自然画像で学習された画像埋め込みにダイアグラムを特徴づける情報は含まれているか？

吉田遥音 [1]　工藤慧音 [1,2]　青木洋一 [1,2]　田中涼太 [1,3]
斉藤いつみ [1]　坂口慶祐 [1,2]　乾健太郎 [4,1,2]
[1] 東北大学　[2] 理化学研究所　[3]NTT 人間情報研究所　[4]MBZUAI
haruto.yoshida@dc.tohoku.ac.jp
{ keito.kudo.q4, youichi.aoki.p2, tanaka.ryota.r7 }@dc.tohoku.ac.jp
{ itsumi.saito, keisuke.sakaguchi }@tohoku.ac.jp　kentaro.inui@mbzuai.ac.ae

概要

ダイアグラムの意味やデザインを考慮して分類や検索，評価を行うための道具として，画像埋め込みがある．しかし，既存の事前学習済み画像モデルから得られる埋め込みに，ダイアグラムを特徴づける情報が十分に含まれているかは明らかでない．本研究では，エッジの向きやノードの形といった要素が異なるダイアグラムの埋め込み分布を比較し，事前学習済みモデルから得られる画像埋め込みがダイアグラムを特徴づける情報を含んでいるかを調べた．既存の事前学習済みモデルから得られる埋め込みはダイアグラムを特徴づける情報を十分には含んでいない可能性があり，ダイアグラムを扱うことができるモデルの必要性が示唆された．

1　はじめに

ダイアグラムは，データや概念を整理し，図示したものである．テキストでの表現に加えて，視覚的な表現を用いることにより理解を促す効果があるため，ビジネス [1] や教育 [2]，学術研究 [3] をはじめとする多くの分野で広く用いられている．また，自然言語からダイアグラムを生成する研究 [4, 5] が行われるなど，ダイアグラムの理解・生成に関する研究の重要度は高まっている．

ダイアグラムの視覚的情報には，ノード間を結ぶエッジやテキストなどで表される記号的な意味情報と，それらをさらに見やすく表現するためのレイアウトや色使いといったデザインの大きく 2 つの要素がある．これら 2 つの要素を適切に用いることで，ダイアグラムは効果的に情報を伝達する．そのため，ダイアグラムの分類や検索といった研究を行うにあたり，ダイアグラムの意味とデザインに焦点を当てることが重要であるといえる．

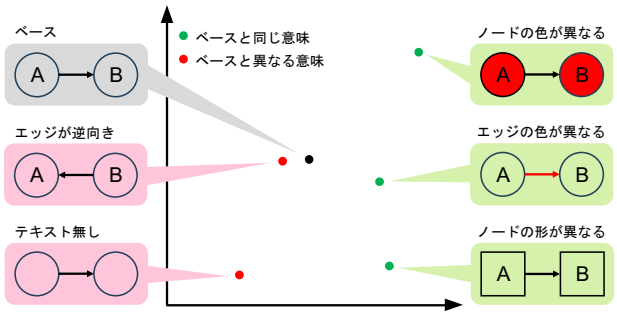

図 1　本研究の概念図．特定の要素が異なるダイアグラムの埋め込み分布の違いを調べた．分布内での距離が大きいほど，異なるダイアグラムとしてモデルが認識していると考えられる．エッジが逆向きのダイアグラムは，意味的には大きな違いがあるものの，埋め込み空間における距離は小さかった．

ダイアグラムが持つ意味やデザインを考慮して分類や検索といった研究を行うための道具として，画像埋め込みがある．実際に画像を扱う研究において，分類 [6] や検索 [7]，生成画像の評価 [8, 9] に至るまで，画像埋め込みが用いられている．つまり，ダイアグラムの特徴を正しく理解可能な埋め込みを構築することで，ダイアグラムの検索や分類といった様々な研究に応用できる．

しかしながら，既存の画像を埋め込むモデルの学習データは自然画像が中心であり，ダイアグラムは少ないと考えられる．[1)] ここでの自然画像は，現実世界の一部を切り取った写真のような画像を指している．またダイアグラムは，特定の情報を表現するために意図的に設計されているため，背景の乱雑さや複雑なテクスチャといった情報は抑制されてお

1）　今回実験に用いる Inception V3 は画像分類を行うために学習されたモデルであり，学習データの中に diagram や figure といったラベルは存在していなかった．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

