集合型視覚言語埋め込み

品川 政太朗[1,a)]　中村 哲[1]

概要：画像と言語を共有の埋め込み空間に射影する視覚言語モデルは，両モダリティ間の検索や生成において幅広く利用されている．しかし，学習済みの視覚言語モデルは，言語側のエンコーダの入力トークン長が限定的であり，長い文に対しては扱いにくいという問題がある．本研究では，大域的な一つの埋め込み同士で画像と言語間の類似度を計算する従来の枠組みを見直し，画像と自然言語の文をそれぞれ複数の埋め込みによる集合で表現して類似度を計算する方法を採用することで，上記の問題の回避を試みる．両モダリティの埋め込み集合間の類似度計算には最適輸送とプーリングによる方法を提案する．本手法は，学習済みの視覚言語モデルに追加の学習を行わずに適用可能である．実験では，代表的な視覚言語モデルである CLIP を対象とし，画像—言語間の検索における検証結果を報告する．

キーワード：視覚言語モデル，CLIP，Visual Semantic Embedding

1. はじめに

視覚言語モデル（Vision-Language Models）は，画像情報と言語情報の二種類のモダリティから，両者の意味的関係を紐づけて問題を解決する深層学習モデルである．視覚言語モデルのうち，画像と言語を共有の埋め込み空間（Visual Semantic Embedding: VSE）に射影するモデルである CLIP [1] は，画像エンコーダと言語エンコーダからなるモデルであり，マルチモーダル大規模言語モデル [2], [3] の画像エンコーダやテキストから画像を生成する Stable Diffusion [4] の言語エンコーダに用いられるなど，基盤モデルの中でも重要な存在として知られている．

CLIP のような視覚言語モデルの課題の一つは，学習時の制約として有限のトークン長で学習されており，推論時にそれよりも長いトークン長には物理的に対応できない点である．本研究では，上記の問題の改善方法として，CLIP のような VSE 手法の多くが従っていた，大域的な一つの埋め込み同士で画像と言語間の類似度を計算する従来の枠組みを見直し，画像と文をそれぞれ複数の埋め込みによる集合で表現して類似度を計算する方法を検討する．

図 1 に，CLIP の通常の VSE（左図）と提案手法である集合型視覚言語埋め込み（右図）の概要図を示す．CLIP には単一の画像もしくは文の入力が与えられ，その埋め込みを出力する．画像と言語の埋め込みは潜在空間を共有しているため，埋め込み間の類似度を測ることで，与えられたサンプルの中で最も近い埋め込み同士を対応するサンプルとして選ぶことができる．CLIP の画像認識やモダリティ間の検索はこの仕組みに基づいている．一方，提案手法の集合型視覚言語埋め込みは，画像もしくは文が集合として与えられ，それぞれを埋め込みに変換して埋め込みの集合を作る点が異なる．ここで，画像の集合とは，図 1 の例では画像を物体ごとの小領域の矩形に分割して得られる集合を想定したが，原理的には動画の連続フレームを画像集合とするなどの応用も考えられる．文の集合とは，図 1 の例では物体ごとの説明文による集合を想定したが，原理的には一つの長い文章を分割したものであるとか，特定の情報についてそれぞれ要約した情報であるとか，質問と応答を結合した文といったデータも適用できる可能性がある．

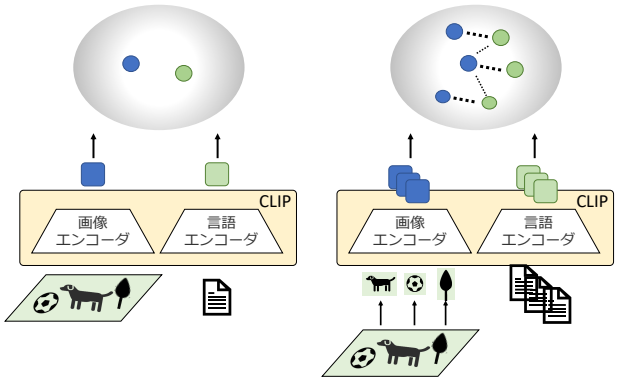

図 1　CLIP の VSE（左図）と集合型視覚言語埋め込み（右図）

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
Takayama-cho 8916-5, Ikoma, Nara, 630-0101, Japan
[a)] sei.shinagawa@is.naist.jp

