構文解析と画像生成の統合による機能語の言語理解

山木良輔 [1]　谷口忠大 [1]　持橋大地 [2]
[1] 立命館大学 [2] 統計数理研究所
{yamaki.ryosuke,taniguchi}@em.ci.ritsumei.ac.jp　daichi@ism.ac.jp

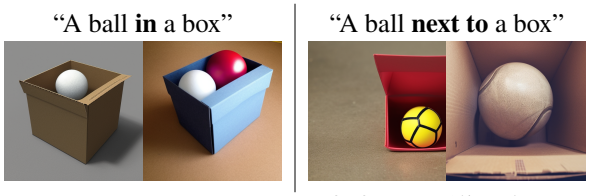

図 1: Stable Diffusion が生成した画像の例.

概要

近年の大規模言語モデルと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一方で，前置詞などの機能語に関しては，それらの意味を正しく捉えられていないことが指摘されている [1]．本研究ではこの問題に対して，CCG に基づく構文解析モデルとテキスト画像生成モデルを統合することで，テキスト中の物体同士の関係性を明示的に捉えた画像生成を可能にし，機能語のような文法的関係を正しく捉える言語理解を実現する．

1　はじめに

自然言語における機能語と実世界情報の対応関係を理解可能な人工知能の実現は，人工知能の実世界応用において重要な課題である．本研究では，構文解析とテキスト画像生成モデルを統合し，機能語が持つ言語情報を画像情報として表出させることで，実世界情報と紐づいた機能語の言語理解の実現を目指す．

近年の大規模言語モデルと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 [2, 3, 4]．一方で言語理解の観点からは，“in” や “on” などの文法的機能に関しては，その意味を正しく捉えられていないことが指摘されている [1]．例えば，図 1 に示すように，Stable Diffusion [4] は “ball” や “box” といった内容語に関しては，これらの物体が出現する画像を生成している一方で，“in” や “next to” といった異なる前置詞 (機能語) に関しては，どちらも “in” の意味に対応する画像を生成しており，これらの意味を捉えられていない．

実世界上において，自然言語を用いた人間とのコミュニケーションが期待される人工知能は，このような機能語に関してもその意味を理解可能であることが望ましい [5]．例えば，“Place a ball next to a box.” のような指示をロボットに与えたときに，ゴール状態として図 1 の右側の画像をロボットが想起してしまうと，適切な行動を期待することはできない．

そこで，本研究ではこのような機能語の言語理解を実現するための方法として，自然言語の統語的情報をテキスト画像生成モデルに取り入れる手法を提案する．具体的には，組み合わせ範疇文法 (Combinatory Categorial Grammar: CCG) [6] に基づく構文解析モデルである Holographic CCG (Hol-CCG) [7] をテキスト画像生成モデルにおけるテキストエンコーダとして採用し，Hol-CCG が出力する統語的及び意味的情報を含んだ単語・句・文の分散表現を画像生成に活用する．これにより，機能語が示す文中の物体同士の関係性を画像情報と対応付けることが可能となり，結果的に機能語の言語理解が可能になると考える．

実験より，テキストから得られる句の分散表現・統語的情報をテキスト画像生成に活用することの有効性を示唆する結果が得られた．なお，後述の実験では機能語が示す物体同士の位置関係に主眼を置いているため，生成画像の写実性は本研究の論点から外れていることに注意されたい．

2　関連研究

2.1　Hol-CCG

本研究では，テキスト画像生成モデルにおけるテキストエンコーダとして筆者らが以前に提案した CCG に基づく構文解析モデルである, Hol-CCG [7] を使用する．Hol-CCG は潜在空間上での単語及び句の分散表現同士の再帰的合成によって，それらの間に存在する階層関係及び依存関係を明示的に考慮した分散表現を計算することが可能であり，機能語が表現する内容語同士の関係性を画像生成に反映す

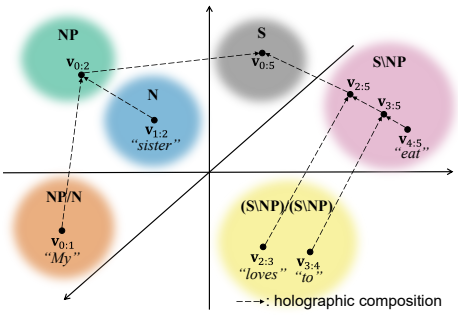

図 2: Hol-CCG のモデル概要図.

