ClipQA: 言語特徴埋め込み空間における 3D 画像質問応答

東慶多 [1]　Edison Marrese-Taylor[1,2]　宮尾祐介 [1]
[1] 東京大学　[2] 産業技術総合研究所
{keitaiazuma33, yusuke}@is.s.u-tokyo.ac.jp　emarrese@weblab.t.u-tokyo.ac.jp

概要

近年, 自然言語処理とコンピュータビジョンの融合分野では, 3 次元質問応答（3D-VQA）などの空間理解に関する研究が関心を集めている. 本研究では既存の 3D-VQA モデル ScanQA [1] に対し, Contrastive Language-Image Pretraining (CLIP) 特徴量を活用した 2 つの改善手法—1. 質問埋め込み及び物体特徴量に CLIP 特徴量を付加する手法, 2. 質問応答に用いるマルチモーダルモデル（Transformer）の注意機構に CLIP 特徴量から計算したバイアスを加える手法—を提案し, その効果を検証する. 実験の結果, 提案手法 1., 2. ともに質問応答の精度改善に有効であり, 両手法を用いることで質問応答の精度がベースモデルと比べて 5 ~ 10% 向上することを示した.

1　背景

2015 年以降, 画像質問応答（VQA）[2] は自然言語処理とコンピュータビジョンの融合として盛んに研究されてきた. ここ数年では関心が 2 次元から 3 次元へと移りつつあり [1, 3, 4], RGB-D スキャンなどの 3D データ（点群）を対象とする 3 次元質問応答（3D-VQA）が注目を集めている [1, 5].

3D-VQA では, 空間中の物体の特性や位置関係について質問応答を行い, そのモデルの多くは主に言語エンコーダ, 空間エンコーダ, マルチモーダルモデルの 3 要素から構成される [1, 5]. 言語エンコーダは質問文の埋め込みを計算し, 空間エンコーダは点群データから物体の位置情報と特徴量を抽出する. これらの物体の特徴量と質問文の埋め込みを Transformer などのマルチモーダルモデルで相互参照することにより質問応答を実現する。本研究では, 既存の 3D-VQA モデル ScanQA [1] において、言語埋め込み, 及び物体特徴量に Contrastive Language-Image Pretraining(CLIP) を活用することで質問応答の精度がどのように変化するかを検証する.

2　関連研究

本論文では既存の 3D-VQA 手法：ScanQA [1] と, 点群に CLIP 特徴量を埋め込む手法：OpenScene [3] とを掛け合わせ, ScanQA の精度を改善する新たな 3D-VQA 手法を提案する. 本節では, ScanQA と OpenScene について, モデルの詳細を解説する.

2.1　ScanQA

ScanQA [1] は点群データ $p \in \mathcal{P}$ と質問 $q \in \mathbb{Q}$ を入力として受け取り, データセットに含まれる全回答集合の中から最適と考えられる回答 $a \in \mathbb{A}$ を出力する 3D-VQA モデルであり, i) 言語エンコーダ, ii) 空間エンコーダ（物体検出モデル）, iii) マルチモーダルモデルの大きく 3 部分から構成される.

i) **言語エンコーダ**　言語エンコーダでは, 質問単語列 $\{w_i\}_1^{n_q}$ の GloVe 埋め込み $Q \in \mathbb{R}^{n_q \times 300}$ を計算し, これを LSTM に入力することで文脈に応じた単語埋め込み $Q' \in \mathbb{R}^{n_q \times d}$ を計算する. これを非線形層に通すことで最終的な質問文の埋め込み表現 $Q_{\mathrm{emb}} \in \mathbb{R}^{n_q \times d}$ を得る. （n_q：質問文に含まれる最大単語数, d：単語埋め込みの次元. また, 点群/物体の特徴量も同じ次元で表現する.（デフォルトでは 256））

$$Q_{\mathrm{emb}} = MLP_{\mathrm{Q}}(Q') \tag{1}$$

ii) **空間エンコーダ**　空間エンコーダでは PointNet++ [6] をバックボーンとする VoteNet [7] を用いて物体検出を行う. まず PointNet++を用いてダウンサンプリングされた点群データ $p_{\mathrm{ds}} \in \mathbb{R}^{n_{\mathrm{ds}} \times d}$ を得る. その後 VoteNet を適用することで物体の位置情報 $V_{\mathrm{box}} \in \mathbb{R}^{n_v \times 8}$ 及び物体特徴量 $V \in \mathbb{R}^{n_v \times d}$ を得る. 更に, 物体特徴量 V を非線形層に通すことで最終的な物体特徴量 $V_{\mathrm{emb}} \in \mathbb{R}^{n_v \times d}$ を得る. （d_p：各頂点の特徴量の次元, n_{ds}：ダウンサンプル後の頂点数（デフォルトでは 1024）, n_v：検出する物体の上限（デフォルトでは 256））

