日本感性工学会論文誌 Vol.18 No.2 pp.151-158 (2019)
doi: 10.5057/jjske.TJSKE-D-18-00071
特集「あいまいと感性」

原著論文

感情推定を利用した感性的な画像説明文自動生成システム

三由 裕也*, 萩原 将文**

* (株)日立製作所, ** 慶應義塾大学

Automatic Affective Image Captioning System using Emotion Estimation

Yuya MIYOSHI* and Masafumi HAGIWARA**

** Hitachi, Ltd., 322-2 Nakazato, Odawara-shi, Kanagawa 250-0872, Japan*
*** Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan*

Abstract : Image captioning has been actively studied these days, however, most of the systems output captions of factual expression. In this paper, an automatic affective image captioning system using emotion estimation is proposed. The proposed system consists of four parts: a base caption generation part composed by the conventional CNN (VGG16), a scene estimation part, an emotion estimation part, and a figurative expression generation part. When a human exists in an image, the emotion is estimated from his/her facial expression and simile is used. When a human does not exist in an image, personification of metaphor is used. Evaluation experiments have been carried out using three kinds of evaluation indexes; BLUE, METEOR, and CIDEr. The experimental results indicate the effectiveness of the proposed system to generate affective captions.

Keywords : *Image captioning, Convolutional neural network, Emotion estimation*

1. はじめに

画像に説明文を自動で付与するImage Captionは，画像処理と自然言語処理の2つの分野にまたがり，単なる物体認識のみならず画像理解に向けた重要なタスクである．インターネットの普及を背景に膨大なデータが手軽に入手可能になり，ニューラルネットワークを利用した高精度なImage Captionが実現可能となっている．例えば近年のImage Captionシステム［1-5］では，Convolutional Neural Network（CNN）［6］を用いて画像の特徴を抽出し，抽出した特徴をRecurrent Neural Network（RNN）［7, 8］に入力することで画像に適切な説明文を生成しており，優れた成果を示している．

Image Captionの応用先として，絵本や漫画の自動生成が考えられる．しかし，実際にはImage Captionシステムを利用する例は少ない．この理由の一つとして，感性表現が不足しているため，生成される文章が無機質になる傾向が高いことがあげれる．従来のImage Captionシステムでは画像内の物体とそれらの動作や物体間の関係を描写することに焦点を当てていたため，画像を見た人が抱く感情の描写には着目していなかった．

人の感情に着目した既存研究として，Yuらの研究［9］が挙げられる．Yuらの研究ではCNNを利用することで，画像中の人物の表情から感情の推定を行っている．CNNはImage Captionのタスクでも用いられる技術であり，感情推定を利用することで，より多くの感性表現を追加可能であることが予想される．しかし，対象は人が存在する画像に限られている．実際，Image Captionのタスクで学習に用いられるMicrosoft COCO［10］やFricker8k［11］，Fricker30k［12］などのデータセットには，人が存在しない画像も非常に多い．このため，Image Captionのタスク向けに人の存在しない画像からも感情推定を行う手法が必要である．

感性に着目して画像の説明文を生成する既存手法として，Mathewsらの研究［13］が挙げられる．これは各物体の質感や状態を表す感性語をその物体の画像から推定するタスクを学習することで，画像の説明文中に感性語を含む説明文を生成する手法である．各物体に感性語を付与することで物体に関する描写力は向上したが，動作には着目していない．文章の基本文型には，例えば英語の場合では，S（主語），V（動詞），O（目的語），C（補語）がある．これらのうちS，V，Cは名詞または形容詞となるため，形容詞である感性語による描写力向上の恩恵を受ける．しかしVは動詞であるため，形容詞である感性語による描写力向上の恩恵を受けない．このため，動作に関する描写力は従来の手法と変わらない．そこで，動詞を修飾する副詞を用いた感性的な説明文生成が大変重要であると考えられる．

そこで本論文では風景画像から感情推定を行い，動詞を修飾する副詞に感性的な語を用いて説明文を生成するシステムを提案する．さらに，気持ちを表現する際に比喩表現が適しているといった知見［14］から，感性的な語として比喩表現となる語を利用する．具体的には，2通りの感情推定を実施し，それぞれ異なる比喩の手法を用いて説明文を生成する．1つ目は人の表情からの感情推定であり，推定した感情を元に直喩の手法で感性的な説明文を生成する．2つ目の風景画像からの感情推定では，推定した感情を元に隠喩の擬人化の手法で感性的な説明文を生成する．

Received: 2018.10.02 / Accepted: 2019.01.13
J-STAGE Advance Published Date: 2019.03.05

Copyright© 2019 日本感性工学会. All Rights Reserved.