るという本研究の目的に適した特性を有している．Hol-CCG はテキスト画像生成モデルに与えられた入力文の構文構造を解析し，それに応じた単語・句・文の分散表現を出力する．そして，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．

Hol-CCG のモデル概要を図 2 に示す． Hol-CCG では，まず入力文を RoBERTa [8] などの Masked Language Model に通すことで，文中の各単語の分散表現を得る．そして，得られた単語分散表現を巡回相関 [9] によって再帰的に合成することによって，句および文の分散表現を計算する．巡回相関の演算は次式で定義される．

$$[\mathbf{c}]_k = [\mathbf{a} \star \mathbf{b}]_k = \sum_{i=0}^{d-1} a_i b_{(k+i) \mod d} \quad (1)$$

ここで，$\star$ は巡回相関の演算子，$\mathbf{a}, \mathbf{b}$ は合成前の分散表現，$\mathbf{c}$ は合成後の分散表現である．例えば，図 2 中の例の場合，最終的な文の分散表現 $\mathbf{v}_{0:5}$ は次のような再帰的な演算によって計算される．

$$\mathbf{v}_{0:5} = (\mathbf{v}_{0:1} \star \mathbf{v}_{1:2}) \star (\mathbf{v}_{2:3} \star (\mathbf{v}_{3:4} \star \mathbf{v}_{4:5})) \quad (2)$$

また，分散表現を合成する過程において，Span-based Parsing [10] に基づく句構造解析アルゴリズムによって，入力文が持つ尤もらしい句構造を探索する．これによって，入力文が持つ構文的曖昧性を解消しつつ，文を構成する単語・句・文の意味的・統語的な情報を含んだ分散表現を構成することができる．Hol-CCG では分散表現を再帰的に合成する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，本モデルは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．

2.2 AttnGAN

本研究では，テキスト画像生成モデルとして Attentional GAN (AttnGAN) [11] を使用する．AttnGAN は Attention 機構によって入力文中の各単語の情報を画像生成に取り込むモデルである．また，Deep Attentional Multimodal Similarity Model (DAMSM) モジュールにおいて，テキストと画像間での対応関係を評価することで，入力テキストに対して意味的に合致するような画像を生成する．

3 提案手法

3.1 手法概要：Hol-CCG と AttnGAN の統合

本研究では，AttnGAN のテキストエンコーダとして Hol-CCG を用いることで，テキスト画像生成に統語的情報を取り入れる手法を提案する．提案手法の概略を図 3 に示す．

まず，Hol-CCG が入力文を構文解析することによって，単語・句・文の統語的・意味的な情報を含む分散表現を出力する．そして，AttnGAN はこれらの分散表現を言語特徴量として画像を生成する．より具体的には, GAN の隠れ層の状態を更新する際に単語及び句の分散表現との間での Attention の計算を行う．ここで，単語の分散表現のみではなく，句の分散表現も取り入れていることが本手法の重要箇所であり，これによって，複数の単語間に存在する階層的な依存関係の情報を画像生成過程に取り込む．さらに，元々の AttnGAN と同様に，最終的に生成された画像は DAMSM モジュールにおいてテキストとの意味的な対応関係が評価される．この評価過程においても，単語・文の分散表現に加えて句の分散表現が言語特徴量として用いられる．

3.2 モデルの学習

本提案手法では，AttnGAN の生成器・識別器の学習に先立って，テキスト及び画像エンコーダの事前学習を行う．この事前学習の方法は基本的には AttnGAN を踏襲しており，統語的情報及び句の分散表現を取り込めるように拡張を行った．まず，元々の AttnGAN における DAMSM モジュールの学習に関する損失関数 $\mathcal{L}_{DAMSM}$ を句の表現が取り込めるものに拡張した方法について説明する．

$$\mathcal{L}_{DAMSM} = \mathcal{L}_{word} + \mathcal{L}_{phrase} + \mathcal{L}_{sentence} \quad (3)$$

