ダイアグラム理解に向けた
大規模視覚言語モデルの内部表現の分析

吉田遥音 [1]　工藤慧音 [1]　青木洋一 [1]　田中涼太 [1,2]
斉藤いつみ [1]　坂口慶祐 [1]　乾健太郎 [1]
[1] 東北大学　[2] 日本電信電話株式会社 NTT 人間情報研究所
yoshida.haruto.p1@dc.tohoku.ac.jp

概要

ダイアグラムを理解できる AI モデルの実現は，学習支援や情報処理の効率化において重要である．しかし，画像理解タスクで顕著な成果を上げている大規模視覚言語モデル（LVLM）であっても，ダイアグラムのような抽象的かつ構造的な画像の理解には限界がある．本研究では，LVLM がダイアグラムのどのような視覚情報を認識しているか，またそれらの情報をどのように保持しているかを明らかにするため，画像エンコーダおよび LLM の隠れ状態を用いてプロービングを行った．その結果，ノードの色や形，エッジの色や有無の情報はどの層でも 10 次元程度の低次元の線形部分空間に保持されていたが，エッジの向きの情報は 10 次元程度の低次元空間には保持されていなかった．また，パッチ単位のプロービングにより，ノードやエッジが描かれていない背景の隠れ状態に，複数のノードやエッジの情報がまとめて保持されていることが示唆された．

1　はじめに

ダイアグラムとは，複雑な概念や関係を図形や線，記号により視覚的に表現したものである [1]．フローチャートや概念マップといったダイアグラムは，教育 [1] やビジネス [2] などの多岐にわたる分野で，効率的な情報伝達のために活用されている．そのため，ダイアグラムを理解できる AI モデルの実現は，教育における学習支援やビジネスにおける情報処理の効率化といった応用に繋がる．

大規模視覚言語モデル（LVLM）は画像理解タスクにおいて顕著な成果を上げているが，ダイアグラムのような抽象的かつ構造的な画像の理解には限界がある．LVLM は，画像から視覚情報を抽出する画像エンコーダと，言語での推論を行う大規模言語モデル（LLM）の組み合わせにより，画像キャプション生成 [3, 4] や視覚的質問応答 [5, 6] などの画像理解タスクで高い性能を達成している．しかし，自然画像の理解においては優れた性能を示す一方で，ダイアグラムのような抽象的かつ構造的な画像の理解，特に要素間の関係の認識には限界がある [7]．

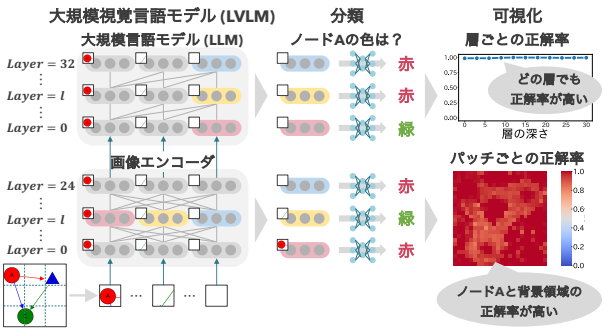

図 1　本研究の概観．LVLM の画像エンコーダおよび LLM の隠れ状態をもとにダイアグラムの属性を分類するように学習した分類モデルの性能から，LVLM の内部表現にどのような属性情報が保持されているかを調べる．

本研究では，LVLM がダイアグラムのどのような視覚情報を認識しているか，またそれらの情報をどのように保持しているかを明らかにするため，画像エンコーダおよび LLM の隠れ状態を用いてプロービングを行った．具体的には，ダイアグラムを入力した際の画像エンコーダおよび LLM の隠れ状態をもとに，ノードの色やエッジの向きといったダイアグラムの属性を分類するモデルを学習し，その性能を評価した．その結果，ノードの色や形，エッジの色や有無の情報は，多くの層で 10 次元程度の低次元の線形部分空間に保持されていたが，エッジの向きの情報は 10 次元程度の低次元空間には保持されていなかった．また，ノードやエッジが描かれていない背景の隠れ状態に，複数のノードやエッジの情報が集約されていることを示唆する結果を得た．

2 関連研究

2.1 LVLM のアーキテクチャ

LVLM は，画像エンコーダと LLM を組み合わせることで，画像と言語情報を統合的に扱っている [8, 9]．画像エンコーダには Vision Transformer [10] が使用され，画像をパッチと呼ばれる固定サイズの小領域に分割して処理している．

