Hol-CCG 構文解析と拡散モデルの統合による 構文構造を陽に考慮した画像生成

山木良輔 [1]　品川政太朗 [2]　持橋大地 [3]　谷口忠大 [1]
[1] 立命館大学
[2] 奈良先端科学技術大学院大学 [3] 統計数理研究所
{yamaki.ryosuke, taniguchi}@em.ci.ritsumei.ac.jp
sei.shinagawa@is.naist.jp　daichi@ism.ac.jp

概要

近年，拡散モデルに基づくテキスト画像生成モデルは高度な画像生成を実現している．しかし，構文的曖昧性を有するテキストに対して，特定の構文的解釈に対応する画像を選択的に生成するという点において課題がある．そこで本研究では，テキストが有する構文構造を分散表現として計算可能な構文解析モデルである Hol-CCG [1] と，画像生成モデルである Stable Diffusion [2] を統合し，特定の構文的解釈に対応する画像を選択的に生成可能なモデルを提案する．

1　はじめに

テキストの内容に沿った画像を生成するテキスト画像生成において，画像生成モデルはテキストに含まれる意味的情報だけでなく，構文的情報も合わせて理解することが求められる．例えば，入力されるテキストに構文的曖昧性が存在する場合，複数存在する構文的解釈の中から，特定の解釈に対応する画像を選択的に生成することが必要となる．このように，テキストが有する構文的情報に基づいた画像生成が可能なモデルの実現は，工学的に有用であると同時に，自然言語の構文構造と実世界情報の関係性を計算論的に捉えるという観点からも重要である．本研究では，構文解析モデルと拡散モデルに基づくテキスト画像生成モデルを統合することで，構文的曖昧性を有するテキストに対して特定の構文的解釈に対応する画像を選択的に生成可能なモデルを提案する．

近年，拡散モデルに基づくテキスト画像生成モデルが高度な画像生成を実現している [3, 2]．ただし，構文的曖昧性を有するテキストに関しては，各構文的解釈に応じて生成すべき画像が大きく異なる．例えば，図 1 に示すように，*“A man enters the room with flowers.”* という文は，前置詞句 *“with flowers”* が動詞句と名詞句のいずれを修飾するのかによって，「花を持った男性が部屋に入る」または「花が存在する部屋に男性が入る」という二つの構文的解釈が可能である．そして，これらの解釈を画像にする場合，画像中の物体同士 (*A man, the room, flowers*) は異なる関係性を有するべきである．

図 1: 構文的曖昧性を有するテキストに対する画像生成の例．

本研究では，テキストが有する構文構造を構文解析手法によって分散表現に変換し，それらを画像生成モデルが使用する言語特徴量とすることで，特定の構文的解釈に対応する画像を選択的に生成するためのモデルを提案する．具体的には，埋め込み空間上で組み合わせ範疇文法 (Combinatory Categorial Grammar; CCG) [4] に基づく構文解析を行うモデルである Holographic CCG (Hol-CCG) [1] と Stable Diffusion [2] を統合することによって，テキストが有する構文構造を分散表現に変換し，それらの情報を元に特定の構文的解釈に対応する画像を選択的に生成する．

実験では，Hol-CCG が計算するテキストの構文構造に関する情報を含んだ分散表現を，画像生成に使用するモデルと使用しないモデルの比較により，本提案モデルの有効性を示す．

2 先行研究: TIED

曖昧性を持つテキストに関する画像生成に関して，Mehrabi らは Text-to-ImagE Disambiguation framework (TIED) [5] を提案している．

TIED では曖昧性を有するテキストに対して，曖昧性を解消するための質問を GPT-2 [6] などの言語モデルによって生成し，質問に対するユーザの回答を元のテキストに付与することで，曖昧性が解消された新たなテキストを作成する．そして，この曖昧性が解消されたテキストを画像生成モデルに入力することによって，特定の解釈に基づく画像の生成を実現しようとしている．

