テキストからの画像生成における意味的な整合性の検討

柳本研究室　　浅野 竣弥

1. はじめに

敵対的生成ネットワーク(Generated Adversarial Networks, GAN)を用いた画像生成の研究が盛んに行われており，高精細な画像が生成されている[1]．また，生成される画像を制御するため，自然言語で書かれた特徴を考慮した画像生成も行われている．多くの画像生成の研究では主観的な評価がなされており，評価に曖昧性があると考えられる．そこで，画像を制約するテキストと画像間の整合性を評価する必要があると考えられる．

本研究では，テキストと生成された画像間に意味的な整合性を持つかどうか検討する．整合性を検討する為，具体性のあるテキスト，例えば個数について言及するものや，空間的位置関係などが記述されたテキストに沿った画像を訓練データとして用いる．実験結果としては，現状の GAN を用いたテキストからの画像生成では，意味的な整合性をもった画像は生成できているとは確認できなかった．

2. 実験内容

2.1. 評価実験のためのデータセット

GAN を用いて画像を生成する際，従来のデータセットでは鳥や花などを主とする自然画像が扱われてきた．しかしながら，これらには背景部などのテキストで述べられていない特徴が多く含まれているため，主観的な評価となり，曖昧性が残る．本研究では，Suhr らが公開している視覚的推論のための自然言語コーパス(Natural Language for Visual Reasoning, NLVR)[2]をデータセットとして利用する．NLVR は自然画像とは違い合成的に作られた画像であり，背景も単色であるため，画像生成に必要な情報がテキストに十分含まれている．例を図 1 に示す．

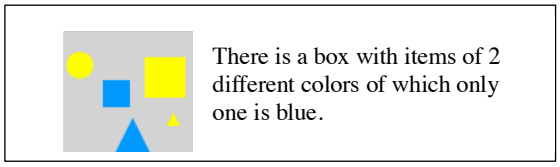

図 1 NLVR の画像とキャプション

2.2. データセットの前処理

学習の前にデータセットからテキストと対応する画像のみを抽出し 100*100 にリサイズした．訓練用データ 115 枚，テスト用データ 32 枚に分け学習する．テキストに対しては Bi-directional LSTM を用いて，最終の隠れ状態をつなぎ合わせた文ベクトル$\bar{e}\in\mathbb{R}^{\hat{D}}$を求める．$\hat{D}$は特徴ベクトルの次元数を表す． 画像に対しては CNN を用いて，画像全体の特徴ベクトル$\bar{f}\in\mathbb{R}^{2048}$ を作成する．次に画像の特徴をテキストの特徴と同じ次元の空間に変換する．

$$\bar{v} = W\bar{f} \tag{1}$$

ここでWは適当な重みを表し，$\bar{v}\in\mathbb{R}^{\hat{D}}$である．

画像I_iがテキストT_iとマッチする事後確率は

$$P(I_i|T_i) = \frac{\exp\left(\gamma\cos(\bar{v}_i,\bar{e}_i)\right)}{\Sigma_j^{115}\exp\left(\gamma\cos(\bar{v}_j,\bar{e}_i)\right)} \tag{2}$$

ここでγは平滑化係数，$\cos(\bar{v}_i,\bar{e}_i)$は画像$I_i$の特徴とテキスト$T_i$の特徴のコサイン類似度を意味する．エンコーダーの損失関数$\mathcal{L}_{encoder}$を下記のように定義する．

$$\mathcal{L}_{encoder} = \Sigma_{i=1}^{115}\log P(I_i|T_i) \tag{3}$$

2.3. 生成モデルの構成

Generator を G, Discriminator を D として，画像生成のための損失関数$\mathcal{L}_G$，$\mathcal{L}_D$を下記のように定義する．

$$\mathcal{L}_G = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}\sim p_G}[\log D(\hat{x})]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}\sim p_G}[\log D(\hat{x},\bar{e})]}_{\text{conditional loss}} \tag{4}$$

$$\mathcal{L}_D = \underbrace{-\frac{1}{2}\mathbb{E}_{x\sim p_{data}}[\log D(x)] - \frac{1}{2}\mathbb{E}_{\hat{x}\sim p_G}[\log(1-D(\hat{x}))]}_{\text{unconditional loss}} + \underbrace{-\frac{1}{2}\mathbb{E}_{x\sim p_{data}}[\log D(x,\bar{e})] - \frac{1}{2}\mathbb{E}_{\hat{x}\sim p_G}[\log(1-D(\hat{x},\bar{e}))]}_{\text{conditional loss}} \tag{5}$$

ここでx は訓練データの画像，$\hat{x}$は Generator が生成した画像を表す．また unconditional loss は画像が本物か偽物かを決定し，conditional loss 画像と文が一致するかどうかを決定する．最終的な GAN の目的関数は次のようになる．

$$\mathcal{L} = \mathcal{L}_G + \mathcal{L}_D + \lambda\mathcal{L}_{encoder} \tag{6}$$

3. 実験結果

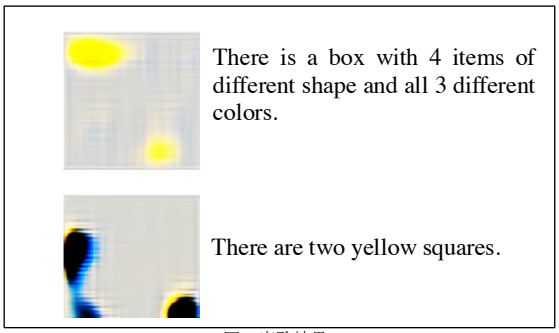

図 2 実験結果

γ = 10としてエンコーダーの学習を行った．生成モデルの学習は 58 epoch 行った．実験結果を図 2 に示す．学習回数について検討の余地はあるが，色と数だけに注目するとどちらも満足しているとは言えない．

4. おわりに

本研究では，GAN を用いたテキストからの画像生成における意味的な整合性の検討を行った．実験より，現状として GAN を用いたテキストからの画像生成では，高解像な画像が生成されたとしても，意味的な整合性を持った画像の生成はできていない．訓練データも少ないため，今後はデータを増やして議論を進めたい．

参考文献

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative Adversarial Text to Image Synthesis, *ICML'16: Proceedings of the 33rd International Conference on Machine Learning,* vol. 48, pp. 1060-1069, 2016.

[2] A. Suhr, M. Lewis, J. Yeh, Y. Artzi. A Corpus of Natural Language for Visual Reasoning, *ACL Anthology*, pp. 217-223, 2017.