上式における $\mathcal{L}_{phrase}$ は句の分散表現を DAMSM モジュールに取り込むために，筆者らが新たに導入した項である．ここで，$\mathcal{L}_{DAMSM}$ を構成する各項はそれぞれ，単語・句・文の分散表現と画像の特徴量を用いて，テキストと画像間での対応関係を評価するためのものである．例えば，M 個の画像情報とテキスト情報のペア $\{(Q_i, D_i)\}_{i=1}^{M}$ に関して，$\mathcal{L}_{word}$ は次

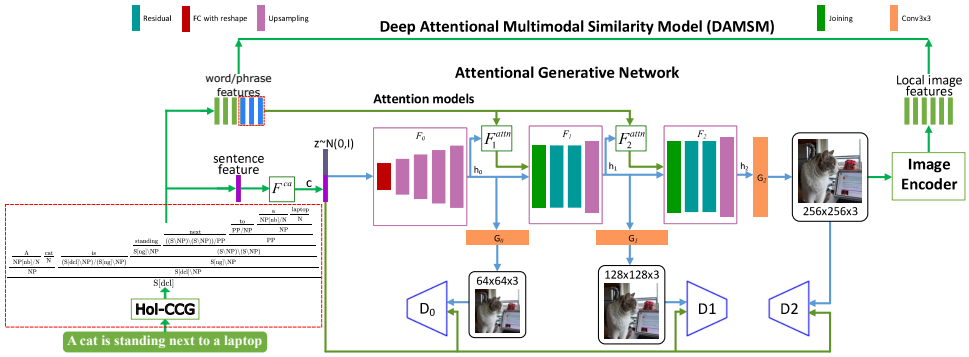

図 3: 提案手法概略図 (AttnGAN の元論文 [11] 中のモデル概略図におけるテキストエンコーダ部分を Hol-CCG による構文解析に置換).

の通りに定義される.

$$\mathcal{L}_{word} = -\sum_{i=1}^{M}\{\log P(Q_i|D_i) + \log P(D_i|Q_i)\} \quad (4)$$

なお, $P(Q_i|D_i), P(D_i|Q_i)$ を計算する過程に関しては本稿では割愛するが, 詳細は AttnGAN の元論文 [11] を参照されたい. また, $\mathcal{L}_{phrase}, \mathcal{L}_{sentence}$ の計算過程は $\mathcal{L}_{word}$ と同様である.

次に, 統語的情報を取り入れるために行った拡張に関して説明する. AttnGAN が画像生成に用いる単語・句・文の特徴量に統語的情報を取り入れるにあたり, 本提案手法では CCG カテゴリの情報を利用する. CCG カテゴリは構文解析の結果として文中の単語や句に対して割り当てられるものであり, 多くの統語的情報を含んでいる. そのため, テキストエンコーダの事前学習において, テキストエンコーダが出力する単語・句・文の分散表現から, それらに対応する CCG カテゴリが正しく予測されるような損失関数を定義することで, 分散表現中に統語的情報が取り込まれるようにする. 具体的には, 次のような $\mathcal{L}_{syntax}$ を定義する.

$$\mathcal{L}_{syntax} = \mathcal{L}_{word-syn} + \mathcal{L}_{phrase-syn} + \mathcal{L}_{span-syn} \quad (5)$$

ここで, $\mathcal{L}_{word-syn}, \mathcal{L}_{phrase-syn}$ はそれぞれ単語・句に対する CCG カテゴリの割り当て誤差, $\mathcal{L}_{span-syn}$ は特定の区間に含まれる単語列が句を構成するか否かの判別誤差である. 詳細に関しては Hol-CCG [7] のモデル学習方法を参照されたい.

