2R-08

深層学習を用いた画像を説明する文生成手法の一考察

松尾映里† 小林一郎‡ 西本伸志§ 西田知史§ 麻生英樹¶
†お茶の水女子大学 理学部情報科学科 ‡お茶の水女子大学 基幹研究院自然科学系
§情報通信研究機構 脳情報通信融合研究センター ¶産業技術総合研究所 人工知能研究センター

1 はじめに

近年，ニューラルネットワークを用いた深層学習 (Deep Learning) によってこれまでの研究成果をさらに飛躍させる技術が次々と開発されている．難解な数値データを可読な文章として自然言語情報に変換するタスクにおいても広く適用され，中でも画像に映る事象を言葉で説明するキャプション付けの手法は、挑戦的な課題ながら既に多くの先行研究が報告されている．

本研究では，Attention Mechanism を導入した画像説明文生成モデルを先行研究 [1] に従って構築し，その有効性を確認するとともに，その手法を画像ではなく様々な定量データからの文生成手法への転用へと繋げていくことを目指す．

2 Encoder-Decoder Network

Encoder-Decoder Network(Enc-DecNet) とは，機械翻訳や情報表現の変換に用いられる深層学習のモデル構築手法である [2]．Encoder，Decoder の役割を果たす 2 つの深層学習モデルを組み合わせることで，入力を中間表現に変換 (encode) し，再び変換 (decode) して別の形に出力するという形で実現される．

本手法では先行研究 [1] と同様に，Encoder に VGGNet，Decoder に LSTM-LM を採用した Enc-DecNet に Attention Mechanism を導入したモデルを構築する．

2.1 CNN (VGGNet)

本手法で Encoder として用いる VGGNet は，画像の特徴量抽出に効果的な深層学習のモデルである Convolutional Neural Network(CNN) の一種である [3].

CNN では，多チャネルの画像に小サイズの二次元フィルタを畳み込む演算を行うことで画像の持つ局所的な特徴を抽出する Convolution 層と，その多チャネル画像の小領域での値を一つの値に集約し解像度を落とすことで抽出された特徴の位置が若干変化しても取り出される特徴はほとんど変化しないという特徴の不変性を獲得する Pooling 層を複数積み重ね、最後に通常の全結合層を数層重ねて出力を計算する．

A Study on Generating a Caption for a Picture by means of Deep Learning
†Eri MATSUO(g1220535@is.ocha.ac.jp),
‡Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
§Shinji NISHIMOTO(nishimoto@nict.go.jp)
§Satoshi NISHIDA(s-nishida@nict.go.jp)
¶Hideki ASOH(h.asoh@aist.go.jp)

2.2 RNN-LM (LSTM-LM)

本手法で Decoder として用いる Long Short-Term Memory-Language Model(LSTM-LM) は，時系列データ対応の深層学習モデル Recurrent Neural Network (RNN) による言語モデル RNN-LM の一種である [2][4].

RNN は隠れ状態 (計算時の変数) の情報を次時刻の入力とすることで過去の履歴を利用した時系列解析を行うモデルで，RNN-LM は過去の文脈 (t-1 個の単語) から t 番目の単語として各語が選ばれる確率を算出するモデルとなる．1 時刻前の隠れ状態 (時刻 1〜t-1 の単語情報)，1 時刻前の予測結果 (時刻 t-1 の単語)，外部情報 (本手法では中間表現に相当) の 3 つを入力とし，逐次的に次の単語の予測を繰り返して文章を生成する．

2.3 Attention Mechanism

Attention Mechanism[2] は，Enc-DecNet に導入することで，出力の各要素ごとに注目すべき入力要素を自動的に学習するシステムである．画像の説明文を生成する手法においては，各語の生成時に画像のどこに注目すべきかを考慮した，より人間の情報処理機構を捉えたプロセスでの文生成を実現する．

従来の Enc-DecNet では Encoder の出力した単一の中間表現をそのまま Decoder の入力として与えるが，Attention Mechanism では，Encoder に複数の中間表現を出力させ，各中間表現に重み係数 (注目度) をかけた重み付き和を Decoder の入力として与える．重み係数は各時刻ごとに 1 時刻前の Decoder の状態と中間表現を入力としたニューラルネットワークで計算され，深層学習のモデルの一部として同時に学習される．

3 実装手法

step 1. Encoder；VGGNet による特徴量の抽出
静止画を入力として VGGNet で特徴量を抽出．Attention Mechanism 適用のため，VGGNet の途中で処理を打ち切り，全結合層直前の 512x14x14 次元のものを Encoder の出力とする．出力された中間表現集合は静止画を重複ありで 512 個に分割した 14x14 小領域の特徴量に相当する．

step 2. Attention Mechanism による重み付き和処理
step 1. において計算された中間表現の集合に対し，1 時刻前の Decoder(LSTM) の隠れ状態を元にニューラルネットで算出した重み係数をかけ，重み付き和を導出．

step 3. Decoder；LSTM-LM による単語予測
step 2. において計算された重み付き和，および 1 時刻前の Decoder(LSTM) の隠れ状態を入力として，LSTM-LM で単語を出力．

step 4. 単語出力の反復による文生成
文末記号が出力されるか設定した最大文長を超えるまで step 2-3 を繰り返し，1 語ずつ出力することで文を生成．

4 実験

システムの実装に際しては，深層学習のフレームワーク Chainer を利用し，train・test 用データセットとして静止画とその説明文のペアからなる Microsoft COCO を使用した．本研究では 414,113 個の train 用データのうち，94,500 個まで学習した結果を提示する．