3 準備

3.1 Stable Diffusion

近年の拡散モデル [7] の顕著な発展により，写実性の高い画像の生成が実現されている [3, 2]．拡散モデルでは，ランダムなノイズに対して段階的にノイズを除去する過程を繰り返すことによって，画像の生成を実現する．ここで，拡散モデルのノイズを除去する過程において，テキスト情報を付与することでテキストの内容に沿った画像の生成が可能となる．

本研究では，これらのテキスト画像生成モデルの中でも特に Stable Diffusion [2] を使用する．Stable Diffusion では，画像の潜在空間上でノイズ除去を行い，得られた画像特徴量を VAE [8] のデコーダによって画像に復号することで画像の生成を行う．

3.2 Hol-CCG

筆者らは CCG [4] に基づく構文解析モデルとして Holographic CCG (Hol-CCG) [1] を提案している．図 2 に Hol-CCG のモデル概要を示す．Hol-CCG は Span-based Parsing [9] に基づく句構造解析アルゴリズムによって，テキストが有する尤もらしい句構造を探索する．ここで，Hol-CCG は句構造の探索と同時に，テキストが有する構文構造を明示的に反映した句の分散表現を計算することができる．また，Hol-CCG では分散表現を計算するための演算自身には学習が必要となるパラメータが存在しないことが特徴であり，大規模かつ複雑なデータセットに対しても適用可能である．

そこで本研究では，構文的曖昧性を有するテキストに対して，各構文構造 (構文的解釈) に応じた句の分散表現を Hol-CCG によって計算し，これらの分散表現を画像生成の際の言語特徴量として用いることで，特定の構文的解釈に対応する画像を選択的に生成するモデルを提案する．

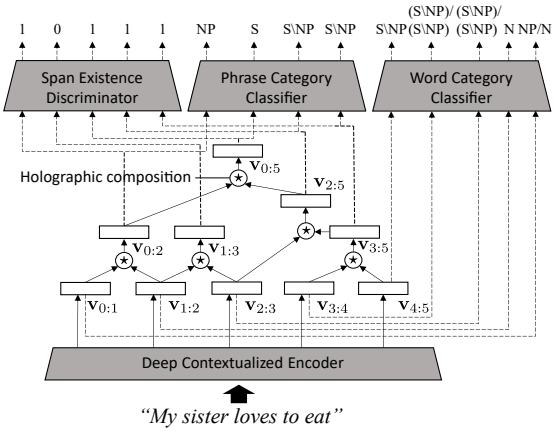


図 2: Hol-CCG [1] のモデル概要図．入力文はテキストエンコーダによって単語分散表現へと変換され，それらの分散表現は句構造に従って句の分散表現へと再帰的に合成される．

4 提案モデル

4.1 モデル概要：Hol-CCG と Stable Diffusion の統合

本研究では，Hol-CCG と Stable Diffusion を統合することによって，構文的曖昧性を有するテキストに対して特定の構文的解釈に対応する画像を選択的に生成するモデルを提案する．本提案モデルの概要を図 3 に示す．

本モデルでは，まず事前学習済みの CLIP [10] によってテキスト中の各単語及びテキスト全体に対する分散表現を得る．そして，Hol-CCG によってこれらの単語分散表現を構文構造に従って再帰的に合成し，構文構造に関する情報を含む句の分散表現を得る．これらの単語・句・テキスト全体の分散表現は，Stable Diffusion に与えられ，特定の構文的解釈に基づいた画像を生成するために使用される．

4.2 モデルの学習

提案モデルの学習はテキストエンコーダの学習と Stable Diffusion の学習の 2 段階に大別される．なお，実験時に使用したハイパーパラメータ等に関しては，付録 A に記す．

