画像キャプションを利用した IconQA タスクへのアプローチ

塩野大輝 [1]　宮脇峻平 [1,2]　長澤春希 [1]　鈴木潤 [1,3]
[1] 東北大学　[2] 株式会社キーウォーカー　[3] 理化学研究所
{daiki.shiono.s1,haruki.nagasawa.s8}@dc.tohoku.ac.jp
{jun.suzuki,shumpei.miyawaki.b7}@tohoku.ac.jp

概要

本研究では，抽象的なダイアグラム画像の読解と多様な推論能力を必要とする Visual Question Answering（VQA）タスクの 1 つである IconQA [1] に取り組む．我々は VQA タスクを解くにあたって，1）視覚情報と言語情報における意味表現の紐付け，2）言語空間上での視覚推論における大規模言語モデルの推論能力の活用，を目的にキャプション生成モデルを用いた視覚情報の拡張を行う．本研究では IconQA で定義された各スキル集合において，視覚情報の拡張が VQA モデルの推論能力にどのような影響を与えるか調査を行う．実験結果より，キャプションによる視覚情報の拡張が VQA タスクを解く手がかりを提供する可能性があることを示す．

1　はじめに

視覚情報と言語情報の意味関係を結びつけたマルチモーダルな知識 [2] を計算機が獲得することは，実社会において人工知能研究が目指す最終目的の 1 つである．この目的の実現に向けた取り組みの 1 つとして，画像中の視覚情報に関連する質問に解答することで計算機の読解能力を定量的に評価する VQA [3, 4] タスクが研究されている．特にアイコンのような抽象的なダイアグラム画像を推論対象とした[1)] IconQA [1] では，VQA タスクを解く上で必要とされる 13 の推論スキル集合を提案しており，例えば物体認識やテキスト読解，常識推論や数値推論などのスキルが含まれる．近年では Transformer [5] を用いて，言語情報と視覚情報の関連する意味表現を紐づけることで質問に解答する VQA モデルが多く提案されているが，豊富な視覚情報を含むダイアグラム画像を対象にした VQA モデルは少なく，その方法論と推論スキル別の関係性については十分に明らかにされていない．

本研究では，キャプション生成モデルを用い，画像を言語として記述する視覚情報の拡張が，様々な推論スキルに対してどのような影響をもたらすのか調査することを目的として，IconQA のベースラインモデルである pyramid patch cross-modal Transformer（Patch-TRM） [1] を用いて入力画像と質問に加えてキャプションを後期接続するモデルを提案する．本研究の貢献は以下の通りである．

- 視覚情報と言語情報の意味表現を紐付けることを目的に，視覚情報の拡張として入力画像のキャプションを VQA モデルに組み込む方法を提案し，推論スキル別にその影響を評価する．
- 大規模言語モデルの推論能力を活用した言語空間上での視覚推論を行うために，言語情報として記述されたキャプションを入力とした GPT-3 [6] の推論性能について調査する．
- キャプション生成を用いた視覚情報の拡張が視覚推論に有効であることを示した．

2　関連研究

2.1　視覚情報と言語情報の融合

VQA を含む Vision and Language 分野において重要な共通事項の 1 つに「視覚情報と言語情報の融合」がある．一般的に Transformer ベースの読解モデル [7, 8, 9, 10] の注意機構を用いて視覚情報と言語情報の意味表現を紐づける．また画像と質問に加えて、画像に関連する付加情報（物体クラス [11]，画像中の文字情報 [12, 13]，視線トレース情報 [14] など）のモデル化により，視覚情報と言語情報の融合が促進されることが明らかとなっている．特に大局的な視覚情報を記述するキャプションは，局所的な読解対象に解答する VQA タスクにおいて視覚情報

1） California Common Core Content Standards の IXL Math Learning より収集したデータに対し，小学校 3 年生までに習う算数の問題をクラウドソーシングで作成している．