り，自然画像とは本質的に異なる [2]．そのため，既存の画像埋め込みにダイアグラムを特徴づける情報が含まれていない可能性があるが，その検証はなされていない．

そこで本研究では，既存の画像埋め込みにダイアグラムを特徴づける情報が含まれているかを調査した．その結果，エッジの色や向きといったピクセル数が小さい要素に関する情報が含まれていない，または取り出しにくい形で含まれている可能性があることが明らかになった．この結果は，ダイアグラムを扱うことができる新たなモデルが必要であることを示唆している．

2 関連研究

2.1 自然画像の埋め込み

自然画像の研究において，分類や検索を目的とした埋め込みが提案されている [10, 11]．これらの研究は，自動車 [10] や建造物 [11] といった一般物体に対する埋め込みモデルの構築を目的としている．そのため，画像中の特定の一般物体についての特徴を含んだ埋め込みを得ることに焦点を当てている．一方で，本研究が対象とするダイアグラムはノードやエッジ，テキストなど記号的な要素が多く，これらの個々の要素に加えて要素間の関係を理解することが重要である．

2.2 自然言語処理における埋め込みの分析

自然言語処理の分野では，単語や文の埋め込みが特定の情報を含んでいるかを分析する研究が行われている [12, 13, 14]．多くの既存研究では，言語的性質の違いを認識する分類器を学習し，その分類精度やモデルの複雑さを見ることで埋め込みに含まれる情報を調べている [12, 13]．ただしこれらの手法では，異なる表現に対して異なる分類器を学習する必要があることや，学習時のハイパラメータによって分類精度が変化してしまうといった問題が指摘されている [14]．本研究では，重心からの距離によって分類する教師なしの分類手法 [15] を用いる．具体的な説明は 3.2 節で述べる．

3 分析の方針

本稿では，ダイアグラムの中でも 2 つのノードとそれらを結ぶ 1 つのエッジからなるシンプルな有向グラフをベースとするダイアグラム対象として，埋め込みの分布を分析する．

3.1 ダイアグラムの準備

A，B の 2 つのノードと，A から B に向かう 1 つのエッジからなる有向グラフをベースのダイアグラムとして，特定の要素を変更したダイアグラムを用意する．変更する要素は，ノードの位置（16 種類），ノードの形（2 種類），ノードの色（2 種類），エッジの向き（2 種類），エッジの色（2 種類），ノード内のテキストの有無（2 種類）であり，基本的にはこれらの直積をとる形で 384 種類のダイアグラムを作成する．ただし，テキストが無いダイアグラムはノードの区別ができないため，そのままでは同一のダイアグラムが複数生成されてしまう．今回はそれらの重複を取り除くため，単純に全ての直積をとった場合の 512 種類とは一致しない．

また本稿では，有向グラフとしての正しさを意味の正しさとし，エッジの向きとテキストの有無を「意味に関わる要素」，それ以外を「デザインに関わる要素」とする．さらに，ノードの位置はバリエーションを出すための要素であり，こちらについての詳細な分析は行わない．

3.2 画像埋め込みの定量分析

特定の要素が異なるダイアグラムの埋め込みを入力とし，どれほどの精度で分類できるかを測ることにより，埋め込みに含まれている情報を分析する．本稿では，既存研究 [15] と同様の方法により分類と定量化を行う．準備として，ダイアグラムの特定の要素についての条件を表す正解クラスを定義する．ここでの正解クラスは，要素の違いを 0 と 1 の二値で表現したものである．例えばノードの形に注目して正解クラスを定義する場合，ノードの形が円であるか否かをそれぞれ 0 と 1 という 2 つの正解クラスとして定義する．

分類手法　図 3 に示す手順により，ダイアグラムを分類する．はじめに，それぞれの正解クラスに属する埋め込みの重心を計算し，それを正解クラスの重心とする．正解クラス $j \in \{0, 1\}$ の重心 $\mathbf{c}_j$ は，j に属する埋め込みの集合を $\mathbb{E}_j$ として式 1 により計算される．

$$\mathbf{c}_j = \frac{1}{|\mathbb{E}_j|} \sum_{\mathbf{e}_k \in \mathbb{E}_j} \mathbf{e}_k \tag{1}$$