上記の $\mathcal{L}_{DAMSM}$ と $\mathcal{L}_{syntax}$ の合計をエンコーダの事前学習の損失関数 $\mathcal{L}_{pretrain}$ とする.

$$\mathcal{L}_{pretrain} = \mathcal{L}_{DAMSM} + \mathcal{L}_{syntax} \quad (6)$$

また, エンコーダの学習完了後に行われる GAN の生成器と識別器の敵対的学習に関しては元々の AttnGAN と同一であるため, 本稿では説明を割愛する. 詳細に関しては AttnGAN の元論文 [11] を参照されたい.

4 実験

4.1 データセット

本研究では, 2 つの性質の異なるデータセットを用いて実験を行った.

COCO テキスト画像生成分野で一般的に用いられている Common Objects in Context (COCO [12]) を客観的な評価・比較のために用いた.

CLEVR 本研究では、AttnGAN を画像生成モデルとして採用したが、COCO (日常生活のあらゆる場面の画像を含む) の画像を高品質に生成することが難しいことから、機能語の言語理解に関する主観的な評価を行うには不適切であると考えた. そのため、本研究では Compositional Language and Elementary Visual Reasoning (CLEVR [13]) データセットを用い、3 個以下の物体を含む画像についてアノテーション情報から自動的にキャプションを生成し, 実験に使用した.

4.2 ベースライン手法

ベースライン 1 AttnGAN のテキストエンコーダに RoBERTa [8] を用いたモデルを, 提案手法において統語的情報が付与されることの有効性を検証するための比較対象として用いる. このモデルではテキスト入力に対して RoBERTa が出力する [CLS] トークンと各サブワードの分散表現を言語特徴量として AttnGAN に与え画像を生成する. 事前学習の損失関数は $\mathcal{L}_{DAMSM} = \mathcal{L}_{word} + \mathcal{L}_{sentence}$, $\mathcal{L}_{pretrain} = \mathcal{L}_{DAMSM}$ となる.

表 1: COCO データセットにおける客観評価の結果.

手法	FID↓	R-prec↑
AttnGAN [11]	35.49	85.47
AttnGAN + VICTR [14]	29.26	86.39
ベースライン 1	60.76	87.07
ベースライン 2	39.55	93.59
提案手法	31.29	93.12

ベースライン 2　提案手法において句の分散表現を画像生成に用いないモデルを 2 つ目のベースライン手法とする．このモデルとの比較によって，提案手法において句の分散表現を画像生成に取り入れることの有効性を検証する．事前学習の損失関数は $\mathcal{L}_{DAMSM} = \mathcal{L}_{word} + \mathcal{L}_{sentence}$, $\mathcal{L}_{syntax} = \mathcal{L}_{word-syn}$ となる．

4.3　評価方法

客観評価　Fréchet Inception Distance (FID[1)]) と R-precision (R-prec[2)]) を評価指標として用いた．

主観評価　16 名の実験参加者に各手法によって生成された 20 組の画像を提示し，各実験参加者はテキストが示す物体同士の位置関係との合致度が最も高い画像を選択する．そして各手法によって生成された画像が選択された割合を比較した．

5　結果・考察

客観評価　客観評価の結果・比較を表 1 に示す．なお，VICTR [14] は AttnGAN のテキストエンコーダ部分にシーングラフから抽出された物体同士の位置関係の情報を明示的に取り入れている点から，本研究との関連があるため比較手法として取り上げた．結果より，提案手法は FID においてベースライン手法に比べて高い性能を発揮している．ただし，ベースライン手法 1 は他の 2 つのモデルに比べて著しく高い FID を示しているため，GAN の学習の不安定性のために学習に失敗している可能性があることには注意されたい．また，既存手法との FID における比較に関しては，提案手法は AttnGAN の性能を上回り，VICTR に競合する性能を発揮している．なお，提案手法では GAN の学習段階におけるハイパーパラメータの設定を改善することで，さらなる性能向上が見込まれる．以上より，提案手法において句の分散表現及び統語的情報をテキスト画像生成に取り込むことの有効性が示唆される．

