# Show and Read, then Tell

川口維文　内田誠一
(九州大学)

## 1　はじめに

本報告は，情景内文字情報を積極的に利用するという新しい着眼点に基づいて，画像からの説明文生成の高精度化を目的としている．画像を説明する上で情景内文字は，文字列自体が持つ意味情報に加えて，その場所を形容する情報となるため重要である．情景内文字情報の導入法として，本報告では最も単純な方式を試みる．すなわち，対象とする画像（原画像）に加え，その中の文字情報を端的に表す画像（付加画像）を別途準備し，画像特徴レベルで両者を統合した上で，説明文生成を行う．ここで画像特徴には様々な物体らしさが反映されていると考えれば，以上の方式により情景内文字が表現する物体らしさが強調された画像特徴が得られ，結果的に説明文には情景内文字情報が反映されると期待できる．

## 2　情景内文字情報を導入した画像説明文生成

### 2.1　ベースとした画像説明文生成手法

画像説明文生成手法の多くは，CNN(Convolutional Neural Network) により出力された特徴ベクトルを用いて，RNN(Recurrent Neural Network) で文章生成を行う [1][2]．本報告では，この枠組みに情景内文字情報を導入して，生成される説明文の高精度化を図る．具体的には，VGG の 16-layer model による 768 次元特徴ベクトルを用いる．その上で，MS COCO の Annotation 付きのデータセットによって，予め RNN を含めた全体の学習を行っておく．

### 2.2　画像特徴レベルでの情景内文字情報統合

本報告では，以下の 2 つの方法で原画像特徴と付加画像特徴の統合を試みる．

第一の方法は，原画像特徴 $f_{\mathrm{img}}$ と付加画像特徴 $f_{\mathrm{txt}}$ の和を用いる．式 (1) に示すように，付加画像特徴 $f_{\mathrm{txt}}$ に重み $\alpha$ を用いることで特徴合成具合を制御する．この方法では，$\alpha = 0, 1$ を除けばどちらの特徴も完全に失われることが無いため，特徴の統合による原画像の説明文情報が失われにくい．

$$f_{\mathrm{total}} = (1-\alpha) f_{\mathrm{img}} + \alpha f_{\mathrm{txt}} \tag{1}$$

第二の方法は，式 (2) に示すように 2 画像特徴の大きい値を用いる．768 次元すべてを比較し，大きい値を選択することで，それぞれの特徴的な部分を反映できる．

$$f_{\mathrm{total}} = \max\{f_{\mathrm{img}}, \alpha f_{\mathrm{txt}}\} \tag{2}$$

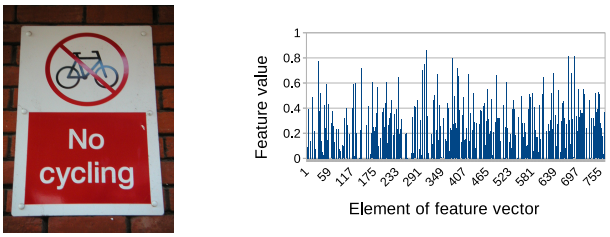
## 3　実験

2 つの方法に基づき画像特徴の統合を行い，得られた画像特徴とそれにより生成される画像説明文を比較した．図 1 は原画像 (a) と原画像特徴 $f_{\mathrm{img}}$(b) である．その原画像特徴 $f_{\mathrm{img}}$ に，図 2 に示すような付加画像 (a) の付加画像特徴 $f_{\mathrm{txt}}$(b) を統合する．本報告の画像特徴レベルでの統合方法は，画像特徴を統合する際に特徴ベクトルが変化し別の物体として認識されてしまうため，付加画像の物体そのものを付加することは難しい．しかし，原画像の説明文に対して，付加画像の性質を統合することは可能である．

表 1 は，それぞれの方法で得られる説明文の出力結果を示している．式 (1) の方法では，$\alpha$ を大きくしていくと，ある地点で説明文が急激に変化する．つまり，この方法ではある $\alpha$ で特徴空間を大きく移動してしまうため，2 画像の特徴が考慮された説明文は生成されない．式 (2) の方法では，$\alpha$ の変動に従って説明文が徐々に変化している．また，$\alpha$=0.8 では自転車の情報が統合されている．つまり，式 (2) の方法は，式 (1) の方法と比較すると特徴空間を急激に移動することはなく，適切な画像説明文が生成され易い．

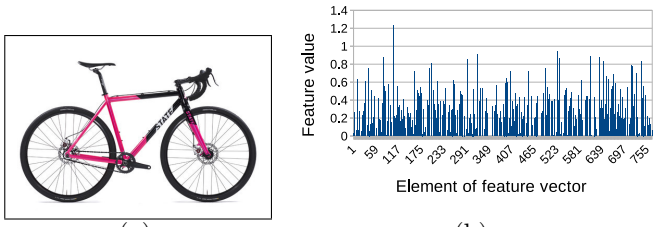
表 1: 画像説明文出力結果

| $\alpha$ | $(1-\alpha) f_{\mathrm{img}} + \alpha f_{\mathrm{txt}}$ | $\max\{f_{\mathrm{img}}, \alpha f_{\mathrm{txt}}\}$ |
|---|---|---|
| 0 | a sign that says UNK UNK on the side of it | a sign that says UNK UNK on the side of it |
| ⋮ | ⋮ | ⋮ |
| 0.7 | a bike parked on the sige of a street | a red and white sign that says UNK |
| 0.8 | a bike parked in front of a red wall | a red and white sign with a bicycle on it |
| 0.9 | a bike parked in front of a red wall | a red bicycle parked in front of a sign |
| 1.0 | a bike parked in front of a window | a bike parked on the side of a street |




(a)　(b)

図 1: 原画像 (a) と原画像特徴 $f_{\mathrm{img}}$(b)




(a)　(b)

図 2: 付加画像 (a) と付加画像特徴 $f_{\mathrm{txt}}$(b)

## 4　まとめ

本報告では，画像特徴同士の統合により高精度な画像説明文の生成を試みた．しかし，本報告の 2 つの方法では付加画像の物体の性質を説明文に統合することはできるが，物体自体を適切に統合できるケースは少ない．そのため，CNN の段階で認識を行えていない場合には，認識したとみなし再学習することで認識が可能になり，適切な文章生成が行えるのではないかと考えている．

## 参考文献

[1] Andrej Karpathy, Li Fei-Fei, ”Deep Visual-Semantic Alignments for Generating Image Descriptions”,CVPR 2015

[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, ”Show and Tell: A Neural Image Caption Generator”, arXiv:1411.4555(cs.CV)