Multimodal LLM を用いた適応的階層分割と段階的統合による ダイアグラム構造認識の検討

内藤美里[1] 宮本健[1] 三輪祥太郎[1]
[1]三菱電機株式会社
Naito.Misato@cb.MitsubishiElectric.co.jp
{Miyamoto.Ken, Miwa.Shotaro}@bc.MitsubishiElectric.co.jp

概要

フローチャートや ER 図のようなダイアグラムの構造認識では，ノードは形状・大きさが多様で局所的に密集する一方，エッジは図全体にまたがる長距離の接続経路を含むため，局所的な詳細理解と大域的な接続関係の把握を同時に行うことが求められる．このような異なるスケールの情報が混在することが，Multimodal LLM での単一視点での認識を困難にしている．本研究では，ノード密度とエッジの配置に応じて領域を柔軟に分割する「適応的階層分割」と，分割で切断された長距離エッジや重複ノードを階層的に復元する「段階的統合」を提案し，GPT-5.1 および GPT-4o による評価を通じて，分割戦略と認識戦略の組み合わせがダイアグラム構造認識性能に与える影響を分析する．

1 はじめに

Multimodal LLM（MLLM）の発展により，文書画像からの情報抽出や質疑応答が現実的になりつつある．ネットワーク構成図から資産情報を抽出する研究[1]や，科学図表の分析の高精度化[2]が進み，図面・ダイアグラムを解釈するニーズは高い．しかし，フローチャートや ER 図のように，多数のノードとエッジが複雑に入り組むダイアグラムでは，MLLM に画像をそのまま入力した場合，ノードやエッジの誤認識が発生する．これにより，実際には存在しない処理ステップやエンティティが検出されたり，本来接続していないノード間に誤った依存関係が導入されたりして，後段の質問応答などのタスクで誤った構造に基づく推論が行われてしまう．

このようなダイアグラムでは，ノードは形状・大きさが多様で局所的に密集する一方，エッジは図全体にまたがる長距離の接続経路を含むため，局所的な詳細理解と大域的な接続関係の把握を同時に行うことが求められる．単一視点で一括認識を行うと，ノード密集領域と長距離エッジの双方を同時に高精度で扱うことが難しい．

そこで本研究では，ノード密度に応じて領域を自動分割して局所的なノード・短距離エッジの認識を目的とする「適応的階層分割」と，分割で切断された長距離エッジや重複ノードを階層的に復元・統合する「段階的統合」を組み合わせることで，ノードとエッジの異質性を考慮したダイアグラム構造認識を目指す．実験では，ルールベースによる一律な画像分割との比較に加え，「階層全体を一括で認識させるフロー」と「階層ごとに段階的に統合するフロー」とを比較し，MLLM による分割戦略および分割後の認識戦略がダイアグラム構造認識性能に与える影響を分析する．

2 関連研究

金子ら[1]は，Segment Anything Model と k-means クラスタリングでネットワーク構成図を分割し，MLLM で資産情報を抽出する手法を提案している．通常入力や格子状分割より F1 と再現率が高いことを示し，適切な画像分割が MLLM の性能を向上させることを報告している．

Li ら[2]は，科学図表を対象に，画像処理と VLM を統合した分割・マージを行う Chain-of-Region を提案した．領域ごとに形状・意味情報を収集してから最終質問に答える逐次的戦略により，従来の VLM を大きく上回る精度を実現している．

本研究は，これらの方向性を踏まえつつ，ノードとエッジからなるダイアグラム構造認識に対して，SAM や複雑な形状検出は用いず，画像処理と MLLM を組み合わせて入力画像に対して柔軟な「適応的階層分割」を行い，その結果を「段階的統合」

によりボトムアップにマージする枠組みを提案する．さらに，「階層全体を一括で認識させるフロー」と「階層ごとに段階的に統合するフロー」を比較することで，MLLM における分割戦略と分割後の認識戦略の違いがダイアグラム構造認識性能に与える影響を明らかにする．