次に，それぞれの埋め込みに対して，全ての正解クラスの重心との L2 距離，またはコサイン類似度を

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

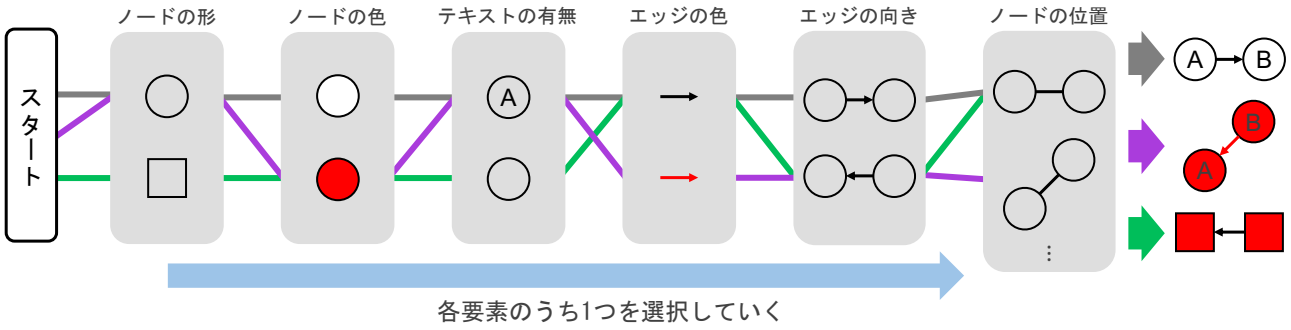

図 2　ダイアグラムの作成過程．すべての要素を 1 つずつ指定することで 1 つのダイアグラムを作成する．

計算し，L2 距離が最小，またはコサイン類似度が最大の正解クラスに分類する．ここで分類されたクラスを予測クラスと呼ぶ．全ての埋め込みの集合を $\mathbb{E}$ とすると，$\mathbf{e}_k \in \mathbb{E}$ の予測クラス $\hat{y}_k$ は式 2 または式 3 により計算される．

$$\hat{y}_k = \underset{j}{\operatorname{argmin}} \|\mathbf{e}_k - \mathbf{c}_j\| \tag{2}$$

$$\hat{y}_k = \underset{j}{\operatorname{argmax}} \frac{\mathbf{e}_k \cdot \mathbf{c}_j}{|\mathbf{e}_k||\mathbf{c}_j|} \tag{3}$$

評価指標　分類精度を定量化するために，クラスタリングの評価に用いられている Purity と Inverse Purity の調和平均である F 値 [16] を用いる．本稿では，Purity と Inverse Purity はそれぞれ「ある予測クラスに分類されたダイアグラムの中に異なる正解クラスのダイアグラムが含まれていないか」と「ある正解クラスのダイアグラムが異なる予測クラスに分類されていないか」を表している．これらを計算するにあたり，予測クラス i に含まれている正解クラス j に属するデータの個数を表す混同行列 $n_{i,j}$ を計算する．$\mathbf{e}_k \in \mathbb{E}$ の正解クラスを y_k とすると，$n_{i,j}$ は式 4 により計算される．

$$n_{i,j} = |\{\mathbf{e}_k \in \mathbb{E} | \hat{y}_k = i \wedge y_k = j\}| \tag{4}$$

$n_{i,j}$ と埋め込みの総数 $|\mathbb{E}|$ を用いて，Purity と Inverse Purity はそれぞれ式 5, 式 6 により計算される．

$$\text{Purity} = \frac{1}{N} \sum_i \max_j (n_{i,j}) \tag{5}$$

$$\text{Inverse Purity} = \frac{1}{N} \sum_i \left(\frac{\max_j (n_{i,j})}{\sum_i n_{i,j}} \sum_j n_{i,j} \right) \tag{6}$$

最後に，式 7 により Purity と Inverse Purity の調和平均をとる．

$$F = \frac{2 \times \text{Purity} \times \text{Inverse Purity}}{\text{Purity} + \text{Inverse Purity}} \tag{7}$$

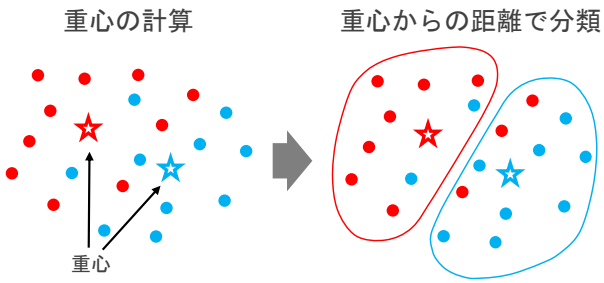

図 3　埋め込みの分類手法．埋め込みと正解クラスの重心との距離により分類する．

3.3　埋め込み分布の可視化

定量分析では分布間の距離や広がりを分析することは難しいため，主成分分析（PCA）[17] を用いて複数の要素における埋め込み分布を 2 次元に圧縮して可視化する．PCA の前処理として標準化を行い，さらに圧縮後に白色化を行う．