提案手法の重要な点は，画像と言語が集合として与えられており，集合間で類似度を計算する点である．これにより，限られたトークン長を持つ視覚言語モデルでも，より長いトークン長を扱えることが期待できる．ただし，集合型視覚言語埋め込みでは埋め込み集合間の類似度を測る必要がある．最適輸送は集合間の類似度尺度として有望であると考えられるため，本研究では最適輸送とプーリング操作を併用したシンプルな方法を提案する．

集合型視覚言語埋め込みの利点は，学習済みの視覚言語モデルをそのまま用いることができるため追加の学習が不要である点である．視覚言語モデルのトークン長は技術の発展によって今後伸長することが期待されるが，視覚言語モデルが有限のトークン長を持つ限りは本研究の知見が役に立つと期待できる．

実験では，CLIP の画像エンコーダと言語エンコーダを直接的に利用するモダリティ間検索において，提案手法の集合型視覚言語埋め込みを従来の CLIP の VSE と比較することで，提案手法の有効性を検証する．

2. 関連研究：画像と言語の共有埋め込み

画像情報と言語情報のように異なる情報源をそれぞれ符号化し，共有の埋め込み空間（Visual Semantic Embedding: VSE） [5], [6] に射影する方法論は Vision and Language 分野の主要なトピックとして広く研究されてきた．初期の VSE 研究では，画像とそのクラスラベルの共有埋め込みを学習する手法 [7] や画像と文の共有埋め込みを学習する手法 [8], [9], [10] が見られた．これらの手法は対応する画像情報と言語情報をそれぞれ大域的に捉えて一つの埋め込みに射影するため，画像中に複数の物体があるような場合に，画像とテキストの局所的な対応関係を捉えることを苦手としていた．この問題に対して，Unified VSE [11] では，画像と文の大域的情報だけでなく，画像の局所領域と文中の名詞句を共有埋め込み空間に射影する．Unified VSE は画像と言語の部分的な埋め込みを考える点で本研究の提案手法に類似する．ただし，Unified VSE は画像中の物体の矩形情報のラベル付けや文の構文解析を伴う学習が必要である．それに対して，提案手法は学習済みの視覚言語モデルを利用するため追加の学習なしに推論できる利点がある．

視覚言語モデルにおいて，画像と言語の埋め込み集合間に最適輸送を適用する先行研究には ViLT [12] がある．ViLT で用いられている Word Region Alignment は最適輸送を用いた損失関数で，類似する単語埋め込みと画像のパッチ埋め込みが近づくように学習を促す働きがある．本研究では，最適輸送を単語ではなく，文の埋め込み集合に適用し，画像集合と言語集合の対応関係の定量化，すなわち埋め込み集合間の類似度計算に利用する点が異なる．

3. 従来手法：CLIP による VSE

CLIP は，図 1 の左図のように，画像エンコーダ E_v により視覚埋め込み $e^v = E_v(v)$ を作成し，言語エンコーダ E_s により言語埋め込み $e^s = E_s(s)$ を作成する．CLIP において，e^v と e^s は一つの VSE 上に存在する制約の下で学習されており，e_v と e_s 間の類似度はコサイン類似度 $\cos(e^v, e^s)$ で定義される．したがって，言語埋め込み e^s をクエリとして，複数の視覚埋め込みから最も近い視覚埋め込みを検索する場合は，$\hat{e^v} = \min_{e_i^v \in E_v} \cos(e^v, e^s)$ で与えられる（ただし，候補の視覚埋め込み集合を E_v とした）．

