1A4-OS-27b-4

キャプションからの画像生成を行うニューラルネットへの対話的修正の導入と検討

Introducing Dialogue Based Modification to Neural Networks for Image Generation from Captions

品川 政太朗 Seitaro Shinagawa　吉野 幸一郎 Koichiro Yoshino　ニュービッグ グラム Graham Neubig　中村 哲 Satoshi Nakamura

奈良先端科学技術大学院大学　情報科学研究科
Nara Institute of Science and Technology, Graduate School of Information Science

Generating or retrieving images from natural language descriptions has potential applications in a number of creative tasks. However, it is not necessarily the case that an retrieved by these systems is sufficiently similar to the target image that the user imagined in advance. In this paper, we try to solve this problem by introducing a framework where the user can iteratively refine their request in a dialogue-like manner. We examine how image retrieval results change over the process of refining the user's query.

1. 本研究の背景と目的

写真や絵はしばしば我々が自分の想像している光景を相手に伝えるための可視化として役に立つ．例えば，見知らぬ海外の土地がどのような場所かをイメージしてもらうために似たような画像を見せたり，プレゼンテーションなどでテキストの代わりに絵を使うことで内容の理解を促すといったことが挙げられる．しかし，意図した通りの写真を撮る，絵を描くといった行為はときに初心者には難易度が高く，労力も大きい．自然言語による指示によってそれと対応する画像を得るシステムを構築することにより，これらの作業のコストを大きく削減することが期待できる．

思い通りの画像を得る方法として，まず画像検索が挙げられる [1]．しかし，検索で得られた既存の画像は著作権で保護されていることも多く，ユーザが利用したい様々な用途に用いることができない場合が多い．これに対して，学習した大量の画像から抽出された特徴量を元に，新たに画像を生成する研究が近年盛んに行われている [2]．しかし，画像検索と画像生成に共通の問題として，一度の検索・生成で必ずしも意図した画像が得られるとは限らないという問題がある．この原因としては，ユーザ側が十分な情報をシステムに与えられていないということが挙げられる．画像検索ではこの問題を解決する方法としてクエリ拡張や適合性フィードバックのような対話的処理の導入 [3] が検討されているが，本稿で検討する画像の説明文を介した対話での有効性は明らかではない．

本研究では自然言語による画像生成のインタラクションを想定する．ユーザはシステムに対してクエリとして指示を与える．指示にはオブジェクトの移動などルールベースで簡単に表現できるものも存在するが，「誰々を椅子に座らせる」などの複数のルールを複雑に組み合わせて初めて表現できるようなものも存在する．そこで本研究では，多様な指示の表現に対応してそれに合った画像を対話的に生成することを目指してニューラルネットを用いた手法を検討する．対話的に指示を行う場合，ユーザの入力する指示は対話履歴となる入力指示と出力画像に依存して変わることが想定され，これを考慮してモデルを構築する必要がある．本稿では既存の，自然言語と画像を共通の潜在空間で扱い，双方向に入出力を扱えるニューラルネット [4] を用いた画像説明文からの画像検索のフレームワークにおいて画像の説明文を介した対話的操作の効果を検証する．

