特性を顕在化する言語の意味を反映した画像生成

渡邊　清子　(指導教員：小林　一郎)

1　はじめに

近年，機械学習や深層学習を用いたテキストからの画像生成に関する研究 (Text-to-Image) が盛んに行われている [1].「茶色と黒の縞模様の羽を持ち, 黄色い嘴の鳥」のように物体の形状を形容する表現に基づく画像生成を行うことは可能になっているが，その形状自体がどのような特性を持っているかという理解に基づいて画像を生成することは出来ていない．本研究では，特性を表現するための言語（特に形容詞を対象）の意味が顕在化する方向性と物体の形状変化の方向性の対応関係を学習し，言語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行うことを目的とする．具体的な試みとして，靴の画像を題材とし Shoes, Boots, Sandals という 3 つの靴カテゴリに対し，open, pointy, sporty, comfortable という 4 種類の形容詞と組み合わせて，“Sporty Boots”などといったテキストから形容詞と靴画像の特性を汲み取り，生成画像に反映させる．

2　物体の特性と形容する言語の対応関係

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，Yu ら [2] はデータセット UT Zappos 50K[1] を構築し，2 つの靴画像に対してどちらの方がより open, pointy, sporty, comfortable かという比較に基づき，特徴量空間におけるそれぞれの方向ベクトルを求めた．方向ベクトルの算出方法としては RankSVM [3] を使用している．具体的には，画像 i より画像 j の方が sporty である場合，画像 i，j の特徴量 $\boldsymbol{x_i}$，$\boldsymbol{x_j}$ に対して，sporty スコア $\boldsymbol{w}_{sporty}^T\boldsymbol{x_i} < \boldsymbol{w}_{sporty}^T\boldsymbol{x_j}$ の関係が常にが成り立つような sporty の方向ベクトル $\boldsymbol{w}_{sporty}$ を求める．比較に基づく形容詞の方向ベクトル 算出の概要図を 1 に示す．

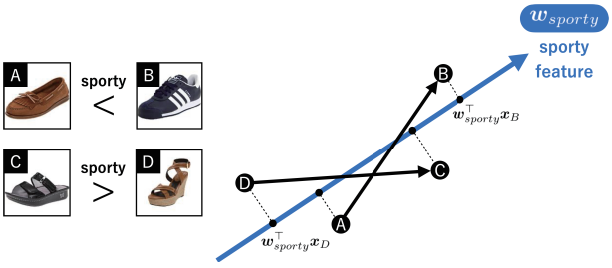

図 1: 形容詞の方向ベクトル算出.

3　Variational Auto-Encoder

本研究では，意味が反映され靴の形状変化をもたらす特定の形容詞を入力情報とする．連続的な形状の変化と形容詞の方向ベクトルとの関係を潜在空間上で捉えやすいことから，Variational Auto-Encoder（VAE）[4] を用いる．これにより形容詞の意味を反映させる程度に応じて生成画像が徐々に変化していく様子が観測可能となる．画像を入出力とする VAE では，入力画像 $\boldsymbol{X}$ の RGB 値を潜在変数 $\boldsymbol{z}$ に符号化し，その潜在変数 $\boldsymbol{z}$ を元に RGB 値へ復号化され，入力画像に近い出力画像 $\boldsymbol{X}'$ を生成するモデルである．VAE の特徴としては，潜在変数に正規分布を仮定することにより確率的なブレが生じ，連続的な画像生成を可能にすることだ．VAE の概要を図 2 に示す．

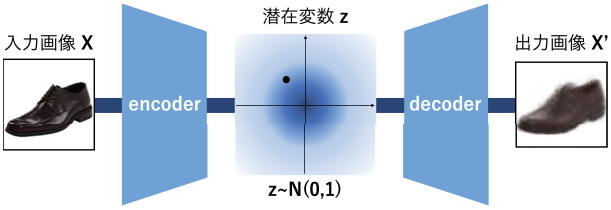

図 2: VAE 概要図.

4　提案手法

本研究では，上記 2 つの先行研究 [2][4] を元に，物体の形状を顕在化する形容詞を入力情報とする画像生成手法を提案する．提案手法の処理の流れを図 3 に示す．

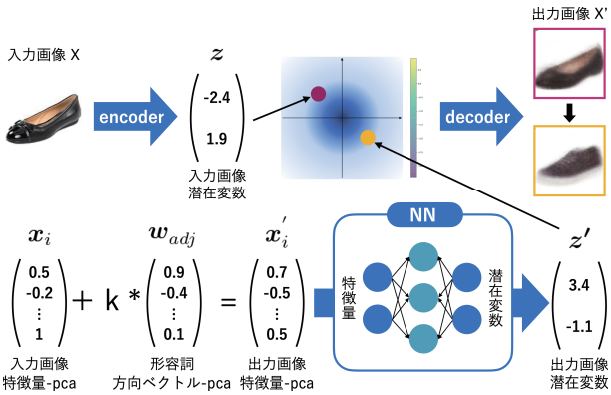

図 3: 提案手法の処理の流れ

Yu らの手法 [2] を用いて UT Zappos 50K の形容詞比較データから，4 種類の形容詞の方向ベクトルを推定する．ここで，使用するデータである UT Zappos 50K では，大域的画像特徴量の 1 つである GIST [5]（960 次元）と Lab 色空間（30 次元）を合わせた 990 次元の情報を画像特徴量としている．その値を主成分分析（PCA）することによって不要なノイズを取り除き，画像特徴量を 150 次元に圧縮した．これにより，靴画像に対して，4 つの情報：(i)RGB 値 (ii) 特徴量（990 次元）(iii) 特徴量（150 次元）(iv) 潜在変数が与えられ，形容詞に対して (ii) 方向ベクトル（990 次元）(iii) 方向ベクトル（150 次元）が得られる．次に，特徴量と潜在変数の関係を求める為に (iii) と (iv) の対応関係を学習するニューラルネットワークを構築する．150 次元に圧縮した画像特徴量 $\boldsymbol{x}$ に，同じく 150 次元に圧縮した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．その値を入力としてニューラルネットワークにより推定した VAE の潜在変数 $\boldsymbol{z'}$ の値を用いて，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/