3 提案手法

3.1 問題設定

入力は，フローチャート，ER 図，マインドマップ図などのダイアグラム画像とする．矩形や円などで囲まれた図形をノードとみなし，それらを結ぶ線分をエッジとする．本研究ではエッジの向きやエッジラベルは扱わず，無向グラフとしてノード間の接続関係を抽出する．出力は，(1)ノード集合，(2)エッジ集合（ノード対の一覧），である．

3.2 適応的階層分割

入力画像を再帰的に 2 分割することで階層的な領域構造を形成する．概要を図 1 に示す．手順は次の通りである．

(1) ノード数推定による分割要否判定 ：各領域画像に対して，MLLM に「ノード数」と「閾値（本実験では 8）を超えるかどうか」を推定させる．ノード数が閾値を超える場合のみ分割を行うことで，ノード密度の高い領域ほど細かく分割される自律的な階層が得られる．

(2) 射影プロファイルに基づく候補線抽出：画像をグレースケール化し，反転した上で行方向・列方向の和を求め，水平・垂直の射影プロファイル（一次元の分布）を得る．このプロファイルに対してピーク検出を行い，値が小さい位置を分割候補線として抽出する．ただし，画像の端に近い候補線は除き，近接する候補線はマージして座標が最小の候補線のみ残す．

(3) 分割方向・位置の決定：分割候補線を重ねて描画した画像を MLLM に提示し，「ノード数のバランス」「エッジ切断の少なさ」などの観点から最適な 1 本を選択させる．選ばれた線に沿って画像を上下または左右に二分し，各部分画像に対して再帰的に同じ処理を行う．

このように，ノード数推定と投影プロファイル解析を組み合わせることで，図面構造に応じた柔軟な適応的階層分割を実現する．

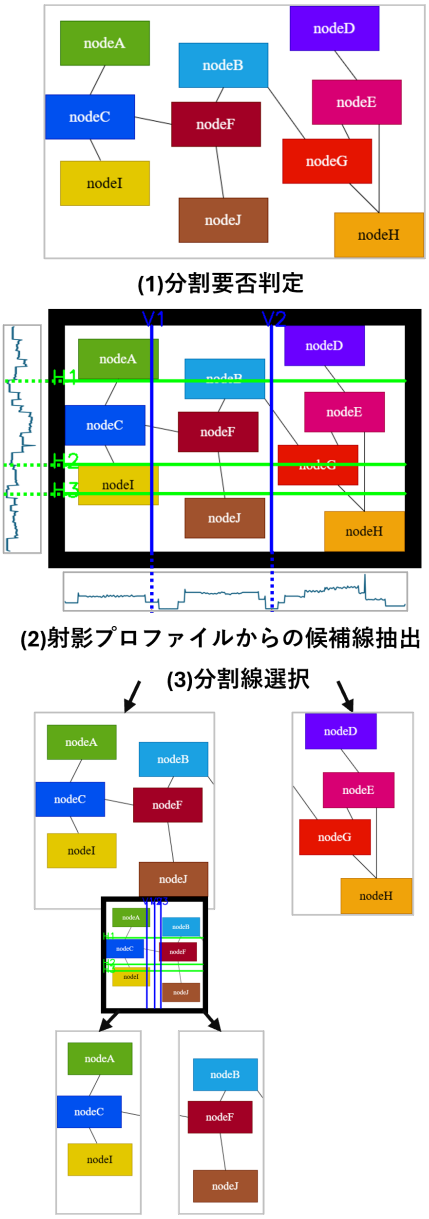

図 1　適応的階層分割手法の概要図

3.3 段階的統合

分割階層の葉から親へボトムアップに認識結果を統合する．まず，分割がこれ以上行われない葉領域ごとに，MLLM に対しノードとエッジ，境界接続候補を認識させる．次に，各親領域について，親画像と子 2 領域の画像，それぞれの局所認識結果および分割情報を MLLM に入力し，重複ノードの統合や分割線付近で途切れたエッジの復元，必要に応じたノード・エッジの追加を行わせる．階層をさかのぼりながらこれらを繰り返し，最終的に統合済みノード・エッジ集合を得る．