以下，第2章では提案システムの概略，第3章では評価実験，第4章では評価実験に関する考察，第5章では結論を述べる．

2. 提案システムの概要

提案する感情推定を利用した感性的な画像説明文自動生成システムは，以下の4つの処理部から構成される．

1）ベースの説明文生成部
感性的な表現を含む前の画像の説明文を生成する．

2）シーン推定部
画像中に含まれる物体から，そのシーンを推定する．

3）感情推定部
画像中に人が存在する場合は，人の表情から感情を推定する．画像中に人が存在しない場合には，風景画像から，その印象に適した感情を推定する．

4）比喩表現生成部
画像中に人が存在する場合は，推定した感情とシーンに適した直喩文を生成し，ベースの説明文に付加する．一方，画像中に人が存在しない場合には画像中に存在する物体を擬人化し，その動作を修飾する副詞を推定した感情を元に，ベースの説明文に付加する．

以降，各処理部について詳細に説明する．

2.1 ベースの説明文生成部

ここではまず，感性的な表現を含む前の画像の説明文を生成する．説明文の生成には，Fuら[5]の使用したモデルを使用する．これはCNNとRNNを組み合わせることで画像に適した説明文を生成するモデルであり，ここではその概要を説明する．

図1に説明文生成モデルの構造を示す．ここで，S_tは時刻tにおける入力単語を表し，Nは単語列の長さを表す．ただし，S_0は文の開始を意味し，S_Nは 文の終端を意味する．また，W_Iは画像の特徴ベクトルに対する重み行列，W_eは入力単語をベクトルへ変換するための重み行列，p_tは時刻tにおける単語の出力確率を表す．

画像の特徴抽出を行うCNNには，VGG16[18]というモデルを使用する．VGG16は，一般物体認識のコンペティションであるILSVRC2014 [19] で高精度を記録したモデルであり，畳込み層が13層，全結合層が3層の，計16層で構成されたCNNである．文生成を行うRNNには，時系列データの長期依存を学習可能なLong short-term memory（LSTM）[20]を使用する．

次に説明文生成モデルの学習について説明する．学習の手順は以下の通りである．

1. 画像IをCNNへ入力し，畳み込みフィルタFを適用し，画像特徴Yを抽出する．

$$Y = FI \tag{1}$$

2. 抽出した画像特徴をLSTMへ入力する．

$$\boldsymbol{x}_{-1} = \boldsymbol{W_I Y} \tag{2}$$

ここで，x_t (t = 0, …, N−1) はLSTMの入力を意味し，単語学習用の単語S_tと重みW_eを掛け合わせたものである．なお，最初のLSTMだけは画像の特徴も考慮させるために，画像特徴$\boldsymbol{Y}$と重み$\boldsymbol{W_I}$を掛け合わせたものを入力とする．この際のX_tのインデックスは単語の系列とは関係ないことから0より前の-1とする．