$$V_{\mathrm{emb}} = MLP_{\mathrm{V}}(V) \tag{2}$$

iii) **マルチモーダルモデル**　マルチモーダルモデルでは, Transformer のエンコーダ・デコーダを用いて, 質問文の埋め込み表現 Q_{emb} 及び物体特徴量 V_{emb} の関連性を計算する. エンコーダでは, Q_{emb} を L(= 2) 層のエンコーダ層に通し, 質問文のエンコード結果 $Q^{\text{enc}} \in \mathbb{R}^{n_q \times d}$ を得る. デコーダでは L 層のデコーダ層に query として物体特徴量 V_{emb} を, key 及び value として言語のエンコード結果 Q^{enc} を与え, 物体特徴量のデコード結果 $V^{\text{dec}} \in \mathbb{R}^{n_v \times d}$ を得る. また, Q^{enc} と V^{dec} を 1 つの MLP に通すことで, 質問と空間情報の統合表現 $f \in \mathbb{R}^d$ を得る.

質問応答　ScanQA では, 点群と質問の入力に対して, 回答集合から回答を選び, 回答が参照する物体の位置とクラスを併せて出力する. 具体的には, 物体特徴量 V^{dec} 及び統合表現 f を複数の MLP に入力し softmax を計算することで, 回答の確率分布 p_{ans}, 回答が参照する物体の確率分布 p_{obj}, 回答のクラスの確率分布 p_{cls} を出力する. 学習時は, これらの出力に対する交差エントロピー損失, 及び VoteNet の物体検出に対する損失の和:$L = L_{\text{ans}} + L_{\text{obj}} + L_{\text{cls}} + L_{\text{det}}$ をモデル全体の損失として学習する.

2.2 OpenScene

CLIP [8] は画像-テキストペア (I, T) の対照学習により, テキストと画像を同一の特徴空間に射影するよう事前学習された画像と言語のマルチモーダルモデルである. CLIP による埋め込みは, 関連度の高いテキスト-画像ペアのコサイン類似度が高くなるよう学習されており様々な下流タスクの zero-shot 性能を高めることができる [4, 9].

OpenScene [3] は, 画像のピクセル毎の CLIP 特徴量を計算するモデル：OpenSeg [10] を用いて CLIP 特徴量を点群に射影することで, 点群に CLIP 特徴量を埋め込むことを可能にした. 実際に, 図 1 では質問の対象である「冷蔵庫」を正しく特定できており, 点群に CLIP 特徴量を埋め込むことで, 空間中の質問文との関連性が高い領域を限定できることが視覚的に確認できる.

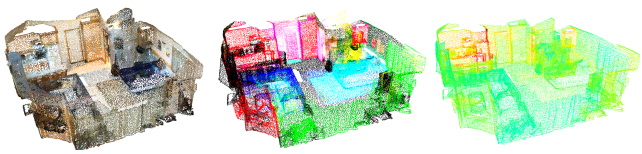

図 1　点群データ（左）, OpenScene により計算された CLIP 特徴量を可視化したもの（中）, 質問文:“What is placed next to the fridge?” に対するヒートマップ（右）

3 提案手法

ScanQA [1] では, 言語の埋め込みと点群に与えられる特徴量には関連性がなく, 言語エンコーダと空間エンコーダの精度が質問応答の精度に直結する. 一方で, OpenScene [3] に示されたように, 初めから点群に CLIP 特徴量を埋め込むことができれば, 質問文との関連性がより明確になり, 質問応答の精度が改善するものと考えられる.

そこで本研究では, ScanQA モデルに OpenScene から計算された点群の CLIP 特徴量を入力として与えることで, 質問応答の精度がどう変化するかについて検証する（ベースモデル）. 更に, 精度改善の為の工夫として 1）物体の特徴量, 及び単語埋め込みに CLIP 特徴量を付加する手法, 2）マルチモーダルモデルの注意機構に CLIP 特徴量から計算したバイアスを与える手法, の 2 点で ScanQA を改良したモデル ClipQA を提案し, その結果について議論する.

