画像キャプション生成における複数形表現の統一

西　友佑　新納 浩幸　古宮 嘉那子　佐々木 稔
茨城大学 工学部 情報工学科
13t4076f@vc.ibaraki.ac.jp, hiroyuki.shinnou.0828@vc.ibaraki.ac.jp,
kanako.komiya.nlp@vc.ibaraki.ac.jp, minoru.sasaki.01@vc.ibaraki.ac.jp

1　はじめに

本論文では，画像キャプション生成における生成文の品質向上を目的とし，その実現のために，訓練データ中の複数形表現を統一することを試みる．具体的には， two dogs や three cars と言った基数を用いた複数形表現を，基数を用いないより単純な表現への統一を行う．これにより，生成文における基数を使用していた部分の誤りが訂正され，品質の向上が期待できる．

画像キャプション生成についての研究は従来より活発に行われてきた [1]．これは，入力された画像のキャプション（説明文）を生成するという研究である．画像キャプション生成は難しい課題であり，多くの問題を抱えている．問題の一つとして，生成したキャプションの定量的な評価が難しい点が挙げられる．定量的に評価する方法として BLEU 評価値 [4] を用いる手法もあるが，この評価方法は単語同士の近さのみを基準とするため，明らかに問題がある．本質的に，画像に対する正しいキャプションは無数にあり，それらの優劣は主観的にしか判断できないことが定量的な評価を困難にしている原因である．正しいキャプションが無数に存在するという一つの理由として，キャプションの詳しさが限定されない点が挙げられる．一つの画像に対して非常に詳しく説明を行っているキャプションでも，簡潔に説明を行ったキャプションでも正しいキャプションとなり得る．本論文では，この点に着目し，生成されるキャプションをより単純化することによって誤りを減らし，生成文の品質を向上させる．具体的な手法としては，基数などを用いた複数形表現をより単純な複数形表現へ統一することによって，生成されるキャプションの誤りを減らすことを目指す．

先に述べたような誤りの解消は，生成された後のキャプションを直接修正することでも解決可能に見える．しかし，深層学習を用いた画像キャプション生成では，深層学習の内部を詳しく検証することが難しいためにどの部分で誤りが発生しているかを特定することが困難である．また，時系列的な影響を受ける文章において，生成された後の一部分を変更しただけでは，文中の変更した箇所以降での部分において更なる誤りが発生する可能性がある．そのため，本研究では学習に使用する訓練データにおける複数形表現を予め書き換えておき，書き換えたデータを用いて学習を行う．

実験では，訓練データに MSCOCO のデータセットを用い，テストデータにはネットなどから収集した 100 枚の画像データを使用した．そして，書き換えられた訓練データで学習したシステムによるキャプションと，書き換え以前の訓練データで学習したシステムによるキャプションとを比較し，主観的に評価を行った．結果，キャプションの大きく変化した画像数は 100 枚中 49 枚となり，そのうち改悪された画像数が 14 枚であるのに対し，25 枚の画像においてキャプションの改善が認められた．

2　画像キャプション生成

本実験で行った画像キャプション生成には論文 [5] で述べられている深層学習アルゴリズムを用いた．このアルゴリズムは，入力画像から畳み込みニューラルネット（CNN）によって特徴量を取り出し，その特徴量を入力として，再帰的ニューラルネット（RNN）が文章を生成するというものである．

ニューラルネットワークの学習では，CNN の部分には以下の URL で公開されている学習済みのモデルを用いる．

```
https://gist.github.com/\
  ksimonyan/3785162f95cd2d5fee77
```

また，RNN の部分の学習には本実験で用意した訓練データを用いる．

3 複数形表現の統一

複数形における問題点とは，複数の対象が写っている画像（犬や人間などが複数写り込んでいるもの）に対して生成された文章において，その物体が単数であったり，複数形になっていても数が間違っている[1]ことである．これらの多くは基数を含む複数形表現に現れる．

これを改善するために，学習に用いられるデータセットにおける複数形の表現をより単純な形に書き換え，複数形の表現方法を統一する．

本実験では，MSCOCO の Annotation 付き画像データセットを用いた．ただし，MSCOCO で直接配布されているデータセットではなく論文 [2] で使用され，以下の URL で公開されているデータセットを用いた．

