The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

言語から画像を生成する深層学習モデルの挙動に関する考察

A Study on Behavior of Deep Neural Text-to-Image Generative Model

藤山 千紘 *1　　小林 一郎 *2
Chihiro Fujiyama　　Ichiro Kobayashi

*1お茶の水女子大学 人間文化創成科学研究科 理学専攻 情報科学コース
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

*2お茶の水女子大学 基幹研究院 自然科学系
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

In this study, we analyze the behavior of the computational mechanism and the structure of the feature representation space in a deep neural text-to-image generative model. This is a fundamental approach with a goal to construct artificial general intelligence reflecting the mechanism of human intelligence. First, we explore whether the model is capable of encoding captions and of generating valid images under the circumstance given input captions without word boundaries. Qualitative and quantitative evaluations demonstrate that it can generate compelling images, but the computational mechanism does not acquire the units of meaning. Secondly, we analyze the semantic compositionality in the embedding space. Our experimental result suggests that the semantic compositionality appears between words indicating positions.

1. はじめに

近年，汎用人工知能の構築を目指して，ヒトの知能に関する知見を取り入れた人工神経回路網の研究が盛んに行われている [Cox 14, Lotter 17]．また，機械学習モジュールが自然言語の意味を理解することを，自然言語から画像へのグラウンディングができることとして捉え，自然言語の意味理解，ひいては自然言語からの概念の獲得を目指して，自然言語で記述されたキャプションを入力とする画像生成モデルの提案が数多くなされている [Mansimov 16, Xu 18, Zhang 18]．一方で，深層学習分野での研究成果は経験的な成果として示されることが多く，モデルの計算機構の挙動や特徴表現空間の構造についての分析はあまり行われていない．

本研究では，ヒトの知能のメカニズムを模倣して構築された，キャプションを入力とする画像生成モデルを対象に，入力の粒度を変更した際の計算機構の挙動や，特徴表現空間の構造を，ヒトの知能との親和性の観点から分析する．

2. alignDRAW

Mansimov ら [Mansimov 16] は，ヒトが絵を描く際の,「特定の言語表現に着目してそれに対応する部分を描く」というプロセスの反復を，深層学習の枠組みで実現することを意図して構築した画像生成モデル alignDRAW を提案している．このモデルは，Variational AutoEncoder[Kingma 14a] を拡張し，生成画像の段階的な高精細化および空間的注意を導入した画像生成モデル Deep Recurrent Attentive Writer(DRAW)[Gregor 15] を，キャプションで条件付けて画像生成できるようにした Encoder-Decoder モデルである．alignDRAW は，単語単位のキャプションを入力にとり，双方向 LSTM[Hochreiter 97] で構成される言語エンコーダで処理を行った後，DRAW を画像デコーダとして，attention mechanism[Bahdanau 15] と合わせて用いることにより，反復的に画像生成を行う．alignDRAW の概要図を図 1 に示す．

連絡先: 藤山千紘，お茶の水女子大学，〒 112-8610 東京都文京区大塚 2-1-1，fujiyama.chihiro@is.ocha.ac.jp

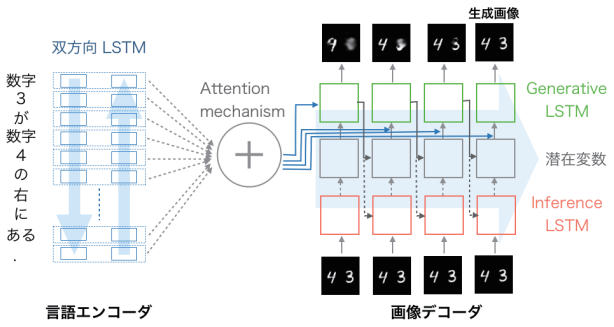

図 1: alignDRAW 概要図

3. 提案手法

3.1 alignDRAW を用いた単語分割タスクを含む画像生成

先行研究 [Mansimov 16] では，単語分割済みのキャプションから画像生成を行っているが，本手法では入力を単語分割されていないキャプションに変更し，単語の境界情報が失われた，すなわち意味の単位の情報が欠落した場合の alignDRAW の言語エンコード能力および画像生成能力を評価する．具体的には，単語分割されていないキャプションに対して，妥当な画像を生成し得るか，またその際の attention mechanism の挙動が言語の意味の単位を表現しているかを考察する．