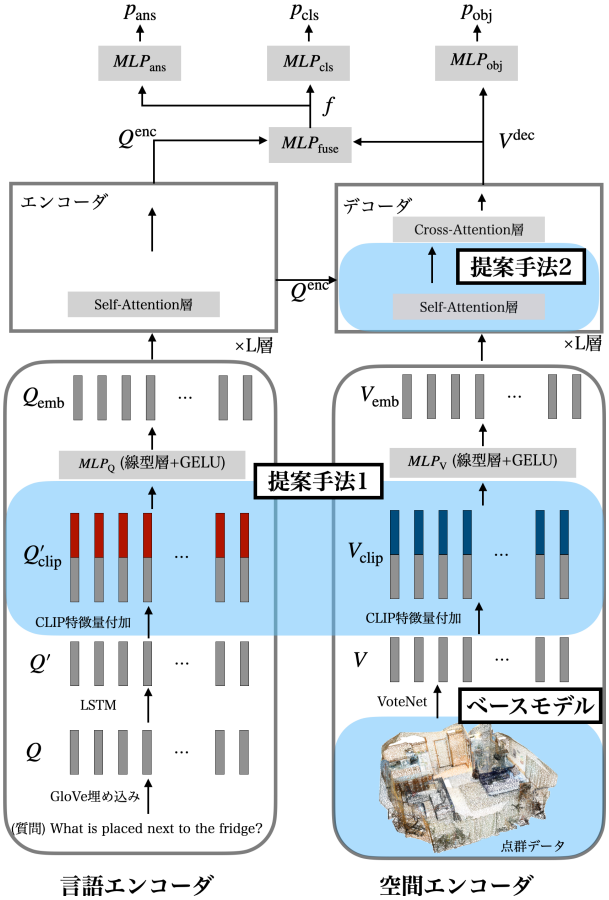

図 2　ClipQA モデル概要：ベースモデルでは ScanQA の入力を OpenScene から得られる点群データ $p \in \mathbb{R}^{n_P \times 778}$ に差し替える. 提案手法 1 では非線形層の直前で言語/物体特徴量に CLIP 特徴量を付加する. 提案手法 2 ではデコーダの self-attention 層の注意機構にバイアスを加算する.

3.1 ベースモデル

ScanQA では, 位置 $x \in \mathbb{R}^{n_p \times 3}$, 色 $c \in \mathbb{R}^{n_p \times 3}$, 各点での法線ベクトル $n \in \mathbb{R}^{n_p \times 3}$, 高さ $h \in \mathbb{R}^{n_p \times 1}$, 及び ENet [11] から得られる特徴量 $f_{\text{multiview}} \in \mathbb{R}^{n_p \times 128}$ を連結し得られる点群データ $p \in \mathbb{R}^{n_p \times 138}$ をモデルの入力として与えている. ベースモデルでは, $f_{\text{multiview}}$ の代わりに, OpenScene [3] から得られる頂点ごとの CLIP 特徴量 $f_{\text{openscene}} \in \mathbb{R}^{n_p \times 768}$ を用い, 点群 $p \in \mathbb{R}^{n_p \times 778}$ をモデルの入力として与えるものとする. (n_p：点群の頂点数)

3.2 提案手法 1：CLIP 特徴量の付加

ベースモデルでは, VoteNet による物体検出の過程で CLIP 特徴量が変換されるため, 物体特徴量 V_{emb} が元々与えられていた CLIP 特徴量から乖離してしまうほか, 言語エンコーダ側で CLIP を用いていない為, 画像-言語間の類似度を表現する CLIP の効果を得られていない. そこで, ベースモデルの精度を改善する工夫として, 物体特徴量 V_{emb} 及び言語埋め込み Q_{emb} に物体と質問の CLIP 特徴量を連結して与える手法を提案する. 具体的には, (1), (4) 式にある 2 つの MLP が物体/質問それぞれの CLIP 特徴量を受け取るよう次のように拡張する.

$$Q_{\text{emb}} = MLP'_{\text{Q}}(Q', Q_{\text{clip}}) \tag{3}$$

$$V_{\text{emb}} = MLP'_{\text{V}}(V, V_{\text{clip}}) \tag{4}$$

ここで, $Q_{\text{clip}} \in \mathbb{R}^{768}$ は質問文の CLIP 特徴量, $V_{\text{clip}} \in \mathbb{R}^{n_v \times 768}$ は検出された物体 bounding box(= V_{box}) 内の CLIP 特徴量の平均である. このように言語/物体の両方に CLIP 特徴量を付加することで, 後続のマルチモーダルモデルにおいても CLIP 特徴量が表現する質問-物体間の関係が有効活用され, 質問応答の精度が改善するものと期待される.