3. 単語S_tをt = 0～N−1 まで順に入力し，各ステップで次単語の出力確率p_{t+1}を取得する．

$$x_t = W_e S_t \qquad (t = 0, \ldots, N-1) \tag{3}$$

$$p_{t+1} = \mathrm{LSTM}(x_t) \qquad (t = 0, \ldots, N-1) \tag{4}$$

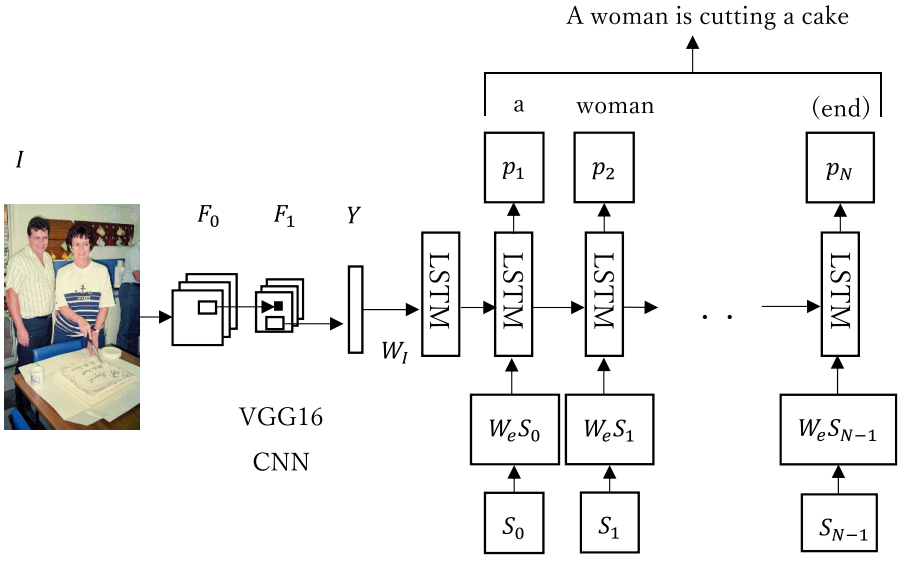

図1 説明文生成モデルの構造

4. 以下の目的関数を最大化するようにパラメータを学習する.

$$L = \frac{1}{N}\sum_{n}\sum_{t=1}^{Tn} \log p(w_t^{(n)} | w_{0:t-1}^{(n)}, I^{(n)}) \quad (5)$$

ここで，T_nはキャプションnの長さである.

モデルの学習には，Microsoft COCO［10］と呼ばれる大規模なImage Captionデータセットを使用した.

使用したモデルには以下の2つの特徴がある.

1. 画像から30個の複数スケールの矩形領域を抽出し，文中の単語にふさわしい切り出し方を学習可能である. すなわち, 文は主語, 動詞, 目的語という順番になるが, どのような矩形画像が主語や動詞になりやすいかを学習可能である. この結果, 文生成のために画像のどのような部分から着目するのが良いかを，細かい粒度で推定する.
2. Latent Dirichlet Allocation（LDA）［21］とMultilayer Perceptron(MLP) を用いて画像とそのシーンを学習する. この結果を用いて，シーンに不適切な表現を回避する.

2.2 シーン推定部

シーン推定部では，後の比喩表現生成部にて比喩表現を生成するためのシーン推定を行う. 前節で説明したようにLDAとMLPを利用することでシーンを学習する. まず，データセット内の各画像について，説明文中の単語からLDAの手法を用いて80次元のトピックベクトルを取得する. これは，対象の画像の説明文中で話題とされている内容を表すことから，画像のシーンを表現していると考えられる. LDAで取得した80次元のトピックベクトルを教師データとして，2層のMLPで学習を行う. 学習モデルの構造を図2に示す. 最初に画像をImage Net［22］を用いて事前学習させたCNNに入力し，画像の特徴を抽出する. なお使用したCNNは，ILSVRC2012［19］で高精度を記録したAlexNet［23］がベースのPlace-205［24］である.

表1 MS COCOから抽出した画像シーンのリスト

Cake	A woman	Vase	Stand
Sitting	Suitcase	Grass	Snow
Road sign	Water	Fly	Jump
Various	Suit, tie	Bear	umbrella
Train	Car	Black and white	Pizza
Banana	Two	Room	Phone
Three	Traffic signal	Skateboard	Platform
Close up	Bike	Cute	Eat
On	A	Computer	Swing
Kite	Ride	Bird	Sandwich
Park	Elephant	Line	Dog
Walk	Window	In	Fruits
Frisbee	Bed	Controller	Tennis
Zebra	Child	Prepare	Head
A man	Bus	Vegetable	Hydrant
Plane	Clock	Cut	Mirror
Bench	Hold	Group	Old
Surfboard	Two people	Tree	Bathroom
Baseball	Donut	Kitchen	Toilet
Horse	Team	Picture	cat