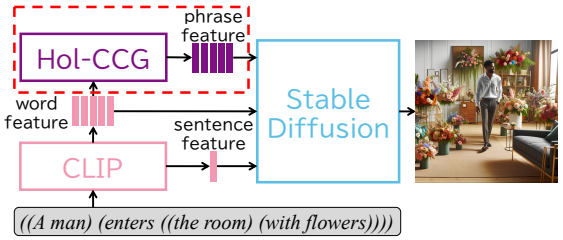


図 3: 提案モデルの概要図．入力テキストは CLIP によって各単語とテキスト全体の分散表現へと変換される．単語分散表現は Hol-CCG によって句の分散表現へと再帰的に合成され，これらの分散表現を Stable Diffusion に与えることで画像を生成する．

テキストエンコーダの学習

まず，第 1 段階目として，テキストを分散表現に変換するテキストエンコーダ部分の学習を行う．前述の通り，本モデルのテキストエンコーダは事前学習済みの CLIP と Hol-CCG によって構成される．これら 2 つのモデルを結合した状態で以下の誤差関数 ($\mathcal{L}_{fine-tuning}$) を最小化するように誤差逆伝播法によるファインチューニングを行う．

$$\mathcal{L}_{fine-tuning} = \mathcal{L}_{contrastive} + \mathcal{L}_{word} + \mathcal{L}_{phrase} + \mathcal{L}_{span}$$

ここで，$\mathcal{L}_{contrastive}$ は CLIP の対照学習に基づく誤差，$\mathcal{L}_{word}, \mathcal{L}_{phrase}, \mathcal{L}_{span}$ は Hol-CCG の学習に使用される CCG のカテゴリ分類に基づく誤差をそれぞれ表す．これらの誤差の詳細については元論文 [10, 1] を参照されたい．$\mathcal{L}_{fine-tuning}$ を最小化するようにテキストエンコーダのファインチューニングを行うことで，意味的情報と構文的情報の両者を分散表現に含めることが可能となる．

Stable Diffusion の学習

次に，Stable Diffusion の学習を行う．Stable Diffusion がノイズを段階的に除去する際に，ノイズ除去後の潜在変数と元の潜在変数との間の平均二乗誤差を計算し，この誤差を最小化するようにモデルの学習を行う．これにより，Stable Diffusion はテキストエンコーダから受け取った分散表現に含まれる構文構造に関する情報を元に，特定の構文的解釈に対応する画像を生成することが可能となる．

5 実験

5.1 データセット

データセットには LAVA Corpus [11] を使用した．本データセットには構文的曖昧性を含むテキストと各構文木及びそれらに対応する画像が格納されている．本データセットに対する前処理の内容に関しては付録 B に記す．

5.2 モデル比較条件

以下の 2 つの比較条件を設けることで，提案モデルの有効性を検証した．

(1) テキストエンコーダの学習の有無

4.2 節で述べた通り，事前学習済み CLIP と Hol-CCG を結合したテキストエンコーダを学習することで，意味的情報と構文的情報を含んだ分散表現を計算することが可能となる．ここで，テキストエンコーダが構文的情報を捉えることの有効性を検証するために，テキストエンコーダをファインチューニングするか否かという比較条件を設けた．

(2) 句の分散表現の使用の有無

Stable Diffusion は Hol-CCG が計算した句の分散表現を画像生成時の言語特徴量として使用することで，特定の構文的解釈に対応する画像を生成することが可能となる．ここで，Hol-CCG が計算する句の分散表現を画像生成時の言語特徴量として使用することの有効性を検証するために，句の分散表現を画像生成に使用するか否かという比較条件を設けた．

5.3 評価方法

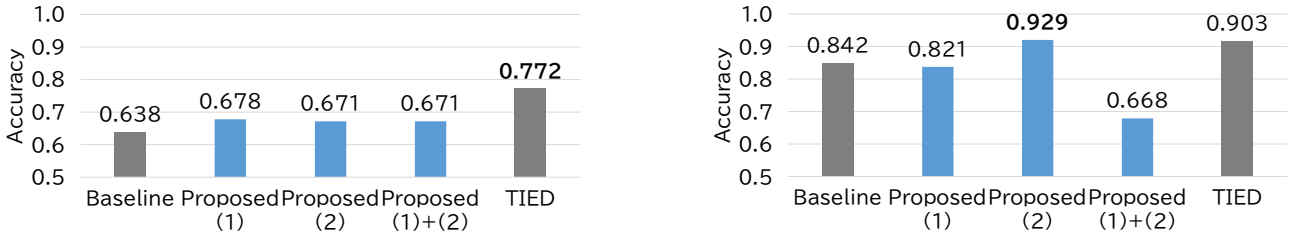
実験結果に関して，VQA による自動評価と人手による評価の 2 種類の評価を行った．

