[論文紹介] 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

水上 和秀 (Kazuhide Mizukani)
u355020@st.pu-toyama.ac.jp

富山県立大学 工学部 電子情報工学専攻

January 19, 2024

1　はじめに

背景

画像に説明文を自動で付与する技術は画像理解にむけた重要なタスクであるが、実際にその技術を利用する例は少なく、理由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られる。感性的な文章の生成には動詞を修飾する副詞を用いた説明文が大変重要であると考えられる。

目的

ニューラルネットワークを用いて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

提案システムの概要

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
   →画像中の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
   →画像中の中に人が存在する場合は人の表情から感情を推定する。
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、ベースの説明文に追加する

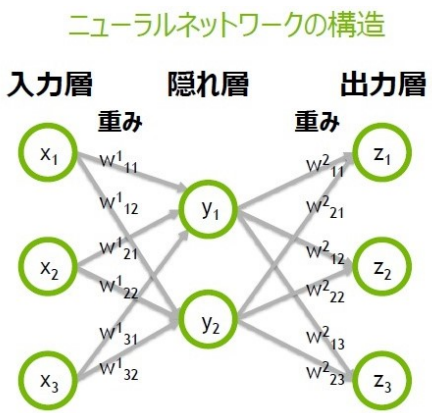
ニューラルネットワークについて

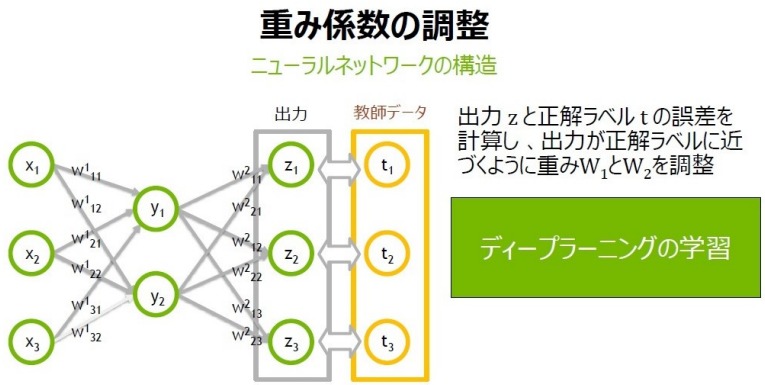
ニューラルネットワーク

機械学習モデルの一種で「入力層」、「中間層」、「出力層」によって構成される

- 入力層: 変数などを入力する層
- 中間層: 入力層と出力層の間にある層。入力層の値と重みの積によって求められる
- 出力層: 中間層の計算結果により予測値が出力される層

それぞれの層の重みは正解データ (教師データ) に近づくように調整し、調整された重みをもとにテストデータの出力層を予測する



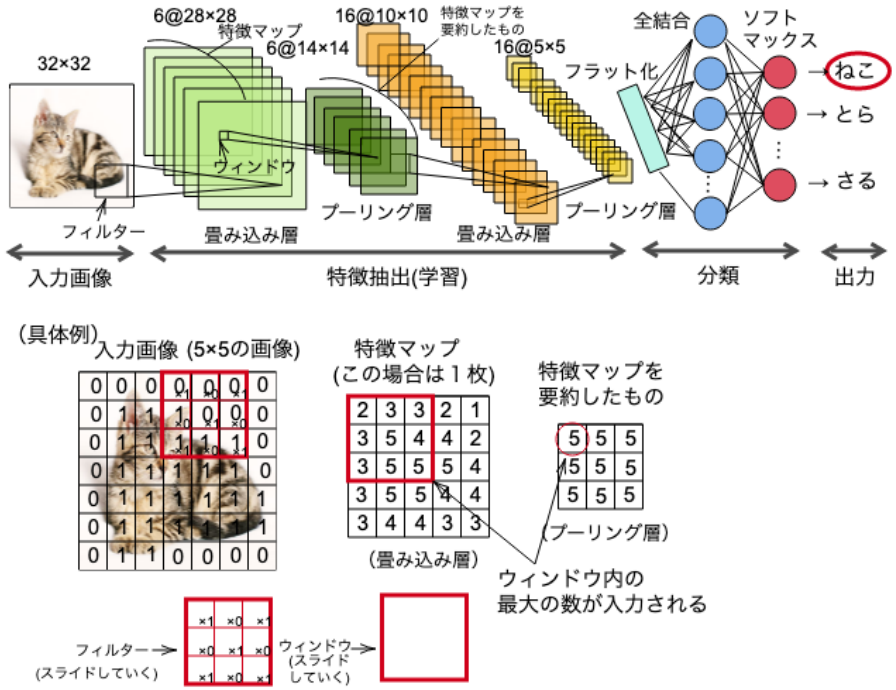


CNN

画像認識に特化したニューラルネットワークで、「畳み込み層」、「プーリング層」、「全結合層」の 3 つの層から構成される

- 畳み込み層: 畳み込みフィルタを用いて画像の特徴を抽出する層
- プーリング層: 畳み込み層で抽出した特徴を圧縮する層
- 全結合層: 畳み込み層やプーリング層での演算で抽出した特徴量から全情報を結合し、最終的な分類や予測を行う層