5　実験

形容詞の方向ベクトル $\boldsymbol{w}_{adj}$ と入力画像の画像特徴量 $\boldsymbol{x}$ を用いて，その方向に沿ったスコア $\boldsymbol{w}_{adj}^T\boldsymbol{x}$（以下，形容詞スコア）を計算し, VAE 潜在空間の分布と形容詞スコアの対応関係を確認する．また，入力画像から形容詞の方向ベクトルに沿って連続的に画像を生成する．

5.1　実験設定

データセットは UT Zappos50K [2] を用い，50,025 枚の靴の画像と各々の画像特徴量，及び，4 種類の形容詞 open，pointy，sporty，comfortable に対する 2 画像間の比較結果（それぞれ 1,000 件程度ずつ）を使用する．比較データは人手によって付けられたものであることから個人差が含まれる．そこで，多くの人が同じ評価をしているデータのみを取り扱う為に, 評価スコアを計算し，上位のものを扱う．

5.2　実験結果

VAE 潜在空間と形容詞スコアの対応関係

靴画像の中でも, 例として Sandals カテゴリについて pointy スコアを色の違いで表現し，各サンプルが対応する VAE の潜在空間の位置にプロットする形で可視化した（図 4 参照）．黄色の部分はスコアが高く，紫色の部分はスコアが低い．生成画像と比べてみると，パンプスのような足先が尖っている靴に対してスコアが高くなっており，ビーチサンダルのようなものについてはスコアが低くなっている．以上より，VAE の潜在空間と Sandals カテゴリにおける pointy スコアに対応関係があることを確認できた．

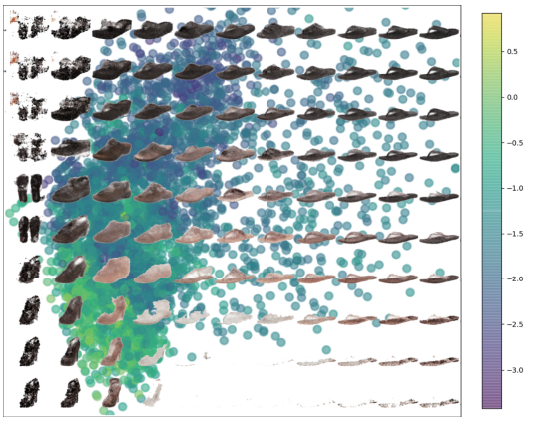

図 4: VAE 潜在空間と Sandals の pointy スコア.

形容詞の特性を顕在化した画像生成

Sandals カテゴリについて sporty の方向ベクトルを反映し, 連続的に画像生成した結果を図 5 に示す．また, 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 6 に示す．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．また，形容詞スコア $(\boldsymbol{w}_{adj}^T\boldsymbol{x}')$ はそれぞれ元画像より高くなっていることを確認した．

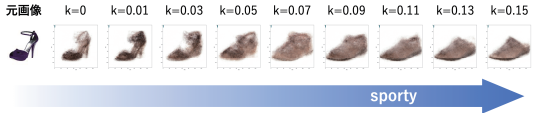

図 5: “sporty”な Sandals の連続的な画像生成.

図 6: 3 カテゴリ ×4 形容詞の画像生成.

5.3　考察

「入力画像の画像特徴量 $\boldsymbol{x}$ + $k\times$ 形容詞の方向ベクトル $\boldsymbol{w}_{adj}$」から画像を生成する際の k の値の範囲について実験結果から 2 つの方法を検討した．

(i) 図 4 を見ると，観測データが存在せず点がプロットされていない箇所については，靴とは言えない何かが生成されている．このことから，k の値は観測データがあるところまでを採用する．

(ii) 実験の結果，k の値がある程度を超えるところで k の変化量に対して潜在変数 z の変化量が減少していた．このことより，それ以上物体の特性を顕在化することが難しいという地点に達すると画像の変化が少なくなると考えられる．よって，$\delta\boldsymbol{z} < \epsilon$ のように $\delta\boldsymbol{z}$ に対して閾値を決め，ϵ を下回らないところまでを採用する．

6　おわりに

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 [2] と連続的な画像生成が可能な VAE を組み合わせ，物体の特性を顕在化する言語の意味を反映した画像生成手法の提案を行い，実験を通じて提案手法の有効性を検証した．

参考文献

[1] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *CoRR*, Vol. abs/1612.03242, , 2016.

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

[3] Kai Chen, Rongchun Li, Yong Dou, Zhengfa Liang, and Qi Lv. Ranking support vector machine with kernel approximation. In *Computational Intelligence and Neuroscience*, Feb 2017.

[4] 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.

[5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, Vol. 42, No. 3, pp. 145–175, 2001.