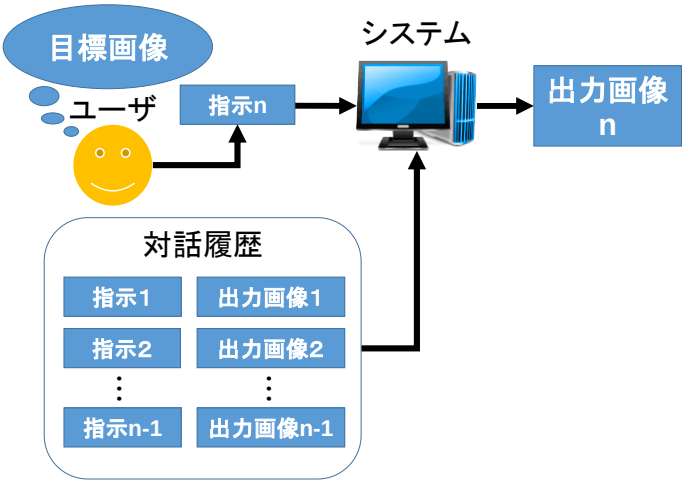

図 1: 対話的画像生成の概念図

2. 意図する画像の生成

2.1 関連研究

本研究の背景となる技術として，ニューラルネットワークによる画像からの説明文生成 [4][5] や説明文からの画像生成の研究 [2] が挙げられる．これらのタスクは画像と説明文の対応関係の紐づけが難しく，人手で特徴量を設計するのは容易ではないことからニューラルネットワークを用いた研究が盛んに行われている．ニューラルネットワークを用いる利点として，異なるモダリティの特徴量を共通の潜在空間にマッピングするよう学習させる類似度計算などの処理が容易にできるという点が挙げられる．ニューラルネットワークに画像内の物体の位置関係 [6] や注意機構 [2][7] を加えることが検討されており，説明文からの画像生成においては説明文と画像の両方に注意機構を導入することで生成される画像の位置や色をコントロールできる

連絡先: 品川 政太朗:shinagawa.seitaro.si8@is.naist.jp
連絡先: 吉野 幸一郎:koichiro@is.naist.jp
連絡先: ニュービッグ グラム:neubig@is.naist.jp
連絡先: 中村 哲:s-nakamura@is.naist.jp

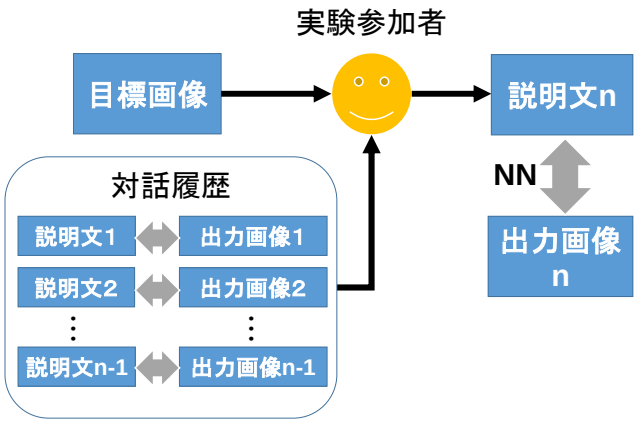

図 2: 対話的画像生成の検討方法

可能性が示唆されている．しかし，ユーザ側に想像している画像があって，それを生成する場合にはユーザ側が十分な情報を一度に与えられるとは限らない．これを解決する方法として，仕様書のようなまとまった情報を用意して与える方法が考えられるが，これは想定される操作項目をあらかじめ用意する穴埋め形式による方法となり，異なる操作を行う際に再度操作に関する定義を行わなければならないという問題がある．

2.2 対話による操作

本稿では上記の問題点を解消するためユーザとシステムとの間で対話的に生成画像を修正していく方法を検討する．概略を図 1 に示す．この対話ではユーザはシステムに自然言語による指示 (クエリ) を与え，システムは対話履歴 (これまでの指示，出力画像) に従い出力画像を生成する．ユーザはその画像を見て指示の変更や修正依頼を行う．この操作を加えることで，システム側は対話履歴を用いて検索に多くの情報を利用し，検索の精度を向上することが期待できる．

