G-14

LLM+Vision システムの幾何問題解答能力向上のためのデータ拡張

小山 陸翔‡

1. はじめに

近年, LLM+Vision システムは教育や自動車など様々な分野に応用されはじめている. しかしながら, その精度には課題が残っており, そのうちの一つが画像内の情報を正しく抽出できていないといった問題である. これは現行の OCR 技術や画像エンコーダでは幾何問題にとって重要である頂点同士の関係やその接続関係といった構造情報や等辺や等角といった数学的な記号を正しく認識し, 抽出できないといったことが原因である.

そこで本研究では, 図として与えられる幾何問題に含まれる条件や構造的情報を, 正確にテキストとして表現する機械学習モデルの構築を試みる. このモデルにより, LLM＋Vision システムにおける幾何問題の理解を支援し, 最終的には正答率の向上を目指す. しかしながら, こうしたモデルの学習には図と対応する正確なテキスト情報を多数含むデータセットが必要であるが, そのような学習データは現状では十分に整備されておらず, データ不足が大きな課題となっている. そこで本研究ではまず、この課題に対処するため, 効果的なデータ拡張手法の検討を行う.具体的には, (1) データ拡張を行わないデータセット, (2) 汎用的な画像認識タスクで用いられる一般的なデータ拡張を施したデータセット, (3) 幾何問題の特性に着目して設計したデータ拡張を行ったデータセットの 3 種類を用いてモデルを学習し, それぞれのケースにおいて幾何問題をどの程度正確にテキスト化できるかを比較・評価する.

2. 関連研究

2.1 AutoGeo

Huang ら[1] は, 大規模で多様な幾何問題データセットを構築することを目的として, 数学的な幾何画像とテキスト説明を自動生成するパイプラインである AutoGeo を提案した. この AutoGeo では, 幾何学的な要素や操作を「Clause」という論理単位にまとめ，それぞれを難易度（easy / medium / hard）に応じて分類するとともに，Clause 間の依存関係をルールベースのアルゴリズムで整合的に選択していくことで，多段階的な幾何図形を構築する仕組みを備えている．さらに，選択された Clause 情報に基づいて自動的に座標や線分を決定し，Matplotlib によって図形を描画したうえで，LLM を活用してテキスト説明を生成・付与するため，画像と説明文の対応付けが容易であり，効率的なデータ拡張を可能にしている． しかしながら, 幾何図形は Matplotlib によって描画されているため, 実際のカメラで撮影された画像とはギャップが存在する. この点は生成されたデータセットの応用範囲を限定する可能性があり, 一つの課題と考えられる.

2.2. G-LLaVA

Gao ら[2] は，幾何学問題に特化した LLM+Vision(Image) システムの学習を目指し，G-LLaVA を提案した．G-LlaVA では既存の幾何問題に対して, 図形を詳細に説明する情報を逆算して生成する仕組みを導入し， 元来不足していた図形構造に関するテキストを補強している．さらに正しい関係と誤った関係を意図的に混在させる「コントラスト QA」を組み合わせることで， 幾何要素（点・線分・角度など）の関連性を厳密に学習させる戦略を用いている． 加えて, 問題文や解答に対し，数値置換やスケーリング， 条件の逆転， 文章表現のパラフレーズなど複数の拡張操作を適用することで単一の論理構造から多様なバリエーションを自動生成し， モデルが幾何学的推論手順をより汎用的に獲得できるように設計されている． しかしながら， この手法においても出力される幾何図形と実際のカメラで撮影された画像との間にはギャップが存在している.

3. 提案手法

3.1 ベースモデル：BLIP-2

本研究では， ベースモデルとして大規模視覚言語モデルである BLIP-2[3]を採用する． BLIP-2 の最大の特徴は， Q-Former と呼ばれる軽量な Transformer を介して、事前学習済みの凍結された画像エンコーダと大規模言語モデル（LLM）を効率的に接続する点にある． この構造により, コストの高い事前学習なしに既存の強力なユニモーダルモデルを活用して最先端の性能を達成する． 本研究では, BLIP-2 の強力な画像理解能力とテキスト生成能力を独自の図形データセットに効率的に適応させることを目指す.

3.2 LoRA (Low-Rank Adaptation)

大規模モデルの全パラメータを更新するファインチューニングは膨大な計算コストを要する． この問題に対処するため， 本研究ではパラメータ効率の良いファインチューニング手法である LoRA (Low-Rank Adaptation) を採用する. LoRA はモデル内の一部の層に低ランク行列を追加し， その行列のみを学習対象とすることで計算資源を大幅に削減しつつ高い性能を達成する． 本研究における LoRA の具体的な設定（適用層、ランク等）について述べる。

3.3 使用するデータ

本研究の対象データは中学校数学の教科書に掲載されている幾何図形をスマートフォンで撮影することで収集した画像 300 枚とその図を表すテキストとの組である. データセットの一例を表 1 に示す.

