人の意図を反映したマルチモーダル情報の生成

理学専攻・情報科学コース
2140673
渡邊 清子

1 はじめに

近年，機械学習や深層学習により大量のデータからルールを学習し，人の脳を模倣したような人工知能に関する研究が盛んに行われている．しかし，人の脳を模倣するということを考えると，ルールには当てはめにくいような人の意図をくみ取ることにも挑戦していく必要がある．そこで本論文では，曖昧性を持つ人の意図に焦点を当てて二つの研究を行った．

2 特性を顕在化する言語の意味を反映した画像生成

例えば,「かわいい靴」といったテキストの場合，人の意図が含まれる曖昧性を持つので一意に画像を生成できない．そこで本研究では，形容詞の意味・特性が顕在化する方向と物体の形状変化の方向の対応関係を学習し，テキストにより物体の特性を強調する形状変化を伴なう画像生成を目的とする．

2.1 比較評価データセット UT Zappos 50K

UT Zappos50K[1]は，Zappos.com[2]から 50,025 個の靴画像を収集した大規模な靴画像のデータセットだ．Yu ら [1] は，物体形状の特性とそれを形容する言語の対応関係を捉えるために，Amazon Mechanical Turk[3]を使って，5 人の作業者に「2 つの靴画像に対してどちらの画像がより open, pointy, sporty, comfortable か」という比較評価を依頼しデータセットを作成した．

2.2 提案手法

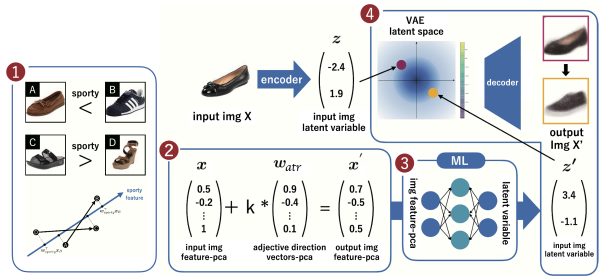

図 1: 提案手法の処理の流れ．

① Yu らの手法 [1] を用いて open, pointy, sporty, comfortable の方向ベクトルを推定する．
② 入力画像の画像特徴量 $\boldsymbol{x}$ に，形容詞の方向ベクトル $\boldsymbol{w}_{adj}$ を k 倍した値を加えて新しい画像特徴量 $\boldsymbol{x}'$ を求める．
③ 画像特徴量と VAE の潜在変数の関係を求める為に画像特徴量と VAE の中間層の潜在変数 $\boldsymbol{z}$ の対応関係を学習する多層パーセプトロンを構築する．
④新しい画像特徴量 $\boldsymbol{x}'$ を入力として多層パーセプトロンにより推定した VAE の新しい潜在変数 $\boldsymbol{z}'$ の値を VAE のデコーダで復号化することで，形容詞の意味を反映し特性を顕在化した画像を生成する．

[1]http://vision.cs.utexas.edu/projects/finegrained/utzap50k/
[2]https://www.zappos.com/
[3]https://www.mturk.com

2.3 実験結果・考察

Sandals カテゴリについて “open” と “sporty”2 つの方向ベクトルを靴画像に反映し，連続的に画像生成した結果を図 2 に示す．最終的な生成画像を見ると，甲が大きく開いた “open” な靴だが，靴のソール部分に厚みがある “sporty” の要素も組み込まれた動きやすそうな靴画像が生成された．また “open+sporty” は，緑マーカで示す “open” と黄色マーカで示す “sporty” の間をとったような方向に生成されていることから，“open” と “sporty” 両方の方向ベクトルを反映していると考えられる．

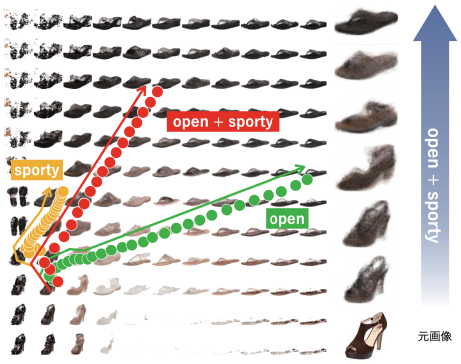

図 2: “sporty+open” なサンダルの連続的な画像生成．

実験結果より,「入力画像の画像特徴量 $\boldsymbol{x}$ + $k\times$ 形容詞の方向ベクトル $\boldsymbol{w}_{adj}$」から生成した出力画像は形容詞の特性を顕在化した画像になった．Shoes, Boots, Sandals の 3 つのカテゴリに対し 4 種類の形容詞の画像生成結果の例を図 5 に示す．

図 3: 3 カテゴリ $\times 4$ 形容詞の画像生成．

2.4 まとめ

本研究では，Yu らの形容詞の方向ベクトルを推定する手法 [1] と連続的な画像生成が可能な VAE [2] を組み合わせ，物体の形状を顕在化する形容詞を入力として画像生成を行なった．従来の text-to-image と異なる点としては，ラベル付けされた画像を学習して画像