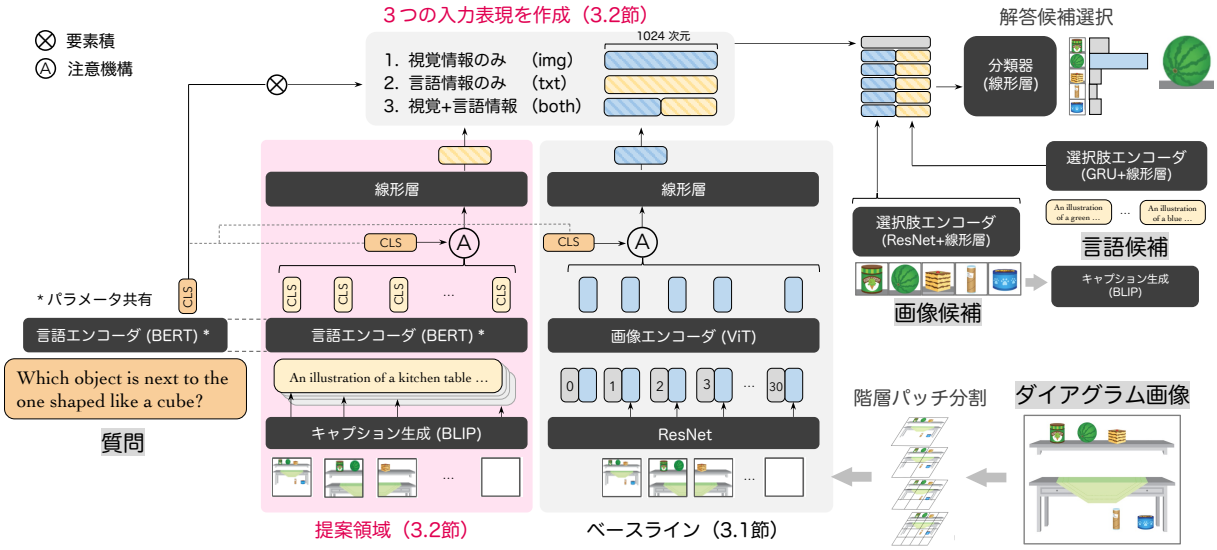

図 1　キャプション生成を用いた視覚情報の拡張による提案モデルにおける画像選択問題の例（3.2 節）

の手がかりを提供する [15, 16, 17]．本研究では 1 つの参照に対して色や形など複数の意味表現を持つダイアグラム画像を対象に，そのキャプションが VQA モデルの推論能力に与える影響を調査する．

2.2　単一モダリティ空間上での視覚推論

視覚情報と言語情報の融合を促進する方法論の 1 つとして，画像情報および言語情報を一方のモダリティとして扱うことが挙げられる [18, 19]．Liu ら [18] は，言語としての文脈と質問をレンダリングし，画像として扱うことで数値推論に取り組んでいる．また Outside Knowledge VQA（OK-VQA）[20] タスクでは，入力画像を言語として記述し，大規模言語モデルの推論能力を利用することで VQA モデルの推論性能を改善する手法が提案されている [21, 22, 23]．本研究では，BERT [24] および GPT-3 [6] の推論能力を活用するために，ダイアグラム画像のキャプションを用いて言語空間上で視覚推論を行い，その推論性能を評価する．

3　キャプション情報を用いた VQA

本章では，抽象的なダイアグラム画像を対象にした IconQA に取り組む．3.1 節では IconQA のベースラインモデルである Patch-TRM [1] を説明し，3.2 節でキャプションの言語情報を Patch-TRM に組み込む方法について提案する．3.3 節では，画像をキャプションとして記述することで大規模言語モデルの推論能力を活用する方法について提案する．

3.1　IconQA ベースラインモデル

IconQA は質問とダイアグラム画像を入力として，質問に対する正解候補の中から解答を 1 つ選択する VQA タスクであり，1）画像候補選択問題 2）テキスト候補選択問題 3）穴埋め問題，の 3 つのサブタスクが選択問題として定義されている[2)]．我々はベースラインモデルとして Lu ら [1] が提案した Patch-TRM を用いる．Patch-TRM は，言語，画像，選択肢の 3 つのエンコーダと，解答を選択する分類器で構成される．言語エンコーダは BERT [24]，画像エンコーダは Vision Transformer [25]，選択肢エンコーダおよび分類器は線形層をそれぞれ用いる．ダイアグラム画像は，階層パッチ領域に分割される[3)]．分割されたパッチ領域は事前に学習された ResNet [26][4)] によって符号化され，画像エンコーダに入力される．質問は，トークンに分割されたのち言語エンコーダによって符号化され，注意機構によってパッチ領域の視覚表現と融合される．また画像候補選択問題では ResNet による埋め込み表現，テキスト候補選択問題では GRU [27] による埋め込み表現を選択肢エンコーダに入力する．最終的に注意機構からの出力および選択肢エンコーダからの出力を連結した中間表現を用いて分類器が解答候補に対するスコアを算出する．なお穴埋め問題においては，選択肢エンコー