本研究では，画像エンコーダに入力した全てのパッチに対応する Transformer の出力を画像特徴量とし，単語埋め込みの代わりとして LLM に入力するアーキテクチャを採用したモデル対象とする．

2.2 LVLM によるダイアグラム理解

GPT-4 [11] をはじめとする LVLM が，ダイアグラムに基づく視覚的推論タスク [12, 13] において高いスコアを達成している．一方で Giledereli ら [7] は，ダイアグラムに基づく質問応答タスクにおいて，LVLM がダイアグラム内のエンティティの関係を理解せず，LLM の知識を利用したショートカットによってタスクを解いている可能性を指摘している．

本研究は，LVLM の内部表現にどのような情報が，どのように保持されているかを分析し，ダイアグラム理解の実態を明らかにすることを目指す．

2.3 LVLM のプロービング

LVLM が画像のどのような視覚情報を認識しているかを明らかにするため，内部表現に含まれる情報を分析する，プロービングという手法がとられている [14, 15, 16]．こうした研究により，自然画像の視覚情報がモデル内でどのようにエンコードされているかが明らかにされつつあるが，ダイアグラムのような抽象的かつ構造的な視覚情報のエンコードに関する分析は限定的である．

本研究では，ダイアグラムを処理する LVLM の内部表現をプロービングし，モデルがノードやエッジなどのダイアグラム特有の視覚情報を認識しているかを明らかにする．

3 データセットの構築

LVLM のダイアグラム理解能力を評価するため，有向グラフを基盤としたダイアグラムデータセットを構築した．各ダイアグラムは，以下の特徴を持つノードとエッジの組み合わせで構成される．

ノードの特徴　全てのダイアグラムは A，B，C の 3 つのノードを有し，各ノードは以下の属性を持つ．

- **色**：赤，緑，青のいずれか
- **形**：円形，三角形，四角形のいずれか

ノードの色と形は独立に決定されるため，各ノードの属性の組み合わせは $3 \times 3 = 9$ 通り存在する．したがって，3 つのノード全体の属性の組み合わせは $9^3 = 729$ 通りである．また，ノードの位置は全てのダイアグラムで共通している．

エッジの特徴　全てのノード間にはエッジが 0 本または 1 本張られ，各エッジは以下の属性を持つ．

- **色**：赤，緑，青のいずれか
- **向き**：どちらのノードから伸びているか（e.g., A → B または B → A）

エッジの色と向きは独立に決定されるため，各ノード間に張られるエッジの属性の組み合わせは $3 \times 2 = 6$ 通り存在する．また，エッジが張られない場合を含めると，各ノード間のエッジの属性は 7 通りである．したがって，3 つのノード間におけるエッジの属性の組み合わせは $7^3 = 343$ 通りである．

以上より，ノードの組み合わせ 729 通りとエッジの組み合わせ 343 通りにより，本データセットは合計 $729 \times 343 = 250,047$ 件のダイアグラムで構成される．データセットに含まれるダイアグラムの例は Appendix の A 節を参照されたい．

4 分析手法

本研究では，LVLM がダイアグラムのどのような属性の情報を内部表現に保持しているか，またどのように保持しているかを明らかにすることを目指す．調査するダイアグラムの属性は，ノードの色，ノードの形，エッジの色，エッジの有無，エッジの向きの 5 つである．画像エンコーダおよび LLM の隠れ状態を入力として，ダイアグラムの属性を分類するモデルに，部分的最小二乗判別分析（PLS-DA）モデル [17] を使用する．本節では，まず PLS-DA モデルについて説明し，その後，画像エンコーダと LLM に対するプロービング手法を詳述する．

4.1 PLS-DA モデル

PLS-DA は，部分最小二乗回帰（PLS 回帰）を分類問題に適用した教師あり学習手法である．PLS-DA モデルは次元圧縮と分類の 2 段階で構成される．

次元圧縮　画像エンコーダまたは LLM の隠れ状態 $\mathbf{h} \in \mathbb{R}^d$ を，k 次元の潜在ベクトル $\mathbf{t} \in \mathbb{R}^k$ に変換する．次元圧縮は以下の式で表される．

$$\mathbf{t} = \mathbf{P}^\top \mathbf{h} \tag{1}$$

ここで，$\mathbf{P} \in \mathbb{R}^{d \times k}$ は隠れ状態を潜在空間にマッピングする射影行列である．この操作では，隠れ状態とクラスラベルとの共分散を最大化する方向にデータを射影することで，分類に有効な特徴を抽出する．

