ニューラルキャプション生成モデルによる画像説明文の選択

Selecting Image Descriptions with Neural Image Caption Generation Models

高里 盛良　　三輪 誠　　佐々木 裕
Seira Takasato　Makoto Miwa　Yutaka Sasaki
豊田工業大学
Toyota Technological Institute

{sd10035, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1　はじめに

現在，深層学習の発展により，画像中のオブジェクトの認識だけではなく，画像において何がどのような状態で何をしているかなどを自然言語文で出力するキャプション生成が可能になりつつある[1]．このキャプション生成の精度は正解キャプションに対する BLEU 値等で評価されているが，応用システムでの有効性の検証は行われていない．

本研究ではこのようなモデルを利用して，図 1 のような TOEIC 等で出題される，1 つの画像に対して与えられた複数の説明文のうち最も合致するものを選択する問題を解くシステムを作成する．これにより，現在のキャプション生成モデルの汎用性を調査することを目的とする．

(A) She is using a computer.
(B) She is reading a book.
(C) She is holding a pen.
(D) She is wearing a T-shirt.

図 1．画像と説明文の例

2　関連研究

Recurrent Neural Network (RNN) は Neural Network (NN) の中間層のユニットを再帰的に接続した，過去の入力を「記憶」できるモデルである．音声や文章などの時系列データをはじめとして様々な構造のデータからの学習に利用できる．Long Short-Term Memory (LSTM) [1] は RNN を構成するユニットの一種であり，単純なユニットを利用した RNN よりも長い範囲の記憶が可能となっている．これを用い，画像からキャプションの対応関係を学習することで，画像からキャプションを生成するモデルを得ることができる[2]．

Xu らの提案した Show, Attend and Tell[2] では，Convolutional Neural Network (CNN) によって画像から取り出した特徴を LSTM ベースの RNN (LSTM-RNN) に入力し，さらにその画像についてのキャプションを入力し，その文が生成される確率が高くなるように学習する．このモデルを画像からの文生成に適用し，Flickr8k, Flickr30k, MS COCO のデータセットにおいて，BLEU, METEOR のそれぞれのスコアで論文発表時点での最高精度を得ている．

LSTM-RNN での文の生成は次のように行われる．最初に画像から抽出された特徴が最初の LSTM ユニットに入力される．それと同時に文の始まりを表す語を LSTM ユニットに入力すると，辞書内の単語がそれぞれ最初に出現する確率が算出される．この辞書は学習時に出現した単語と 1 つの未知語を表す語から成る．その中で最も出現確率の大きいものを最初の単語とする．その単語を LSTM ユニットへ入力すると，さらにその次に各単語が出現する確率が求まる．再度最も出現確率が高いものを 2 番目の単語とし，次の LSTM ユニットへの入力とする．これを繰り返し，文の終わりを表す語が出力さ

れるまで繰り返すことで文を生成することができる．学習時に未知語は出現しないが，Xu ら[2]は出現確率の低い単語は未知語に置き換えて学習している．

3　提案手法

本論文では 3 つの手法を提案する．それぞれの手法では，設問文の生成確率を求めることで判定を行っている．設問文の生成確率は LSTM-RNN から文を生成する際に求まる．2 つ目と 3 つ目の手法は 1 つ目の手法をベースとしたものになっている．2 つ目の手法では，学習時に出現しなかった単語が設問文に出現した場合の対応を行ったもの，3 つ目の手法は設問の文を言い換えて判定を行ったものになっている．

3.1　各選択肢の生成確率

各提案手法では，LSTM-RNN が算出する単語の出現確率から，特定の単語の出現確率を求めていき，設問文がどれほどの確率で生成されるのかを見積もる．画像 I に対する選択肢の 1 つに，文 S：They are leaving the room. とあった場合の動作を図 2 に示す．図 2 のように，選択肢の文の単語を順番に LSTM ユニットに入力し，その際に計算される各単語の出現確率の対数の和をその文の生成確率と考え，その値を単語数 N で平均をとった値の正負を反転したものをその文の生成コスト(以下コスト)と考え，その値が小さいものを選択する．画像 I から S_1 から S_N の N 個の単語を持つ文 S を生成する確率 $\log p(S|I)$ とコストは次のように計算する．