3.2 alignDRAW における言語の意味の構成的特性の分析

本手法では，モデルの特徴表現空間において言語の意味の構成的特性が表現されるかを評価する．先行研究 [Mansimov 16] では，キャプションに含まれる各単語を one-hot ベクトルとして双方向 LSTM に入力しているが，本手法では各単語を分散表現に埋め込んだ後，双方向 LSTM に入力するよう，埋め込み層を追加する形でモデルの拡張を行う．モデルは単語分割済みのキャプションを用いて学習し，埋め込み層の分散表現について，単語の意味の構成的特性を分析する．

4. 実験

4.1 実験設定

データセットは，表 1 に示すテンプレートと手書き数字画像のデータセット MNIST[*1] を用いて作成した．キャプションはプレースホルダーを含むテンプレートを各実験 8 種類用意し，MNIST から無作為に抽出した画像および正解ラベルの組のうちラベル情報を埋め込む形で作成した．画像は，ラベルと対応する MNIST 画像をキャプション内容に適合する領域に，無作為に 4 ピクセルのゆらぎを持たせて配置した 60×60 ピクセルのグレースケールの画像とした．両実験ともに，学習データ 40,000 事例，開発データ 4,000 事例，評価データ 4,000 事例を用いた．モデルの実装には深層学習のフレームワーク TensorFlow[*2] を用い，学習は各実験，表 2 の設定で行った．

表 1: キャプション作成時のテンプレート

単語分割タスクを含む画像生成	構成的特性の分析
すうじ＿がすうじ＿のひだりにある．	数字 ＿ が画像の左にある．
すうじ＿がすうじ＿のみぎにある．	数字 ＿ が画像の右にある．
すうじ＿がすうじ＿のうえにある．	数字 ＿ が画像の上にある．
すうじ＿がすうじ＿のしたにある．	数字 ＿ が画像の下にある．
すうじ＿がぞうのひだりうえにある．	数字 ＿ が画像の左上にある．
すうじ＿がぞうのひだりしたにある．	数字 ＿ が画像の左下にある．
すうじ＿がぞうのみぎしたにある．	数字 ＿ が画像の右下にある．
すうじ＿がぞうのみぎうえにある．	数字 ＿ が画像の右上にある．

表 2: alignDRAW 学習時のハイパーパラメータ

	単語分割タスクを含む画像生成	構成的特性の分析
入力語彙サイズ	33	28
言語エンコーダ	one-hot ベクトル → 128 ユニット双方向 LSTM	32 次元分散表現 → 128 ユニット双方向 LSTM
attention mechanism	512 ユニット Bahdanau Attention	256 ユニット Bahdanau Attention
デコーダ	300 ユニット DRAW LSTM	
描画反復回数	32 ステップ	
潜在変数 z	150 次元	
最適化アルゴリズム	RMSProp [Tieleman 12]	
学習率	初期学習率: 0.001 110 エポック以降 0.0001 に減衰	初期学習率: 0.001 75 エポック以降 15 エポック毎に 0.5 倍
パラメータ初期値	平均: 0, 分散: 0.1 の正規分布乱数	
学習エポック数	200	150

生成画像の定量評価としては，生成された画像を入力キャプション毎に分類する分類器を用い，弁別性の自動評価を行った．分類器の学習は alignDRAW の学習に用いたデータセットと同一のデータセット上で行い，ハイパーパラメータは表 3 に示す設定とした．なお，キャプションは単語分割タスクを含む画像生成で 440 種類，構成的特性の分析で 80 種類であるので，各々440 クラス分類，80 クラス分類となる．

表 3: 定量評価のための分類器学習時のハイパーパラメータ (両実験とも共通)

アーキテクチャ	(CNN+最大値プーリング) × 4 層 + 全結合 × 1 層
畳み込みに関する設定	フィルタサイズ: 5 × 5 (全層) チャネル数: 入力層から順に 32, 64, 128, 256 0 パディング ストライド: 1
全結合層	1024 ユニット
活性化関数	ReLU 関数
損失関数	交差エントロピー損失
勾配クリッピング	[-10.0, 10.0]
最適化アルゴリズム	Adam [Kingma 14b]
初期学習率	0.0001
パラメータ初期値	平均: 0, 標準偏差: 0.1 の $\pm 2\sigma$ の切断正規分布乱数
学習エポック数	100

*1 http://yann.lecun.com/exdb/mnist/
*2 https://www.tensorflow.org/

4.2 単語分割タスクを含む画像生成

単語分割されていないキャプションからの生成画像例を図 2 に示す．単語分割済みのキャプションからの画像生成と比較して，単語の境界情報が失われている，つまり意味の単位の情報が欠落している点で，画像生成タスクとしてはより困難になっていると考えられるが，生成結果からキャプション内容に適合する画像を生成できていることが確認できる．

