キャプション生成ゲームを通じた複数の視覚言語モデルのベイズ的統合

松井悠太 [1] 山木良輔 [1] 上田亮 [2] 品川政太朗 [3] 谷口忠大 [4]
[1] 立命館大学 [2] 東京大学 [3] 奈良先端科学技術大学院大学 [4] 京都大学
{matsui.yuta, yamaki.ryosuke}@em.ci.ritsumei.ac.jp, ryoryoueda@is.s.u-tokyo.ac.jp
sei.shinagawa@is.naist.jp, taniguchi@i.kyoto-u.ac.jp

概要

本研究では，複数の視覚言語モデル（VLM）を記号創発の枠組みで統合する手法である，メトロポリスヘイスティングスキャプション生成ゲーム（MHCG）を提案する．MHCG では，異なるデータで事前学習した VLM エージェント間で画像に対するキャプションを提案・受容・更新するプロセスを通じて，モデル間の知識 (本稿では画像に対する言語表現) を統合する．この手法は，既存のモデル統合の手法が持つ推論コストやモデル構造の一致の成約を受けない．実験では，COCO と CC3M でそれぞれ事前学習した 2 体のエージェントが MHCG を行い，相手のエージェントの学習データに対するキャプション生成性能が向上することを示した．

1 はじめに

複数の学習済みモデルを統合することで，汎化性能や予測精度がより高いモデルを開発する試みは広く行われている．既存のモデルの統合の手法として，それぞれのモデルの出力を統合するアンサンブル学習 [1] や，パラメータをある重みに基づいて平均化することで統合する重み平均化 [2] などの手法があり，大規模言語モデルや VLM に対しても適用されている [3]．しかし，アンサンブル学習は複数のモデルを同時に推論させるため推論にかかるコストが高くなる，重み平均化は同じ構造のネットワークを持つことが前提となるため，異なるアーキテクチャで学習された VLM に適応することができないという制約がある．特に LLM の学習に使うコーパスに比べて，VLM の学習に使う画像キャプションペアは比較的少ないため，異なるデータで学習した VLM を追加の画像キャプションペアなしで統合する手法が求められる．

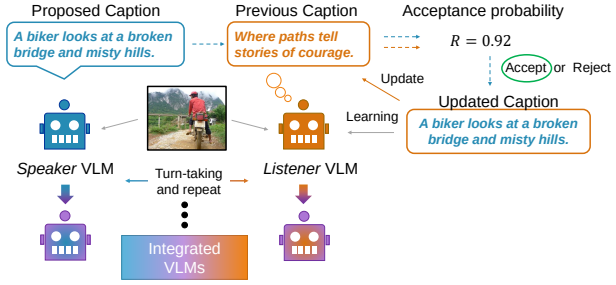
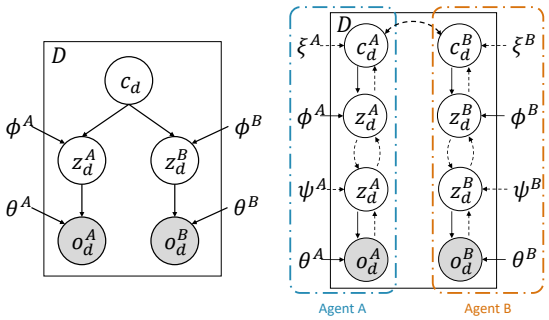


図 1: 研究の概要．2 体の VLM エージェントがキャプション生成ゲームを通して知識を統合する．

そこで，マルチエージェントによる記号創発をモデル化したメトロポリスヘイスティングス名付けゲーム（Metropolis-Hastings Naming Game: MHNG）[4] を VLM に導入し，これらの制約を排除した統合手法について検討する．MHNG では，物体の名前を提案し，それを受容・棄却することによる名付けと，各エージェントの学習とが分散的に行われるベイズ推論として定式化されている．これに事前学習済みの VLM を適用することで，キャプションを通じて各々の事前知識を共有し，それらを統合した VLM が得られると考えられる．しかし，MHNG はカテゴリカルなラベルを物体に割り当てる名付けを仮定した推論手法であるため，VLM を導入するためには，VLM が扱う言語情報であるキャプションを生成しあう推論手法として定式化する必要がある．