2）　穴埋め問題では，事前に収集したデータセットの各質問に対する解答セットを選択肢として利用する
3）　入力画像に対して $(1\times1), (2\times2), \ldots, (k\times k)$ と分割する．
4）　アイコン画像で構成される Icon645 [1] データセットに対する分類タスクで学習される．

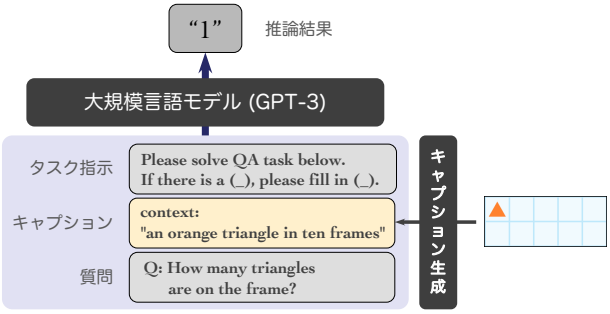

図 2　GPT-3 [6] による視覚推論（3.3 節）

ダは使用せず，注意機構からの出力表現を分類器に入力する．学習時は，分類器が出力したスコアに対して二値交差エントロピーを用いた最適化を行う．

3.2　キャプションを用いた視覚情報の拡張

本節では，ダイアグラム画像と質問における視覚および言語情報の意味表現の融合を促進することを目的として，ダイアグラム画像をキャプションとして記述し，その言語情報を Patch-TRM に組み込む手法を提案する（図 1）．キャプションの記述には，BLIP [28] のキャプション生成モデルを用いる[5]．キャプションの組み込みを実現するため，Patch-TRM の 3 つのエンコーダに加えてキャプションを符号化するための言語エンコーダを導入する[6]．具体的には，3.1 節で説明したベースラインの画像エンコーダを踏襲し，階層的にパッチ分割された画像に対して BLIP を適用し，言語として記述されたキャプションを言語エンコーダで符号化する[7]．以降は，3.1 節と同様に，質問および選択肢の表現と組み合わせることで分類器に入力するための中間表現を作成する．本研究では分類器に入力する中間表現として，画像エンコーダによる視覚表現（img），言語エンコーダからの言語表現（txt），両エンコーダから出力された視覚および言語の連結表現（both）の 3 種類を用いて比較評価を行う．なお画像候補選択問題については，分類器に入力される選択肢の中間表現において，画像候補に加えて BLIP で記述した言語候補も使用し，言語候補選択問題で使用する GRU と線形層で符号化された言語選択表現を視覚選択表現に連結したものを使用する[8]．

5）　大規模コーパス [29, 30] を用いて事前学習されており，ダイアグラム画像にも適当なキャプションを生成する．https://github.com/LAION-AI/BLIP

6）　モデルのパラメータ数増加を避けるため，質問とキャプションの符号化には同一の言語エンコーダを用いる

7）　パッチ間の意味関係を考慮する同一画像に対して記述されたパッチ画像のキャプションを連結する方法が考えられるが，ここでは各キャプションを独立して符号化する．

表 1　IconQA [1] データセットの質問数

サブタスク	訓練	開発	評価
画像候補選択	34,603	11,535	11,535
テキスト候補選択	18,946	6,316	6,316
穴埋め問題	10,913	3,638	3,638
合計	64,462	21,489	21,489

3.3　GPT-3 による言語推論

Lu ら [1] が提案した Patch-TRM は，数値推論など高度な推論能力が要求される問題に対して改善の余地がある．我々は Schwenk ら [20] の研究に倣い，大規模言語モデルの推論能力を活用することで，IconQA における予測性能の改善を目指し，数値推論などのスキル別に評価を行う．具体的には，GPT-3 [6] を用いて言語空間上での推論を行うために，1 枚のダイアグラム画像全体に対して，人手および BLIP によって記述されたキャプションを使用する．記述されたキャプションは，図 2 で示すタスク指示および質問のプロンプトと連結され，ゼロショットの設定で GPT-3 に入力される．