4　実験

画像を埋め込むモデル　画像を埋め込むモデルには，Inception V3 [18] と CLIP [19]（ViT-B/32 [20]）を用いた．また，計算される埋め込みの次元は Inception V3 が 2048，CLIP が 512 である．以降は，画像を埋め込むモデルを「埋め込みモデル」と呼ぶ．

データセット　埋め込みモデルの入力には，3.1 節で述べた 384 種類のダイアグラムを使用した．

4.1　定量分析

エッジの向き，テキストの有無，ノードの形，ノードの色，エッジの色の 5 つの観点について F 値を測定した結果を表 1 に示す．

意味に関わる要素の中でもエッジの向きでの分類は困難：　表 1 より，意味に関わる要素を比較すると，どちらの埋め込みモデルにおいてもテキストの有無に対する F 値はエッジの向きに対する F 値より

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1　各要素に対する F 値．F_{L2}，F_{cos} はそれぞれ L2 距離，コサイン類似度をもとに計算した．

	意味に関わる要素				デザインに関わる要素					
	エッジの向き		テキスト		ノードの形		ノードの色		エッジの色	
	F_{L2}	F_{cos}	F_{L2}	F_{cos}	F_{L2}	F_{cos}	F_{L2}	F_{cos}	F_{L2}	F_{cos}
Inception V3	0.60	0.60	0.99	0.99	0.99	0.99	0.91	0.99	0.76	0.77
CLIP	0.52	0.51	1.00	1.00	1.00	1.00	0.97	0.97	0.79	0.79

も高かった．これは，テキストの有無が異なるダイアグラムを埋め込みの違いで容易に分類できるが，エッジの向きが異なるダイアグラムは埋め込みの違いで分類することが困難であることを意味する．つまり，意味に関わる要素の中でも，テキストの有無についての情報は埋め込みに含まれるが，エッジの向きについての情報は埋め込みに含まれない，または取り出しにくい形で含まれる可能性がある．

デザインに関わる要素の中でもエッジの色での分類は困難：　表 1 より，デザインに関わる要素を比較すると，どちらの埋め込みモデルにおいてもノードの形や色に対する F 値はエッジの色に対する F 値よりも高かった．これは，ノードの形や色が異なるダイアグラムを埋め込みの違いで容易に分類できるが，エッジの色が異なるダイアグラムを埋め込みの違いで分類することが困難であることを意味する．つまり，デザインに関わる要素の中でも，ノードの形や色についての情報は埋め込みに含まれるが，エッジの色についての情報は埋め込みに含まれない，または取り出しにくい形で含まれる可能性がある．

意味・デザインを問わずエッジの違いでの分類は困難：　表 1 より，エッジについての情報は埋め込みに含まれない，または取り出しにくい形で含まれる可能性がある．今回使用したダイアグラムにおいてエッジは他の要素に比べてピクセル数が少なく，既存の画像埋め込みがピクセル数の多い要素についての情報を取り出しやすい形で含んでいる可能性がある．一方で，今回のエッジのような，ピクセル数は少ないが，ダイアグラムの本質的な意味に関わる要素も多くあるため，そういった要素の情報が取り出しやすい形で含まれる埋め込みが望まれる．

4.2　定性分析

意味に関わる要素における埋め込み分布を図 4 に示す．ただし，PCA を行った後の第二主成分までの累積寄与率は Inception V3 で埋め込んだ場合が 0.20，CLIP で埋め込んだ場合が 0.33 であった．

CLIP の埋め込み分布は定量分析の結果と整合：　表 4 より，CLIP の埋め込み分布については，基準となるダイアグラムの埋め込みとエッジが逆向きのダイアグラムの埋め込みは混ざり合って分布しているが，テキストの無いダイアグラムの埋め込みはそれらとは混ざり合わずに分布していた．これは表 1 の結果と整合している．一方，Inception V3 の埋め込み分布は，CLIP の埋め込み分布に比べてテキストの無いダイアグラムの埋め込みがそれ以外の埋め込みと混ざり合って分布していた．これは表 1 の結果と整合しているとは言えない．Inception V3 の PCA における第二主成分までの累積寄与率が CLIP の寄与率よりも低いことから，埋め込み全体を説明するのに不十分であることが原因だと考えられる．