分類　得られた低次元の潜在ベクトル $\mathbf{t}$ を用いて，ダイアグラムの属性クラスラベル y を予測する．分類は以下の式で表される．

$$\hat{y} = f(\mathbf{t}) \tag{2}$$

ここで，f は潜在ベクトルからクラスラベルを予測する分類器である．各クラスの中心からのマハラノビス距離をもとに，最も近いクラスに分類する．

4.2　画像エンコーダのプロービング

画像エンコーダの各層がダイアグラムのどのような属性情報を含んでいるか，またその情報がどのように分布しているかを調べるため，層ごとかつパッチごとにダイアグラムの属性の分類を行った．

具体的には，ノードの色，ノードの形，エッジの色，エッジの有無，エッジの向きの 5 つの属性に着目し，各属性に対する二値分類タスクを設定した．ノードの色，ノードの形，エッジの色は 3 種類の値を取り得るため，その中から 2 つの値を選択し，それらの値を持つデータを用いて二値分類を行った．また，各属性の分類では，特定のノードやエッジを対象とし，それ以外のノードやエッジの属性は考慮しなかった．つまり，例えばノードの色の分類では，対象となるノードを A，B，C から 1 つ選び，さらにそのノードの色を赤，緑，青から 2 つ選び，対象のノードがどちらの色かを分類した．

分析対象の層は，Transformer 層を適用後の全 24 層のうち，第 1 層，第 12 層，第 23 層の 3 つである．これらの層はそれぞれ，初期，中期，後期の層であり，層に進むごとにダイアグラムの属性情報がどのように表現され，変化するかを調べるために選択した．

先行研究 [16] と同様に，層ごとに各パッチに対応する隠れ状態に対して分類を行った．パッチ単位のプロービングにより，属性情報が画像内のどの位置に強くエンコードされているかを明らかにできる．これにより，属性情報の局所性や全体性を明らかにし，画像エンコーダがダイアグラムの属性情報をどのように内部表現に保持しているかを分析した．

4.3　LLM のプロービング

LLM の各層がダイアグラムのどのような属性情報を保持しているか，またその情報がモデル内でどのように伝達されるかを調べるため，層ごとに分類を行った．分類の方法は画像エンコーダのプロービングと同様であり，特定のノードやエッジを対象として，その属性について二値分類を行った．

分析対象の層は，LLM の全ての層である．ダイアグラムの全てのパッチのベクトルを入力し，最後のパッチ（時刻）に対応する各層の隠れ状態ベクトルを用いて分類を行った．これは，単方向性を持つ LLM では最後のパッチ (時刻) でのみ画像全体の情報にアクセスできるためである．また，LLM に入力されるのはダイアグラムの情報のみであり，テキストは入力しなかった．

5　実験

LVLM として LLaVA-1.5-7B [8] を選び，その画像エンコーダおよび LLM に対してプロービングを行った．このモデルの画像エンコーダは CLIP-ViT-L-336px [18]，LLM は Vicuna-7B [19] であり，隠れ状態はそれぞれ 1024 次元，4096 次元である．

5.1　画像エンコーダのプロービング結果

低次元空間にエンコードされている属性とそうでない属性があった　図 2 から，ノードの色，ノードの形，エッジの色，エッジの有無の正解率はチャンスレートよりも高く，エッジの向きの正解率はチャンスレートと同程度であった．つまり，ノードの色，ノードの形，エッジの色，エッジの有無の情報は隠れ状態に含まれており，これらの情報は 10 次元程度の低次元の線形部分空間に保持されていた．一方で，エッジの向きの情報は 10 次元程度の低次元の線形部分空間には保持されておらず，LVLM にとってエッジの向きの認識が他の属性の認識に比べて難しいことを示唆している．

背景のパッチは複数のノードやエッジの情報を含んでいた　図 2 において正解率が高かった 4 つの属性はどれも，背景のパッチに対する正解率が高かった．ここで，ノード A の色を分類するためには，3 つのノードのうちどれがノード A であるかを認識したうえで，そのノードの色を分類する必要がある．