を生成するのではなく，比較データから形容詞の意味をベクトルの形で表現し，形容詞の特性を顕在化出来る方向性を与えた点だ．そうすることで物体の形状変化の方向性との対応関係を学習し，生成画像に形容詞の意味を反映させることが出来た．その結果，“より○○な靴”という指示から画像生成出来るようになったり，意味を一意に定義することが難しくラベル情報のつけにくいテキストからの新しい画像生成のアプローチ方法を提案出来たと考える．

3　トレースから説明者の意図を反映した画像キャプション生成

近年，画像キャプション生成の研究は画像から得られる情報だけでなく，コントロールシグナルと呼ばれる追加情報を与えることにより，制御可能な画像キャプション生成へと発展している．人は一般的に画像の内容を説明する際，その説明対象を差しながら発話することに着目し，本研究では，画像をなぞることをコントロールシグナルとみなし，なぞった軌跡（‘トレース’と呼ぶ）から推定されるユーザの意図を反映するキャプション生成手法を提案する．

3.1　Localized Narratives

Pont-Tuset ら [3] は，トレースデータセット Localized Narratives(LN) を構築した．LN は，マウスで画像をなぞりながら画像の内容を音声で説明するという実験によって収集されたデータセットである．

3.2　提案手法

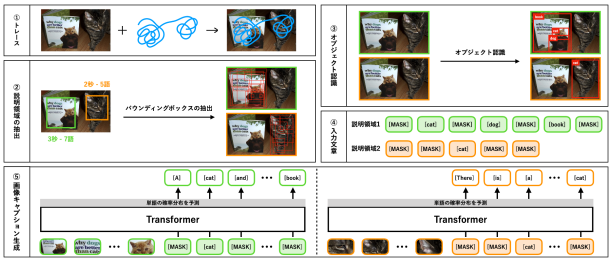

図 4: 提案手法の概要

① 画像中の説明したい箇所をトレースする．
② トレースの描画範囲から説明領域を抽出し，各領域のトレースの滞在時間から文長を推定する．また，各領域のバウンディングボックス (B.Box) を抽出する．
③ 各領域内でオブジェクト認識を行う．
④ ②で推定した文長と③で認識されたオブジェクトの単語から，画像キャプションを生成モデルに入力する文章を準備する．
⑤ ②によって抽出した各 B.Box の特徴量と④で準備した文章を入力とし，Deng ら [4] による文長を制御可能な画像キャプション生成モデル LaBERT を用いて，それぞれの領域の画像キャプションを生成する．

3.3　実験

画像とトレースを入力として，トレースの意図を反映した画像キャプションを生成する．結果の例を図 5 に示す．

説明領域 3 では，女性のバックを強くトレースしており，正解キャプションも “bag”について言及している．物体認識の結果，“bag”が挿入対象単語に設定され，Step1 での入力として “bag”が中央に追加された．その結果，生成キャプションに “bag”が含まれ，original-LaBERT よりもトレースの意図をよく捉えた結果になった．また，トレースの滞在時間からキャプションの単語数を予測した結果，説明領域 2,3,4 では，単語数の指定範囲に正解単語数が収まっており，正しく予測出来た．説明領域 1 は，正解キャプション 12words に比べ，25.62words と大幅に多く推定された．実際はトレースの滞在時間程，長い文による説明は求められていないことがわかる．しかし，トレースを見ると入念にトレースがされており詳しい説明を求めているような特性が見られる．生成されたキャプションは，suitcase というオブジェクトを捉えるだけではなく，スーツケースに貼ってあるステッカーを pictures と捉え詳細な説明ができている．説明者の意図には沿えていないが，トレースの意図を汲み取った生成結果となっている．

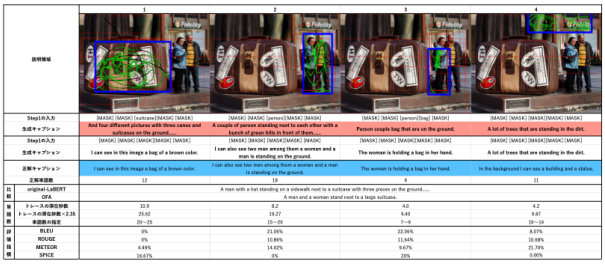

図 5: トレースを入力とした画像キャプション生成．

3.4　まとめ

本研究では，画像に対してトレースを用いながら説明するデータセット Localized Narratives と文長制御が可能な非自己回帰型テキスト生成を行う LaBERT のデコーダを組み合わせ，トレースからユーザの説明意図を汲み取り反映する画像キャプション生成手法を提案した．非自己回帰型のモデルを用いてトレースの滞在時間からユーザの興味度合いをトレースの滞在時間を文長に比例させ，トレースの順番を保持したまま情報を漏らさずに生成文に反映させた．説明対象となる領域の選択やキャプションの長さは LN の統計量から求めた値を採用したが，今後の課題として，ユーザによって個人差があるのでユーザ毎にパラメータを変更したい．また，トレースから説明範囲を B.Box として取り出すのではなく，線の情報として扱えるようにし，より自由度の高いユーザの意図に沿ったキャプション生成を行いたい．

参考文献

[1] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition (CVPR)*, Jun 2014.

[2] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[3] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020.

[4] Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. Length-controllable image captioning. *CoRR*, Vol. abs/2007.09580, , 2020.