本研究では，MHNG を VLM を導入するために拡張したメトロポリスヘイスティングスキャプション生成ゲーム (Metropolis-Hastings Captioning Game: MHCG) を提案する．図 1 に示すように，VLM エージェントが相手の知識が含まれたキャプションを確率的に受容し，それに基づいて自身の認識を更新する．この過程を繰り返すことで，MHCG に参画す



(a) Inter-ProbVLM のグラフィカルモデル　(b) 分割後のグラフィカルモデル

図 2: 確率的生成モデル: Inter-ProbVLM

表 1: Inter-ProbVLM のパラメータ

表記	説明
D	画像観測の総数　$d \in \{1, \cdots, D\}$
o_d^*	d 番目の画像観測
w_d	d 番目のキャプション
z_d^*	d 番目の観測に対する潜在変数
ξ^*	テキストデコーダのパラメータ (ClipCap [7])
ϕ^*	テキストエンコーダのパラメータ (ProbVLM [8])
ψ^*	画像エンコーダのパラメータ (ProbVLM [8])
θ^*	画像デコーダのパラメータ

る VLM エージェントが持つ事前知識を統合したモデルが得られると期待される．本稿は，VLM エージェントに MH 法を適用することによってキャプションを推論する筆者らの基礎検証 [5] を基盤としている．従来の検証では扱っていない VLM のパラメータ更新を含む MHCG の推論を対象とし，その過程で両者の知識が統合されるかを検証する．

2　提案手法

2.1　確率的生成モデル：Inter-ProbVLM

まず，MHCG に参加する 2 つの VLM エージェントを共通の事前分布を持つ確率的生成モデルとして定義し，Inter-ProbVLM を構築する．Inter-ProbVLM の生成過程を式 (1)-(3) に，グラフィカルモデルを図 2a に，推論のために Neuro-SERKET [6] の枠組みで分割したグラフィカルモデルを図 2b に，各パラメータの説明を表 1 に示す．$* = \{A, B\}$ は各エージェントを指すインデックスである．

$$c_d \sim p(c_d) \qquad d = 1, \ldots, D \tag{1}$$

$$z_d^* \sim p(z_d^* | c_d, \phi^*) \qquad d = 1, \ldots, D \tag{2}$$

$$o_d^* \sim p(o_d^* | z_d, \theta^*) \qquad d = 1, \ldots, D \tag{3}$$

ここで，テキストエンコーダ ϕ^* 及び画像エンコーダ ψ^* の VLM に ProbVLM [8] を導入している．ProbVLM は，CLIP [9] などの事前学習済み VLM が出力する決定論的な埋め込み表現に対して一般化ガウス分布を仮定し，そのパラメータを推定するアダプターを導入することで，潜在変数の確率分布を推定する確率的 VLM である．

2.2　Metropolis-Hastings Captioning Game (MHCG)

MHCG では参加する 2 体のエージェントが事前学習を通して言語知識を獲得した後に，画像に対して両エージェントにとって尤もらしいキャプションを共有することを目的としてコミュニケーションを行う．まず，話し手のエージェントは画像に対するキャプションを提案する．それを受け取った聞き手は自身の信念に基づいて相手のキャプションを受容するか棄却するか判断し，受け入れる場合は自身の信念を更新する．この過程を繰り返すことで，キャプションを通して両者が持つ知識を共有できる．

MHCG において事前学習済みの VLM のパラメータで初期化された 2 体のエージェント聞き手（Sp）と話し手（Li）の役割を入れ替えながら (1) 知覚，(2) 提案，(3) 判断，(4) 更新の 4 ステップを繰り返すことでキャプションの推論を行う．

(1) 知覚では，両エージェントが観測 o_d^* から潜在表現 z_d^* を，画像エンコーダ ξ^* に基づき式 (4) に従ってサンプリングすることで得る．この潜在表現は，観測に対する各エージェントの認識を反映し，キャプションの提案や受容・棄却の判断に利用される．

$$z_d^{Sp} \sim p(z_d^* | o_d^*, \psi^*) \tag{4}$$

(2) 提案では，話し手エージェントが潜在表現 z_d^{Sp} に基づき，画像エンコーダのパラメータ ξ を用いて式 (5) に従いキャプション $c_d^\star$ をサンプリングする．