図 2: 単語分割されていないキャプションからの生成画像例

分類における評価データ上での正解率と alignDRAW による生成画像データ上での正解率を表 4 に示す．生成画像データ上での正解率が評価データ上での正解率を上回っており，alignDRAW によって生成された画像がどのキャプションに由来するかを十分に識別できる質を達成できていることが分かる．これは，生成モデル学習の結果，データセットに存在するノイズや手書き文字の個人差が吸収され，正規化された数字を生成できるようになったためと考えられる．

表 4: 評価データおよび alignDRAW による生成画像データ上での分類正解率

評価データ	生成画像データ
0.719	0.737

またテンプレートに従わないキャプションからの生成画像例を図 3 に示す．テンプレートに含まれるキャプションからの生成結果 (図 2) と比較して，描かれる数字の質が劣化している事例が見受けられるが，おおよそキャプション内容に適合する画像を生成できており，「すうじ」という言語表現の省略や主語の位置変更に対して，モデルがある程度頑健に動作していることが認められる．

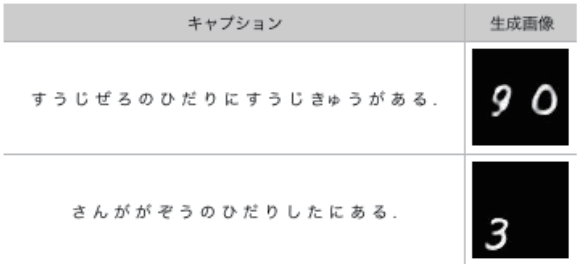

図 3: テンプレートに従わないキャプションを入力とする生成画像例

続いて，画像を生成する際の attention mechanism の挙動を図 4 に示す．図 4 上段については，「いち」という言語表現付近に attention が当たっているときに画像空間上で 1 に相当する部分の生成が進んでおり，続いて数字 2 を表す「に」という言語表現付近に attention が移って画像空間での 2 に相当する

部分の鮮明化が進んでいる様子が見てとれる．2 箇所のプレースホルダーを持つテンプレートに由来するキャプションでは，一方の数字のアイデンティティを示す言語表現およびその周囲に attention が当たっているときにそれに対応する部分が画像空間で鮮明化し，attention が他方の数字に移ると生成部位も移るという傾向が共通して見られ，数字のアイデンティティを特定する言語表現と画像空間での表現に大まかな対応がとれていることが確認された．一方で，この場合には空間を表す言語表現についてはほとんど attention が当たっていなかった．また図 4 下段については，生成画像には 3 を描くことができているが，数字のアイデンティティを表現する「さん」にはほとんど attention が当たっていない．この傾向はプレースホルダーが 1 箇所のみのテンプレートに由来するキャプションに共通して見られ，数字のアイデンティティを表す言語表現と画像空間での表現に対応関係が認められなかった．

