Metropolis-Hastings Captioning Game による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] 立命館大学大学院 情報理工学研究科 [2] 東京大学
[3] 奈良先端科学技術大学院大学 [4] 立命館大学 情報理工学部
{matsui.yuta, yamaki.ryosuke}@em.ci.ritsumei.ac.jp, ryoryoueda@is.s.u-tokyo.ac.jp
sei.shinagawa@is.naist.jp, taniguchi@em.ci.ritsumei.ac.jp

概要

本論文では分散的なベイズ推論に基づく記号創発のモデル化法である Metropolis-Hastings (MH) naming game をカテゴリカルなラベルを割り当てる名付け (naming) から画像に関するキャプションの生成（captioning）へと拡張し，これが複数の視覚言語モデル（VLM）がその知識をベイズ的に統合する能力に関して基礎的な検証を行う．本稿では VLM を確率モデルとして拡張した ProbVLM を導入し，２つの ProbVLM を結合した Inter-ProbVLM を定義することで，キャプション生成結果を表す共有ノードの推論アルゴリズムとして MH captioning game (MHCG) を提案する．本実験では ProbVLM が潜在空間上の確率分布に基づき画像とキャプションの一致度を判定できること，また，異なるデータセットで学習された２つの ProbVLM が MHCG を通してより妥当なキャプションを生成し，２体のエージェントの視覚情報表現に対する尤度を高めることを示し，VLM のベイズ的に統合に関する基礎的検証を行う．

1 はじめに

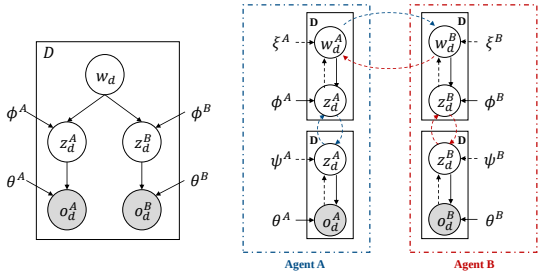
Metropolis-Hastings Naming Game (MHNG) は確率的生成モデルのための推論手法として提案され，マルチエージェントによる記号創発のモデル化に用いられてきた [1, 2]．MHNG はマルコフ連鎖モンテカルロの一種である Metropolis-Hastings 法の理論に基づき，適切なサインの使用に収束することが保証されている．この過程において互いのエージェントは互いの内部変数に一切アクセスせずにサインを共有する．このことから，MHNG はサインに対する分散的ベイズ推論であるとみなすことができる [3]．谷口らの研究では確率的生成モデルである Inter-GMM+VAE に MHNG を適用し，二体のエージェントがサインを共有することが示された [1]．

しかし，MHNG は観察対象をカテゴリカルなサイン（たとえば単語）を用いて名付けることを前提としており，画像に対するキャプションのようにより複雑な対象をより複雑なサイン（単語系列としての文）を用いて表現する記号創発を構成できていなかった．

本論文では分散的なベイズ推論としての理論的基盤を持つ Metropolis-Hastings (MH) naming game をカテゴリカルなラベルを割り当てる名付け（naming）から画像に関するキャプションの生成（captioning）へと拡張する．このために視覚言語モデル（VLM）を確率モデルとして拡張した ProbVLM[4] を導入し，これに基づき二つの VLM を結合した Inter-ProbVLM を定義し，その共有ノードの推論アルゴリズムとして MH captioning game (MHCG) を提案する．本稿の実験では VLM を構成するネットワークの訓練自体は行わず，MHCG において相手のサインを受容する際に用いられる受容確率の妥当性についての検証と，２体のエージェントが得る視覚情報を共に観測として持つ潜在変数として定義されるサイン（キャプションに相当）が分散的ベイズ推論としての MHCG を通して妥当な形で推論されることの検証を行った．

2 確率的生成モデル：Inter-ProbVLM

MHNG では，記号創発に参画する各エージェントを確率的生成モデル（PGM）で表現した上で，２つの PGM に共通の事前分布を持たせた Inter-PGM を定義する．本稿ではこの PGM を確率モデル化した VLM を用いて構成することで MHNG を MHCG へと拡張する．



(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 1: Inter-ProbVLM

表 1: Inter-ProbVLM のパラメータ

表記	説明
D	画像観測及び単語列の総数　$d \in \{1, \cdots, D\}$
o_d^*	d 番目の画像観測
w_d	d 番目のサイン (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
ϕ^*	テキストエンコーダのパラメータ
ψ^*	画像エンコーダのパラメータ
θ^*	画像デコーダのパラメータ