$$c_d^\star \sim q(c | z_d^{Sp}, \xi^{Sp}) \tag{5}$$

(3) 判断では，聞き手エージェントが自身の認識に基づいたキャプション c_d^{Li}，提案されたキャプション $c_d^\star$ と潜在表現 z_d^{Li} に基づいて，$c_d^\star$ を受容するか否かを判断する．この過程は事後分布 $p(c_d | z_d^A, z_d^B, \phi^A, \phi^B)$ を，式 (5) の提案分布に基づいた MH アルゴリズムでサンプルすることに相当する．このとき，聞き手エージェントは式 (6) で導出される受容確率 $r = min(1, R)$ で $c_d^\star$ を受容する．

$$R = \frac{p(c_d^\star \mid z_d^{Sp}, z_d^{Li}, \phi^{Sp}, \phi^{Li}) q(c_d^{Li} \mid z_d^{Sp}, \xi^{Sp})}{p(c_d^\star \mid z_d^{Sp}, z_d^{Li}, \phi^{Sp}, \phi^{Li}) q(c_d^\star \mid z_d^{Sp}, \xi^{Sp})} \approx \frac{p(z_d^{Li} \mid \phi^{Li}, c_d^\star)}{p(z_d^{Li} \mid \phi^{Li}, c_d^{Li})} \tag{6}$$

(4) 更新では，(3) 判断後のキャプション c_d^{Li} に基づいて，聞き手エージェントが自身のパラメータ $\xi^{Li}, \phi^{Li}, \psi^{Li}, \theta^{Li}$ を更新する．更新に使用する損失関数はそれぞれ式 (7) から (10) のようになる．

$$L_\xi = -\mathbb{E}_{p(c|z^{Li}, z^{Sp}, \phi^{Li}, \phi^{Sp})}\left[\log q(c^{Li} \mid z^{Li}, \xi^{Li})\right] \tag{7}$$

$$L_\phi = -\mathbb{E}_{p(c|z^{Li}, z^{Sp}, \phi^{Li}, \phi^{Sp})}\left[\log p(z^{Li} \mid c^{Li}, \phi^{Li})\right] \tag{8}$$

$$L_\psi = -\mathbb{E}_{p(z^{Li}|c, \phi^{Li})}\left[\log q(z^{Li} \mid o^{Li}, \psi^{Li})\right] \tag{9}$$

$$L_\theta = -\mathbb{E}_{p(z^{Li}|c, \phi^{Li})}\left[\log p(o^{Li} \mid z^{Li}, \theta^{Li})\right] \tag{10}$$

$\xi^{Li}, \phi^{Li}, \psi^{Li}, \theta^{Li}$ には事前学習で得た知識が含まれるが，MHCG の観測に過学習すると破滅的忘却 [10] が生じる可能性がある．そこで，特定の層に低ランク行列を追加する Low Rank Adaptation (LoRA) [11] と，継続学習手法である Dark Experience Replay ++ (DER++) [12] を導入する．これにより，それぞれのエージェントの事前知識を保持しつつ，相手が提案したキャプションに適応することが可能となる．

3 実験

3.1 実験目的

本稿では異なる知識を持つ 2 体の VLM エージェントによる MHCG に関して，以下の 2 項目を目的として実験を行う．

(1)MHCG を通して共有されるキャプションは両エージェントにとって尤もらしいか．

(2)相手の事前学習データに対するキャプション生成性能が上昇するか．

目的 (1) では，MHCG は確率的生成モデルの推論として機能し，エージェント間で推論されるキャプション c が尤もらしいかを検証する．目的 (2) では，MHCG を通して相手の知識を獲得できるかどうかを検証する．[1)]

1) 本研究では MHCG による言語的な知識の統合に注目するため，パラメータ更新において式 (9), (10) で示す画像エンコーダ・デコーダは学習しない．

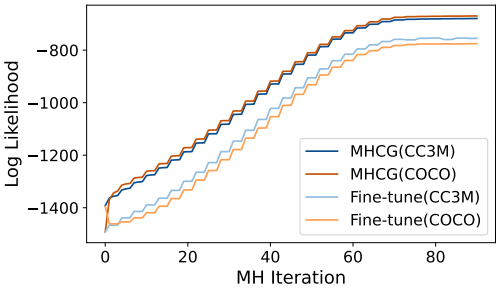


図 3: キャプションの尤度 $\log p(z^A, z^B | c, \phi^A, \phi^B)$

3.2 データセット

データセットには，Conceptual Captions 3M (CC3M) [13] と MSCOCO (COCO) [14] を使用する．エージェント A は CC3M を，エージェント B は COCO を用いて事前学習する．MHCG には事前学習には使用していない COCO と CC3M から抽出した合計 15,000 枚の画像を使用する．MHCG 後の検証には，それぞれの検証データを使用する．