VQA による自動評価

各モデルによって生成された画像が指定された構文的解釈に対応したものになっているかを自動的に評価するために，Mehrabi らが提案した VQA による自動評価手法 [5] を採用した．本評価手法では，生成された各画像が指定された構文的解釈に対応している場合に，その答えが “Yes” となるような質問文を用意する．そして，質問文と生成された画像の組に対して VQA が “Yes” と回答した割合を評価指標として使用する．本評価手法の詳細に関しては，元論文 [5] を参照されたい．

人手評価

人手による評価では，各モデルによって生成された 20 組の画像とそれぞれに対応する質問文を 19 名



(a) VQA による自動評価の結果.

(b) 人手評価の結果.

図 4: VQA 及び人手による評価の結果.

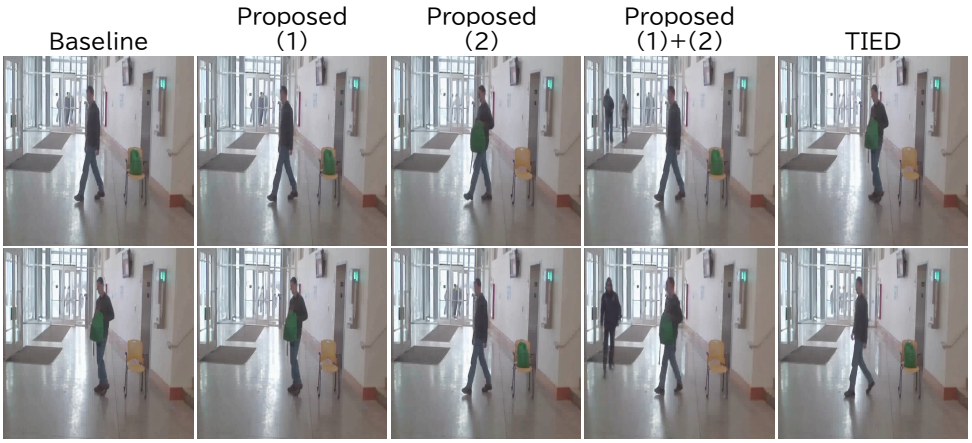


図 5: *"Danny left the chair with a green bag."* に対する生成画像の例．上段は「*Danny* が *bag* を持っている」解釈，下段は「*bag* が *chair* に置かれている」解釈にそれぞれ対応する．

の実験参加者に提示し，VQA による自動評価と同様のプロセスを実行する．すなわち，各実験参加者は提示された画像の内容に基づいて，質問文に対する回答を "Yes" か "No" から選択する．そして，実験参加者による質問文への回答が "Yes" となった割合を各モデルごとに比較した．

6 結果・考察

図 4 に VQA による自動評価及び人手評価の結果を示す．なお，グラフ中の値は指定された構文的解釈に対して各モデルが生成した画像の Accuracy を表している．まず，VQA による自動評価に関しては，5.2 節で示した比較条件の (1)・(2) をいずれも適用しなかった Baseline モデルが最低の性能となっており，提案モデルの方が高い性能を発揮している．また，比較条件 (1)・(2) の一方もしくは両者を適用したモデル間での性能差はほとんど見られない．

人手による評価に関しては，(2) の比較条件のみを適用した場合に最も性能が高くなっており，TIED を上回る性能を発揮している．ただし，比較条件 (1)・(2) を両方とも適用したモデルの性能が他のモデルに比べて著しく低くなっており，これらの条件を同時に適用することがモデルの学習に対して悪影響を与えている可能性を示唆している．