表 1　比較手法

比較フロー名	適応的階層分割の有無	段階的統合の有無
Baseline	×	×：全体画像のみを 1 回 MLLM に入力して認識
段階的統合のみ	×：深さ 2 まで分割．分割線は，3.2 適応的階層分割(2)で得られた分割候補線のうち，対応する射影プロファイル値が最小となる 1 本を選択する．	○
適応的階層分割のみ	○	×：全階層の領域画像と分割情報をまとめて 1 回の MLLM 呼び出しで認識
適応的階層分割＋段階的統合	○	○

表 2　平均値（F1=F1 Score, P=Precision, R=Recall）．Top は太字，次点は下線で示す．

使用モデル	GPT-5.1						GPT-4o					
対象	edge			node			edge			node		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Baseline	0.57	0.58	0.57	0.90	0.90	0.90	<u>0.49</u>	**0.52**	<u>0.47</u>	0.88	0.91	0.86
段階的統合のみ	0.58	0.61	0.57	0.91	0.91	0.92	0.42	0.44	0.41	0.90	0.91	0.90
適応的階層分割のみ	**0.69**	**0.70**	**0.69**	**0.97**	**0.97**	**0.97**	**0.52**	<u>0.51</u>	**0.52**	**0.97**	**0.97**	**0.97**
適応的階層分割＋段階的統合	<u>0.62</u>	<u>0.63</u>	<u>0.62</u>	<u>0.96</u>	<u>0.95</u>	<u>0.97</u>	0.45	0.46	0.44	<u>0.92</u>	<u>0.94</u>	<u>0.91</u>

4 実験・考察

4.1 実験設定

データセットとして，Mermaid [3]を用いてダイアグラム画像 6 枚を作成した．フローチャート，ER 図，マインドマップ図の 3 種について，それぞれ 30 ノードのランダムグラフを 2 通りで生成した(図 2)．比較するフローは，表 1 に示す 4 種類である． 評価指標は，Precision，Recall，F1 Score とし，ノード検出・エッジ検出それぞれについて算出した．MLLM には当初, Azure OpenAI の GPT-5.1 のみを用いた（reasoning_effort = "medium", max_completion_tokens = 100000）．この設定下で，階層全体一括認識である「適応的階層分割のみ」が最も高い精度を示したが, Reasoning による補完が強力であるため，段階的統合の有効性を純粋に評価しにくいことが分かった．そこで，明示的 Reasoning 制御を行わない GPT-4o を追加し比較を行った（max_tokens = 16384）．

4.2 結果・考察

表 2 に全体の結果を, 図 3 に GPT-5.1 による評価データ別のノード認識 F1 を示す．いずれの手法・モデルにおいてもノードの認識精度は高いが，エッジの認識精度はそれより大きく低下している．入力データ別では，いずれの手法でも ER 図とフローチャート図のスコアがマインドマップ図より大きく低下している．ER 図およびフローチャート図では，画像を大きくまたぐ長距離のエッジが含まれ，このような構造を含む画像では現状の汎用 MLLM によるエッジ認識が特に困難であることが示唆される．

表 2 より，F1 に関して「適応的階層分割のみ」が最も高い．「適応的階層分割＋段階的統合」は「適応的階層分割のみ」と比べてノード・エッジともに F1 が低下している．「適応的階層分割のみ」では，特に長距離エッジを含む ER 図やフローチャート図

において，柔軟に分割を行いながらも，全階層の領域画像と分割情報を一括して提示することで全体構造を失わずに認識できたためと考えられる．一方，「適応的階層分割＋段階的統合」では，局所領域ごとの認識結果を後段で統合する際に，分割によって切断されたエッジを十分に補完できず，結果としてエッジの誤りが階層的に蓄積した可能性がある．このことから，長距離エッジを含むダイアグラムに対しては，分割後の段階的な統合に過度に依存するよりも，初期段階で MLLM に全体構造を明示的に与えることが重要であると示唆される．

一方で，「段階的統合のみ」と比較すると，「適応的階層分割＋段階的統合」の方がノード・エッジの F1 が高い．固定深度かつ画一的な分割よりも，MLLM による柔軟な分割選択によって分割領域間のバランスやエッジ切断の少なさを考慮できたためであると考えられる．