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

$$\text{コスト} = -\log p(S|I)\,/N$$

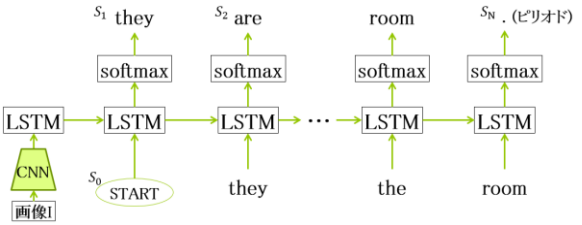

図 2．LSTM-RNN への画像の入力と文生成

3.2　基本手法

1 つ目の手法は次のように行う．3.1 節で紹介したように，画像に対して与えられた設問文それぞれのコストを求め，最も小さいものを選択する．未知語が含まれていた場合は辞書内の未知語を表す語を当てはめ，その語の出現確率を使用する．これを基本手法と呼ぶこととする．

3.3　未知語の言い換え

今回使用するモデルによって生成される文は学習中に出現した単語から構成される．学習中に出現しなかった単語が設問文に出現した場合は，3.2 節で述べたように，辞書内の未知語を表す語を当てはめる．これは学習時には出現しないため，出現確率が 0 である単語と同義であり，未知語を含む文にこれを当てはめた時のコストは高くなってしまう．これを解決するために学習の際に 1%の確率で文中の単語のいずれか 1 つを未知語に変換して学習することで，未知語を含む文でも生成確率を求められるようにする．

3.4　設問文の言い換え

同じ意味の文でも，形式によって生成確率は異なる．学習時に多く出現した表現で書かれた文は生成確率が大きくなり，逆にあまり出現しなかった表現で書かれた文の生成確率は小さくなる．この問題を解決するために，データセット中と設問で主に使用されている文の形式が異なる場合に，コストを計算する際に設問文をデータセット中に多い表現に言い換えることで，文の形式によるコストの変化を抑える．

4　実験結果

4．1　実験設定

学習にはFlickr8k，Flickr30kとMS COCOのデータセットを合わせたものを使用した．画像の特徴はCNNの実装の1つであるOxford VGGnet[3]を用い，8×512のサイズに抽出したものを用いた．CPUはCore i7-5960X，GPUはNVIDIA TITAN Xを使用し，22時間学習を行った．出現頻度が低い単語も未知語に変換せずに学習を行った．問題には，TOEIC第1問の画像と選択肢のセットを610問[4]用いた．入力された画像から，どの選択肢が正解かを予測した．

次節では3節で紹介した3つの手法の結果を比較する．3.2節で紹介した基本手法とあわせて，3.3節で紹介した学習時に単語の一部を未知語に変換する手法を未知語学習有の手法，3.4節で紹介した判定時に設問文をデータセット中に多い表現に言い換える手法を設問文言い換え有の手法と呼ぶこととする．

設問文言い換え有の手法では，Enju[5]によってデータセットと設問に含まれる文を品詞の並びで分類した結果から設問文を言い換えた．表1は分類した結果の上位5つのパターンを示している．結果より，データセットには名詞句のみの表現が多いことが分かった．また，設問中には名詞句のみの形の文は含まれていなかった．言い換えの規則は表2のように設定した．

4．2　実験結果

各手法の正解数は表3のようになった．

表1．データセットと設問に多い品詞の並びと例

データセット(771,837文中)		
文の数	文の割合[%]	品詞の並びと例
2417	0.31	DT NN VBG IN DT NN IN DT NN a man riding on the back of a motorcycle
2003	0.26	DT NN VBG DT NN IN DT NN a man riding an elephant in a river
1576	0.20	DT NN VBZ VBG IN DT NN IN DT NN this cat is sitting on a porch near a tire
1525	0.20	DT NN IN DT NN IN DT NN a woman in a room with a cat
1488	0.19	DT NN VBZ VBG DT NN IN DT NN a man is doing a trick on a skateboard