4. 提案手法：集合型視覚言語埋め込みと最適輸送による類似度計算

提案手法では，サンプルあたりの画像情報と言語情報が一対一であった点が拡張され，画像集合 $V = \{v_1, v_2, \ldots, v_N\}$ と言語集合 $S = \{s_1, s_2, \ldots, s_M\}$ で与えられる．画像集合を作成する方法として，画像一枚を部分画像に分割する場合は，物体検出器や Semantic-SAM [13] のように画像の部分領域を抽出するモジュールが必要になる．文集合を作成する方法として，長文を分割する場合は，大規模言語モデルなどであらかじめ短文に分割する必要がある．画像集合と言語集合は，各要素を画像エンコーダと言語エンコーダにより，それぞれ視覚埋め込み集合 $E_v = \{e_1^v, e_2^v, \ldots, e_N^v\}$ と言語埋め込み集合 $E_s = \{e_1^s, e_2^s, \ldots, e_M^s\}$ に変換される．

埋め込み集合同士の類似度計算には，集合同士が全体的に似ている点を満たす手法が望ましい．そこで，本研究では最適輸送とプーリングによる手法を提案する．**図 2** に提案手法の概要図を示し，以下，詳細に説明する．

4.1 IPOT による各埋め込み間の類似度計算

最適輸送には，Inexact Proximal Point Method for Optimal Transports (IPOT) [14] を用いる．IPOT は，generalized KL Bregman divergence と呼ばれる制約項を最適輸送の目的関数に導入することで，行列の並列計算により近似的に輸送量（埋め込み集合間の各埋め込み同士の類似度）を求めることができる．この輸送量が平均的に高ければ，全体的に対応しているサンプルだと判断できる．

まず，IPOT が前提とする，最適輸送の問題設定について説明する．埋め込み集合間における最適輸送問題は，二つの埋め込み集合を分布として捉え，一方の埋め込み集合の分布を他方の埋め込み集合の分布へ最小のコストで変形させる問題とみなせる．ここでは，視覚埋め込み集合の分布 $\mu = \frac{1}{N}\sum_{n=1}^{N} \delta[e_n^v]$ を言語埋め込み集合の分布 $\nu = \frac{1}{M}\sum_{m=1}^{M} \delta[e_m^s]$ に変形する問題を考える[*1]（ただし $\delta[e_n^v]$ はデルタ関数で，$\delta[e^v = e_n^v] = 1$，$\delta[e^v \neq e_n^v] = 0$

[*1] 原理的にはどちらからの変形でも構わない．

を満たす関数であり，$\delta[e_m^s]$ についても同様である）．ここで，視覚埋め込み集合の各埋め込みは $\frac{1}{N}$ ずつの質量を持つと考え，これを言語埋め込み集合の各埋め込みに輸送する．視覚埋め込み e_n^v から言語埋め込み e_m^s への単位当たり質量の移動にかかるコスト C_{nm} を定義し，この二点間で生じる輸送量行列 T を定義すると，輸送にかかる総コストは $\sum_{n=1}^{N}\sum_{m=1}^{M} T_{nm}\cdot C_{nm}$ と定義できる．この総コストを最小化するように輸送量を最適化するのが最適輸送であり，IPOT はこの最適化を行う手法の一つである．IPOT は ViLT [12] の Word Region Alignment でも用いられている手法であり，本研究は ViLT を参考にコストとハイパーパラメータを設定した．すなわち，画像と言語の埋め込み集合間での単位当たり質量の移動にかかるコストは埋め込み間の距離であると考えられることから，$C_{nm} = 1 - \cos(e_n^v, e_m^s)$ とおいた．制約項の強さに対応する変数 β は $\beta = 0.5$ とした．また，行列計算を繰り返す回数 N_{iter} は $N_{iter} = 50$ とした[*2]．

4.2 輸送量行列のプーリングによる埋め込み集合間の類似度計算

モダリティ間の検索を行うためには，複数の画像と言語のサンプル間について輸送量を計算して最終的な類似度を求める．図 2 の下図は画像サンプル集合 $\{E_{v1}, E_{v2}\}$ と言語サンプル集合 $\{E_{s1}, E_{s2}, E_{s3}\}$ がある場合の輸送量行列の概念図である．ただし，$E_{vi} = \{e_1^{vi}, e_2^{vi}, e_3^{vi}\}$ は i 番目の画像サンプルの埋め込み集合が三つの視覚埋め込みで構成されていることを示し，$E_{sj} = \{e_1^{sj}, e_2^{sj}, e_3^{sj}, e_4^{sj}\}$ は j 番目の言語サンプルの埋め込み集合が四つの言語埋め込みで構成されていることを示す．上記の場合，$N = 3$，$M = 4$ である．