ProbVLM は事前学習済み VLM が出力する埋め込み表現に対する確率的アダプターを導入し確率的な VLM を実現する [4]．CLIP[5] などの画像エンコーダとテキストエンコーダを持つ事前学習済み VLM が出力する埋め込み表現は決定論的である．ProbVLM では，ここで得られ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ように追加学習することで潜在変数の確率分布を推定することができる．

従来の Inter-PGM の構成と同様に確率的 VLM である 2 つの ProbVLM を結合することで Inter-ProbVLM を構成することが出来る．Inter-ProbVLM の生成過程を式 (1)-(3) に，グラフィカルモデルを図 1a に，各パラメータの説明を表 1 に示す．

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

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

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

3 MHCG

MHCG は，観測に対してカテゴリカルなラベルを割り当てる (naming) MH naming game[1] を，観測に対するキャプションの生成（captioning）へと拡張したものである．MHCG に参加する 2 体のエージェントは事前学習を通して言語知識を獲得した後に，それぞれの言語表現を揃え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自身の信念に基づいて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け入れるか受け入れないかを判断し，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によって，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ができる．

まず，Inter-ProbVLM は Neuro-SERKET[6] に基づいて，図 1b に示すようにエージェント A とエージェント B に分割される．このとき，z^* と w^* の近似事後分布のパラメータとして ψ^* と ξ をそれぞれ導入する．MHCG では，グローバルパラメータ ξ, ϕ, ψ, θ を事前学習済みの VLM のパラメータで初期化する．つまり，各エージェントは事前に獲得した言語知識を持った状態からコミュニケーションを行う．本研究では，ϕ, ψ は ProbVLM[4] の枠組み，ξ は ClipCap[7] の枠組みに従って事前学習する．

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) 判定，(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：話し手のエージェントは式 (4) に示すように，自身の観測 o_d^{Sp} に基づいて潜在変数 z_d^{Sp} を得る．

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

(2)提案：次に，話し手のエージェントは式 (5) に示すように自身の潜在変数 z_d^{Sp} に基づいてキャプション w_d^{Sp} を提案する．

$$w_d^{Sp} \sim p(w_d^{Sp} | z_d^{Sp}, \xi^{Sp}) \tag{5}$$

(3)判定：聞き手エージェントは，話し手エージェントから受け取った w_d^{Sp} と自身が持つキャプション w_d^{Li} に基づいて，式 (6) に示す受容確率を計算する．

$$r = \min\left(1, \frac{p(z_d^{Li} | \phi^{Li}, w_d^{Sp})}{p(z_d^{Li} | \phi^{Li}, w_d^{Li})}\right) \tag{6}$$

この確率に基づいて相手のサインを受け入れるかを判定する．受け入れる場合は w_d^{Li} を w_d^{Sp} で更新する．

(4)更新：聞き手エージェントは判定した後のキャプション w_d^{Li} に基づいて，グローバルパラメータ ξ, ϕ, ψ, θ を更新する[1)]．更新には VLM の事

1)　本稿の基礎的検証ではグローバルパラメータは固定する．

正確性	キャプション
Lv.4	A rider in a red jacket pauses to enjoy the mountain scenery.
Lv.3	Rider at mountain viewpoint.
Lv.2	A person on a scooter on a city street with buildings perceived as mountains.
Lv.1	A teacher giving a lecture in a crowded classroom.

図 2: 実験 1 に用いる画像とキャプションの例

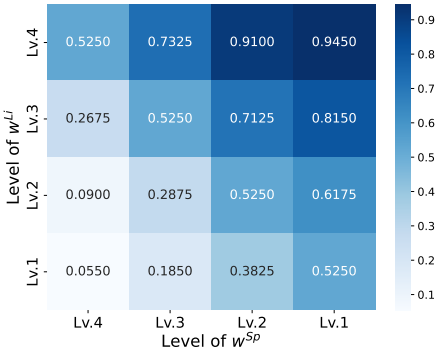


図 3: 正確性の異なるキャプションによる受容確率の違い

前学習と同様の方法が用いられる．

MHCG が MHNG と異なる点はサイン w に対する仮定とグローバルパラメータの更新方法である．MHNG ではカテゴリカルなサインを仮定していたのに対して，MHCG では画像に対するキャプションを仮定している．

4 実験

4.1 実験 1: 受容確率の妥当性検証

実験 1 では，ProbVLM が推論する確率分布に基づいて計算される受容確率の妥当性について検証する．Inter-ProbVLM において，ProbVLM は相手が提案したキャプションを受容するか棄却するかを判断するための受容確率の計算に用いられる．ここで計算される受容確率は，自身が持つキャプションより正確なものを提案された時には高くなり，正確でないものを提案された時には低くなるべきである．そこで，式 (6) に示す受容確率において，w^{Sp} と w^{Li} に様々な正確性のキャプションを適用することで，w^{Sp} と w^{Li} の正確性にどの程度差がある時にどの程度キャプションが受容されるかという受容割合を計算することでその妥当性を検証する．