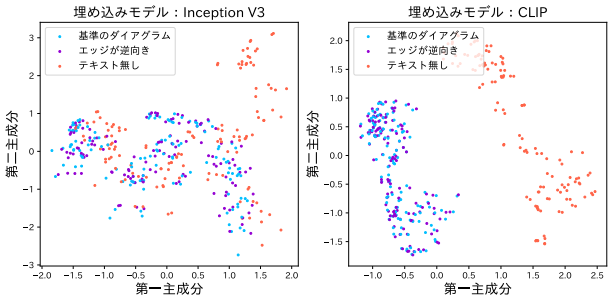

図 4　意味に関わる要素の埋め込み分布．次元圧縮には PCA を用いた．

5　おわりに

本稿では，既存の画像埋め込みに有向グラフをベースとしたダイアグラムのエッジに関する情報が含まれていない，または取り出しにくい形で含まれている可能性があることを明らかにした．この結果は，ダイアグラムを特徴づける情報を取り出しやすい形で含んでいる画像埋め込みを構築する必要性を示唆している．

今後は，ダイアグラムを特徴づける情報を取り出しやすい形で含む埋め込みを構築し，ダイアグラムの分類や検索といった研究に応用することを目指す．また，テキストとダイアグラムにおける埋め込みの対応関係を明らかにする方向性も興味深い．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は JSPS 科研費 JP21K21343，JP22H00524 の助成を受けたものです．また，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羽根田賢和氏，松﨑孝介氏をはじめとする Tohoku NLP グループの皆様に感謝いたします．

参考文献

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

[2] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram Is Worth A Dozen Images. **arXiv preprint arXiv:1603.07396**, 2016.

[3] Helen C. Purchase. Twelve years of diagrams research. **Journal of Visual Languages & Computing**, Vol. 25, No. 2, pp. 57–75, 2014.

[4] Jonas Belouadi, Anne Lauscher, and Steffen Eger. AutomaTikZ: Text-Guided Synthesis of Scientific Vector Graphics with TikZ. **arXiv preprint arXiv:2310.00367**, 2023.

[5] Abhay Zala, Han Lin, Jaemin Cho, and Mohit Bansal. DiagrammerGPT: Generating Open-Domain, Open-Platform Diagrams via LLM Planning. **arXiv preprint arXiv:2310.12128**, 2023.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In **Advances in Neural Information Processing Systems**, Vol. 25. Curran Associates, Inc., 2012.

[7] Bingyi Cao, André Araujo, and Jack Sim. Unifying Deep Local and Global Features for Image Search. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, **Computer Vision – ECCV 2020**, Lecture Notes in Computer Science, pp. 726–743. Springer International Publishing, 2020.

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7514–7528. Association for Computational Linguistics, 2021.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. **arXiv preprint arXiv:1706.08500**, 2018.

[10] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In **2013 IEEE International Conference on Computer Vision Workshops**, pp. 554–561. IEEE, 2013.

[11] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 – A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In **2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. IEEE, 2020.

[12] Yadollah Yaghoobzadeh, Katharina Kann, T. J. Hazen, Eneko Agirre, and Hinrich Schütze. Probing for Semantic Classes: Diagnosing the Meaning Content of Word Embeddings. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5740–5753. Association for Computational Linguistics, 2019.

[13] Katarzyna Krasnowska-Kieraś and Alina Wróblewska. Empirical Linguistic Study of Sentence Embeddings. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5729–5739. Association for Computational Linguistics, 2019.

[14] Yichu Zhou and Vivek Srikumar. DirectProbe: Studying Representations without Classifiers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5070–5083. Association for Computational Linguistics, 2021.

[15] 坂田将樹, 横井祥, Benjamin Heinzerling. 事前学習済み言語モデルによるエンティティの概念化. 言語処理学会, pp. 1310–1315, 2023.

[16] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. **Information Retrieval**, Vol. 12, No. 4, pp. 461–486, 2009.

[17] Jonathon Shlens. A Tutorial on Principal Component Analysis. **arXiv preprint arXiv:1404.1100**, 2014.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. **arXiv preprint arXiv:1512.00567**, 2015.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. **arXiv preprint arXiv:2103.00020**, 2021.

[20] 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**, 2020.

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. **Journal of Machine Learning Research**, Vol. 9, No. 86, pp. 2579–2605, 2008.