4　実験設定

5.1 節では，3.2 節で提案した Patch-TRM の読解能力を評価するため，13 の推論スキル[9]が定義された IconQA を用いて，画像候補選択問題（Img.），テキスト候補の選択問題（Txt.），穴埋め問題（Blank.），の 3 つのサブタスクで評価を行う（表 1）．評価指標には正解率を用いる．また Patch-TRM の分類器への 3 つの入力表現に対して適切な比較を行うため，視覚および言語の連結表現（both）に関して，直前の線形層における出力次元数を，他 2 つの入力表現の次元数の半分になるように設定する（図 1）．Patch-TRM における学習設定は表 3 を参照されたい．

5.2 節では，言語モデルによる推論能力を活用するために，3.3 節で説明した 2 つのキャプション生成を導入して，IconQA の推論スキル別に評価を行う．評価対象の言語モデルは，3.2 節で説明した Patch-TRM（txt）に加えて GPT-3（表 4）を対象とする．評価対象のタスクとして選択肢の条件に依存し

8）　事前調査より txt における画像候補選択の正解率が低かったため画像候補に加えることとした．

9）　Geometry, Counting, Comparing, Spatial, Scene, Pattern, Time, Fraction, Estimation, Algebra, Measurement, Commonsense, Probability．詳細は Lu ら [1] の論文を参照されたい．

表 2　IconQA 評価セットにおけるサブタスクおよび推論スキル別の Patch-TRM の正解率

モデル	サブタスク			推論スキル												
	Img. (11535)	Txt. (6316)	Blank. (3638)	Geo. (8575)	Cou. (7493)	Com. (2976)	Spa. (2143)	Sce. (2013)	Pat (1827)	Tim (1803)	Fra (1567)	Est. (1530)	Alg. (1456)	Mea. (1287)	Sen. (1158)	Pro (1077)
img	78.71	65.92	86.44	80.87	76.64	74.73	53.49	59.85	56.07	69.90	81.37	96.97	61.03	96.43	79.39	76.26
txt	69.07	62.40	47.16	72.05	52.36	68.20	49.98	59.95	54.92	66.94	44.33	65.42	38.94	60.14	82.13	83.94
both	79.60	64.90	87.39	81.10	76.73	74.90	54.69	62.36	55.74	67.68	81.28	98.98	60.90	98.78	78.04	75.18
Human	95.69	93.91	93.56	94.63	97.63	94.41	93.31	92.73	95.66	97.94	97.45	87.51	96.29	86.55	97.06	85.67

ないタスク設定に限定するため，IconQA の穴埋め問題のみを使用し，この中から無作為に抽出した 57 件のデータを評価セットとする．また GPT-3 の生成結果に対して適切な評価を行うため，意味的に一致する解答を正解として人手による判断を行う[10]．

5　実験結果

5.1　キャプションを考慮したモデルの評価

3.2 節で説明した，キャプション生成を用いた視覚情報の拡張による Patch-TRM の読解性能の評価結果を表 2 に示す．表 2 より，画像候補選択（Img.）および穴埋め問題（Blank.）タスクでは，視覚情報と言語情報を共同でモデル化する both モデルの正解率が単一情報を用いるモデルの正解率を凌駕した．また推論スキル別の評価においても，7 つの推論スキルにおいて both モデルが単一情報を用いるモデルの正解率を上回ったことから，キャプション生成モデルによる視覚情報の拡張が IconQA において効果的であることを示した．この結果から，視覚情報の拡張であるキャプションが，VQA タスクを解く上で必要となる画像と質問の読解に関する手がかりを提供できることが示唆される．

5.2　GPT-3 を用いた視覚推論

3.3 節で説明した，人手および BLIP によるキャプションに対する言語モデルの推論結果に対して，IconQA の推論スキル別に評価を行った（図 3）．また言語モデル別の性能差を調査するため，3.2 節で提案した Patch-TRM（txt）でも同様に評価を行った．図 3 から，人手キャプションを用いた場合に GPT-3 の推論結果の正解率が Patch-TRM に対して大きく上回った．これにより，GPT-3 を使用することで，VQA モデルの読解性能が向上する可能性があることが示唆された．また GPT-3 と Patch-TRM（txt）で