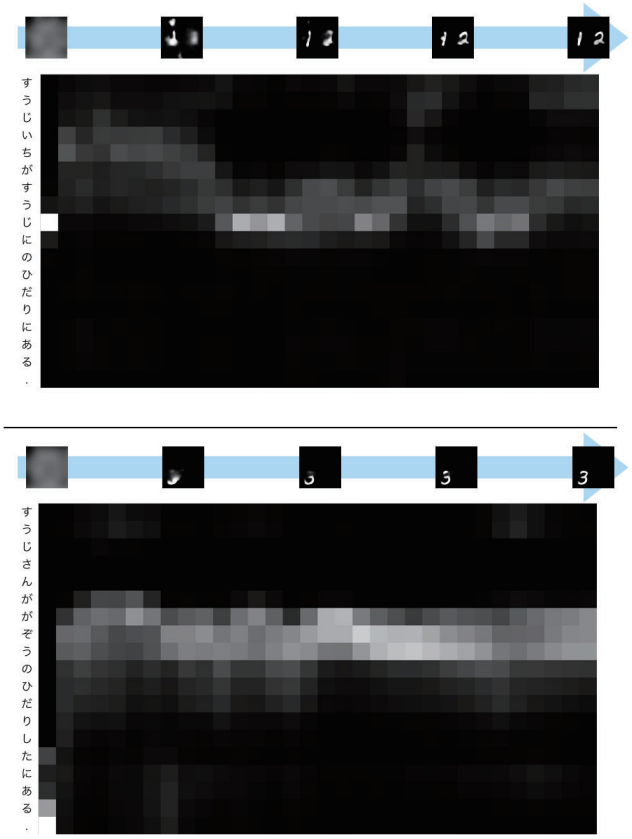

図 4: attention mechanism の挙動：左に入力キャプション，上に画像の生成過程 (左から t=0, 7, 15, 23, 31) を示す．

単語分割されていないキャプションから妥当な画像を生成できたことから，attention mechanism の挙動が意味の単位を表現していることが期待されたが，図 4 のように，本設定下ではそのような傾向は見られなかった．これは，言語エンコーダとして用いた双方向 LSTM の表現能力が高いことが一因であると考えられるため，言語エンコーダを簡素化した追加実験を行って分析を加える．本稿では系列処理を行わず，one-hot ベクトルを直接 attention mechanism の入力として画像生成を行った場合について述べる．なおこの場合，系列処理を行わないため入力の各文字は集合として扱われ，並びの順序は無視される．したがって，キャプション内の主体と対象の関係を画像空間上に適切に反映させ得ない場合があることが想定される．

キャプションの系列処理を行わない場合の生成画像例を図 5 に示す．図 5 上段のように妥当な画像を生成できる場合もある一方で，下段のように位置関係が反転する事例があることが確認された．キャプションを系列処理しなかった場合の attention mechanism の挙動を図 6 に示す．系列処理を行わない，つまりキャプションを文字の集合として扱った場合には，数字のアイデンティティと描画位置を一意に特定するために必要な言語表現に attention が集中する傾向があったが，これは必ずしも単語単位での表現とは一致しないため，attention mechanism の挙動が意味の単位を表現している様子は見られなかった．一方で，双方向 LSTM を用いた場合と比較すると，数字のアイデンティティや空間を表す言語表現の一部に確実に attention が当たっていることが確認できた．

図 5: キャプションを系列処理しない場合の生成画像例

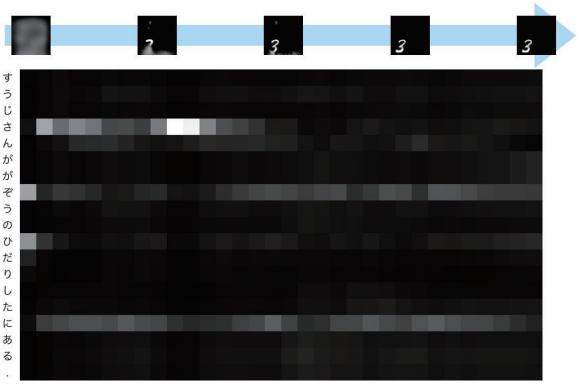

図 6: キャプションを系列処理しない場合の attention mechanism の挙動：左に入力キャプション，上に画像の生成過程 (左から t=0, 7, 15, 23, 31) を示す．

alignDRAW では画像生成過程に数字を描く順序の制約がなく自由度が高いため，このことに起因して attention の学習が困難である．また現在のモデルには，言語表現の連続したセグメントが意味の単位となり得るという情報が一切与えられていないため，attention mechanism の挙動で言語の意味の単位を表現することが困難であったと考えられる．attention mechanism で意味の単位を捉えつつ，ヒトが絵を描く過程を計算機構に反映して画像生成を行うためには，損失関数の検討を含めてモデルの拡張が必要であると考えられ，これは今後の課題の一つである．

4.3 言語の意味の構成的特性の分析

学習したモデルを用いた生成画像の例を図 7 に示す．本実験においても，キャプション内容に適合する画像が生成されていることが確認できる．

分類における評価データ上での正解率と alignDRAW による生成画像データ上での正解率を表 5 に示す．本実験では，生成画像データ上での正解率が評価データ上での正解率を下回る結果となり，学習した生成モデルがデータセットの分布を同質といえる程には近似できていなかったことが分かる．図 8 に

図 7: 構成的特性の分析を目的とした実験での生成画像例

正しく分類できなかった事例を示す．間違って分類された生成画像は，入力キャプションに含まれる数字のアイデンティティを正しく画像内に表現できていない事例が多かった．本分析では空間を意味する言語表現に着目するため，数字が描かれる位置のみを分類する，すなわち数字のアイデンティティの区別を分類の対象としない 8 クラス分類を表 3 の設定で学習した結果，評価データ上，生成画像データ上ともに分類正解率が 1.00 となり，数字を描く位置を間違えた事例はなかったと推測される．