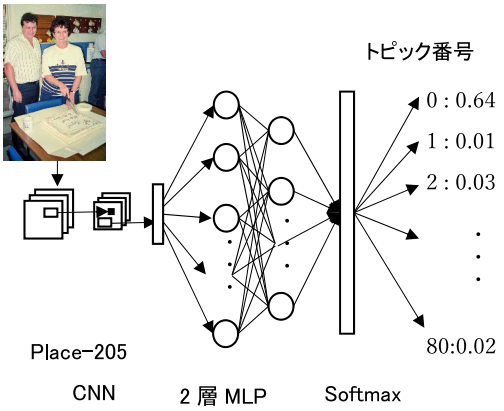

図2 画像からのシーン推定

次にPlace-205で抽出した画像特徴を2層のMLPに入力し, トピックベクトルを予測する学習を行う. 図2の例では, 入力した画像からトピックベクトル0である確率が64%と推定されている. 表1にMS COCOデータセットに対してLDAを適用し, 抽出した80個の画像シーンのリストを示す. 各トピックベクトルは，80個の中の1つと対応している.

2.3 感情推定部

感情推定部では，以下のような2通りの方法で感情推定を行う. まず画像中に人が存在する場合は，人の表情からの感情推定を行う. 人が存在しない場合は，風景画像からの感情推定を行う. 以下, それぞれの方法について詳しく説明する.

2.3.1 人の表情からの感情推定

人の表情からの感情推定にはMicrosoft社のEmotion API［5］を利用する. Emotion APIでは，人の表情を元に，Happiness（幸福), Sadness（悲しみ），Angry（怒り），Disgust（嫌悪），Contempt（侮辱），Fear（恐れ），Surprise（驚き），Neutral（無感情）の8種類の感情の中から最適な感情を推定可能である.

2.3.2 風景からの感情推定

風景からの感情推定は，4層の畳込み層と2層の全結合層を持つCNNを学習させることで実現した. 学習に使用する画像データについては，インターネットの画像検索を利用し，Happiness（幸福），Sadness（悲しみ），Disgust（嫌悪），Fear（恐れ）を表す風景画像を5,000枚収集した. 収集後, 学習データの画像枚数を増加させる目的で，画像の平滑化, ガウス分布に基づくノイズの付加，画像の左右反転を行った. 最終的に，30,000枚の画像データを学習に使用した.

2.4 比喩表現生成部

比喩表現生成部では，画像中に人が存在する場合は直喩で, 人が存在しない場合は隠喩の擬人化で比喩表現を生成する.

以下，それぞれの方法について説明する.

2.4.1 画像中に人が存在する場合

ベースの説明文生成部で生成された文に対し，以下2種類の追加文を付加する.

・感情を表す副詞
・直喩表現を含む文

以下，それぞれの追加文の付加方法を説明する．

a　感情を表す副詞

画像中の人の動作を感性的に修飾するために，感情を表す副詞を説明文に付加する．感情を表す副詞は，以下の方法で選定する．

1）副詞の中でも動作の状態を表す副詞の選定
副詞には，状態の副詞，程度の副詞，陳述の副詞，指示の副詞が存在するが，本論文では動作を修飾する目的で副詞を利用するため，状態の副詞からの選定を行う．

2）Emotion APIで推定可能な8種類の感情ごとに副詞のクラスタ分け
1）で選定した状態の副詞についてword2vec [25, 26] を用いてベクトル化し，そのベクトル値と8種類の各感情の名詞をベクトル化した値とのcos類似度を比較し，最も類似する感情に副詞をクラスタ分けした．似た単語ほど近い単語ベクトルになるという性質があるため，状態の副詞を各感情にクラスタ分けする際に適していると考えられる．感情を表す名詞ごとに各副詞とのcos類似度を計算し，感情ごとに副詞のクラスタ分けを行った．表2にその結果の一部を示す．

3）動詞と感情の組み合わせごとに使用する副詞の決定
説明文中の動詞と，表情から推定した感情，また感情の推定値の組み合わせごとに，実際に文に付加する副詞を決定する．動詞，感情の推定値の組み合わせごとの副詞は表3に示すように定めた．ここで動詞は，“stand”のような静的な動詞と，“run”のような動的な動詞の2種類に分類した．感情は2）で説明したHappiness（幸福），Sadness（悲しみ），Angry（怒り），Disgust（嫌悪），Contempt（侮辱），Fear（恐れ），Surprise（驚き），Neutral（無感情）の8種類で分類した．感情の推定値は①0%~25%，②25%～50%，③50%～75%，④75%～90%，⑤90%～100%の5種類に分類した．各副詞は2）でのクラスタ分けにもとづいて，各感情に分類されている．感情の推定値による副詞の使い分けに関しては，cos類似度の大きさに基づき，感情推定値が大きい値の場合は，cos類似度の大きな副詞を利用した．