10）　例えば，推論結果が one で正解が 1 であった場合，これらは同じ内容を指しているものとし正解とする．

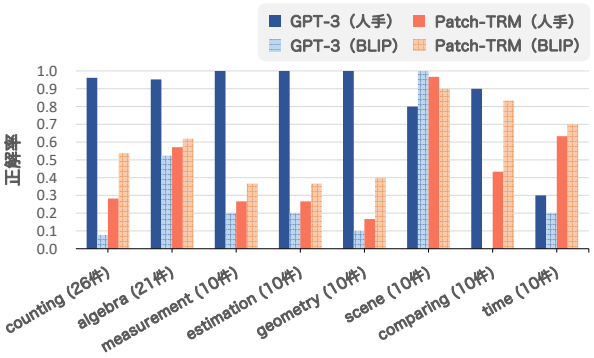

図 3　IconQA 穴埋め問題における推論スキル別の正解率

キャプション別に比較を行うと，GPT-3 では人手キャプションを使用した方が正解率が高い一方で，Patch-TRM（txt）では BLIP キャプションを使用した方が正解率が高い結果となった．これは Patch-TRM（txt）の学習データが BLIP キャプションを使用していることに起因していると考えられる．また，A.2 節では GPT-3 による実際の推論結果を示す．この結果より，入力画像の大域的な視覚情報を記述したキャプションが，VQA の質問に答えるための適切な手がかりを GPT-3 に提供していることが示唆された．

6　おわりに

本研究では，キャプション生成を用いた視覚情報の拡張が VQA タスクに与える影響を調査した．5.1 節では，視覚情報と言語情報を共同で学習する手法が多くの推論スキルで有効であることを示した．5.2 節では，キャプションを用いた GPT-3 の推論性能を評価し，多様な推論スキルが要求される VQA タスクにおいてキャプションが推論に有効な情報を提供する可能性があることを示した．今後の展望として，推論スキルに対する精緻分析，適切な言語情報の探索 [31]，モデル構造の改善に取り組みたい．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011 (fundamental research) の助成を受けて実施されたものである．

参考文献

[1] Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In **NeurIPS**, 2021.

[2] C.K. Odgen and I.A. Richards. **The Meaning of Meaning A Study of the Influence of Language upon Thought and of the Science of Symbolism**. Routledge & Kegan Paul Ltd., 1923.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In **ICCV**, pp. 2425–2433, 2015.

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In **CVPR**, pp. 6325–6334, 2017.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NeurIPS**, pp. 5998–6008, 2017.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In **NeurIPS**.

[7] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In **CVPR**, pp. 5579–5588, 2021.

[8] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In **ICML**, pp. 5583–5594, 2021.

[9] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In **ICLR**, 2022.

[10] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. **CoRR**, 2021.

[11] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In **ECCV**, pp. 121–137, 2020.

[12] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In **CVPR**, pp. 8751–8761, 2021.

[13] 田中涼太, 西田京介, 許俊杰, 西岡秀一. テキストと視覚的に表現された情報の融合理解に基づくインフォグラフィック質問応答. 言語処理学会, 2022.

[14] Zihang Meng, Licheng Yu, Ning Zhang, Tamara L. Berg, Babak Damavandi, Vikas Singh, and Amy Bearman. Connecting what to say with where to look by modeling human attention traces. In **CVPR**, pp. 12679–12688, 2021.

[15] Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for VQA are image captions. In **NAACL-HLT**, pp. 1947–1963, 2022.

[16] Radhika Dua, Sai Srinivas Kancheti, and Vineeth N Balasubramanian. Beyond vqa: Generating multi-word answers and rationales to visual questions. In **CVPR**, pp. 1623–1632, 2021.

[17] Jialin Wu, Zeyuan Hu, and Raymond Mooney. Generating question relevant captions to aid visual question answering. In **ACL**, pp. 3585–3594, 2019.

[18] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel H. Collier, and Julian Martin Eisenschlos. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. **CoRR**, 2022.

[19] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Image-and-language understanding from pixels only. **CoRR**, 2022.

[20] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In **ECCV**, pp. 146–162, 2022.

[21] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of GPT-3 for few-shot knowledge-based VQA. In **AAAI**, pp. 3081–3089, 2022.

[22] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In **CVPR**, pp. 5067–5077, 2022.

[23] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. KAT: A knowledge augmented transformer for vision-and-language. In **NAACL-HLT**, pp. 956–968. Association for Computational Linguistics, 2022.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[25] 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 **ICLR**, 2021.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **CVPR**, pp. 770–778, 2016.