1)　実画像と生成画像の特徴量空間上での分布の距離を測る指標であり，値が小さいほど生成画像が高品質とみなす．
2)　画像特徴量とテキスト特徴量の類似度を評価する指標であり，値が 100 に近いほど生成画像が入力テキストの条件に合致しているとみなす．

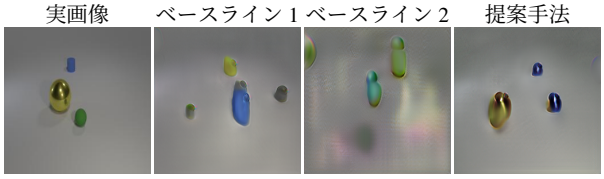

図 4: CLEVR における生成画像例 (入力テキスト：“large yellow sphere is to the left front of small blue cylinder, and large yellow sphere is to the left behind of small green sphere”). 提案手法が “yellow sphere” と “blue cylinder” の位置関係 (“to the left front of”) を最も正確に反映している．

主観評価　各手法によって生成された画像が実験参加者によって選択された割合はそれぞれ，ベースライン 1: 39.1%，ベースライン 2: 18.1%，**提案手法: 42.8%**となった．実験参加者に提示した画像の一例を図 4 に示す．図の例の様に，提案手法がベースライン手法に比べて，入力テキストが説明する物体同士の位置関係をより正確に捉えた画像を生成している例が複数見られた．一方で，ベースライン 1 と提案手法の画像が選択された割合が僅差であることを踏まえると，大域的には手法間の差は小さいと考えられ，提案手法の有効性は十分には示されていないと考える．そのため，入力するテキストの複雑度や GAN の学習におけるハイパーパラメータの設定などにさらなる改良が必要であると考えている．なお，ベースライン 2 の画像が選択された割合は他の 2 つの手法に比べて著しく低く，この手法においては GAN の学習に失敗している可能性があることには注意されたい．

6　おわりに

本研究では構文解析とテキスト画像生成モデルを統合することによって自然言語における機能語の言語理解を実現するための手法を提案した．入力テキストに対して構文解析モデルが出力する統語的・意味的情報を含んだ単語や句の分散表現を画像生成に活用することで，機能語が示す物体同士の関係性を画像情報と対応付けた画像の生成に取り組んだ．実験より，提案手法の有効性を示唆する客観的な結果が得られた．

今後の展望としては，画像生成部分を Stable Diffusion [4] などの拡散モデルベースのものに置換し，より高品質な画像の生成を可能にすることや，物体同士の位置関係に関する表現以外の前置詞・機能語に関して，より詳細な実験・評価を行うことなどが挙げられる．

謝辞

本研究は，JST，ムーンショット型研究開発事業，JPMJMS2033 の支援を受けたものです．

参考文献

[1] Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022.

[2] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[3] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[4] 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.

[5] Amir Aly, Tadahiro Taniguchi, and Daichi Mochihashi. A bayesian approach to phrase understanding through cross-situational learning. In *International Workshop on Visually Grounded Interaction and Language (ViGIL), in Conjunction with the 32nd Conference on Neural Information Processing Systems (NeurIPS), 2018.*, 2018.

[6] Mark Steedman. *The Syntactic Process*, Vol. 24. MIT press Cambridge, MA, 2000.

[7] 山木良輔, 谷口忠大, 持橋大地. Holographic Embeddings による CCG 構文解析. 言語処理学会第 28 回年次大会 (NLP2022), 2022.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[9] Tony A Plate. Holographic reduced representations. *IEEE Transactions on Neural networks*, Vol. 6, No. 3, pp. 623–641, 1995.

[10] Mitchell Stern, Jacob Andreas, and Dan Klein. A minimal span-based neural constituency parser. *arXiv preprint arXiv:1705.03919*, 2017.

[11] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324, 2018.

[12] 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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

[13] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

[14] Caren Han, Siqu Long, Siwen Luo, Kunze Wang, and Josiah Poon. Victr: Visual information captured text representation for text-to-vision multimodal tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3107–3117, 2020.