深層学習による日本語キャプション生成システムの開発

Development of Japanese caption generation system by deep learning

小林豊[1]　鈴木諒[1]　谷津元樹[2]　原田実[2]

Yutaka Kobayashi[1], Ryo Suzuki[1], Motoki Yastu[2] , Minoru Harada[2]

[1] 青山学院大学大学院　理工学研究科
[1] Graduate school of Science and Engineering, Aoyama Gakuin University
[2] 青山学院大学理工学部情報テクノロジー学科
[2] Department of Integrated Information Technology, College of Science and Engineering, Aoyama Gakuin University

Abstract: For the purpose of developing a dialogue system to dialogue after visually understanding the surrounding situation. We developed Japanese Caption generation system *Deep Watcher* and image datasets with captions. We used the Show and Tell model using CNN and LSTM to generate captions. We also evaluated the coincidence rate of caption content and five feature items manually. As a result the coincidence rate of the contents of the generated caption was 41.6%, the highest characteristic item was gender and was 86.9%. The coincidence rate of the caption contents were not high by over learning, but we could show the possibility of application to the dialog system for the feature item of gender.

1　はじめに

近年，深層学習を用いた画像キャプション生成が発展しており[1,2]，コンピュータに入力された画像に映る状況を説明文（キャプション）として出力することが可能となっている．例として，Microsoft の Web 上で画像のキャプションを生成するボット Caption Bot[1]が挙げられる．

深層学習を利用した画像キャプション生成技術を，周辺状況を視覚的に理解した上で対話する対話システムに応用することを考えた．これは，対話システムが対話相手の性別や人数，服装，持ち物，動作などを中心に対話に応用するというものである．また，過去の深層学習によるキャプション生成の研究[2,3]では，英語のみを使用しているので，日本語キャプションの生成が可能かどうかを試みる為に，日本語キャプション生成システム Deep Watcher を作成した．対話内容に応用することができるキャプション生成について調査を行った．

本研究では日本語でのキャプション生成を考える．しかし，既存のデータセット[4,5]では，付与されているキャプションは英語のみである．また，本目的に沿っていない画像も含まれてしまうため新たに人物を中心とした日本語キャプションが付与された学習データセットを作成した．

2　Deep Watcher

この節では作成したキャプション生成システム Deep Watcher について説明する．Deep Watcher の実行イメージを図 1 に示す．Deep Watcher は GUI を用いたシステムであり同時に複数の画像についてキャプション生成を行うことができる．

2.1　キャプション生成モデル

Deep Watcher には，Show and Tell Model[3]をキャプション生成モデルとして採用する．このモデルは畳込みニューラルネットワーク(CNN)[6]と長短期記憶ニューラルネットワーク(LSTM)[7]の組み合わせで構成されている．モデルの構成を図 2 に示す．入力画像から CNN によって画像の特徴量を抽出したものと，画像の説明文を Word Embedding により単語ベクトルに変換したものを LSTM に入力して，次の単語の出現確率を計算し，最大確率の語をつなげて画像のキャプションを生成する．

モデルの構築には，Deep Learning のフレームワークである Chainer を用いる．CNN では ILSVRC-2014

[1] https://www.captionbot.ai/

model with 19 weight layers(VGG19) の学習済みモデルを用いる．LSTM では，本研究で作成した学習データセットの訓練データから言語モデルを生成する．

図 1　Deep Watcher の実行イメージ

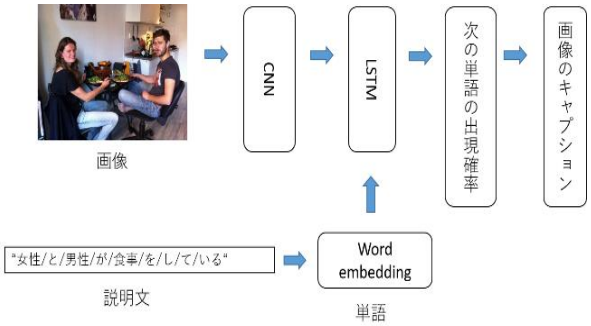

図 2　キャプション生成モデルの構成

2.2　モデルの学習

モデルの学習には，学習データに画像 1 枚に対し 5 文の日本語説明文をペアとして使用する．

順伝播では，まず画像を CNN に入力し画像特徴量を抽出し，それを LSTM に入力する．次に，日本語説明文の単語を Word Embedding に入力し単語ベクトルを生成し，それを LSTM に入力して次の単語の単語出現確率を出力する．