表 5: 評価データおよび alignDRAW による生成画像データ上での分類正解率

評価データ	生成画像データ
0.985	0.882

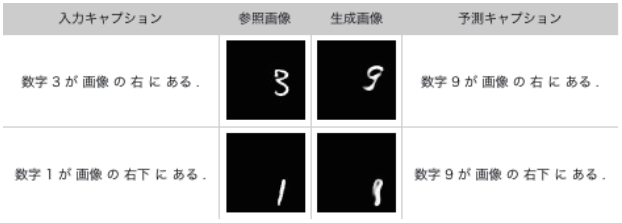

図 8: 正しく分類できなかった生成画像の例：左から生成モデルへの入力キャプション，参照画像，生成画像，分類器によって予測されたキャプションを示す．

言語の意味の構成的特性の分析では，キャプションに含まれる空間を意味する単語「左」「右」「上」「下」「左上」「左下」「右下」「右上」の 8 単語を対象とした．我々は例えば「左上」を，意味的に「左」と「上」を足したものとして解釈していると考えられる．この加法構成性を，埋め込み層の分散表現において獲得できるかを評価するため，まず学習された分散表現のうち「左」「右」「上」「下」に対応するものの和をとり，各々「左上」「左下」「右下」「右上」の推定分散表現を作成した．例えば，「左上」の推定分散表現は，「左」と「上」それぞれに対応する実分散表現の和として構成した．このようにして推定した分散表現に対して，学習で得た実分散表現それぞれとの cos 類似度を算出した結果を表 6 に示す．また，各推定分散表現と cos 類似度が高かった分散表現に対応する単語上位 5 件を表 7 に示す．本設定下では，「右上」以外の 3 単語については高い cos 類似度が得られ，表 7 においても推定分散表現と実分散表現が比較的近くに確認されることから，埋め込み空間において言語の意味の構成的特性が表現される可能性を示唆する結果を得たと言える．

表 6: 推定分散表現と実分散表現の cos 類似度

	左上	左下	右下	右上
cos 類似度	0.89	0.80	0.92	0.29

表 7: 推定分散表現との cos 類似度が高い分散表現上位 5 件

	左上	左下	右下	右上
1	左上	左	右	右
2	上	左上	1	1
3	左	左下	右下	右下
4	0	0	の	の
5	2	2	7	4

5. おわりに

本研究では，キャプションからの画像生成モデルについて，計算機構の挙動や特徴表現空間の構造の分析を行った．入力のキャプションをヒトが言語を獲得する過程を一段階遡る形で，単語分割されたキャプションから単語の境界情報の欠落したキャプションに変更した場合には，定性的にも定量的にもキャプション内容に適合する画像を生成するモデルを学習できた．しかし，その際の attention mechanism の挙動において意味の単位を獲得している様子は確認されなかった．また，特徴表現空間における言語の意味の構成的特性の分析については，埋め込み空間の分散表現について空間を意味する単語間で，意味の加法構成性を得られる可能性を示唆する結果を得た．

今後の課題としては，より精緻な分析を行うことや，ヒトの知能のメカニズムを反映して動作する生成モデルの構築が挙げられる．

参考文献

[Bahdanau 15] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate.”, In ICLR, 2015.

[Cox 14] D. D., Cox, and T., Dean, “Neural networks and neuroscience-inspired computer vision.”, Current Biology, 24(18), R921-R929, 2014.

[Gregor 15] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, “DRAW: A recurrent neural network for image generation.”, In ICML, 2015.

[Hochreiter 97] S. Hochreiter, and J. Schmidhuber, “Long short-term memory.”, Neural Computation, 9(8), pp. 1735-1780, 1997.

[Kingma 14a] D. P., Kingma, and M., Welling, “Auto-encoding variational bayes.”, In ICLR, 2014.

[Kingma 14b] D. P., Kingma, and J., Ba, “Adam: A method for stochastic optimization.”, arXiv preprint arXiv:1412.6980, 2014.

[Lotter 17] W., Lotter, G., Kreiman, and D., Cox, “Deep predictive coding networks for video prediction and unsupervised learning.”, In ICLR, 2017.

[Mansimov 16] E. Mansimov, E. Parisotto, J. L. Ba., and R. Salakhutdinov, “Generating images from captions with attention.”, In ICLR, 2016.

[Tieleman 12] T., Tieleman, and G., Hinton, Lecture 6.5-RMSProp, COURSERA: Neural Networks for Machine Learning. University of Toronto, Tech. Rep, 2012.

[Xu 18] T., Xu, P., Zhang, Q., Huang, H., Zhang, Z., Gan, X., Huang, and X., 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.

[Zhang 18] Z., Zhang, Y., Xie, and L., Yang, “Photographic text-to-image synthesis with a hierarchically-nested adversarial network.”, In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.