3. 対話的画像生成の検討

対話的画像生成を行うにあたって，クエリとしてどのような指示を与えればよいかという点は明らかではない．例えば画像内の物体を「上に移動」「赤色に変える」といった簡単に思いつくものについてはテンプレートの穴埋めによる方法が有効だと言えるが，テンプレートの種類をどの程度用意すればよいかの検討がつかないという問題がある．また，ユーザが指示を自由記述する場合にどのような指示を与えるか，目的の画像が得られなかった場合にどのように指示を変更するかが明らかではない。これらを調査することにより，対話的画像生成を行う上での問題点，およびどのような指示変更依頼をユーザに出せばよいかが明らかとなる。そこで調査として，今回の実験ではユーザである実験参加者には指示の代わりにシステムに出力して欲しい画像の説明文を入力してもらった．実験の概略を図 2 に示す．ここで実験参加者はユーザには「指示を出す」という対話行為を「現在までに入力した画像説明文と出力画像を元に，目標画像により近い画像を出力すると推測される画像説明文の入力」を行ってもらった．これにより，実験参加者に入力可能なクエリの自由度を担保しつつ，画像とその説明文により学習された既存のニューラルネットワークをそのまま用いて実験を行った．ここで，図 1 におけるユーザの時刻 n での入力指示 n は画像の説明文 $\{1,\cdots,n-1\}$ と説明文 n の差分で表現されると仮定した．

また，画像生成はタスク自体の難しさから．生成できている画像は生成される物体の種類が人の顔や屋内などの限定されたデータセットであるか，生成できても多少ぼやけた画像になるという問題がある．そこで本稿では実験の簡単化と対話の導入に主眼に置いて検討するため，Kiros ら [4] が公開している学習済みニューラルネット *1 による画像検索，説明文検索手法を用いて実験を行った．各時刻での入力説明文と出力画像はニューラルネットによってそれぞれ共通潜在空間に同次元の特徴量としてマッピングされ，コサイン類似度による計算によって相互にデータベース上の画像，説明文を近い順からランク付けして検索する．このマッピングを用いて，対話中の共通潜在空間上の説明文の特徴量の動きを追跡する．

3.1 実験

実験は英語を日常的に不自由なく扱える（TOEIC800 点以上)20 代の男女 5 名を対象に行った．扱う画像は MSCOCO[8] を用いた．目標画像は train2014(82,783 画像,413,915 キャプション) からランダムに選び，検索対象となる画像は validation2014(40,504 画像,202,520 キャプション) を用いた．各実験参加者は 1 タスク 10 ターンの対話を 10 タスク行った．ここで，1 ターンとは「説明文の入力→出力画像生成」を 1 ターンとして最大 10 ターンを 1 タスクと定義する．各実験参加者は各ターンの終了ごとに 2 種類の主観評価を 5 つの評価項目について行った．各項目は 5 段階評価で数字が大きいほど程度が高いとした．各項目について実際のインストラクションの原文を下に示す．

- Existence: Whether the appropriate objects exist in the output image
- Color: Whether object color is similar
- Position: Whether the object absolute position in the image is similar
- Related position: Whether the every relative position of objects is similar
- Naturalness: the objects in the image look realistic and natural

Existence は画像内の物体や背景中のラベルの種類と物体の数の一致度合，Color は画像内の全体的な色合い，Position は画像内の物体の絶対位置の近さ，Related position は画像内の物体同士の相対位置の近さ，Naturalness は出力画像の実画像への近さである．Naturalness は画像生成を行う場合と将来的に比較するために用意したもので，画像検索手法をとる今回の実験では基本的には 5 となる．ただし，手ブレや加工された画像もデータセットには存在するので，その場合は低い評価値をつけるよう実験参加者に依頼した．主観評価はこれらの 5 項目について「出力画像と目標画像の近さ」,「出力画像と入力説明文の近さ」の 2 種類について行った．また，実験参加者は以下の条件に該当する場合，タスク実行中に途中でタスクを打ち切ることができることとした．

- 出力画像が目標画像に十分近い画像であると実験参加者が判断した場合
- これ以上対話を続けても過去の履歴よりも目標画像に近い画像が得られないと実験参加者が判断した場合画像検索では複数画像を検索結果として出力することも可能だ