CNN により画像の特徴量を抽出し、その特徴量をもとに画像の分析を行う



RNN

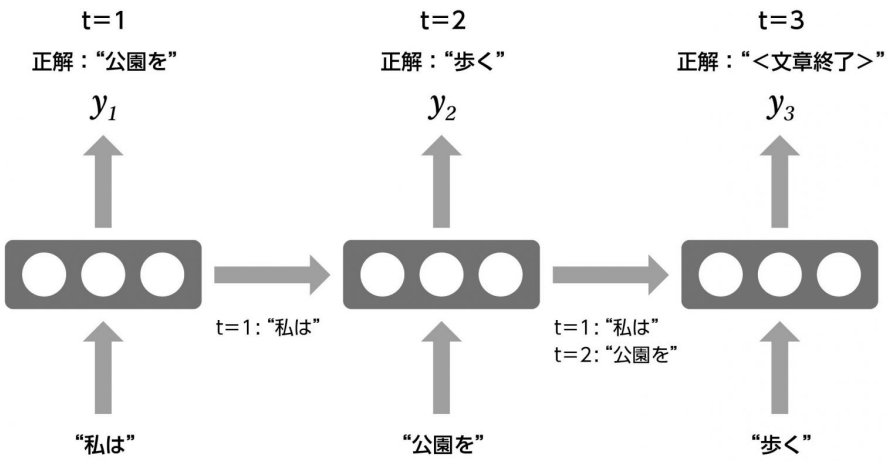
時系列データを解説するニューラルネットワークで、自然言語処理などに利用されている。

- 入力層: 入力データが出力される層
- 再帰層: 過去の時点の情報を保持する層
- 出力層: 最終的な出力を生成する層

RNN は中間層の演算結果を出力層に出力すると同時に、同じ演算結果を次の時系列の再帰層に入力して再演算を行う。

RNN は過去の情報を保持し、それを未来の入力と組み合わせることで、系列データのパターンを学習する

図表 11 RNNによる言語モデル



2.1 ベースの説明文生成部 1

説明文の生成には CNN と RNN を組み合わせて画像に適した説明文を生成する。画像の特徴抽出を行う CNN には折り込み層が 13 層、全結合層が 3 層の計 16 層で構成された VGG16 を用いる。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 LSTM を用いる。

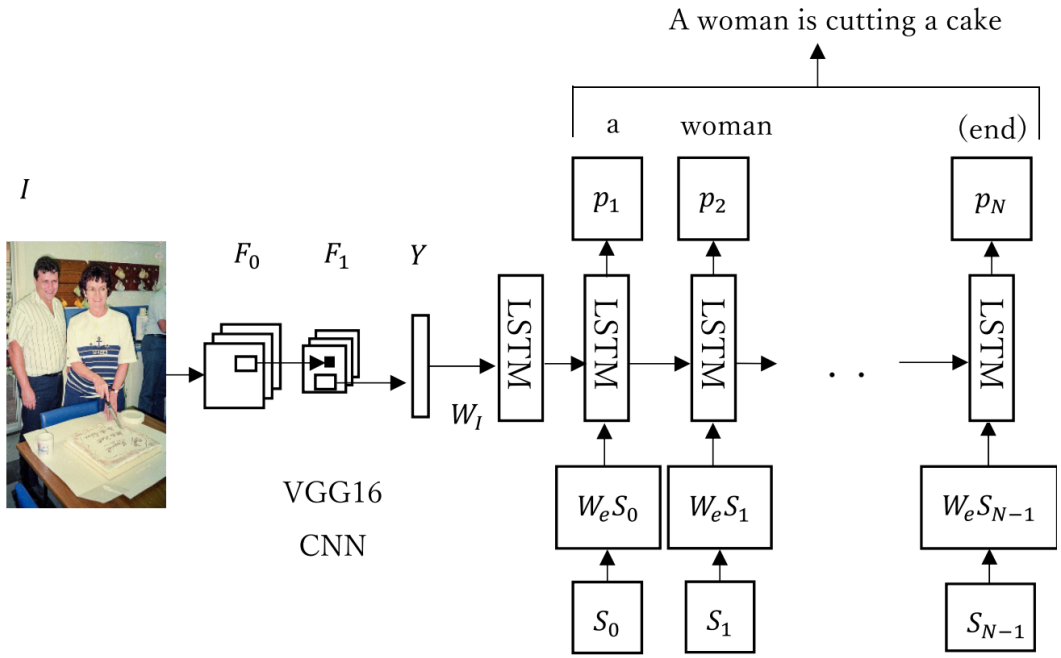


図 1　説明文生成モデルの構造

2.1 ベースの説明文生成部 2

学習の手順

1. 画像 I を CNN へ入力し、畳み込みフィルタ F を適応し、画像特量 Y を抽出

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

2. 抽出した画像特徴を LSTM へ入力する

$$x_{-1} = W_1 Y \tag{2}$$

ここで x(t=0,1,...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 をかけ合わせたものである