以上の3段階の方法で決定した副詞を，ベースの説明文に付加することで，画像中の人の動作を感性的に修飾する．

b　直喩表現を含む文

直喩の比喩表現生成部では，シーン推定部で推定したシーンと，感情推定部で推定した感情の組み合わせによって付加するたとえを決定する．今回は比喩文の作りやすさを考慮して，感情はHappiness（幸福）とNeutral（無感情）のみ利用することとした．また同様の理由から，シーンについても80個の中から9個のみを利用することとした．

副詞と“as if ~”という特定の表現を用いることで，「まるで～のように」といった直喩表現を含む感性的な説明文を生成する．

2.4.2　画像中に人が存在しない場合

画像中に人が存在しない場合は，ベースの説明文生成部で生成された文に対し，隠喩の擬人化を表現する追加文を付加する．付加する文は，画像中の物体の動作を修飾する副詞とし，その物体が感情を持っているかのようにベースの説明文に副詞を付加する．付加する副詞は感情推定部で風景からの感情推定を実施した結果と，推定した感情値の組み合わせで決定する．

2.4.1の場合と同様に，感情と感情の推定値ごとの副詞を定めた．今回は比喩文のつくりやすさを考慮して，感情はHappiness（幸福），Sadness（悲しみ），Fear（恐れ），Disgust（嫌悪）の4種類を利用することとした．なお，隠喩文は擬人化の対象となるような物体が存在する画像に対してのみ実施した．擬人化の対象となる物体は作りやすさを考慮して，バス，電車，バイク，車の4種類とした．

表2　副詞のクラスタ分け結果（上位3単語のみ）

感情の名詞	1位の副詞	2位の副詞	3位の副詞
Happiness	happily	really	honestly
Sadness	sadly	frankly	painfully
Angry	angrily	rightfully	violently
Contempt	contemptibly	justly	sternly
Disgust	calmly	irritably	wearily
Fear	fearfully	eventually	easily
surprise	surprisingly	unexpectedly	exactly
Neutral	sharply	straight	carefully

表3　動詞，感情の推定値の組み合わせごとの副詞

感情	動詞	① 0%～25%	② 25%～50%	③ 50%～75%	④ 75%～90%	⑤ 90%～100%
Happiness	動的	cheerfully		joyfully		happily
	静的	brightly		gladly		happily
Sadness	動的	sadly				
	静的	sadly				
Angry	動的	nervously	wildly	violently	fiercely	fiercely
	静的	nervously			angrily	
Contempt	動的	contemptibly				
	静的	sternly			contemptibly	
Disgust	動的	awkwardly	wearily	irritably		disgustingly
	静的	awkwardly	wearily	irritably		disgustingly
Fear	動的	anxiously	deliberately		fearfully	
	静的	anxiously	fearfully			
Surprise	動的	surprisingly				
	静的	surprisingly				
Neutral	動的	normally		seriously		
	静的	calmly		ordinary		normally

3. 評 価 実 験

図3に提案システムで出力した直喩を付加した画像説明文を，図4に隠喩（擬人化）を付加した画像説明文の例を示す．

提案システムの評価のために，生成した画像説明文の定量的評価と主観評価の2つの実験を行った．以下，各実験の概要と結果を示す．

3.1　生成した画像説明文の定量的評価

提案システムで生成した画像説明文について定量的に評価した．評価には，画像説明文の定量的評価で一般的に使用されているBLEU [15]，METEOR [16]，CIDEr [17] を使用した．

a man swinging a tennis racquet on a tennis court seriously as if he were professional tennis player.

a man in a suit and a tie happily as if he were success at business.

図3　提案システムで出力した直喩を付加した画像説明文例

a truck is driving down the road happily.

a cat sitting on the back of a car disgustingly.