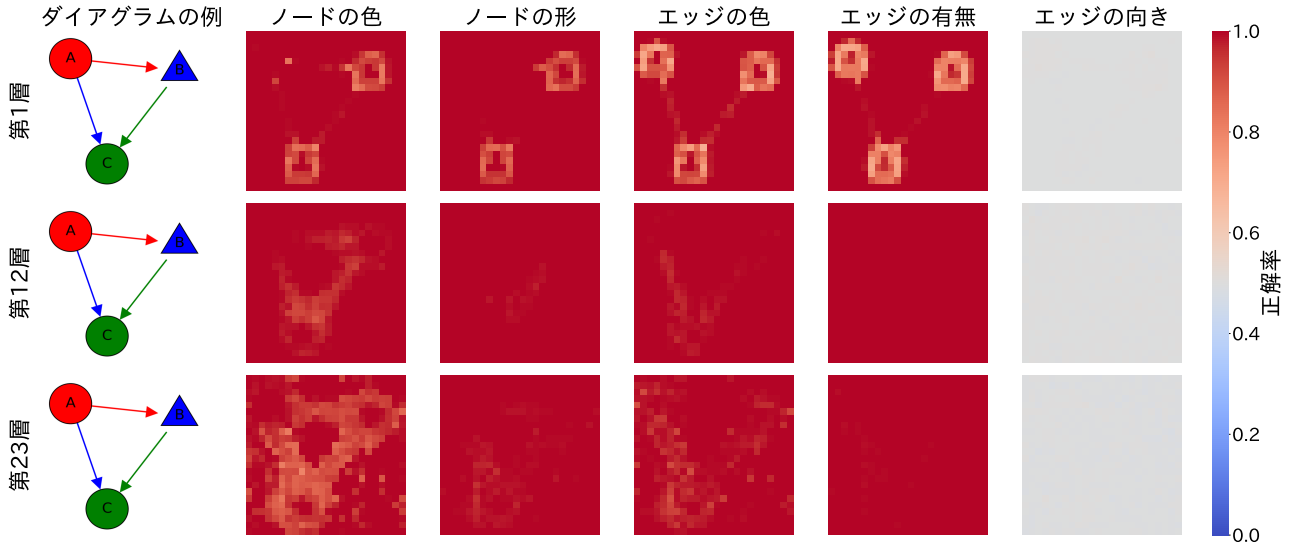

図 2　各属性の層ごとの分類結果．10 次元に圧縮した隠れ状態を用いて分類を行った．チャンスレートは 0.5 である．

つまり，背景のパッチ画像からノード A の色を分類できることは，背景のパッチに対応する隠れ状態に，ノード A の絶対位置と色の情報が一体として保持されていることを示している．加えて，他のノードやエッジの属性の分類でも同様の傾向が見られたことから（Appendix の C 節），モデルは背景のパッチに対応する隠れ状態に，複数のノードやエッジの情報をまとめてエンコードしていることを示唆している．

序盤の層で認識できている属性はそれ以降も認識できていた　図 2 において正解率が高かった 4 つの属性はどれも，全ての層で正解率が高かった．この結果は，これらの属性情報は序盤の層で抽出され，それ以降の層でも 10 次元程度の低次元空間に保持されていることを示している．

5.2　LLM のプロービング結果

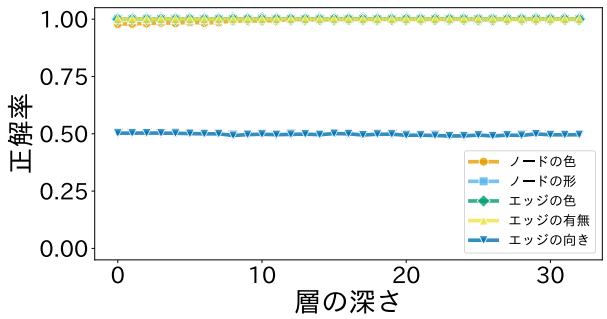

図 3　各属性の層ごとの正解率．ノードの色や形，エッジの色や有無の正解率はどの層でも高く，エッジの向きの正解率はどの層でも低かった．

エッジの向きはどの層でも分類できなかった　図 3 から，ノードの色，ノードの形，エッジの色，エッジの有無の正解率はどの層でも高く，エッジの向きの正解率はどの層でも低かった．この結果は画像エンコーダのプロービング結果と整合し，LVLM がエッジの向きなどの要素間の関係を認識できないのは，画像エンコーダが必要な情報を抽出できないためであることを示唆している．