誤差逆伝播法を用いてパラメータの調整を行うため Softmsax-cross entropy 関数を用いて，次に LSTM に入力する単語と単語の出現確率から誤差を算出する．誤差を LSTM と Word Embedding の各層に伝播させて,勾配降下法に基づいてパラメータを調整する．順・逆伝播を文末の単語まで繰り返し，学習させていき，テキスト及び画像を入力とする言語モデルを生成する．

入力には，CNN を用いて画像から抽出した画像特徴量ベクトル(4096×2000)と，日本語説明文の各単語(1×単語種類数次元の one-hot ベクトル)を Word Embedding により変換させた分散表現ベクトルを用いる．出力では，各単語の出現確率を持つ単語出現確率ベクトル(1×単語種類数次元)を得る．

2.3　キャプション生成

まずキャプション生成させたい画像を CNN に入力し画像の特徴量を抽出して LSTM に入力する．次に Word Embbeding を用いて文開始記号<S>を単語ベクトルにしたものを LSTM に入力し，LSTM の出力を Softmax 関数に入力し，単語出現確率ベクトルを求め，文開始記号<S>における次の単語の出現確率を求める．BeamSearch アルゴリズムより，単語の出現確率が高い上位 M 位と確率値の対数を保持する．次に単語出現確率が上位 M 位の単語 M 個を，LSTM に入力し出力を Softmax 関数に入力し，得られた確率を対数に変換後、和を計算して M×M 個の単語列を求める．そして，再び BeamSearch アルゴリズムより，単語列の出現確率の和が高い上位 M 位と確率値の対数を保持する．次に単語列の出現確率が高い上位 M 位の単語列の最後の単語を LSTM に入力し，上記と同じ処理を繰り返す．最後に，BeamSearch アルゴリズムより求めた，出現確率の和の平均が高い上位 M 位の単語列全てに文終端記号</S>が出現したら処理は終了する．作成された M 個の単語列の中から確率の和が高い上位 5 件をキャプション生成として出力する．

図 3 学習データセットの画像例

女性/と/男性/が/食事/を/し/て/いる
女性/は/左側/に/いる
男性/は/椅子/の/上/に/胡座/を/かい/て/いる
女性/は/右手/に/フォーク/を/持っ/て/いる
顎髭/を/生やし/た/男性/が/いる

図 4 学習データセットの日本語説明文の例

3 学習データセット

本研究で作成した学習データセットについて説明する．作成したデータセットは画像とその画像の日本語説明文から構成されている．

3.1 画像

画像は，人物が中心に映っているものを MPII Human Pose Dataset より 1,400 枚, 室内で人物が中心に映っているものを MS-COCO より 600 枚の計 2,000 枚を人手により選択した. 画像のサイズは最大 1,920×1,080 画素から最小 300×168 画素と制限はない．学習データセットの画像の例を図 3 に示す.

3.2 日本語説明文

日本語説明文は，計 20 名で作成した．1 人 1 画像に対して 5 文を担当し，何をしているのか，持ち物は何か，何を着ているのか，人物の数，性別，場所に着目して 1 人あたり 100 枚分作成し合計 10,000 文を作成した.

また，考案した日本語説明文は，日本語形態素解析::無料 WEB 便利ツール[2] により，形態素解析を行い，形容詞，形容動詞，感動詞，副詞，連体詞，接続詞，接頭辞，名詞，動詞，助詞，助動詞，特殊の 12 の形態素ごとに分かち書きを行った. 分かち書きされた語を 1 単語単位として使用する．日本語説明文の例を図 4 に示す

4 実験

4.1 ハイパーパラメータ

実験を行う前に，より精度の良い文章を生成するためのハイパーパラメータとデータセットの調整を行った.