図4　提案システムで出力した隠喩（擬人化）を付加した画像説明文例

BLEUは，生成キャプションと教師キャプションとの類似度を*n*-gram一致数をもとに算出する手法である．

METEORはBLEUの欠点を補った評価指標である．まず，BLEUの欠点の一つとして，評価の際に適合率のみを考慮しており，再現率を考慮できていないという点があげられる．METEORでは，適合率と再現率の調和平均であるF値を用いてキャプションの評価を行っている．BLEUのもう一つの欠点として，単語の同義語や語形変化を考慮できていないという点がある．そこでMETEORでは，WordNet［27］を用いた同義語の考慮や語形変化を考慮した評価を行い，より人の評価と相関の高い指標になっている．

BLEUとMETEORは，機械翻訳のために考案された評価指標であり，画像のキャプション生成の評価に最適化されていない．CIDErは画像の説明文生成のために考案された評価指標である．説明文中の単語の重要度を考慮している点が特徴である．Microsoft COCOなどの画像の説明文データセットには1枚の画像に複数の教師説明文が付与されている．CIDErでは任意の画像の説明文について，他の画像の説明文にも出現している単語は重要度が低く，同じ画像の説明文にも出現している単語は重要度が高いという仮定をTF-IDF［28］を用いて定式化している．

表4　データセットの内訳

学習データ（枚）	交差検証データ（枚）	テストデータ（枚）
82,783	40,504	1,000

以上の3つの評価指標を用いて生成した画像説明文の定量評価を行い，以下の手法と比較を行った．

・Mathewsら［13］の研究
・比喩文を付加しない提案システム（ベースの説明文生成部のみ）

データセットは，Microsoft COCOと呼ばれる画像説明文データセットを使用した．また，表4に使用したデータセットの内訳を示す．

表5に生成した説明文の定量評価の結果を示す．すべての指標においてMathewsらの研究の精度を上回り，感性的な表現を含みながらもより正確な画像説明文の生成が可能であることが示されている．また，やや劣るもののベースの説明文生成部に近い数値を示しており，感性的な表現を含まないモデルと同程度に正確な画像説明文の生成が可能であることも示唆されている．

3.2　生成した画像説明文の主観評価

生成した画像説明文について，人手による主観評価実験を行った．画像中に人が存在する場合と，人が存在しない場合で実験を行った．

3.2.1　画像中に人が存在する場合

画像中に人が存在する場合については，以下の3つの手法の比較を行った．

・ベースの説明文生成部のみ
・ベースの説明文生成部＋副詞付加
・ベースの説明文生成部＋副詞付加＋直喩付加

20代の男性11名，女性3名の計14名の被験者に対して1名あたり10枚の画像を以下の評価項目について5段階で評価してもらった．

- 各画像説明文について
  - 文の正しさ　　1（正しくない）～5（正しい）
  - 描写の適切さ　　1（適切でない）～5（適切である）
  - 文の表現力　　1（低い）～5（高い）
- 比喩生成部の出力を含む説明文のみについて
  - たとえの妥当性　1（妥当でない）～5（妥当である）

使用したデータセットは前の実験と同様にMicrosoft COCOである．

表5　生成した説明文の定量評価実験の結果

手法	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	CIDEr
Mathewsら［14］	50.0	31.2	20.3	13.1	16.8	61.8
ベースの説明文	**70.9**	**54.1**	**40.2**	**29.7**	**23.7**	**89.8**
提案システム（直喩）	70.1	53.4	39.7	29.2	23.6	85.4
提案システム（隠喩）	70.6	53.9	40.1	29.5	23.6	85.2

表6　直喩表現を含む説明文の主観評価実験の結果

手法	文の正しさ	描写の適切さ	文の表現力	例えの妥当性
ベースの説明文生成部	**4.4**	**3.8**	3.1	–
ベースの説明文生成部 +副詞付加	4.3	3.7	3.7	–
ベースの説明文生成部 +副詞付加 +直喩付加	4.2	**3.8**	**4.4**	3.9

表7　隠喩（擬人化）表現を含む説明文の主観評価実験の結果

手法	文の正しさ	描写の適切さ	文の表現力	推定感情の妥当性
ベースの説明文生成部	**4.5**	**4.0**	3.6	–
ベースの説明文生成部 +隠喩（擬人化）付加	4.2	3.9	**4.2**	3.4