6　おわりに

本研究では，LVLM がダイアグラムのどのような属性の情報を認識しているか，またそれらの情報をどのように保持しているかを明らかにするため，画像エンコーダおよび LLM の隠れ状態を用いてプロービングを行った．その結果，画像エンコーダ，LLM ともにノードの色や形，エッジの色や有無の情報は多くの層で 10 次元程度の低次元の線形部分空間に保持しているが，エッジの向きの情報は 10 次元程度の低次元空間には保持していなかった．さらに，背景のパッチに対応する隠れ状態には，複数のノードやエッジの属性の情報が，絶対位置とともに保持されていることを示唆する結果を得た．

今後の課題は，画像エンコーダが抽出した情報を LLM がどのように推論に利用しているかの分析である．抽出した情報を LLM がどのように推論に利用しているかを明らかにすることで，性能向上のための新たなアプローチに繋がると考える．

謝辞

本研究は，JST CREST JPMJCR20D2 及び JST 博士後期課程学生支援 JPMJBS2421 及び JST 次世代研究者挑戦的研究プログラム JPMJSP2114 及び JSPS 科研費 JP21K21343 の支援を受けたものです．また，本研究は九州大学情報基盤研究開発センター研究用計算機システムの一般利用を利用しました．本研究を進めるにあたり多大なご助言，ご協力を賜りました Tohoku NLP グループの皆様に感謝いたします．

参考文献

[1] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In **Computer Vision – ECCV 2016**, Lecture notes in computer science, pp. 235–251. Springer International Publishing, Cham, 2016.

[2] Emelie Havemo. A visual perspective on value creation: Exploring patterns in business model diagrams. **European Management Journal**, Vol. 36, No. 4, pp. 441–452, August 2018.

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. **arXiv [cs.CV]**, April 2015.

[4] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In **2015 IEEE International Conference on Computer Vision (ICCV)**, pp. 2641–2649. IEEE, December 2015.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In **2015 IEEE International Conference on Computer Vision (ICCV)**, pp. 2425–2433. IEEE, December 2015.

[6] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 1988–1997. IEEE, July 2017.

[7] Buse Giledereli, Yifan Hou, Yilei Tu, and Mrinmaya Sachan. Do vision-language models really understand visual language? **arXiv [cs.CL]**, September 2024.

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 26296–26306, 2024.

[9] Abhimanyu Dubey and et al. et al. The llama 3 herd of models. **CoRR**, Vol. abs/2407.21783, , 2024.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In **International Conference on Learning Representations**, 2021.

[11] OpenAI. GPT-4 technical report. **arXiv [cs.CL]**, March 2023.

[12] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATH-VERSE: Does your multi-modal LLM truly see the diagrams in visual math problems? In **Lecture Notes in Computer Science**, Lecture notes in computer science, pp. 169–186. Springer Nature Switzerland, Cham, 2025.

[13] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In **The Twelfth International Conference on Learning Representations**, October 2023.

[14] Mingxu Tao, Quzhe Huang, Kun Xu, Liwei Chen, Yansong Feng, and Dongyan Zhao. Probing multimodal large language models for global and local semantic representations. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 13050–13056, 2024.

[15] Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. Cross-modal projection in multimodal LLMs doesn't really project visual attributes to textual space. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 657–664, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

[16] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. **arXiv [cs.CV]**, October 2024.

[17] Matthew Barker and William Rayens. Partial least squares for discrimination. **J. Chemom.**, Vol. 17, No. 3, pp. 166–173, March 2003.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, A Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I Sutskever. Learning transferable visual models from natural language supervision. **ICML**, Vol. 139, pp. 8748–8763, February 2021.

