画像と対応する言語の意味の単位および構成的特性の分析

藤山 千紘 †
† 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

小林 一郎 ‡
‡ お茶の水女子大学
基幹研究院 自然科学系

{fujiyama.chihiro, koba}@is.ocha.ac.jp

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．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．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，自然言語による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内部計算機構の挙動や，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

alignDRAW[1] 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した部分を描く」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された画像生成モデルである．alignDRAW は，単語単位のキャプションを入力にとり，双方向 LSTM[2] で構成される言語エンコーダで処理を行った後，Deep Recurrent Attentive Writer[3] をデコーダとして，attention mechanism[4] 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．

3 提案手法

3.1 単語分割タスクを含む画像生成

先行研究 [1] では，単語分割済みのキャプションを入力として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．

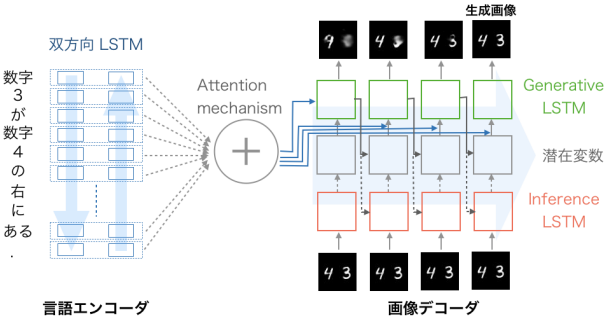

図 1: alignDRAW 概要図

3.2 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 [1] 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，モデルに埋め込み層を追加する．モデルの学習には単語分割済みのキャプションを用い，学習後，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．具体的には，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 単語を対象に分析を行う．例えば，我々は「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．具体的な手順としては，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する．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分散表現の和として構成する．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 類似度を算出して評価を行う．

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 種類ずつ用意し，MNIST から無作為にサンプリング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．各実験において，モデルの学習は表 2 の設定で行った．なお，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	300 ユニット DRAW LSTM
描画反復回数	32 ステップ	32 ステップ
潜在変数 z	150 次元	150 次元
最適化アルゴリズム	RMSProp	RMSProp
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	平均: 0, 分散: 0.1 の正規分布乱数
学習エポック数	200	150

[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を入力とした際の生成画像例を図 2 に示す．単語分割済みのキャプションを入力とする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の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．

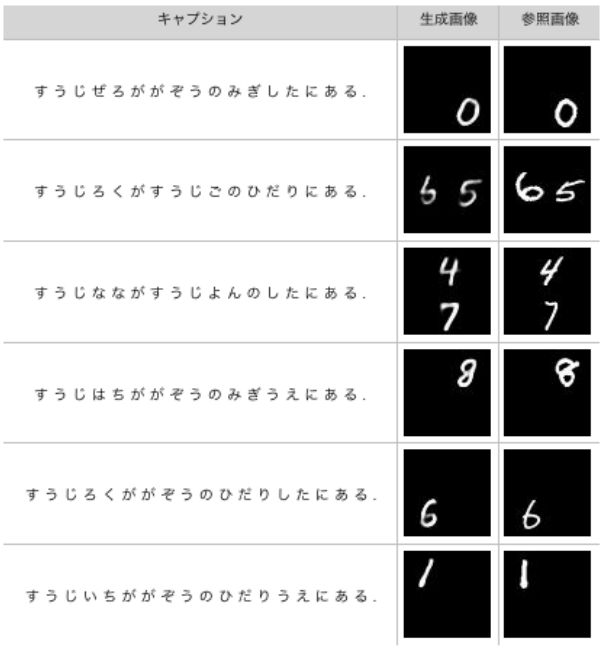

図 2: 単語分割されていないキャプションからの生成画像例

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の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の変更に対して，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．

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

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

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．

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．従って，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ることができ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