表6に実験結果を示す．文の表現力において提案システムが感性表現を含まないベースの説明文生成手法を上回っている．文の正しさや描写の適切さについても，提案システムは感性表現を含まないベースの手法と同程度の評価であった．文の表現力においては，どの組み合わせにおいても$p<0.05$で有意差が確認された．また，直喩のたとえの妥当性についても高く評価をした被検者が多く，提案システムによって画像に対して妥当なたとえを生成できていることが示唆された．

3.2.2　画像中に人が存在しない場合

画像中に人が存在しない場合については以下の2手法を比較した．実験方法は画像中に人が存在する場合と同様である．新たな評価項目については以下の通りである．

– 推定された感情の妥当性　1（妥当でない）～5（妥当である）

表7に実験結果を示す．文の表現力において提案システムが感性表現を含まない既存手法を上回り，$p<0.05$で有意差が確認された．また，文の正しさや描写の適切さについても，提案システムは感性表現を含まない既存手法と同程度の評価であった．また，風景画像からの感情推定の妥当性についても高く評価をした被験者が多かった．

3.3　実験結果の考察

直喩文と比喩文の付加を行うことで，通常の画像説明文生成のモデル［5］と比較して表現力の高い画像説明文を自動生成できることが示唆された．ここでは，提案システムにおいて主観評価実験の結果が低かった項目に着目して，その原因の考察を行う．

3.3.1　直喩表現付加に関する実験結果の考察

評価実験において提案システムでは，ベースの説明文生成部のみのモデルと比較して描写の適切さの評価値が低くなることがあった．また，描写の適切さが低い説明文については，たとえの妥当性についても低くなる傾向が見られた．このことより，妥当なたとえができていない場合には描写力が損なわれることが示唆される．この場合，たとえになっていない場合と，たとえが間違っている場合がある．

前者に関しては，子供に対して「まるで子供のように」といった表現や，プロのテニス選手に対して「まるでプロテニス選手のように」といった表現を付加することがあった．これは，ベースの画像説明文生成部と比喩表現部が独立して動作しているためと考えられる．

後者のたとえが間違っている場合については，カジュアルな恰好の男性に対して「まるで会社員のように」といった表現や，サッカーをしている男性に対して「まるでプロテニス選手のように」といった表現を付加することが原因であった．これは，シーン推定部での問題があると考えられる．シーン推定部ではLDA［20］に基づいて得たトピックベクトルをMLPで推定しているが，その際に推定結果に誤りが生じたためと考えられる．

3.3.2　隠喩表現付加に関する実験結果の考察

評価実験において提案システムでは，ベースの説明文生成部のみのモデルと比較して文の正しさの評価値が低くなることがあった．正しい文を出力できていない場合の大きな要因の一つとして，ベースの説明文において動詞の無い文章の存在があげられる．無理に副詞が付加されることにより，文の正しさが損なわれたと考えられる．

また，感情推定の誤りによって描写力が損なわれている場合もあった．評価の低い結果例の検討を行った結果，物体自体の色合いや，画像全体の色合いが感情推定に大きく影響していることがわかった．例えば，黒い車の画像に“fearfully”，全体的に黒っぽい画像に対して，“sadly”などの単語が付加されていた．

4.　結　論

本論文では，感情推定を利用した感性的な画像の説明文自動生成システムを提案した．提案システムでは，最初にCNNとLSTMを用いたモデルで感性表現を含まないベースの画像説明文を生成する．次に画像中に人が存在する場合と，存在しない場合で2通りの感情推定を行う．画像中に人が

存在する場合は画像中の人の表情から感情推定を行い，存在しない場合は，画像の背景となる風景画像から感情推定を行う．その後，画像中に人が存在する場合は，直喩表現を生成する．画像中に人が存在しない場合は，隠喩の擬人化表現を生成する．

2種類の評価実験を行った．まず定量的評価実験では，感性的な表現を含む既存研究より正確な画像説明文の生成が可能であることが示された．また，感性的な表現を含まないモデルと同程度に正確な画像説明文の生成が可能であることも示された．次に主観的評価実験では，提案システムによって描写の正確さや文の正しさを損なうことなく，表現力の高い画像の説明文生成が可能であることが示唆された．

参 考 文 献

［1］ Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2625-2634, 2015.

［2］ Vinyals, O., Toshev, A., Bengio, S., and Erhan, D.: Show and tell: a neural image caption generator, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3156-3164, 2015.