ハイパーパラメータの調整には，3 節で作成した学習データセット(画像:2,000 枚，日本語説明文:10,000 文)を使用した. ハイパーパラメータは，学習に使用する単語の最低出現回数の Word_mini_count，勾配降下法の最適化，過学習抑制の DropOut，LSTM の入出力と Word Embedding の出力との共通した次元数の com_dim，ミニバッチの学習サンプルのサイズ数の batch_size の 5 つに関して調整を行い，データセットでは，データセット内の訓練データとテストデータの比率を調整した.

調整の結果，本実験では，各ハイパーパラメータの値は， Word_mini_count=1， 最適化は Adam， DropOut=0.5，com_dim=512，batch_size=256 とし，データセットに対しては，訓練データを 1,200 枚，テストデータを 600 枚(2:1)に振り分けたものを使用した.

4.2 実験設定

今回の実験では, 3.1 節と同条件で新たに 100 枚の画像を集め，100 枚の画像に対して 1 画像につき 5 文，合計 500 文のキャプションを生成し，以下の 2 つの評価実験を行った. 評価実験に使用する画像は，学習データセットに使用した画像と同様に集めた.

- **内容一致**

キャプション生成システム DeepWatcher で生成した文章と画像の内容がどのくらい合っているかを「完全に合っている」「一部分合っている」「間違っている」の 3 項目に第 1 著者が人手で分類し評価を行った.

- **特徴一致**

キャプション生成システムで生成した文章が画像に写っている人間の，性別，人数，持ち物・容姿，動作，場所の特徴を出力することができているのか，正しく出力しているかを第 1 著者が人手で次のように分類し評価を行った.

① キャプション生成システムの生成した各文において，人物に関する性別，人数，持ち物・容姿，動作，場所についての特徴の出現頻度を数える

② ①の特徴のうち画像と特徴が一致した正しいものの数を数え，その割合を正解率とする．例えば「男性と女性が食事をしている」という文章では性別について出力された特徴数は「男性」と「女性」という単語から 2 つ．動作については「食事をしている」から 1 つ．持ち物・容姿，場所については特徴の出力なしとし，画像とキャプションの特徴が一致したものを数える.

[2] http://tools.metro-bb.com/api/keitaiso

女性はキッチンにいる
彼女はキッチンで土台の焦げを落としている
彼女は右手にスプーンを持っている
彼女はカーディガンを羽織っている
彼女は誕生日ケーキの蝋燭に息を吹き掛けようとしている

図 5　テストデータ内の画像およびキャプションの具体例

4.3　実験結果

図 5 に生成されたキャプションの具体例を示す.

- 内容一致

表 1 に結果を示す.

Deep Watcher により生成されたキャプション自体は自然な日本語文という印象を与えるものであった. しかし文章の内容としては 500 文中，完全に合っていたものが 137 文，一部合っていたものが 71 文，間違っているものが 292 文となり 41.6%が内容に合っている又は一部内容に合っており, 58.4%が内容と間違っているキャプションを生成した

- **特徴一致**

表 2 に結果を示す.

性別の特徴は 90%以上，動作の特徴は 60%以上の文で出力されたが,人数，持ち物・容姿，場所については，50%を下回り，出現頻度は低かった. 正解率としては性別，場所，人数，行動の順に高く，持ち物・容姿が一番低く 45%となった. 生成されたキャプションの具体例のように性別について 5 文出力している画像が 63 枚あれば，性別，人数，持ち物・容姿，行動，場所すべての特徴を 1 つ以上出力した画像は 3 枚あった.