図の種類は対頂角や同位角, 錯覚などの基本的な角の関係に関する図, 平行線と角に関する図, 三角形の内外角に

‡ 関西学院大学, Kwansei Gakuin University

関する図, 四角形以上の多角形の内外角に関する図, 円周角や円に内接する多角形に関する図となっている.

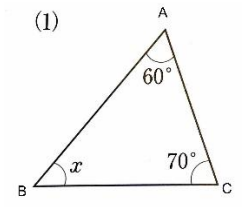
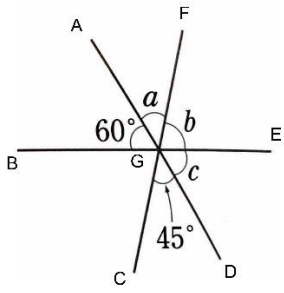
幾何問題の画像	図を表現したテキスト
	△ABC ∠BAC=60° ∠ACB=70° ∠ABC=x
	line(A,D) line(B,E) line(C,F) G=line(A,D)∩line(B,E)∩ line(C,F) ∠AGB=60° ∠CGD=45° ∠AGF=a ∠EGF=b ∠DGE=c

表 1 使用したデータセットの一例

4. 実験

4.1 実験の目的

本実験では, データ拡張を適用しないベースラインモデルと汎用的な画像処理である回転拡張のみを適用したモデルの性能を比較検証する. この比較実験は汎用的な拡張アプローチの有効性と限界を調査し, より高度な幾何問題に特化したデータ拡張手法の必要性を明らかにすること, またそのような手法のための知見を得ることを目的としている.

4.2 実験の設定

本実験は, 提案手法の有効性を客観的に評価するため以下の設定で行った.

・比較対象: 提案手法の有効性を明確にするため, 以下の 2 つの条件下でモデルを学習・評価し, その結果を比較する.

・ベースラインモデル : データ拡張を一切適用せずに学習したモデル

・ベースライン + 回転拡張 : 汎用的なデータ拡張の代表例として幾何学的変換である回転処理のみを適用したモデル. このモデルの学習時には各画像を入力するたびに{0°, 90°, 180°, 270°}の中からランダムに一つの角度が選択され、動的に回転が適用されるようにした.

・共通パラメータ: 両モデルともに実験の公平性を保つため, 以下の共通のパラメータを用いた.

・ベースモデル: Salesforce/blip2-opt-2.7b

・ファインチューニング手法: LoRA (ランク: 16, alpha: 32)

・学習プロセス: Blip2ForConditionalGeneration の言語生成部をオートレグレッシブクロスエントロピー損失で学習し, この損失を最小化するようにモデルを更新した. パラメータ最適化には学習率 5e-5 の AdamW オプティマイザ[4]を用い, LoRA による効率的ファインチューニングを採用したため更新対象は事前学習モデル本体ではなく, 追加した LoRA 層のパラメータのみに限定した.

・バッチサイズ: 4

・エポック数: 10 (各 Fold)

・評価方法: 評価の信頼性を担保するため, 5-Fold Cross-Validation を採用した.

4.3 評価指標

本研究では, データ拡張手法の有効性を定量的・定性的に評価する.

・定量的評価: 客観的な性能評価のため, 交差検証によって得られた**交差エントロピー誤差関数**を主要な指標として用いる. この値が低いほどモデルが未知のデータに対して高い汎化性能を持つことを示す.

・定性的評価 : 検証データからいくつか生成例を抽出し, モデル出力の傾向や生成パターンを比較・考察することで, 数値指標では捉えきれない出力の違いを分析する.

5. 結果と考察

本節では, 実験によって得られた定量的および定性的な結果を基にデータ拡張の有効性と限界について考察する.

5.1 定量的結果

データ拡張の有無がモデルの汎化性能に与える影響を比較するため, 両モデルの平均交差エントロピーを表 2 に示す.

モデル	平均交差エントロピー損失
ベースライン	0.5993
ベースライン+回転拡張	0.6202

表 2 データ拡張の有無による
平均交差エントロピー損失の比較

5.2 定性的結果

次に, 検証データに対してモデルが生成したテキストの具体例を示すことでその定性的な振る舞いを評価する. 本実験では各画像に対して最大 15 件のテキスト候補を生成させ, その内容を分析した.

ここでは代表的な一例として、表 3 に入力画像とその正解ラベルを示し, 続く表 4 には、この画像に対してベースラインモデルと回転拡張を適用したモデルがそれぞれ生成したテキストの代表例を示す.