`http://cs.stanford.edu/people/\`
`karpathy/deepimagesent`

複数形表現を統一した場合の品質の変化を調べるためにデフォルトのデータセットと合わせて 5 パターンの訓練データを用意した．内訳は以下のとおりである．

1. デフォルトのデータセット
2. 1 について，頭文字の表記ゆれを取り除くために全ての文章を小文字にしたもの
3. 2 について，複数形の文章について，基数の部分を some に置換したもの
4. 3 について，some となっている部分を a group of にしたもの
5. 4 について，a couple of となっている表現を a group of に統一したもの

これら 2 から 5 のパターンのデータセットを用いて学習を行ったモデルによって生成されたキャプションと，デフォルトのデータセットを用いて学習を行ったモデルによって生成されたキャプションとを比べることによって品質が向上したかを調べる．

4 実験

4.1 実験設定

今回の実験では全ての訓練データパターンに対して RNN の学習の回数を 100 回とし，100 回学習したモデルから生成されたキャプションを結果として評価を行った．学習したモデルからキャプションを生成させる際に使用する画像は，学習時に使用されたことのない画像を新しく 100 枚用意した．50 枚の画像には複数の対象が写っている画像が含まれており，もう 50 枚の画像には，複数形以外の表現にも変化が起こるのかを調べるために，対象が 1 つであったり風景のみの画像などを用意した．

実験に用いた訓練データの調整は以下のような手順で行った．

1. 訓練データの全キャプションを小文字に変換
2. TreeTagger を用いての品詞の確認・基数の置換
3. 単語 some の置換
4. 慣用句 a couple of の置換

4.2 MSCOCO データセット

MSCOCO データセットとは画像に対してキャプションが付与されているデータセットであり，Microsoft によって提供されている．今回使用したデータセットは MSCOCO のデータセットを元に変更が加えられたものを使用している．データセットの内容は，画像枚数が 123,287 枚，それらに付与された総キャプション数が 616,767 文となっている．デフォルトデータセットのうち，複数形が含まれている文章は 284,549 文あり，さらにそのうちの 65,643 文が基数を含むキャプションであった．この基数を含むキャプションの基数部分を some に置き換えたものがパターン 3 のデータセットとなる．ただし，複数形が含まれており基数も存在するが，その基数が one のみである場合は some に置き換えることができないため，省いている．そのため，実際に置換を行ったキャプション数は 63,794 文であった．

次に，パターン 4 を作成するためにキャプション中に some を含むキャプションに対して置換を行った．some を a group of に置換したキャプション数は 86,173 文であった．これはデフォルトデータセットの時点で含まれていた some を含むキャプションとパターン 3 で置換されて生成された some を含むキャプションになっている．

また，デフォルトデータセットのうち a couple of が使用されていたキャプションが 9,433 文，a group of が使用されていたキャプションが 19,359 文あった．パターン 5 を作成するためにこの 9,433 文に含まれている a couple of を全て a group of に置換した．

[1] 対象が 3 つなのに生成文では 2 つとされるなど．

以上から，パターン 5 における a group of を含むキャプション数は，113,072 文となった．置換したキャプションの総和よりも最終的な a group of を含むキャプション数が少ないのは，パターン 2 からパターン 5 まで順に変換していく間に一つのキャプションに変換すべき語がいくつかあるキャプション[2]があったため，重複してカウントされたものと考えられる．

4.3 実験結果

実験の結果として，キャプション生成に用いた画像と，それぞれの 3 章で説明した各パターンのモデルにおける生成文を 3 文ずつ以下に示す．

図 1: 生成文がより良くなった場合

図 1 において，それぞれのパターンでの生成結果を以下に示す．

パターン 1 を用いた場合

a bird that is standing in the water
a flock of birds standing on top of a beach
a couple of birds standing on top of a beach

パターン 2 を用いた場合

a bird that is standing in the water
a bird flying over a body of water
a bird standing on top of a sandy beach

パターン 3 を用いた場合

a bird flying over a body of water
a bird that is standing in the water
a couple of birds that are standing in the water

パターン 4 を用いた場合

a bird flying over a body of water
a bird that is standing in the water
a couple of birds that are standing in the water

[2]some が使用されているかつ a couple of が使用されているキャプションなど．

パターン 5 を用いた場合

a group of ducks swimming in a body of water
a group of ducks are swimming in the water
a group of birds that are standing in the water