［3］ Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., and Zweig, G.: From captions to visual concepts and back, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1473-1482, 2015.

［4］ Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., and Sienkiewicz, C.: Rich image captioning in the wild, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.434-441, 2016.

［5］ Fu, K., Jin, J., Cui, R., Sha, F., and Zhang, C.: Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), pp.2321-2334, 2017.

［6］ Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86(11), pp.2278-2324, 1998.

［7］ Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.1724-1734, 2014.

［8］ Sutskever, I., Vinyals, O., and Le, Q.V.: Sequence to sequence learning with neural networks, Advances in Neural Information Processing Systems, 27, pp.3104-3112, 2014.

［9］ Yu, Z., and Zhang, C.: Image based static facial expression recognition with multiple deep network learning, In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp.467-474, 2015.

［10］ Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, C.L.: Microsoft COCO: common objects in context, European Conference on Computer Vision (ECCV) 2014, pp.740-755, 2014.

［11］ Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J.: Collecting image annotations using Amazon's Mechanical Turk, Proc. NAACL HLT Workshop Creating Speech Language Data Amazon's Mech. Turk, pp.139-147, 2010.

［12］ Young, P., Lai, A., Hodosh, M., and Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics, 2, pp.67-78, 2014.

［13］ Mathews, A., Xie, L., and He, X.: SentiCap: generating image descriptions with sentiments, AAAI, 2016.

［14］ 岡隆之介，楠見孝：感情の“字義と比喩”表現および“気持ちと行動”記述の差異が感情評価に与える影響，日本感性工学会論文誌，16(3)，pp.307-313，2017.

［15］ Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pp.311-318, 2002.

［16］ Denkowski, M.J. and Lavie, A.: Meteor universal: language specific translation evaluation for any target language, Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.376-380, 2014.

［17］ Vedantam, R., Zitnick, C.L., and Parikh, D.: CIDEr: consensus based image description evaluation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4566-4575, 2015.

［18］ Simonyan, K., and Zisserman, A.: Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556, 2014.

［19］ Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., and Fei-Fei, L.: ImageNet large scale visual recognition challenge, International Journal of Computer Vision, 115(3), pp.211-252, 2015.

［20］ Hochreiter, S., and Schmidhuber, J.: Long short-term memory, Neural Computation, 9(8), pp.1735-1780, 1997.

［21］ Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation, Journal of Machine Learning Research, 3, pp.993-1022, 2003.

［22］ Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L.: ImageNet: a large-scale hierarchical image database, IEEE

Conference on Computer Vision and Pattern Recognition, pp.248-255, 2009.
［23］Krizhenvshky, A., Sutskever, I., and Hinton, G.: ImageNet classification with deep convolutional networks, Advances in Neural Information Processing Systems, 25, pp.1097-1105, 2012.
［24］Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A.: Learning deep features for scene recognition using places database, Advances in Neural Information Processing Systems, 27, pp.487-495, 2014.
［25］Mikolov, T., Yih, S. W.-T., and Zweig, G.: Linguistic regularities in continuous space word representations, In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013), pp.746-751, 2013.
［26］Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, arXiv:1301.3781 [cs.CL], 2013.
［27］Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.: WordNet: an on-line lexical database, International Journal of Lexicography, 3, pp.235-244, 1990.
［28］Robertson, S.: Understanding inverse document frequency: on theoretical arguments for IDF, Journal of Documentation, 60(5), pp.503-520, 2004.

三由　裕也（非会員）
2011 年 株式会社日立製作所入社．ストレージシステムの設計，開発に従事．2017 年度慶應義塾大学理工学部萩原研に共同研究員として滞在．1 年間深層学習や感性工学に関する研究に従事．

萩原　将文（正会員）
1982 年 慶大・工・電気卒．1987 年 同大学院博士課程修了．工博．同年同大助手．現在，同大教授．1991-92 年度スタンフォード大学訪問研究員．視覚・言語・感性情報処理とその融合の研究に従事．1990年 IEEE Consumer Electronics Society 論文賞，1996 年 日本ファジィ学会著述賞，2004 年，2014 年 日本感性工学会論文賞，2013 年 日本神経回路学会最優秀研究賞，2018 年 日本知能情報ファジィ学会論文賞受賞．