3.3 提案手法 2：マルチモーダルモデル（デコーダ）へのバイアス付加

ベースモデルでは, マルチモーダルモデルのデコーダ部分において, 注意機構における attention は学習中に End-to-End で学習される. 一方で, 物体及び質問の CLIP 特徴量が既知である場合, そのコサイン類似度を計算することで質問と各物体の関連度が分かると考えられる. ここでは, 質問応答において i) 物理的に近接した物体同士は同時に参照されやすいこと, ii) 質問と関連度の高い物体により注目すべきであること, の 2 つの観点に基づいて, 次のような注意行列 $M_{\text{Bias\_Attn}}$ を考える.

$$M_{\text{dist}} = \text{softmax}(-\text{dist}(V_{\text{box}})) \tag{5}$$

$$M_{\text{relativity}} = \text{softmax}(Q_{\text{clip}} \cdot V_{\text{clip}}) \tag{6}$$

$$M_{\text{Bias\_Attn}} = \text{softmax}(M_{\text{dist}} * M_{\text{relativity}}) \tag{7}$$

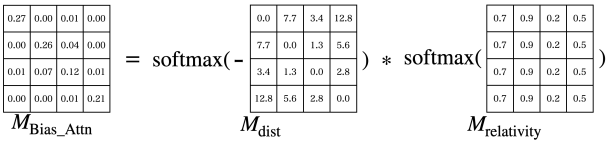

図 3 距離及び CLIP 関連度を用いた注意行列計算例

ここで dist 関数は検出された物体の中心間のユークリッド距離を計算する関数, $M_{\text{Bias\_Attn}} \in \mathbb{R}^{n_v \times n_v}$ は $M_{\text{dist}} \in \mathbb{R}^{n_v \times n_v}$ と $M_{\text{relativity}} \in \mathbb{R}^{n_v \times n_v}$ の要素ごとの積により計算するものとする. 直感的には M_{dist} は物理的に近い物体はより相互に注目されるべきことを表し, $M_{\text{relativity}}$ は質問と高い関連度を持つ物体に, より注目すべきことを表している.

学習時は, $M_{\text{Bias\_Attn}}$ とデコーダが計算したアテンションを 0 ~ 1 の割合で内分して用いるものとし, 内分の割合はヘッド毎に変えられるよう学習可能なパラメータを設定した.

$$Final\_Attn = interp(Decoder\_Attn, M_{\text{Bias\_Attn}}) \tag{8}$$

これにより質問応答の際にモデルが注目すべき物体が明確化され, 精度が改善するものと期待される.

4 データセット

本研究では, ScanNet [12] が公開する屋内スキャンデータセット及び, これに対する質問応答のデータセットである ScanQA（データセット[1)]）[1] を用いる. また, 点群に与える CLIP 特徴量は ScanNet に含まれる RGB-D データに OpenScene[2)] を適用し, 事前計算したものを用いた.

5 実験

ここでは ScanQA データセットを用いて, ScanQA と ClipQA（ベースモデル）, 及びベースモデルに上記変更を施した ClipQA(提案手法 1 のみ), ClipQA（提案手法 2 のみ）, ClipQA(提案手法) をそれぞれ 20 epoch ずつ学習させ, 提案手法の有効性を検証する. 学習時の最適化アルゴリズムには Adam を適用し, パラメータは次のように設定した. 学習率:$lr = 0.0005$, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 1.0 \times 10^{-5}$.

1) ScanQA（モデル）を提案した Azuma ら [1] がモデルと同時に公開した 3D-VQA 用のデータセット. なお, 本研究では test_w/o_objects 以外のデータを用いて実験を行った.
2) CLIP モデルには OpenSeg（ViT-L/14@336px）を用いた.