表 1　内容一致の評価実験

評価項目	一致率
完全に合っている	137(27.4%)
一部合っている	71(14.2%)
間違っている	292(58.4%)

表 2　特徴一致の評価実験

特徴	出力された特徴数	正解数	正解率
性別	459	399	86.9%
人数	107	66	61.6%
持ち物・容姿	201	77	38.3%
動作	329	153	46.5%
場所	72	49	68.0%

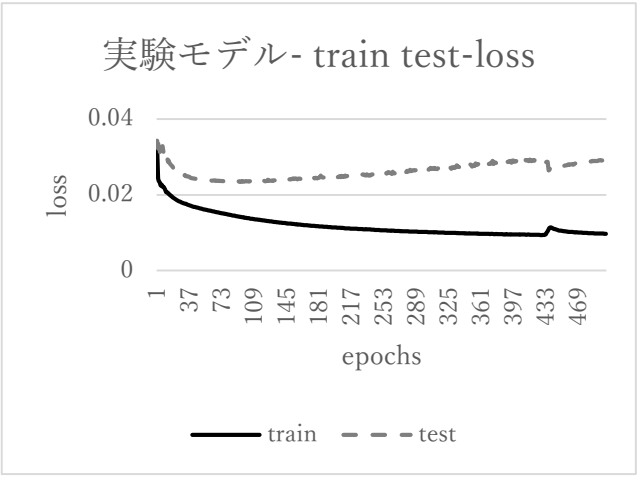

図 6　実験モデルの epoch 数による train と test の loss 値の変化

5　考察

Deep Watcher が生成したキャプションは，文章と画像の内容が一致する確率が低く，Deep Watcher には改善の余地がある．しかし，”カヌー”，”釣り”，”キッチン”，”体育館”など正確に画像の特徴をとらえているものもあった.

キャプションの内容の正確性を阻んでいる要因のの一つに，過学習が考えられる．図 6 に実験モデルの train と test の loss 値を示す．図 6 より，train と test の loss 値が乖離しているため過学習を起こしていたと考えられる．評価画像のうち水辺の画像は 10 枚（50 文）であり，内容に合っている・一部内容に合っている割合は，水辺の画像が 52%，水辺以外の画像は，40.4%と 11.6%の差があったため，水辺の画像に対しての特徴を強く学習してしまい過学習が起きたといえる.

特徴一致では多くの文に性別について出力された. これは学習データセットのキャプションのほとんどに性別についての記述があったためであると考えら

れる．学習データセットのキャプション 10,000 文中性別の記述があったものは 8,180 文であった．

6 おわりに

本研究では，周辺状況を視覚的に理解した上で対話するロボット搭載型対話システムの開発を目的として，日本語キャプション生成システム Deep Watcher を開発した. また対話に応用できるキャプションについて調査を行った．

今後の課題は，キャプション生成した文章と画像の内容がどのような画像を選んだ場合にも合うように，また正確に特徴をとらえるように，さらなる文章出力精度の向上が挙げられる．そのためには，学習データセットの増加による過学習の改善と CNN に対して，より特徴量の認識精度が高い Deep Residual Learning(ResNet)[8]による画像特徴量の抽出を検討する．また人物が中心的な周辺状況を視覚的に理解する対話システムの開発目的に沿うよう,既存の学習済み CNN を使用するのではなく，人物認識に特化・限定した独自の学習済み CNN の作成も検討する．

謝辞

データセット構築に協力していただいた青山学院大学の学生有志に感謝致します.

参考文献

[1] A.Farhadi,M.Hejrati,M.A.Sadeghi,P.Young,C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV. 2010.

[2] A.Karpathy,L.Fei-Fei:DeepVisual-SemanticAlignments for Generating Image Descriptions. In CVPR,pages 3128-3137, 2015.

[3] O.Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell:A neural image caption generator. In CVPR,pages 3156-3164, 2015.

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar,andC.L.Zitnick.MicrosoftCOCOcaptions: Datacollectionand evaluationserver. arXivpreprint,1504.00325,2015..

[5] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions,of,theAssociationforComputationalLinguistics,2:67?78,2014.

[6] K.Simonya and A.Zisserman:Very Deep Convolutional Networks,for,LargeScale,Image,Recognition.arXiv:1409.1556.2014

[7] F. A. Gers, J. Schmidhuber, and F. Cummins.Learning to forget: Continual prediction withLSTM. Technical Report IDSIA-01-99, IDSIA,Lugano, CH, 1999.

[8] H. Kaiming, Z.Xiangyu, R.Shaoqing, and S.Jian . "Deep Residual,Learning,for,Image,Recognition."InCVPR,pages 770-778, 2016.