[22] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **arXiv preprint arXiv:1802.03426**, 2020.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　埋め込み分布の可視化

A.1　t-SNE，UMAP による意味の正しさに関わる要素についての可視化

t-SNE [21]，UMAP [22] を用いて次元圧縮した場合の意味の正しさに関する要素についての埋め込み分布をそれぞれ図 5，図 6 に示す．どちらの次元圧縮手法を用いた場合でも，CLIP の埋め込み分布はテキストの有無によってクラスタが分かれていることが確認でき，これは表 1 の結果と整合している．Inception V3 の埋め込み分布は，t-SNE を用いて次元圧縮した場合にはテキストの有無によってクラスタが分かれていることが確認でき，これは表 1 の結果と整合している．

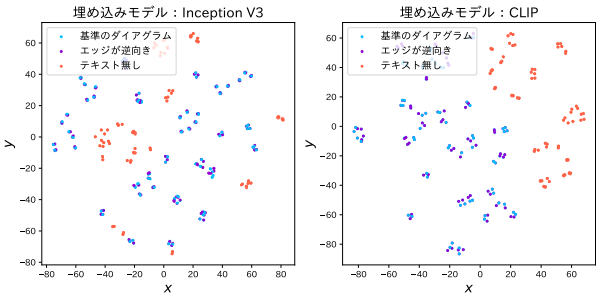

図 5　意味の正しさについての埋め込み分布．次元圧縮には t-SNE を用いた．

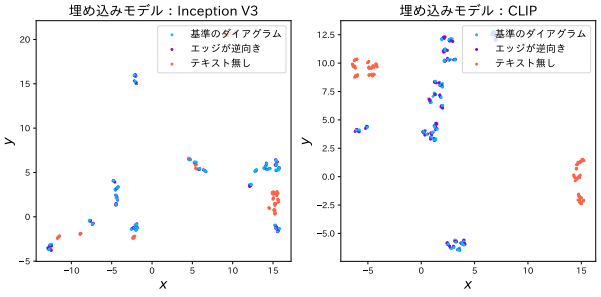

図 6　意味の正しさについての埋め込み分布．次元圧縮には UMAP を用いた．

A.2　PCA，t-SNE，UMAP によるエッジの色についての埋め込みの可視化

PCA，t-SNE，UMAP を用いて次元圧縮した場合のエッジの色についての埋め込み分布をそれぞれ図 7，図 8，図 9 に示す．どの次元圧縮手法を用いた場合でも，両モデルの埋め込み分布がエッジの違いによって明確に分かれている様子は見られなかった．一方で，t-SNE を用いて次元圧縮した場合には小さなクラスタレベルでは分かれているものもあり，その中にはペアになっているように見えるものもあった．

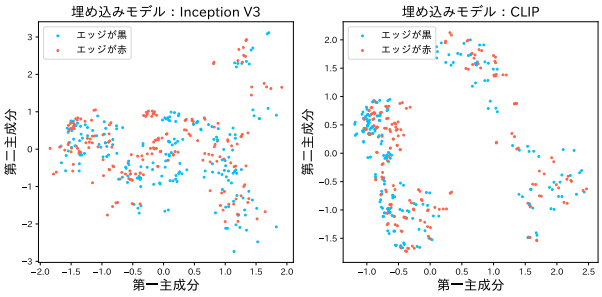

図 7　エッジの色についての埋め込み分布．次元圧縮には PCA を用いた．

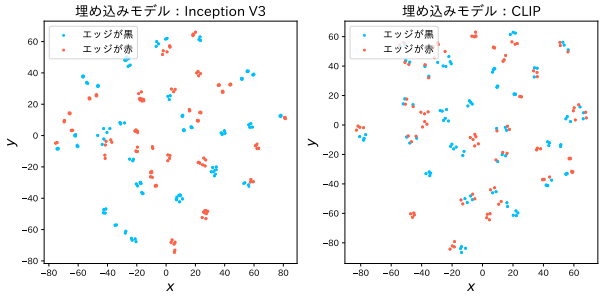

図 8　エッジの色についての埋め込み分布．次元圧縮には t-SNE を用いた．

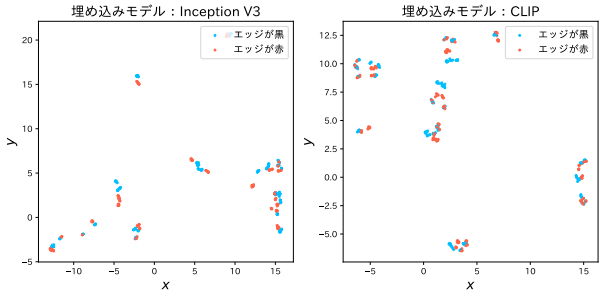

図 9　エッジの色についての埋め込み分布．次元圧縮には UMAP を用いた．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).