3. 単語 S_t を $t = 0$ $N-1$ まで順に入力し、各ステップで次の単語の出力確立 p_{t+1} を取得する

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

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

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

$$L = \frac{1}{N} \sum_{n} \sum_{t=1}^{Tn} logp(w_i^{(n)} | w_{0:t-1}^{(n)}, I^{(n)}) \tag{5}$$

T_n はキャプション n の長さ

2.2 シーン推定部

のち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部分

推定方法

データセットから LDA(トピックモデルを分析する手法) を適応し、80 個の画像シーンを抽出する。
最初に画像を Image Net(画像データベース) 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。次に、CNN で抽出した画像特徴を 2 層のニューラルネットワークに入力し、どのシーンに分類されるか予測する学習を行う。

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

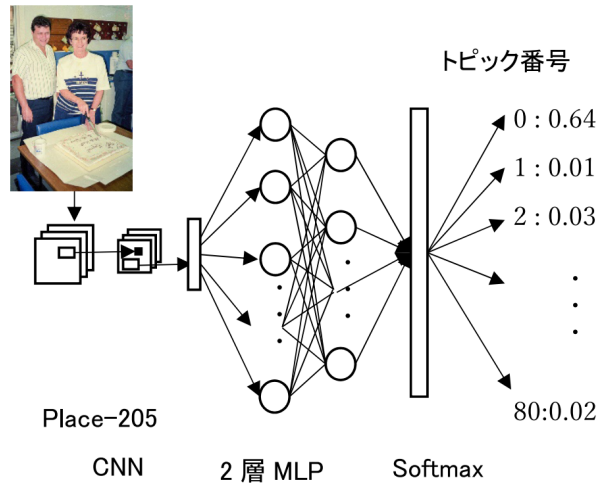


図2　画像からのシーン推定

2.3 感情推定部

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 を利用する。それでは人の表情から幸福、悲しみ、怒り、嫌悪侮辱、恐れ、驚き、無感情の 8 種類の感情の中から最適な感情を推定することができる

風景からのからの感情推定

風景からの感情推定は 4 層の畳み込み層と 2 層の全結合層を持つ CNN を学習させることで実現した。データはインターネットの画像検索を利用し、幸福、悲しみ、嫌悪、恐れを表す画像 30000 枚を学習に使用した。

2.4 比喩表現推定部 1

画像中に人が存在する場合

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。

1. 副詞の中でも動作の状態を表す副詞の選定
   →本研究では動作を修飾する目的で副詞を利用するため、状態の副詞からの推定を行う
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 で選定した副詞について word2vec 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 類似度を計算し、感情ごとに副詞のクラスタ分けを行う
3. 動詞と感情の組み合わせごとに使用する副詞の決定
   →説明文中のどうしと、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

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、画像の中の人の動作を感性的に修飾するする。

2.4 比喩表現推定部 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分に対し、隠喩の擬人化を表現する説明文を追加する。
- 付加する分は、画像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与する
- 本研究で付加する感情は「幸福」、「悲しみ」、「恐れ」、「嫌悪」の 4 種類を利用することとした。

3. 評価実験の概要

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。

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

3.1 定量評価実験の概要

生成した画像説明文の定量的評価

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以下の 3 つを使用した。

- BLEU
  →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
- METEOR
  →調和平均 (生成文に含まれる正しい単語数を全体の単語数で割ったものと参照訳に含まれる正しい単語数を全体の単語数で割ったものの平均) を用いた評価。単語の同異義語や語形変化を考慮した評価をすることができる
- CIDEr
  →単語の重要度を考慮した評価手法。任意の画像の説明文についてほかの画像の説明文にも出現している単語は重要度が低く、ほかの説明文にも出現している単語は重要度が高いという過程を定式化し評価する

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、学習データを 82783 枚、交差検証データを 40504 枚、テストデータを 1000 枚用意した。

3.1 定量評価実験の結果

生成した画像説明文の定量的評価

- すべての指標において Mathews らの研究 (既存研究) の制度を上回り、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。

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

3.2 主観評価実験の概要

生成した画像説明文の定量的評価

人が存在する場合については、以下の文章を日本語に翻訳してください 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のみ+副詞付加
- ベースの説明文生成部のみ+副詞付加+直喩付加

20 台の男性 11 名、女性 3 名の計 14 名の被験者に対して 1 名当たり 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

- 各画像の説明文について
  -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)
- 比喩性西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 を用いた。

3.2 主観評価実験の結果 1

結果

- 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。
- 直喩のたとえの妥当性についても高く評価したした被験者が多く、提案システムによって画像に大した妥当なたとえを生成できていることが示唆された

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

3.2 主観評価実験の結果 2

人が存在しない場合

画像中に人が存在しない場合は先の実験に以下の項目を追加する

-推定された感情の妥当性　1(妥当でない)～5(妥当である)

結果

- 文の表現力において提案システムが感情表現を含まない既存手法を上回った
- 文の正しさや描写の適切者についても提案システムは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した被験者が多かった

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

まとめ

- 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。
- 定量評価実験では正確な画像説明文の説明が可能であることが示された。
- 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だけでなく、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。