3.3 評価指標

目的 (1) の検証に，尤度関数 $\log p(z^A, z^B | c, \phi^A, \phi^B)$ を使用する．これは，確率的生成モデルの生成分布の尤度であり，両者の潜在表現 z_d^A, z_d^B に対するキャプションの尤もらしさを示す．目的 (2) のキャプション生成性能の評価には，BLEU [15], METEOR [16], BERT-Score [17], CLIP-Score [18], PAC-Score [19]，RefPAC-Score [19] を使用する．

3.4 比較手法

比較手法として，Pretrain，Fine-Tune，MHCG の手法をそれぞれデータセットごとに 2 体ずつ使用する．Pretrain は事前学習したのみのエージェントを指す．Fine-Tune は 2 体の Pretrain エージェントが互いに生成しあったキャプションを使ってファンチューニングした後のエージェントを指す．MHCG は 2 体の Pretrain エージェントが MHCG を通して学習した後のエージェントを指す．

3.5 実験結果

3.5.1 共有されるキャプション

図 3 に，エージェントごとのキャプションの，両エージェントにとっての尤度 $\log p(z^A, z^B | c, \phi^A, \phi^B)$ の遷移を示す．MHCG を行ったエージェントが持

表 2: CC3M と COCO の検証データに対するキャプション生成性能

Agent		CC3M Validation Data						COCO Validation Data					
Pretrain Data	Method	BLEU@4	METEOR	BERT-S	CLIP-S	PAC-S	RPAC-S	BLEU@4	METEOR	BERT-S	CLIP-S	PAC-S	RPAC-S
CC3M	Pretrain	**7.42**	**11.14**	**0.881**	0.306	0.612	**0.686**	6.32	12.93	0.886	0.299	0.598	0.695
	Fine-tune	1.64	7.16	0.868	0.287	0.574	0.635	**26.52**	**24.39**	**0.911**	0.309	0.619	0.723
	MHCG	3.50	9.09	0.877	**0.307**	**0.614**	0.683	17.50	19.19	0.901	**0.313**	**0.625**	**0.726**
COCO	Pretrain	1.30	6.61	0.868	0.288	0.575	0.631	**31.61**	**27.37**	**0.918**	**0.323**	**0.645**	**0.750**
	Fine-tune	**4.19**	**8.22**	**0.875**	0.282	0.563	0.644	9.59	15.00	0.892	0.299	0.599	0.703
	MHCG	2.30	7.74	0.874	**0.294**	**0.588**	**0.660**	20.77	20.72	0.904	0.315	0.629	0.732

表 3: CC3M の画像に対して各エージェントが生成したキャプション

Input image	Agent		Caption
	Pretrain Data	Method	
	CC3M	Pretrain	honeybee with a flower in the beehive.
		Fine-tune	a picture of flowers in a vase with some bees.
		MHCG	a painting of a bee with flowers.
	COCO	Pretrain	A picture of a white bird with flowers on it.
		Fine-tune	vector illustration of a honeybee.
		MHCG	vector illustration of a honeybee with the name and address.

つキャプションは，イテレーションを繰り返すごとに尤度が向上し，Fine-tune より高い値となっている．これは MHCG のプロセスを通して推論されるキャプションが，別のデータで学習した 2 体の両エージェントにとって尤もらしいものが得られることを示している．

3.5.2 キャプション生成性能

表 2 に各データセットに対するキャプション生成性能の定量評価の結果を示す Pretrain エージェントは自身の事前学習データに対応する性能は高いが，相手のデータに対する性能は低い傾向にある．MHCG 後のエージェントはすべての指標において，相手のデータセットに対して性能が上昇している．これは，相手が事前学習で得た知識を，MHCG 時のキャプションを通して獲得できていることを示す．

MHCG と Fine-tune を比較すると，生成キャプションの画像との意味的類似度と自身の学習データの忘却を防ぐ点で優れていることがわかる．Fine-tune エージェントはアノテーションのキャプションとの類似度を計測する指標においては高いが，画像の意味的類似度を計測する指標は MHCG に劣っている．また，MHCG は Fine-tune よりも自身の事前学習データに対する性能の低下を防いでいる，つまり忘却を低減していることがわかる，これらの結果は，MHCG の (3) 判断のステップで，自身の画像観測に対して尤もらしいキャプションを受容して学習することが要因であると考えられる．