輸送量行列は全てのサンプル間について同時に計算されるため，$2N \times 3M$ の行列となる．この中で，i 番目の画像サンプルと j 番目の言語サンプル間の輸送量行列 T^{ij} はそれぞれ $N \times M$ 行列となる．最終的な類似度スコア $\mathrm{Similarity}(i, j)$ はこの T^{ij} の平均値プーリングによって与える（式（1））．

$$\mathrm{Similarity}(i,j) = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M} T_{nm}^{ij} \qquad (1)$$

5. 実験

提案手法の有効性を確認するため，CLIP の画像エンコーダと言語エンコーダを直接的に利用するモダリティ間検索において，図 1 のように，画像と文からそれぞれ物体ごとの局所画像とその説明文の集合を得られる状況を想定し，従来の CLIP の VSE と提案手法を比較する．

[*2] このパラメータの妥当性については，検討する余地があると考えられる．

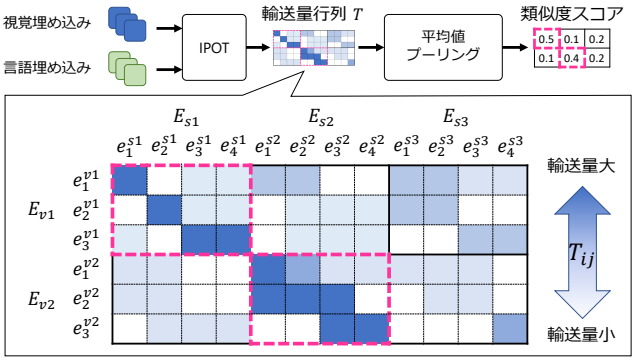

図 2 視覚埋め込み集合 $\{E_{v1}, E_{v2}\}$ と言語埋め込み集合 $\{E_{s1}, E_{s2}, E_{s3}\}$ における集合型視覚言語埋め込みの類似度計算の手続き．枠内の各要素の青色の濃さが輸送量の大きさを示し，点線赤枠が類似度スコア最大のサンプル同士を示す．

5.1 実験設定

評価用データセットには，Densely Captioned Images (DCI) データセット [15] を用いた．このデータセットは，8,012 枚の画像（**大域画像**）に対して，大域画像の説明文の他，画像中に含まれる物体の**局所画像**とその説明文の集合が付与されている．各説明文は人手で付与されたものだが，各説明文には CLIP のトークン長 77 が収まるように Llama2 [16] で要約した説明文も付与されている．本研究では CLIP を対象とするため，要約済みの説明文を用いた．DCI データセットのうち，実際に評価に用いたのは，テスト集合 112 サンプルである．なお，テスト集合中の各画像に付与されている局所画像は平均 12.05 枚である．

評価対象とするモデルには学習済みの CLIP の ViT-B/32[*3]を用いた．評価指標には，モダリティ間検索でよく用いられる R@K を用いた．R@K とは，上位 K 件の候補の中に正解が含まれているサンプルの割合であり，高いほど良い．本実験では $K = 1, 5, 10$ を用いた．実験環境には，AMD Ryzen9 3900X（12 コア/24 スレッド, 3.8GHz），RTX A6000，メモリ 128GB の計算機一台を用いた．

5.2 結果：従来の CLIP の VSE と提案手法の比較

全体の結果を**表 1** に示す．左列から，手法名，画像と言語の条件，言語をクエリとした画像の検索性能，画像をクエリとした言語の検索性能を表す．提案手法は，IPOT と平均値プーリングによる処理に分かれるため，各処理の影響を条件 (a)(b) で検証した後，条件 (c)-(f) を検証した．

VSE+IPOT は，VSE と同様，大域画像とその説明文を扱う条件下で VSE に IPOT のみを適用した場合である．R@1 はそれぞれ 0.964, 0.973 と，従来手法の VSE を +0.15 以上上回った．集合型の埋め込みを前提とせずとも，IPOT を適用するだけで VSE の性能向上に寄与するといえる．