[27] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In **EMNLP**, pp. 1724–1734. ACL, 2014.

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In **ICML**, pp. 12888–12900, 2022.

[29] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In **CVPR**, pp. 3558–3568, 2021.

[30] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. **CoRR**, 2021.

[31] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In **CVPR**, 2016.

A　参考情報

A.1　Patch-TRM の学習設定

4 章で説明した Patch-TRM [1] の学習設定を表 3 に，GPT-3 [6] の推論時の設定を表 4 に示す．

表 3　Patch-TRM（3.2 節）の実験設定

サブタスク	Img.	Txt.	Blank.
学習率（img）	1e − 4	9e − 4	1e − 3
学習率（txt）	1e − 5	1e − 4	1e − 4
学習率（both）	1e − 4	1e − 3	1e − 3
シード値	3 種類		
エポック数	50		
バッチサイズ	64		
画像エンコーダ	Vision Transformer [25]		
画像パッチサイズ	79（$1^2+2^2+3^2+4^2+7^2$）		
中間表現次元数	2,048		
出力表現次元数	1,024	1,024	512
Transformer 層数	1		
注意機構ヘッド数	4		
言語エンコーダ	BERT [24]		
キャプション数	14（$1^2+2^2+3^2$）		
中間表現次元数	768		
出力表現次元数	1,024	1,024	512
Transformer 層数	12		
注意機構ヘッド数	12		
分類器	2 層線形層		
入力表現次元数	1,024		
中間表現次元数	2,048		
Transformer 層数	12		
注意機構ヘッド数	12		
画像選択肢エンコーダ	ResNet [26]		
言語選択肢エンコーダ	GRU [27]		
キャプション生成	BLIP w/ CapFilt-L [28]		
入力画像サイズ	384		
ビーム幅	3		
出力トークン数	[5, 20]		

表 4　GPT-3（3.3 節）の実験設定

評価セット	IconQA [1] 評価セット
キャプション生成モデル	BLIP [28]（表 3 参照）
推論エンジン	text-davinci-003
最大出力トークン数	256
温度	0.0
top-p	0.05

A.2　GPT-3 の推論結果の定性評価

5.2 節では，人手および BLIP [28] で生成されたキャプションを用いて GPT-3 [6] の推論結果を評価したが，その際に生成された 2 つのキャプションを表 5 に示す．表 6 では，人手キャプションを使用した際の GPT-3 の推論結果を示す．推論結果より，キャプションを参照せずに質問文から解答可能な問題もあれば，キャプションを記述しないと解答不可能な問題も含まれており，適切な記述が GPT-3 の視覚推論に有効な手がかりを提供する可能性があることが分かった．

表 5　人手および BLIP [28] のキャプション例

	（質問）	There are 3 cars in the top row. How many cars are in the bottom row?
	（人手）	seven cars in ten frames
	（BLIP）	a yellow car with the number one on it
	（質問）	Use dice to measure the line. The line is about (_) dice long.
	（人手）	9 red dice in the gray area and a black line for 5 red dice
	（BLIP）	a red dice on a white background

表 6　人手キャプションを用いた際の GPT-3 の推論例

	（質問）	What time is shown? Answer by typing a time word, not a number. It is (_) after six.
	（記述）	a clock showing 6:15
	（予測）	Six fifteen
	（正解）	quarter
	（質問）	There are 3 cars in the top row. How many cars are in the bottom row?
	（記述）	seven cars in ten frames
	（予測）	4 cars in the bottom row
	（正解）	4
	（質問）	Read the story. There are 5 kangaroos. Then, 4 kangaroos hop away. Subtract to find how many kangaroos stay.
	（記述）	ten kangaroos in the meadow
	（予測）	5 - 4 = 1 kangaroo stays
	（正解）	1

A.3　本研究の限界

3.2 節では，階層的にパッチ分割された画像に対してキャプションを生成していたため，分割境界に位置する物体に対して適切なキャプションを生成できない．また Patch-TRM [1] において，選択肢の情報は，分類器に入力される前に質問と視覚情報の中間表現に連結されるだけであり，選択肢と入力画像との関係性を十分に学習できていない可能性がある．また提案法では，キャプション生成モデルとして BLIP [28] を用いたが，VQA モデルの読解能力が BLIP の生成能力に依存してしまう．