*1 https://github.com/ryankiros/visual-semantic-embedding

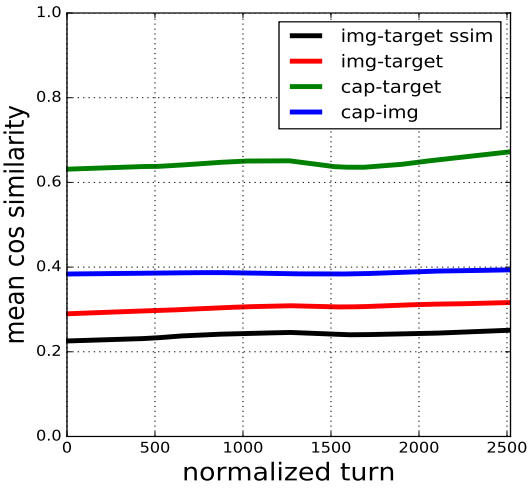

図 3: 対話中の正規化されたターンに対する客観評価の推移

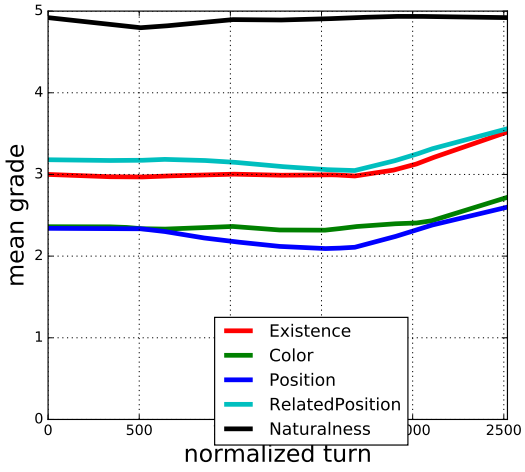

図 5: 対話中の正規化されたターンに対する主観評価 (目標画像) の推移

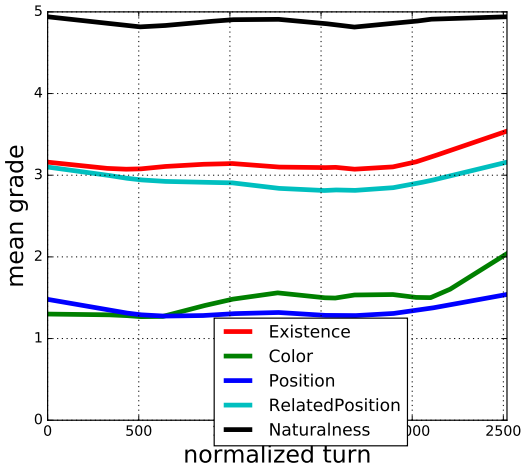

図 4: 対話中の正規化されたターンに対する主観評価 (説明文) の推移

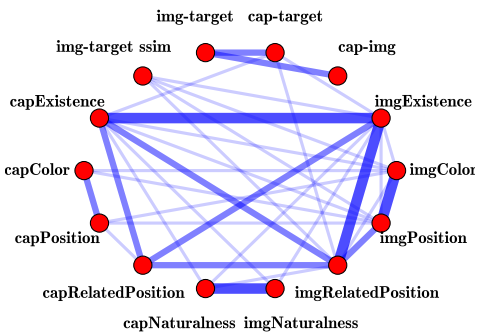

図 6: 各評価項目の相関関係

が，簡単化のため本稿では出力結果をランク 1 位の結果のみとした．

4. 今回得られた結果に対する考察

4.1 客観評価

本稿では得られた結果に対して 4 つの項目について客観評価を行った (図 3)．img-target ssim は [2] で用いられている SSIM[9] という画像評価尺度によって出力画像と目標画像との類似度を評価する．img-target は出力画像の特徴ベクトルと目標画像の特徴ベクトルのコサイン類似度である．cap-target は実験参加者の入力説明文の特徴ベクトルと目標画像の特徴ベクトルのコサイン類似度である．cap-img は実験参加者の入力説明文の特徴ベクトルと出力画像の特徴ベクトルである．各特徴ベクトルは [4] のニューラルネットワークによって得られたものであり，各特徴ベクトルの長さは 1 に近似されている．横軸はターン数である．ただし，終了ターン数が異なる各タスクについてターンの経過における評価の平均的な推移を評価するためターン数に正規化処理を行った．具体的には最大 10 ターンを正規化するのに 1 から 10 までの最小公倍数である 2520 を正規化されたターン数として，各タスクのターン数を線形補間を用いて 2520 ターンに拡張した．