表 1 テストデータセットでの実験結果

Model	EM@1	EM@10	BLEU 1	BLEU 4	METEOR	ROUGE	CIDEr
ScanQA	**21.945**	**54.843**	28.991	8.444	12.649	32.291	61.815
ClipQA（提案手法）	21.764	54.562	**31.229**	**11.417**	**13.073**	**33.022**	**64.203**

表 2 開発データセットでの実験結果

Model	EM@1	EM@10	BLEU 1	BLEU 4	METEOR	ROUGE	CIDEr
ScanQA	19.294	49.155	28.340	**10.559**	11.939	31.136	58.283
ClipQA（ベースモデル）	19.037	48.813	27.056	8.768	11.735	30.367	56.651
ClipQA（提案手法 1 のみ）	19.444	49.668	**29.743**	8.102	**12.401**	31.839	60.167
ClipQA（提案手法 2 のみ）	19.487	49.711	28.304	7.319	12.222	31.599	57.775
ClipQA（提案手法）	**19.765**	**50.866**	28.883	9.953	12.254	**31.861**	**60.800**

5.1 実験結果

ClipQA(提案手法) と既存モデル ScanQA[3] のテストデータでの評価結果を表 1, 開発データでの学習結果を表 2 に示す. 各表において, 全体を通した最高精度を太字で, 表 2 においては提案手法のみで比較した場合の最高精度に下線を引いて記載した.

5.2 ScanQA との比較

テストデータによる評価からは, ClipQA では回答の完全一致を示す EM スコアは ScanQA に僅かに劣るものの, その他の評価指標においては ClipQA が ScanQA の精度を平均 1.75 ポイント上回った. また, 開発データによる結果においても, BLEU 4 を除くすべての指標で ClipQA が ScanQA の精度を上回っている. 特に, テストデータでの BLEU1, BLEU 4, CIDEr, 及び開発データでの CIDEr では ClipQA が ScanQA の精度を 3 ポイント近く上回っており, これは提案手法が ScanQA に対して優位であることを示していると考えられる.

5.3 提案手法同士の比較

開発データによる評価では, ベースモデルでは全ての指標で ScanQA を下回ったが, 両者ではモデルが同一である為, これは CLIP と ENet の特徴量の埋め込み空間の質の違いによるものと考えられる.

一方で, 提案手法同士の性能をベースモデルを基準として見た場合には, 片方の手法のみを用いた場合の BLEU4 を除くすべての指標で精度が向上しており, これは提案した 2 手法がいずれも質問応答に有効であることを示していると考えられる. 特に, 提案手法 1 は単体でも一部の評価指標で最高精度を達成しているほか, 提案手法 2 のみを用いたモデルよりも, 両提案手法を用いたモデルの方が全評価指標において精度が改善していることから, 2 つの提案手法の中では提案手法 1 がより精度改善に貢献していると考えられる.

また, 提案手法 1,2 の一方のみでは BLEU 4 においてベースモデルよりも精度が悪化する一方で, 2 つの手法を組み合わせたモデルはベースモデルと比較して, 全評価指標を通して精度が 3.8 ~ 15.0% 向上しており, 提案手法 1,2 を同時に用いることの有効性を示唆していると考えられる.

6 結論

本論文では 3D-VQA モデルへの入力として点群に CLIP 特徴量を与えることにより, 質問応答の精度がどのように変化するかを検証した. その結果, CLIP 特徴量をモデルに単純に入力するだけでは既存手法：ScanQA の精度を下回ったが, i) 質問/物体特徴量に CLIP 特徴量を付加する, ii) マルチモーダルモデルの注意機構にバイアスを付加をするよう改善した提案手法では ScanQA よりも高い精度を達成することが示された. これは CLIP 特徴量が質問応答に有用な情報を含むことを示唆しており, 今後 3D-VQA 以外の VL タスクへの CLIP の応用が期待される.

今後の展望としては, 本研究で手動で設計した注意機構のバイアスの改良, またはデコーダの attention とバイアスの差分に対して罰則値を設けるなどの別手法について検討したい.

3) ベースモデルに合わせて GloVe と LSTM を用いた ScanQA のモデルを比較対象とした. これにより ScanQA の元論文の精度と差異があることに注意されたい.

参考文献

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, 2022.

[2] 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, 2015.

[3] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, 2023.

[4] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In **International Conference on Computer Vision (ICCV)**, 2023.

[5] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. **IEEE Transactions on Visualization and Computer Graphics**, pp. 1–16, 2022.

[6] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. **arXiv preprint arXiv:1706.02413**, 2017.

[7] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In **Proceedings of the IEEE International Conference on Computer Vision**, 2019.

[8] 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. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[9] Nur Muhammad (Mahi) Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. **ArXiv**, Vol. abs/2210.05663, , 2022.

[10] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In **ECCV**, 2022.

[11] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation, 2016.

[12] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In **Proc. Computer Vision and Pattern Recognition (CVPR), IEEE**, 2017.