また，表 3 に CC3M に含まれる画像に対して各エージェントが生成したキャプションの例を示す．Pretrain の COCO エージェントは ”white bird” としており，誤り (赤) を含んでいる．MHCG 後の COCO エージェントは ”honeybee” と正しく (青) 説明できている．この ”honeybee” は CC3M の事前学習データには含まれるが，COCO の事前学習データには含まれない単語であった．つまり，MHCG を通して，COCO エージェントは CC3M データセットの直接的な参照を必要とせず，CC3M エージェントのキャプションを通して知識を獲得したことがわかる．

4 おわりに

本論文は，分散的ベイズ推論に基づく記号創発のモデルである MHNG を，画像に対するキャプション生成へ拡張した Metropolis-Hastings captioning game を提案した．2 つの ProbVLM を結合した Inter-ProbVLM を定義し，MHCG の理論について示した．実験を通して，MHCG を通して 2 つの VLM エージェントにとって尤もらしいキャプションが共有されること，相手の事前学習データへのキャプション生成性能が向上することを示した．

今後の展望として，Inter-ProbVLM の画像エンコーダ θ^* とデコーダ ψ^* のパラメータの更新を含めたモデル全体の推論を行うことが考えられる．これによって，片方のエージェントが観測した画像を，キャプションを通して相手に想起（生成）させることができるかを検証することができる．

謝辞

本研究は JSPS 科研費 JP21H04904 および JP23H04835 の助成を受けたものです.

参考文献

[1] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 6691–6706, 2021.

[2] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In **International conference on machine learning**, pp. 23965–23998. PMLR, 2022.

[3] Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. **arXiv preprint arXiv:2408.07666**, 2024.

[4] Tadahiro Taniguchi, Yuto Yoshida, Yuta Matsui, Nguyen Le Hoang, Akira Taniguchi, and Yoshinobu Hagiwara. Emergent communication through metropolis-hastings naming game with deep generative models. **Advanced Robotics**, Vol. 37, No. 19, pp. 1266–1282, 2023.

[5] 松井悠太, 山木良輔, 上田亮, 品川政太朗, 谷口忠大. Metropolis-hastings captioning game による複数の視覚言語モデルのベイズ的統合. 言語処理学会第 30 回年次大会発表論文集, pp. 1925–1930, 2024.

[6] Tadahiro Taniguchi, Tomoaki Nakamura, Masahiro Suzuki, Ryo Kuniyasu, Kaede Hayashi, Akira Taniguchi, Takato Horii, and Takayuki Nagai. Neuro-serket: development of integrative cognitive system through the composition of deep probabilistic generative models. **New Generation Computing**, Vol. 38, pp. 23–48, 2020.

[7] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. **arXiv preprint arXiv:2111.09734**, 2021.

[8] Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 1899–1910, 2023.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International conference on machine learning**, pp. 8748–8763. PMLR, 2021.

[10] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In **Psychology of learning and motivation**, Vol. 24, pp. 109–165. Elsevier, 1989.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. **arXiv preprint arXiv:2106.09685**, 2021.

[12] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. **Advances in neural information processing systems**, Vol. 33, pp. 15920–15930, 2020.

[13] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2556–2565, 2018.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In **Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13**, pp. 740–755. Springer, 2014.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[16] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. **ACL 2007**, p. 228, 2007.

[17] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7514–7528, 2021.

[19] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for image and video captioning evaluation. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 6914–6924, 2023.

表 4: エージェント間で共有されるキャプションの例

Input image	Agent		Caption
	Pretrain Data	Method	
	CC3M	Pretrain	the fruits are in season.
		Fine-tune	a man and woman are shopping at a fruit stall.
		MHCG	**a market with people shopping for fruit and vegetables.**
	COCO	Pretrain	A group of people standing around a fruit and vegetable market.
		Fine-tune	people at a market selling fruit and vegetables.
		MHCG	**a market with people buying fruit and vegetables.**

表 5: エージェント間で共有されるキャプションの類似度

Models	BLEU@4	METEOR	BERT score
Pretrain	3.03	15.28	0.888
Fine-tune	7.36	23.53	0.906
MHCG	**33.93**	**49.11**	**0.934**