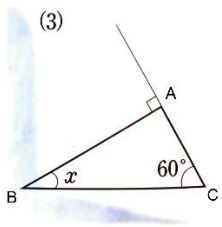
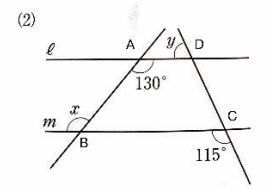
入力画像	正解ラベル
	△ABC ∠ACB=60° exte(∠BAC)=90° ∠ABC=x
	l=line(A,D) m=line(B,C) line(A,B) line(C,D) A=l∩line(A,B) B=m∩line(A,B) C=m∩line(C,D) D=l∩line(C,D) ∠(l,line(A,B))=130° ∠(m,line(C,D))=115° ∠(m,line(A,B))=x ∠(l,line(C,D))=y

表 3　入力画像と正解ラベル

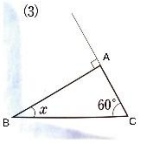
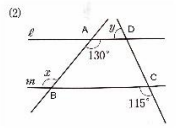
入力画像	ベースライン	ベースライン +回転拡張
	△ABC □ABCD □ABC ∠ABC=60° 以下略	△ABC △ABCD △ABE □ABCD 以下略
	△ABC l line(A,D)/line(B,E)/ line(C,F)/line(D,G) 以下略	△ABC △ABCD △ABD l 以下略

表 4　入力画像と各モデルの出力結果

定性的評価の結果においても, ベースラインモデルがより正解ラベルに近いテキストを出力していることが確認できる. ただし，依然として出力の精度には大きな改善余地が残されている.

5.3 考察

本節では 5.1 節, 5.2 節で示した実験結果に基づき、データ拡張が幾何問題解答タスクに与える影響について考察する.

5.3.1 汎化性能の低下

本実験において, 回転拡張を適用したモデルがベースラインモデルよりも低い汎化性能を示した. 本来, この拡張は「画像の向きが変わっても、その幾何学的な意味は不変である」という頑健性をモデルに学習させることが目的であった. しかしながら, 実験結果はこのアプローチが意図とは逆の効果をもたらした.「頂点と直線の関係などの幾何学的特徴」をノイズとして学習してしまったことが原因だと考えられる.

5.3.2 局所最適解

5.2 節で示した定性的な結果からモデルが画像の形状を無視して最も頻出するラベル「△」や「ABC」などを生成する強いバイアスを示したということがわかった. この結果からモデルが局所最適解に陥っていることが考えられる.

6. 結論と今後の展望

6.1 結論

本研究では, 大規模視覚言語モデルを幾何問題解答タスクに適応させるにあたり, データ拡張手法の有効性を検証した. 実験ではデータ拡張を適用しないベースラインモデルと汎用的な回転拡張を適用したモデルの性能を比較した.

その結果, 回転拡張を適用したモデルは, ベースラインモデルよりも検証損失が悪化するといった仮説とは逆の結果が得られた. さらに定性的な分析から, この拡張手法ではモデルが画像の幾何学的特徴を正確に解釈するのではなく, 入力画像の形状に関わらず最も頻出するラベルを出力するといった表面的なパターンに依存した学習を行ってしまうことが明らかになった.

以上の結果から, 本研究では回転をはじめとする汎用的なデータ拡張はこのタスクにおいて有効に機能せず, むしろモデルの学習を阻害し性能を低下させる可能性があると結論付ける.

6.2 今後の展望

今後の展望として, 回転以外の汎用的なデータ拡張が精度にどのような影響を与えるか実験するとともに, 幾何問題に特化したデータ拡張手法の構築に取り組む. この手法は本実験で明らかになった課題, すなわち, 画像の幾何学的関係とテキストラベルとの関係性を整合的に扱い, モデルを局所最適解から脱却させることを目的とする.

参考文献

[1] Huang, Z., T. Wu, W. Lin, S. Zhang, J. Chen, and F. Wu (2024). Autogeo: Automating geometric image dataset creation for enhanced geometry un derstanding. arXiv preprint arXiv:2409.09039.

[2] Gao, J., R. Pi, J. Zhang, J. Ye, W. Zhong, Y. Wang, L. Hong, J. Han, H. Xu, Z. Li, et al. (2023). G-llava: Solving geometric problem with multi-modal large language model. arXiv preprint arXiv:2312.11370.
[3] Li, J., D. Li, S. Savarese, and S. Hoi (2023). Blip 2: Bootstrapping language-image pre-training with frozen image encoders and large language mod els. International conference on machine learning, 19730–19742
[4] Loshchilov, I., and F. Hutter (2017). Decou-pled weight decay regularization. arXiv preprintarXiv:1711.05101.
[5] 中学校数学 2. 学校図書株式会社. 2009, 2012
[6] 中学数学 2. 日本文教出版. 2009, 2012
[7] 新版 中学校数学 2. 大日本図書. 2009, 2012
[8] 楽しさ広がる数学 2. 啓林館. 2009, 2012
[9] 新しい数学 2. 東京書籍. 2009, 2012
[10] 未来へひろがる数学 2. 啓林館. 2009, 2012
[11] 中学校数学 3. 学校図書株式会社. 2012
[12] 中学数学 3. 日本文教出版. 2012
[13] 新版 中学校数学 3. 大日本図書. 2012
[14] 楽しさ広がる数学 3. 啓林館. 2012
[15] 新しい数学 3. 東京書籍. 2012
[16] 未来へひろがる数学 3. 啓林館. 2012