[19] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

A　使用したダイアグラムの例

構築したデータセットに含まれるダイアグラムの例を図 4 に示す．

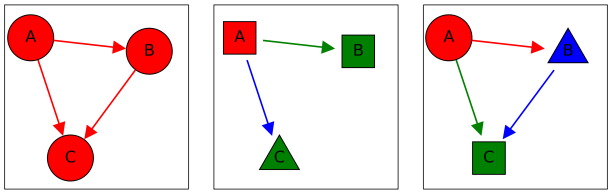

図 4　構築したデータセットに含まれるダイアグラムの例．ノードの位置はすべてのダイアグラムで共通である．

B　分類モデルの学習

評価データ　分類タスクごとに個別の評価データを用意した．具体的には，「ノード A の色が赤か青か」の分類と「ノード A の色が赤か緑か」の分類では，それぞれ別の評価データを使用した．評価データは，各分類タスクに関連するダイアグラム集合から，ランダムに 10,000 件を抽出することで作成した．例えば，「ノード A の色が赤か青か」の分類では，ノード A が赤または青のいずれかであるダイアグラムの集合から，10,000 件をランダムに抽出した．

学習データ　評価データと同様に，分類タスクごとに個別の学習データを作成した．学習データは，各分類タスクに関連するダイアグラム集合から評価データに含まれる 10,000 件のダイアグラムを除き，残ったダイアグラム集合からランダムに 40,000 件を抽出することで作成した．

C　画像エンコーダのプロービング

分類対象を変えた場合の結果　分類対象をノード B，エッジ AC にした場合の結果を図 5 に示す．図 2 と図 5 を比較すると，対象とするノードやエッジによらず同様の傾向が見られることがわかる．

次元数を変えた場合の結果　図 6 から，次元数が小さいほど，正解率が高い領域が分類対象のノード（エッジ）の位置に集中し，次元数が大きいほど正解率が高い領域がダイアグラム全体に広がることがわかる．この結果は，ノードやエッジの情報は，それが描かれている領域においては低次元の線形部分空間に保持されているが，それ以外の領域においては，より高次元の部分空間に保持されていることを示唆している．加えて，分類対象とは異なるノードやエッジが存在する領域と背景領域の正解率を比較

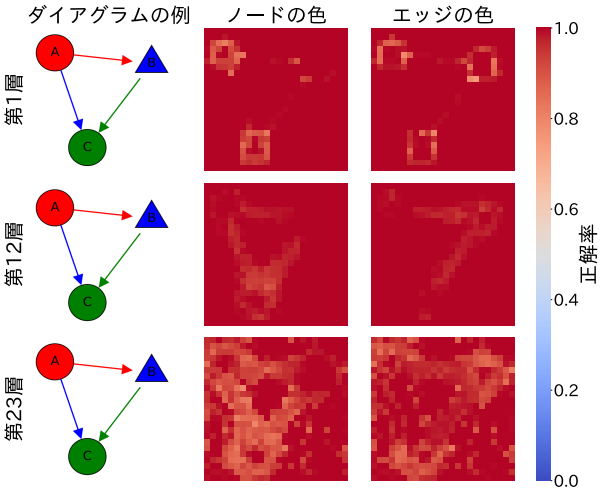

図 5　分類対象のノードやエッジを変えた場合の結果．

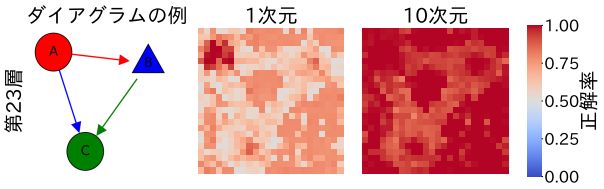

図 6　次元数を 1，10 にした場合の結果の比較．

すると，背景領域の正解率が高いことがわかる．これは，もともと別の情報が含まれている領域において，注意機構によって集められる情報は，もともと含まれている情報に比べて高次元の部分空間に保持されることを示唆している．

D　LLM のプロービング

分類対象をノード A，エッジ AC にした場合の結果を図 7 に示す．図 3 と図 7 を比較すると，対象とするノードやエッジによらず同様の傾向が見られたことがわかる．

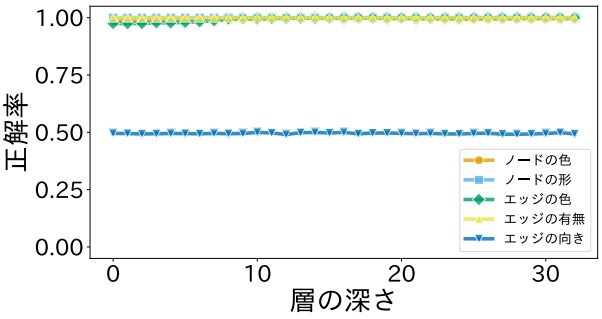

図 7　分類対象のノードやエッジを変えた場合の結果．