VSE+平均値プーリングは，物体ごとの局所画像とその説明文が得られる状況下で，IPOT を用いず直接 VSE の

[*3] https://github.com/openai/CLIP

表 1 条件ごとのモダリティ間検索の性能（大域：大域的な画像または説明文，集合：物体レベルの局所画像またはその説明文の集合）．

手法		条件		言語 → 画像			画像 → 言語		
		画像	言語	R@1	R@5	R@10	R@1	R@5	R@10
VSE (従来手法)		大域	大域	0.750	0.902	0.946	0.804	0.955	0.991
VSE+IPOT	(a)	大域	大域	0.964	1.000	1.000	0.973	0.991	1.000
VSE+平均値プーリング	(b)	集合	集合	0.036	0.125	0.205	0.036	0.125	0.205
VSE+IPOT+平均値プーリング	(c)	集合	集合	0.964	0.991	0.991	0.964	0.991	0.991
	(d)	大域	集合	0.714	0.893	0.973	0.794	0.929	0.991
（提案手法）	(e)	集合	大域	0.893	0.982	1.000	0.875	0.973	0.991
	(f)	大域＋集合	大域＋集合	0.964	0.991	0.991	0.982	1.000	1.000

類似度によって得られる行列に平均値プーリングを施した場合である．結果は大きく悪化しており，まともに問題が解けていないことが分かる．この点は，複数の埋め込みの単純なプーリングはうまくいかない [15] と述べている先行研究と同じ結果を得られたといえる．

VSE+IPOT+平均値プーリングは，提案手法である集合型視覚言語埋め込みによる結果である．VSE+平均値プーリングの場合 (b) と同条件の場合 (c) を比べると，性能は大きく向上しており，大域画像とその説明文を用いた VSE+IPOT とほぼ同程度の結果が得られた．この結果から，大域的な画像や説明文が得られない状況下でも，提案手法は有効に働く可能性が示唆される．

一方のみが大域的情報である条件 (d)(e) では，画像が大域的である条件 (d) の方が低い性能となった．これは，大域画像に対して画像中の局所領域を表す説明文を対応づけるのが難しいことを示唆している．類似の報告として，CLIP は画像付き質問応答のように画像中の局所的な物体に関する情報抽出を苦手とすることが知られている [17]．これは，CLIP が大域画像とその説明文の対照学習で訓練されたモデルであるためだと考えられており，本研究の結果も同様の要因によって得られた可能性がある．対して，局所画像集合に対して大域画像の説明文を対応させる場合は R@1 で +0.1 程度良好な結果が得られた．経験的に CLIP の言語エンコーダは名詞形容詞に強く反応することが知られており，これが性能向上に寄与したと考えられる．

大域画像とその説明文，局所画像とその説明文の集合の両方を用いる条件 (f) では，条件 (c)-(f) の中で最良の結果を示した．制約が無ければ，利用できる全ての情報源を用いれば良いことが示唆された．ただし，条件 (a) と比べると条件 (f) が低い項目もみられた．これは集合中の埋め込み間の対応関係に全体の結果が引きずられた可能性がある．

5.3 輸送量行列の埋め込み同士の対応関係の可視化

5.2 節条件 (d)(e) の結果の考察を裏付けるため，得られた輸送量行列から埋め込み同士の対応関係の一部（30 サンプル）を**図 3** に可視化した．上図が条件 (d)，下図が条件 (e) に対応する．図 3 の明るい部分が，輸送量が大きい，すなわち，より類似している関係だと解釈される．上図，下図において縦の白線で区切られた区間は同一サンプルに属する埋め込み集合に関する輸送量を示す．縦にはサンプルごとに大域的埋め込みが一つずつ並んでいるため，左上から右下にかけて輸送量が大きく，区間内は横一列に並んでいるほど埋め込みの対応関係を正しく得られているといえる．上図よりも下図の方が正しく対応関係を捉えていることが確認できることから，大域的な視覚埋め込みは言語埋め込みよりも集合型埋め込みとの相性が悪いといえる．