ただし，フローごとのプロンプトの違いにより，わずかながら精度に影響が生じている可能性もある．例えば，基本的な認識プロンプトではノードラベルの認識において「読み取れる範囲で」と明示しているのに対し，段階的統合のプロンプトではこの文言を含めていないなど，細部の記述が異なる．

まとめ

本研究では，MLLM によるダイアグラム構造認識のために，ノード数推定と投影プロファイルに基づく柔軟な「適応的階層分割」と，葉領域から親領域へボトムアップに結果を統合する「段階的統合」を設計し，4 種の認識フローを比較した．GPT-5.1 と GPT-4o を用いた実験の結果，適応的階層分割はノード・エッジの F1 を一貫して向上させるのに対し，段階的統合は長距離エッジを含む ER 図やフローチャート図では，むしろ精度を悪化させる場合があることが分かった．

今後は，長距離エッジの多寡やノード密度に応じて，「分割のみ」「分割＋統合」「無分割」を MLLM の判断で切り替えるメタ戦略や，より柔軟な分割線生成を導入し，多様なダイアグラムに対して頑健な構造認識を実現したい．

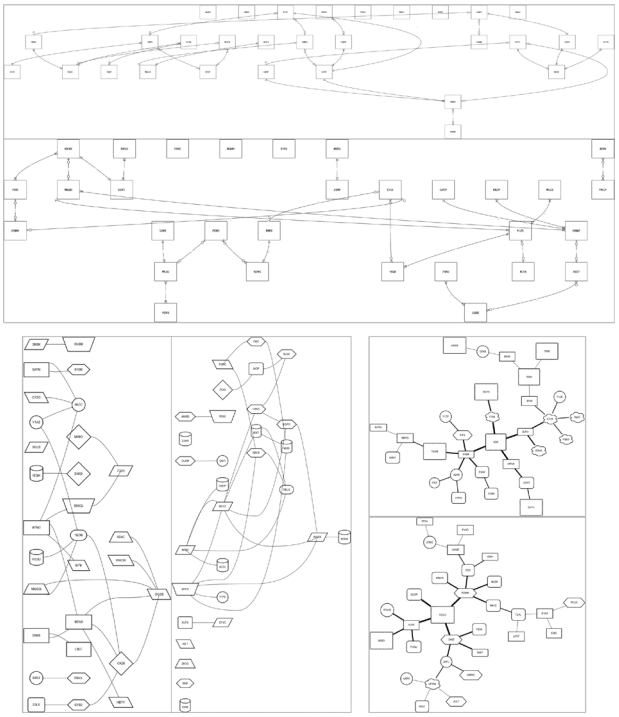

図 2　全評価用データ．上：ER 図，左下：フローチャート図，右下：マインドマップ図．

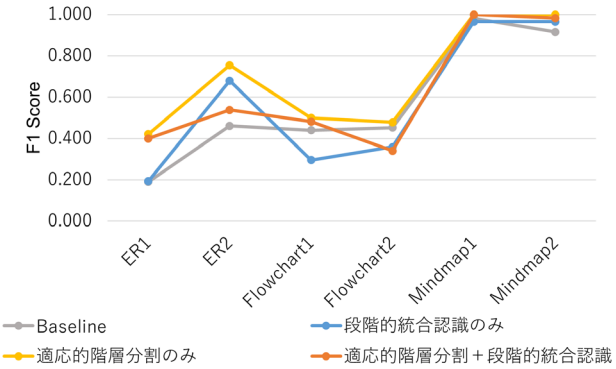

図 3　評価データ別の GPT-5.1 ノード認識 F1

参考文献

[1] 金子 武史, 安達 貴洋, 岡田 裕幸. MLLM を用いた図面活用による資産管理手法の提案. 暗号と情報セキュリティシンポジウム（SCIS2025), 2025.
[2] Xue Li, Yiyou Sun, Wei Cheng, Yinglun Zhu, Haifeng Chen. Chain-of-region: Visual Language Models Need Details for Diagram Analysis. The International Conference on Learning Representations (ICLR 2025), 2025.
[3] Mermaid. (オンライン) https://mermaid.js.org/.