設問文(2,240文中)		
文の数	文の割合[%]	品詞の並びと例
55	2.46	DT NN VBZ VBG DT NN the girl is reading a book
44	1.96	EX VBP CD NNS IN DT NN there are two cranes in the picture
42	1.88	DT NN VBZ VBG DT JJ NN the girl is wearing a red top
42	1.88	DT NN VBZ VBG IN DT NN a motorcycle is driving behind the truck
38	1.70	DT NN VBZ RB JJ the water looks very rough

表2．設問の言い換え規則

変更前	変更後
There is/are ～	～
主語～ is/are 動詞ing	主語 動詞ing
He/She	A man/woman
The ～	Theをとった自然な形に

表3．各手法の正解数とその割合

	各手法の正解数(610問中)		
手法	基本手法	未知語学習あり	設問言い換えあり
正解数	168問 (28%)	194問 (32%)	149問 (24%)

表4．単語を未知語に変換して学習した場合の未知語を含んだ75文に対するコストの平均

	未知語変換	
	なし	あり(1%)
コストの平均	1.38	1.17

表5．言い換えた設問の生成確率の平均

	言い換え(1375文)	
	なし	あり
生成確率の平均	-9.26	-7.90

表6．言い換えた設問のコストの平均

	言い換え(1375文)	
	なし	あり
コストの平均	1.09	1.06

未知語学習有の手法を用いることで，表 4 のように学習文の単語をランダムに未知語変換しない場合より文のコストが約 15%低くなった．未知語を含む文は 2,240 文中 75 文，3.3%であった．

設問文言い換え有の手法において変更された文の生成確率は表5のように約15%高くなった．変更された文は 2,240 文中 1,375 文，61%であった．そのうちの 96%は生成確率が高くなり，4%が低くなった．しかし，表 6 のように，コストは約 3%しか低くならなかった．また，全体でコストが高くなった文は 515 文，37%あった．コストが高くなった文が多くなった原因は，言い換えにより be 動詞が削られたためであると考えられる．be 動詞は，データセットの多くの文に含まれており，出現確率が著しく高くなっている．よって，be 動詞が含まれると文の生成確率の単語平均は大きくなるため，コストは小さくなる．この be 動詞が言い換えで削られたことでコストが高くなったと考えられる．

5 考察

学習文での出現頻度の違いにより単語によって出現確率が異なるので，基本手法では出現確率が高くなりやすい単語を多く含む文は有利になってしまう．出現頻度が一定以下の単語は未知語に変換したほうが良いと考えられる．

未知語学習有の手法では，学習後の未知語の出現確率は学習時の未知語への変換確率に依存するので，適切な変換確率を考えなければならない．また，現在の手法は未知語を全てまとめて 1 つの未知語にしているので，品詞ごとに分ける等の工夫が必要となる．

設問文言い換え有の手法では，一部の文の言い換えしか行われていないため，正答率の向上にはつながらなかったと考えられる．現在の手法では品詞の並びにのみ注目しているので，文の構造も考えることで学習データと問題文の形式をより詳細に解析し，学習データの表現を全て設問文と同じ形式に変える，もしくは文の形式に依存しない判定を行う必要があると考えられる．

6 おわりに

本論文では，キャプション生成モデルの汎用性を調査するために，画像に関する選択問題を解き，最大で 610 問中 194 問，32%正解という結果を得た．学習に用いたキャプションと設問の文の形式が異なることが不正解の 1 つの主な原因であると考えられるが，単純な言い換えでは精度が下がってしまう結果となった．構文解析をするなどし，2 つのデータセットにおける文の差異をより詳細に埋めることが今後の課題である．

参考文献

[1] Sepp Hochreiter, Jurgen Schmidhuber. Long short-term memory. Neural Computation. 1997

[2] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho , Aaron Courville , Ruslan Salakhutdinov , Richard S. Zemel , Yoshua Bengio. Show, Attend and Tell : Neural Image Caption. 2015

[3] Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014

[4] <http://www.english-test.net/toeic/listening/#photographs> (2015/12/27 アクセス)

[5] Yusuke Miyao, Jun'ichi Tsujii. Feature Forest Models for Probabilistic HPSG Parsing. Computational Linguistics. 34(1). pp. 35-80, MIT Press. 2008