対話開始時 (左端) と対話終了時 (右端) を比べると，対話的操作を通して各評価尺度について向上が確認できることから対話的操作が画像説明文を用いた画像検索に有効であることが示唆される．

4.2 主観評価

主観評価についても客観評価と同様に横軸を正規化されたターン数として対話中の平均的な推移の評価を行った．縦軸は平均の 5 段階評価値である．出力画像と実験参加者の入力した説明文を比較した主観評価を図 4，出力画像と目標画像を比較した主観評価を図 5 に示す．これらを比べると，図 4 の Color と Position は図 5 のものよりも対話中を通して低い値をとっている．この理由としては「実験参加者の入力した説明文をニューラルネットワークが出力画像に反映していない」「実験参加者が物体の色や絶対位置情報を説明文として入力していない」の 2 つの可能性が考えられる．また，対話中は対話開始時と終了時と比べて全体的に下がる傾向が見てとれることから，

実験参加者は色々な説明文を試して最終的に良い結果を選ぶ傾向にあると考えられる．

4.3 客観評価と主観評価の相関関係

客観評価 (図 3) と主観評価 (図 4，図 5) の関係を考察するため各ターンごとの各評価項目における相関係数を算出しグラフ化を行った (図 6)．相関係数 c を 4 段階の太さで表しており，(太:$\{0.7 < c\}$，中:$\{0.4 < c \leq 0.7\}$，細:$\{0.2 < c \leq 0.4\}$，無:$\{-0.2 < c \leq 0.2\}$) である．$\{c \leq -0.2\}$ は存在しなかった．主観評価と客観評価間には強い相関が確認できなかったことから，今回の実験では実験参加者の主観評価のどの項目が客観評価の向上と関係しているかを確認することはできなかった．客観評価項目間では Existence と RelatedPosition が相対的に強い相関関係にあった．これは，画像の説明文と出力画像が物体とその位置関係によって表現されている傾向にあることを示唆していると考えられる．

5. 結論

本稿では画像検索のフレームワークで入力説明文を対話的に修正することで最終的に客観評価の高い画像を検索できる傾向があることを示した．今後の課題として，本論文で得られた結果に基づき，対話的な画像生成への応用を検討していく予定である．

6. 謝辞

本研究の成果の一部は SCOPE の支援によるものである．

参考文献

[1] Datta, R., Joshi, D., Li, J., Wang, J. Z., Image retrieval: Ideas, influences, and trends of the new age, ACM Computing Surveys (CSUR), 40(2), 5, 2008

[2] Mansimov, Elman, Emilio Parisotto, Jimmy Lei Ba, Ruslan Salakhutdinov, Generating Images from Captions with Attention, ICLR, 2016.

[3] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schtze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[4] Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, TACL, 2015.

[5] Vinyals, Oriol and Toshev, Alexander and Bengio, Samy and Erhan, Dumitru, Show and Tell: A Neural Image Caption Generator, CVPR, 2015.

[6] Elliott, Desmond, and Arjen P. de Vries, Describing Images using Inferred Visual Dependency Representations, Annual Meeting of the Association for Computational Linguistics, 2015.

[7] Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y, Show, attend and tell: Neural image caption generation with visual attention, ICML, 2015.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C Lawrence Zitnick. Microsoft coco: Common objects in context. arXiv preprint, arXiv:1405.0312, 2014.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, Apr, 2004.