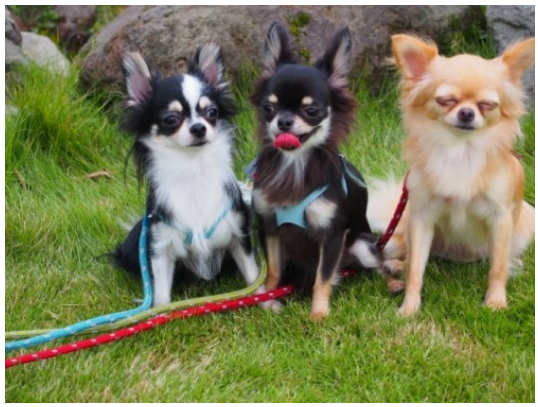

図 2: 生成文が悪くなった例

図 2 において，それぞれのパターンでの生成結果を以下に示す．

パターン 1 を用いた場合

a brown and white dog laying on a bed
a dog laying on top of a bed
a black and white dog laying on a bed

パターン 2 を用いた場合

a close up of a dog wearing a hat
a close up of a dog laying on a couch
a close up of a dog wearing a santa hat

パターン 3 を用いた場合

a dog that is sitting in the grass
a dog that is laying down in the grass
a dog that is sitting on a couch

パターン 4 を用いた場合

a dog that is sitting in the grass
a dog that is laying down in the grass
a dog that is sitting on a couch

パターン 5 を用いた場合

a close up of a cat laying on a bed
a close up of a cat laying on a couch
a close up of a dog laying on a couch

100 枚の画像に対して，デフォルトデータセットでの生成文と各訓練データのパターンでの生成文を比較した結果を表 1 に示す．

この結果より，置換したキャプション数が多くなっていくに連れて変化したキャプション数が増加してい

表 1: デフォルトデータセットとの比較

	良くなった生成文数	悪くなった生成文数
パターン 2	13	11
パターン 3	10	12
パターン 4	15	13
パターン 5	25	14

ることがわかる．また，パターン 4 までは良くなったキャプション数と悪くなったキャプション数にほとんど差がないが，パターン 5 では良くなったキャプションが多くなっている．パターン 3 では良くなったキャプション数よりも悪くなったキャプション数のほうが多くなっているが，生成されたキャプションを見ると，置換して増加したはずの単語 some がほとんど使用されていなかった [3]．

パターン 4 とパターン 5 では，置換した a group of や a couple of の表現が多く出現していた．その影響でキャプションが良くなった場合が多く見られたが，副作用として主語となる対象の認識が誤っているものも見られた [4]．

5 考察

画像キャプション生成でおかしなキャプションが生成された場合，その原因がどこにあるのかを特定するのは容易ではない．これは本実験でも確認できている．つまり本実験では訓練データの書き換えを行うことで，書き換えた部分はもちろん，書き換えられていない部分に置いても誤りが改善された．これは，深層学習における文章生成の部分において，前の語を踏まえた上で次の語の確率を求めるという仕組みに深く関係があると考えられる．例えば，深層学習での文章生成では，訓練データにおいて two という単語の次に来る単語が dogs と cars しかなかった場合には，two のあとには dogs か cars が続くと学習してしまうため，二匹の羊が写っている画像を入力として与えた場合に，two という個数まで合っていたとしてもその次に sheep が出現しづらくなる．今回の実験では two に当たる部分を統一したことによって，次に来る語の選択肢が大幅に増えたことで，書き換えを行った部分以外でも 誤りが発生しにくくなったと考えられる．

また本研究はキャプションの言語が英語であることを前提としている．日本語の場合，単数や複数の表現は曖昧であるため，本研究で取ったアプローチは全く使えない．本研究ではキャプションの内容の粒度を少し粗くすることで，キャプションの品質を向上させることを狙っている．日本語の画像キャプション生成 [3] では，どのような形で内容の粒度を少し粗くできるのかを考案する必要がある．

[3] 500 文中 1,2 文程度．
[4] 数匹の猫が写っている画像を羊と表現するなど．

6 おわりに

本論文では，画像キャプション生成の複数形表現に注目することで，生成されるキャプションの品質の向上を行った．具体的には，訓練データのキャプションの複数形表現を統一した．MSCOCO データセットを訓練データに用いた実験では，生成されるキャプションの品質の向上が確認できた．

画像キャプション生成がおかしなキャプションを生成した場合，その原因がどこにあるのかを特定するのは容易ではない．本実験の結果はそれを示唆している．今後はこの点を考察したい．また本研究での訓練データのキャプションの内容の粒度を粗くするというアイデアを日本語の画像キャプション生成でも試したい．

参考文献

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*, pp. 15–29. Springer, 2010.

[2] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

[3] Takashi Miyazaki and Nobuyuki Shimizu. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1780–1790, 2016.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.