現物体数が増えるほど、また、形状や大きさが異なるものが増えるほど、正解率が低下する傾向が見られた。

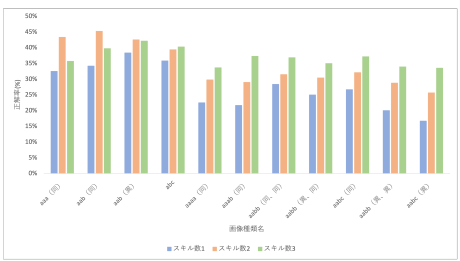

図 7　画像種別の違いに対する正解率

5.4 考　　察

スキル数が増加するにつれ、正解率が増加するという結果は、意外なものであった。今回、スキル数が増加するにつれて、参照表現の種類数も多くなっているが、画像中の物体数が 3 から 4 と少なく、正解となる物体に偏りがないか検証する必要がある。詳細な誤り分析を進めるのと並行して、スキル数が少ない場合の参照表現についても種類を増やすことを検討する。

所与での正解率が、要推論の場合の正解率を上回ったことから、モデルは数え上げ対象を並び替えるといった推論を行うことが困難であることが示唆される。ただ、今回は画像中の物体数が 3 から 4 と少なく、推論すべき状態の物体数がかなり限定的であったことも結果に関係している可能性がある。今後、物体数をより増やした画像を作成し、同様の比較実験を進めていく必要がある。

起点・向きを指定すると正解率が増加したことから、モデルは起点や向きが指定されていない場合、多くの場合左から数えるという人間の特性とは異なる数え方をしている可能性が示唆される。今回、指定ありの内容としては、“ 左から ” あるいは “ 右から ” のいずれかであるため、両者で正解率に違いがあるかなど分析を進めていく必要がある。

画像中の出現物体数が増えるほど、また、形状や大きさが異なるものが増えるほど、正解率が低下する傾向があることから、画像中の物体の属性が多様となるほど正しく遂行することは難しくなることが確認できた。一方で、今回は属性の組合せ数が限定的であったため、今後はより属性の組合せ数を増やした画像を用いた上で同様の分析を進めていく予定である。

6 ま と め

本稿では、Transformer に基づく画像・言語モデルが、順序数の概念をどの程度的確に把握し活用する能力があるのかを明らかにすることを目的としたプロービングタスクを提案した。具体的には、いくつかの物体が写った画像に対し、順序数を用いた的確な数え上げが必須となる参照表現理解タスクとして設計した。CLEVR データセットで提供されている画像集合から、物体間の順序に曖昧性が見られない画像を収集し、それら画像に対して、順序数を用いた数え上げが必要となる参照表現を付与することでデータセットを構築した。

実験の結果、与えた参照表現に対するモデルの正解率は、必要な数え上げ数が増えると増加する傾向にあることが示された。また、起点と向きの指定は、指定なしの場合より指定ありの方が正解率が高くなる傾向にあることがわかった。数え上げ対象の範囲については、所与の場合より要推論の場合の方が、また、画像中の物体の属性が多様となるほど、正解率は低下する傾向にあることが確認できた。より詳細な誤り分析を進める必要がある一方で、今回用いた画像中の物体数が限定的であることは、より一般的な活用能力を把握する調査としては十分でないとも考えられる。今後、これらを改善した形で同様の分析を行っていく必要がある。また、モデル自体の内部状態の挙動の分析や可視化についても進めていく予定である。

謝　　辞

本研究の一部は科研費 18K11557 の助成を受けたものである。

文　　献

[1] Ashish Vaswani et al. Attention Is All You Need, 2017.
[2] Kai HaH et al. A Survey on Vision Transformer, 2023
[3] Ze Liu et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, 2021
[4] Xu Hana et al. Pre-trained models: Past, present and future, 2021.
[5] David Saxton et al. ANALYSING MATHEMATICAL REASONING ABILITIES OF NEURAL MODELS, 2019.
[6] 平井安久 et al. 数の概念の捉え方について 2013.
[7] Aishwarya Kamath et al. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding, 2021.
[8] Eric Wallace et al. Do NLP Models Know Numbers? Probing Numeracy in Embeddings, 2019.
[9] 青木洋一 et al. 多段の数量推論タスクに対する適応的なモデルの振る舞いの検証, 2022.
[10] 有山 知希 et al. Transformer モデルのニューロンには局所的に概念についての知識がエンコードされている, 2022.
[11] 松本 悠太 et al. 四則演算を用いた Transformer の再帰的構造把握能力の調査, 2022.
[12] Sahar Kazemzadeh et al. Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP. pp. 787–798, 2014.
[13] Junhua Mao et al. Generation and comprehension of unambiguous object descriptions. In: CVPR, 2016.
[14] Justin Johnson et al. ”Clevr: A diagnostic dataset for compositional language and elementary visual reasoning.” Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
[15] Runtao Liu et al.CLEVR-Ref+: Diagnosing Visual Reasoning with Referring Expressions, 2019.