ProbVLM の事前学習には MSCOCO[8] を使用する．MSCOCO は人間が生成したキャプションと画像のペアで構成されるデータセットである．実験では 566,435 個の画像キャプションのペアを使用する．

また，受容確率の妥当性の検証には，画像に対して 4 段階の正確性を持つキャプションを GPT-4[9] で生成したものを使用する．Lv.4 は正しく詳細なキャプション，Lv.3 は正しいが不十分なキャプション，Lv.2 は誤った表現を含むキャプション，Lv.1 は全く無関係なキャプションである．これらのキャプションは GPT-4[9] を用いて，各正確性ごとに 20 文ずつ，計 80 文を生成する．実験ではこれら 80 文のすべての組み合わせで w^{Sp} と w^{Li} に適用し，各正確性ごとに受容された割合を計算する．使用した画像と各正確度のキャプションの例を図 2 に示す．

結果　正確性の異なるキャプションによる受容割合の違いを図 3 に示す．w^{Sp} が Lv.4，w^{Li} が Lv.1 のキャプションを ProbVLM に入力して受容確率を計算した場合，受容された割合は 0.945 となった．また w^{Sp} が Lv.2，w^{Li} が Lv.4 では受容確率が 0.185 となっている．これらのことから，w^{Sp} の正確性が高く，w^{Li} の正確性の低い時に受容割合が高くなり，正確性の差が大きいほど受容確率も高くなる傾向が見られることがわかる．また逆に w^{Sp} の正確性が低く，w^{Li} の正確性が高い場合は，受容割合が低くなる．以上のことから，ProbVLM は妥当な受容確率を計算できる確率的 VLM を構成できていると考えられる．

4.2 実験 2: VLM のベイズ的統合

実験 2 では，Inter-ProbVLM に対して MHCG の (4) 更新を除いた過程を適用し，図 4 異なるデータセットで事前学習したエージェントが持つサインが，最終的に両エージェントにとって尤度が高くなるかについて検証する．MHCG を通して共有されるサインは両エージェントにとって尤度が高くなることが望まれる．ここでは，各エージェントは異なるデータセットで事前学習したパラメータで初期化し，MHCG の (4) 更新を除いた過程を行う．

MHCG の前段階としての事前学習において，エージェント A は Conceptual Captions[10] を，エージェ

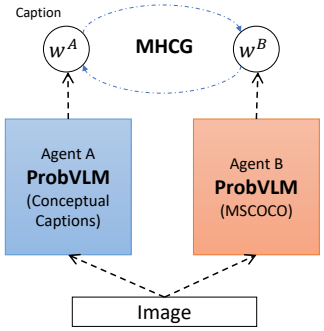


図 4: 実験 2 における MHCG の概要

表 2: MHCG の前後での各エージェントのキャプションと対数尤度の違い（付録参照）

	エージェント	キャプション	$\log p(z^A\|w)$	$\log p(z^B\|w)$
	A (MHCG 前)	the best part of the meal is the fried chicken.	-59232.76	-64790.60
	A (MHCG 後)	a white plate topped with a sandwich and a salad.	-46825.36	-57566.79
	B (MHCG 前)	a sandwich and a salad are on a plate.	-51686.87	-61723.88
	B (MHCG 後)	my friend made this for me.	-58469.59	-37546.62

ント B は MSCOCO のデータセットを使用する．MSCOCO は実験 1 のデータと同様のものを使用する．Conceptual Captions は Web で収集した画像と，それに対応する代替テキストがペアとなったデータセットであり，画像に対する抽象的なキャプションが多く含まれる．各エージェントは異なるデータセットにより事前学習を行うため，それぞれのデータセット特有の言語表現を獲得する．MHCG による推論で使用するデータは MSCOCO から抽出した 100 枚の画像で構成される．これらは事前学習で使用されなかったデータである．

各エージェントが持つサイン w の尤もらしさは以下の式に示す対数尤度に基づいて評価する．

$$\log p(z^A, z^B|w) = \log p(z^A|w) + \log p(z^B|w) \quad (7)$$

z^A, z^B は各エージェントが観測に基づいて推論した視覚言語表現の潜在変数である．両エージェントにとって尤もらしいサインに収束していれば対数尤度が高くなる．

結果　MHCG を 100 イテレーション行ったときのエージェント A とエージェント B のサインに対する対数尤度をそれぞれ図 5 の上図と下図に示す．両図において，MHCG のイテレーションを通して両エージェントの対数尤度が高くなっている．このことから，MHCG の (1)-(3) の過程を経て推論されたサインは，両エージェントの視覚言語表現に対して尤度を高めるものになっていることがわかる．