学習に関するハイパーパラメータの数値設定については，学習率を 0.001 とする先行研究 [1] の設定と，深層学習の効率化手法を取り入れ，学習率を 1.0(パラメータ更新毎に ×0.999)，勾配閾値 5，L2 正則化項 0.005 とした設定の，2 通りについて実験を行った．その他のハイパーパラメータは VGGNet の出力次元に揃え，各語は 512 次元のベクトルで表現し，LSTM のユニット数は各層 14×14=196 次元に設定した．また，train 用データの中で 50 回以上出現した 3469 語を語彙とした．

学習するパラメータは Attention Mechanism 及び Decoder(LSTM) の重み係数とし，[-0.1,0.1] でランダムに初期化した．Encoder(VGGNet) は事前学習し，更新を行わない．学習アルゴリズムは確率的勾配降下法，誤差関数は交差エントロピーを用いている．

4.1 実験結果

設定した 2 通りのハイパーパラメータ (先行研究／手法導入) について，test 用画像からランダムに抽出した 2 つの画像に対して生成した説明文，およびその主語生成時の Attention の重みを，それぞれ図 1，図 2 に示す．また，表 1 のように train データ数毎に perplexity を記録し，その減少により学習の進度を確認した．

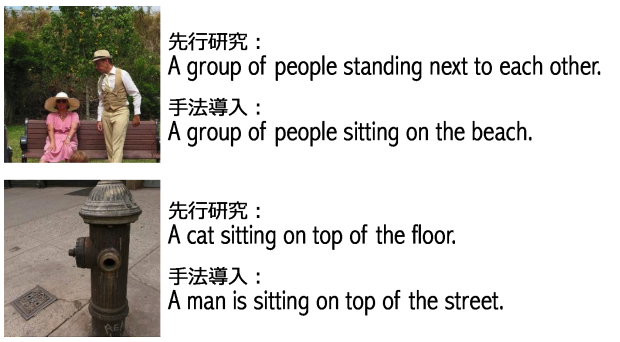

図 1: 生成した説明文の例．画像はランダムに抽出．

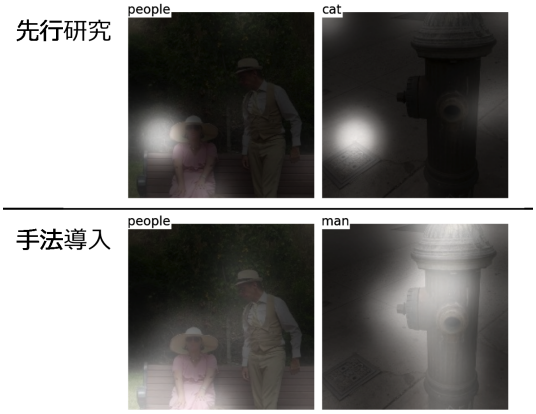

図 2: 主語生成時の Attention を白く可視化した例．

表 1: training 時の perplexity の変化

データ数	先行研究	手法導入
7000	147.83	240.17
24500	66.52	69.47
42000	50.87	66.24
59500	42.96	79.74
77000	37.77	64.35
94500	35.04	61.59

4.2 考察

出力された説明文は逐次出力の文章としては文意を読み取ることが十分可能であり，画像を正確に説明できていない要素も見受けられるものの，おおむね画像の大意を認識し表現していると評価できる．興味深いのは，どちらのハイパーパラメータ値設定でも人間は認識できているが，ポストは cat あるいは man と誤認している点である．これは，train データに人物画像が多く含まれるのに対し，ポストの画像は数個しか存在しないことから，ポストという概念の獲得にはデータ量が不十分であったことが原因の一つと考えられる．

ハイパーパラメータ値の設定による差が顕著に現れたのは，Attention の学習結果である．先行研究の設定では Attention の学習が不十分だが，効率化手法による設定では画像中の注目すべき部分を的確に捉えており，導入手法が深層学習の学習効率を向上させたと推測される．一方、生成文および perplexity は先行研究の方が優れており、今後学習が進んで Attention が獲得されれば、値が適切に調整されている先行研究の方が全体として良い結果となる可能性が考えられる。

5 おわりに

本稿では，Encoder-Decoder Network に Attention Mechanism を導入した深層学習モデルを構築し，画像からの説明文生成を行ってその有効性を確認した．また，ハイパーパラメータ値設定による学習結果および学習効率への影響を観察した．

今後の課題として，train データの追加や数値設定の見直しによる精度向上，BLEU や METEOR などの評価指標を用いての実験結果の更なる考察および他手法との比較などが挙げられる．また本手法の転用により，functional Magnetic Resonance Imaging(fMRI) を用いて記録した脳神経活動 (BOLD 信号) データを分析し自然言語文として理解する手法の構築を検討している．

参考文献

[1] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, ”Show, attend and tell: Neural image caption generation with visual attention,” in ICML ’2015, 2015.

[2] K. Cho, A. Courville, Y. Bengio. “ Describing Multimedia Content using Attention-based Encoder-Decoder Networks. ” CoRR, abs/1507.01053, 2015.

[3] K. Simonyan, A. Zisserman,“ Very deep convolutional networks for large-scale image recognition, ” in ICLR, 2015.

[4] S. Hochreiter, J. Schmidhuber, “ Long short-term memory, ”Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.