また，図 5 に実際に生成された画像の例を示す．これより，提案モデルの内，比較条件 (2) のみを適用したモデルは，各構文的解釈に対応する画像を選択的に生成できていることが分かる．

以上より，特定の構文的解釈に対応する画像を選択的に生成するという本研究の目的に関して，Hol-CCG が計算した句の分散表現を画像生成時の特徴量として使用することは有効であるといえる．

7 おわりに

本研究では，構文解析モデルである Hol-CCG と Stable Diffusion を統合することで，構文的曖昧性を持つテキストに関して，特定の構文的解釈に対応する画像を選択的に生成するためのテキスト画像生成モデルを提案した．実験より，Hol-CCG がテキストに対して計算する句の分散表現を画像生成に使用することの有効性を示す結果が得られた．今後の展望としては，より大規模なデータセットを使用した提案モデルの有効性の検証や，構文的曖昧性の種類ごとでのより詳細な結果の分析などが挙げられる．

謝辞

本研究は JSPS 科研費 JP23H04835 の助成を受けたものです.

参考文献

[1] Ryosuke Yamaki, Tadahiro Taniguchi, and Daichi Mochihashi. Holographic CCG parsing. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 262–276, Toronto, Canada, July 2023. Association for Computational Linguistics.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 10684–10695, 2022.

[3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 36479–36494, 2022.

[4] Mark Steedman. **The syntactic process**. MIT press, 2001.

[5] Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. Resolving ambiguities in text-to-image generative models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 14367–14388, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. **Advances in neural information processing systems**, Vol. 33, pp. 6840–6851, 2020.

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. **arXiv preprint arXiv:1312.6114**, 2013.

[9] Mitchell Stern, Jacob Andreas, and Dan Klein. A minimal span-based neural constituency parser. In Regina Barzilay and Min-Yen Kan, editors, **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 818–827, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[10] 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.

[11] Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. Do you see what I mean? visual resolution of linguistic ambiguities. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 1477–1487, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019**. OpenReview.net, 2019.

[13] 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.

表 1: テキストエンコーダの学習時に使用したハイパーパラメータ．

Hyperparameters	Values
Pretreined CLIP	openai/clip-vit-large-patch14
Optimizer	Adam [12]
Max Training Epochs	100
Batch Size	4
Learning Rate	1e-5

表 2: Stable Diffusion の学習時に使用したハイパーパラメータ．

Hyperparameters	Values
Pretreined Stable Diffusion	CompVis/stable-diffusion-v1-4
Optimizer	Adam [12]
Max Training Steps	50000
Batch Size	1
Learning Rate	1e-5
Denoising Steps	100

A　モデルの学習に関するハイパーパラメータ

提案モデルの学習時に使用した事前学習済みモデル及びハイパーパラメータを表 1, 2 にそれぞれ示す．なお，実験には NVIDIA A100 GPU [80GB] を 1 台使用した．

B　データセットの前処理

本研究における実験では LAVA Corpus [11] に対して以下の 2 つの前処理を適用し，データセット全体を 9:1 の割合で学習用と評価用のサブセットに分割して使用した．

1. **構文木の変換**：LAVA Corpus では構文的曖昧性を有するテキストの各構文的解釈に対して PCFG による構文木のアノテーションが付与されている．これらの PCFG によるアノテーションを CCG によるものに変換する．
2. **評価用質問文の作成**：VQA による自動評価で使用するための質問文を各テキストの各構文的解釈に関して作成する．例えば，*“Danny approached the chair with a yellow bag”* というテキストに対して「*Danny* が *bag* を持っている」という解釈をする場合，*“Is the man carrying a yellow bag?”* という質問文を作成し，同じテキストに対して「*bag* が *chair* に置かれている」という解釈をする場合，*“Is the chair accompanied by a yellow bag?”* という質問文を作成する．なお，これらの質問文の作成には GPT-4 [13] を使用した．