また，MHCG の前後で各エージェントが持つサインの例と，その時の対数尤度を表 2 に示す．この例ではエージェント B が MHCG 後に持つサインのキャプションは “my friend made this for me.” であった．これは，COCO で事前学習したエージェントが，エージェント A が事前学習に用いた Conceptual captions の持つ抽象的な表現を受け入れていることを示している．

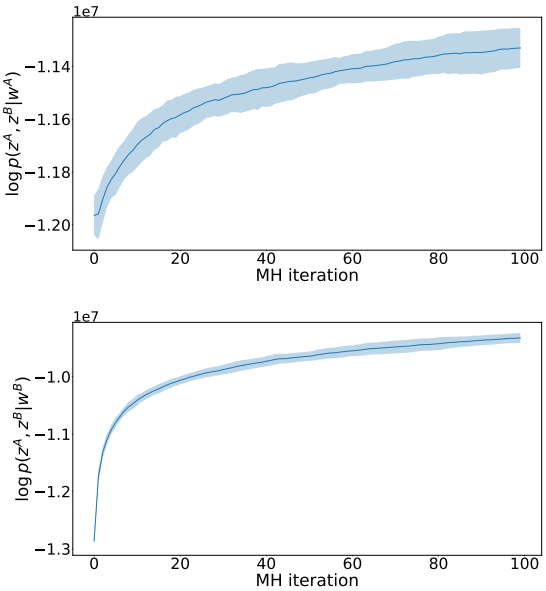


図 5: 上: 対数尤度 $\log p(z^A, z^B|w^A)$ の遷移．下: 対数尤度 $\log p(z^A, z^B|w^B)$ の遷移．

5　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化法である MHNG を，画像に対するキャプション生成へと拡張した 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，その基本的な検証を行った．

今後の展望として，グローバルパラメータの更新まで含めた MHCG による推論を実現し，記号創発（コミュニケーション創発）の枠組みにおいて，異なる VLM がコミュニケーションを通して動的かつ相互にお互いの言語表現と理解をアラインメントしていく言語創発モデルの構成を行なっていく．またこれを通して分散的な VLM のベイズ的統合を実現する手法の構成を目指す．

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] 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.

[2] Yoshinobu Hagiwara, Hiroyoshi Kobayashi, Akira Taniguchi, and Tadahiro Taniguchi. Symbol emergence as an interpersonal multimodal categorization. **Frontiers in Robotics and AI**, Vol. 6, p. 134, 2019.

[3] 谷口忠大. 分散的ベイズ推論としてのマルチエージェント記号創発. 日本ロボット学会誌, Vol. 40, No. 10, pp. 883–888, 2022.

[4] 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.

[5] 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.

[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] 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.

[9] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[10] 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.

表 3: MHCG の前後での各エージェントのキャプションと対数尤度の違い

画像	エージェント	キャプション	$\log p(z^A \mid w)$	$\log p(z^B \mid w)$
	A (MHCG 前)	a kitchen in the early morning kitchen of a typical small town in the south of the country.	-54459.75	-32795.40
	A (MHCG 後)	a kitchen with a sink and some yellow towels.	-49466.35	-57429.73
	B (MHCG 前)	a kitchen has a sink , dishwasher and window.	-59262.30	-60266.11
	B (MHCG 後)	the kitchen of a modern home in the early 1 9 7 0 s.	-56021.66	-53117.23
	A (MHCG 前)	a zebra in the snow.	-50668.16	-61060.09
	A (MHCG 後)	a zebra in the snow.	-50668.16	-61060.09
	B (MHCG 前)	a zebra standing in the dirt near some trees.	-60535.46	-57902.65
	B (MHCG 後)	a photo of the zeans in the winter season.	-55840.02	-51635.70
	A (MHCG 前)	person kicks soccer balls in the game.	-64488.96	-54213.11
	A (MHCG 後)	a man playing soccer in front of a green jersey.	-51727.23	-57839.92
	B (MHCG 前)	a man playing soccer on a field with a soccer ball.	-58999.56	-62202.45
	B (MHCG 後)	person kicks a goal during the tournament.	-52826.02	-56172.42

A 付録

A.1 実験２：MHCG 前後の各エージェントのキャプションと対数尤度

表 3 に実験２で行った MHCG 前後での各エージェントのキャプションと対数尤度の違いを示す。一枚目の例では Conceptual Captions で事前学習したエージェント A が COCO のようなキャプション”a kitchen with a sink and some yellow towels.” を受容している．またこのキャプションは MHCG 前のキャプションよりも正確な表現を含んでいる．二枚目はエージェント A が相手のサインを受容せずに棄却し続けた例である．三枚目は MHCG を通して相手のデータセット特有の表現を受け入れることは出来ているが，誤った表現を含んでいる例である．