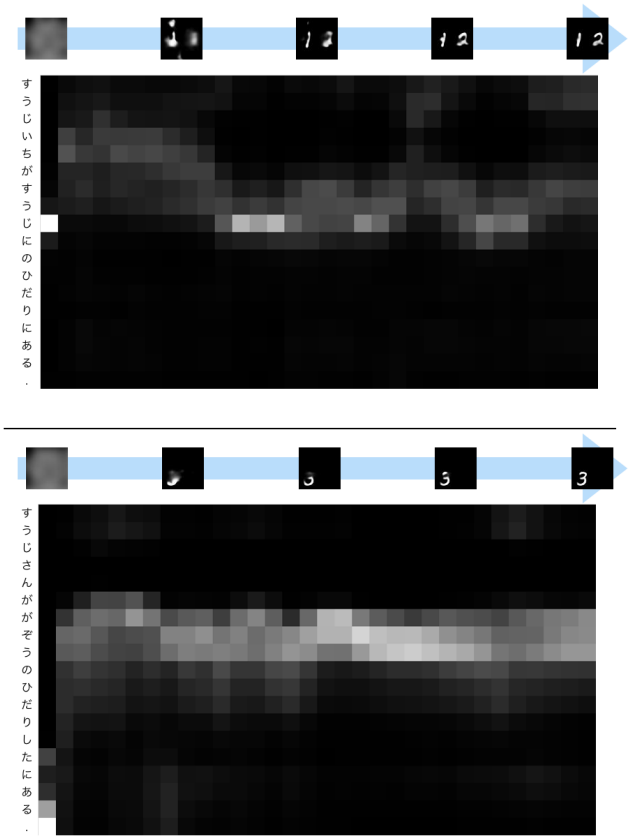

図 4: attention mechanism の挙動．左に入力キャプション，上に画像の生成過程 (左から t=0，7，15，23，31) を示す．

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．

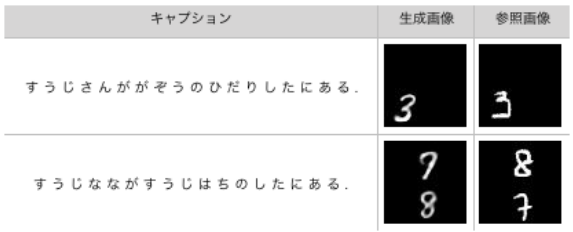

図 5: キャプションを系列処理しない場合の生成画像例

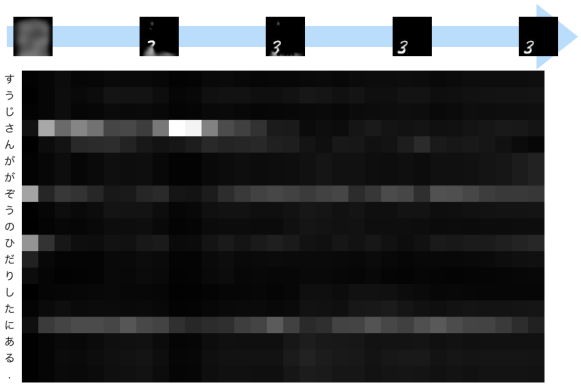

図 6: キャプションを系列処理しない場合の attention mechanism の挙動．左に入力キャプション，上に画像の生成過程 (左から t=0，7，15，23，31) を示す．

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を表現することが困難であったと考えられる．attention mechanism で言語の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，キャプションの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．

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

空間を意味する単語「左上」「左下」「右下」「右上」の 4 単語について，推定分散表現と実分散表現の cos 類似度を表 3 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 4 に示す．

表 3: 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 4: 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

本設定下では，「右上」以外の 3 単語については比較的高い cos 類似度が得られ，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．

5 おわりに

本研究では，自然言語を用いて記述されたキャプションを入力とする画像生成モデルについて，内部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，キャプションの内容に適合する画像を生成するモデルを学習することはできたが，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

[1] E. Mansimov, E. Parisotto, J. L. Ba., and R. Salakhutdinov, “Generating images from captions with attention.” in ICLR, 2016.
[2] S. Hochreiter, and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
[3] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, “DRAW: A recurrent neural network for image generation.” in ICML, 2015.
[4] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate.” in ICLR, 2015.