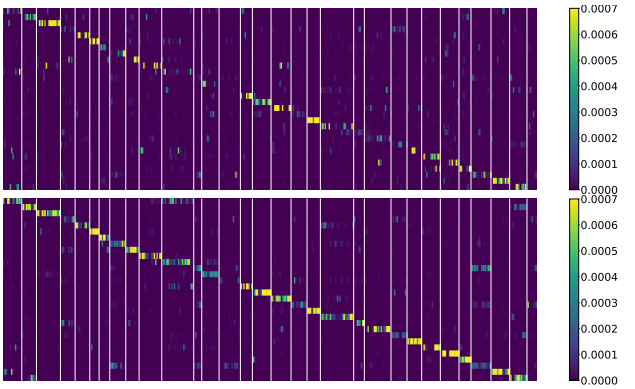

図 3 輸送量行列の一部の可視化．上図：大域画像の視覚埋め込み（縦軸）に対する言語埋め込み集合（横軸），下図：大域画像の説明文の言語埋め込み（縦軸）に対する視覚埋め込み集合（横軸）．縦の白線で区切られた区間内は同一サンプルに属する．

6. おわりに

本研究では，画像と言語の共有の埋め込み空間を持つ視覚言語モデルにおいて，視覚言語埋め込みを集合として扱い画像と言語の対応関係を計算する方法を提案した．実験では，最適輸送の一種である IPOT が画像と言語を対応づけるのに有効であり，集合型視覚言語埋め込みによる類似度計算が実際に可能であることを示した．また，新たな知見として，大域的埋め込みと集合型埋め込みの組み合わせによって性能が変化し得ることも示した．今後は，他のタスクやドメインへの適用可能性の検討が考えられる．

謝辞 本研究は JSPS 科研費 JP21K17806 の助成を受けたものである．

参考文献

[1] Radford, A. et al.: Learning Transferable Visual Models From Natural Language Supervision, *Proc. of ICML*, Vol. 139, pp. 8748–8763 (2021).
[2] Li, J. et al.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models, *Proc. of ICML* (2023).
[3] Liu, H. et al.: Visual Instruction Tuning, *Proc. of NeurIPS* (2023).
[4] Rombach, R. et al.: High-Resolution Image Synthesis With Latent Diffusion Models, *Proc. of CVPR*, pp. 10684–10695 (2022).
[5] Guo, W. et al.: Deep Multimodal Representation Learning: A Survey, *IEEE Access*, Vol. 7, pp. 63373–63394 (2019).
[6] Ueki, K. et al.: Survey of Visual-Semantic Embedding Methods for Zero-Shot Image Retrieval, *arXiv preprint arXiv 2105.07391* (2021).
[7] Frome, A. et al.: DeViSE: A Deep Visual-Semantic Embedding Model, *Proc. of NIPS*, Vol. 26 (2013).
[8] Kiros, R. et al.: Unifying visual-semantic embeddings with multimodal neural language models, *arXiv preprint arXiv:1411.2539* (2014).
[9] Vendrov, I. et al.: Order-Embeddings of Images and Language, *Proc. of ICLR* (2016).
[10] Faghri, F. et al.: VSE++: Improving Visual-Semantic Embeddings with Hard Negatives, *Proc. of BMVC* (2018).
[11] Wu, H. et al.: Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations, *Proc. of CVPR* (2019).
[12] Kim, W. et al.: ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision, *Proc. of ICML* (2021).
[13] Li, F. et al.: Semantic-SAM: Segment and Recognize Anything at Any Granularity, *arXiv preprint arXiv 2307.04767* (2023).
[14] Xie, Y. et al.: A Fast Proximal Point Method for Computing Exact Wasserstein Distance, *arXiv preprint arXiv 1802.04307* (2018).
[15] Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V. and Romero-Soriano, A.: A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions (2023).
[16] Touvron, H. et al.: Llama 2: Open Foundation and Fine-Tuned Chat Models, *arXiv preprint arXiv 2307.09288* (2023).
[17] Shen, S. et al.: How Much Can CLIP Benefit Vision-and-Language Tasks?, *Proc. ICLR* (2022).