A　参考情報

A.1　パラメータ更新の詳細

ここでは 2.2 の (4) 更新で説明した更新式について詳細に述べる．式 (7)-(10) に継続学習手法である DER++ [12] を適用した損失関数は，以下の式 (11)-(14) の通りとなる．

$$
\begin{aligned}
L_\xi = &- \mathbb{E}_{p(c|z^{Li},z^{Sp},\phi^{Li},\phi^{Sp})}\left[\log q(c^{Li} \mid z^{Li},\xi^{Li})\right] \\
&+ \alpha\, \mathbb{E}_{(z',c',h')\sim M_\xi^{Li}}\left[\|h' - h_{\xi^{Li}}(z')\|_2^2\right] \\
&- \beta\, \mathbb{E}_{(z'',c'',h'')\sim M_\xi^{Li}}\left[\log q(c'' \mid z'',\xi^{Li})\right] \qquad (11)\\
L_\phi = &- \mathbb{E}_{p(c|z^{Li},z^{Sp},\phi^{Li},\phi^{Sp})}\left[\log p(z^{Li} \mid c^{Li},\phi^{Li})\right] \\
&+ \alpha\, \mathbb{E}_{(z',c',h')\sim M_\phi^{Li}}\left[\|h' - h_{\phi^{Li}}(c')\|_2^2\right] \\
&- \beta\, \mathbb{E}_{(z'',c'',h'')\sim M_\phi^{Li}}\left[\log p(z'' \mid c'',\phi^{Li})\right] \qquad (12)\\
L_\psi = &- \mathbb{E}_{p(z^{Li}|c,\phi^{Li})}\left[\log q(z^{Li} \mid o^{Li},\psi^{Li})\right] \\
&+ \alpha\, \mathbb{E}_{(z',o',h')\sim M_\psi^{Li}}\left[\|h' - h_{\phi^{Li}}(o')\|_2^2\right] \\
&- \beta\, \mathbb{E}_{(z'',o'',h'')\sim M_\psi^{Li}}\left[\log q(z'' \mid o'',\phi^{Li})\right] \qquad (13)\\
L_\theta = &- \mathbb{E}_{p(z^{Li}|c,\phi^{Li})}\left[\log p(o^{Li} \mid z^{Li},\theta^{Li})\right] \\
&+ \alpha\, \mathbb{E}_{(z',o',h')\sim M_\theta^{Li}}\left[\|h' - h_{\phi^{Li}}(z')\|_2^2\right] \\
&- \beta\, \mathbb{E}_{(z'',o'',h'')\sim M_\theta^{Li}}\left[\log p(o'' \mid z'',\phi^{Li})\right] \qquad (14)
\end{aligned}
$$

ここで，$M_\xi^*, M_\phi^*, M_\psi^*, M_\theta^*$ は各エージェントの事前学習データによって初期化されるバッファであり，サンプルされる z', z'' は事前学習データの画像から得られる潜在表現を，c', c'' は事前学習データのキャプションを，o', o'' は事前学習データの画像観測を，$h_{\xi^*}, h_{\phi^*}, h_{\psi^*}, h_{\theta^*}, h', h''$ は各パラメータを通して得られるの出力される分布のパラメータの値を示す．バッファに含まれるサンプルに基づいた項を損失関数に追加することで，事前学習で得た知識の破滅的忘却を防ぐことができる．

表 6: 実験設定

Parameter	Value
MHCG の繰り返し回数	30
DER++ の α	0.05
DER++ の β	0.05
DER++ の バッファに含まれるサンプル数	5000
パラメータの更新のエポック数	10
テキストデコーダ ξ^* の学習率	1e-4
テキストエンコーダ ϕ^* の学習率	1e-6
LoRA における r	8
LoRA における α	16
LoRA におけるドロップアウト率	0.1
バッチサイズ	40

A.2　実験設定の詳細

本研究の学習に関する実験設定を表 6 に示す．

A.3　実験結果：サインの類似度

表 5 に，MHCG を通して 2 体のエージェントが推論した観測画像に対するキャプションの類似度を示す．MHCG は，Pretrain 及び Fine-tune よりもすべての類似度指標において高い値を示している．これは，両エージェントが類似したキャプションを生成するようになったことのみならず，両エージェントが類似したキャプションを尤もらしいと判断するように学習されていることを示している．

また，表 4 に共有されるキャプションのサンプルを示す．MHCG 後の両エージェントが持つキャプションに ”a market with people” や ”fruit and vegetables” が含まれており，MHCG を通して類似した言語表現を行うようになったことがわかる．