深層学習における注意機構に対する一考察

廣田　梨那　(指導教員：小林一郎)

1　はじめに

近年, 医療分野など人の命に直接関わる場面での AI の活用が期待されている. AI は高い精度での予測や認識ができる一方, 結果に対する説明が困難であるため, 実用化の面で大きな課題を抱えている. これらの背景を踏まえて, 本研究では, 深層学習において重要な役割を果たしている注意機構 (Attention Mechanism) について, 先行研究にて提案された画像説明文生成のメカニズム [2] を再考することで, 実世界で信頼できる説明可能な AI（Explainable AI）の開発を目指す.

2　深層学習における注意機構の役割

本研究では, 機械翻訳やメディア変換に用いられる深層学習のモデルである Encoder-Decoder Network（Enc-DecNet）[1] を用いる. Enc-DecNet は, Encoder と 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, 再び復号（decode）して別の表現を出力する.

Xu ら [2] は, 同様の Enc-DecNet に注意機構を導入したモデルを提案し, 生成文の精度向上を示した. 注意機構は, Enc-DecNet に導入することで出力の各要素ごとに着目すべき入力要素を自動的に学習するシステムであり, 画像の説明文を生成する手法においては, 注目すべき画像の箇所を考慮した人間の情報処理に近いプロセスでの文生成を実現する.

3　注意機構の基本性能検証

まず, 先行研究 [2] における, 深層学習を用いた画像説明文生成プロセスを以下に示す. 図 1に概要を示す.

step 1. Encoder ; VGGNet による特徴量の抽出
静止画を入力として VGGNet で画像特徴量を抽出. VGGNet の途中層から 512 個の 14×14 次元データを Encoder の出力とする.

step 2. 中間表現の処理
step 1. において計算された中間表現集合の重み付き和を Decoder に渡す入力として算出. 重み係数は 1 単語前の Decoder（LSTM）の隠れ状態と 512 個の中間表現から計算される.

step 3. Decoder ; LSTM による単語予測
step 2. で計算された中間表現および 1 単語前の Decoder の隠れ状態を入力として, LSTM 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 語ずつ出力して文章を生成.

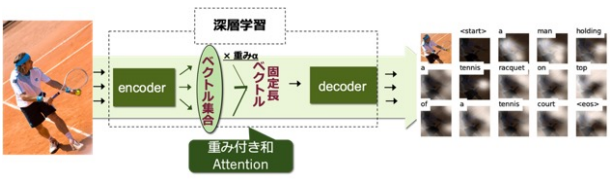

図1　画像データからの文章生成

Encoder の中間表現を $\boldsymbol{a}$, Decoder の隠れ状態を $\boldsymbol{h}$ とする時, 注意機構は以下のように計算される. 中間表現と 1 単語前の Decoder の隠れ状態を入力し注意機構を計算することで, 次の単語を出力すると同時に, その単語に対応する画像箇所が可視化可能となる.

$$e_{ti} = f_{att}(\mathbf{a_i}, h_{t-1}),$$
$$\alpha_{ti} = \frac{exp(e_{ti})}{\sum_{k=1}^{L} exp(e_{tk})},$$
$$Z_t = \sum_{i=1}^{L} \alpha_{ti}\mathbf{a_i}$$

本研究では，上記の画像説明文生成プロセスにおける注意機構の基本性能を様々な実験設定を通じて検証する.

3.1　先行研究に基づく画像説明文生成モデル

train 用データセットとして 82,783 ペアの静止画とその説明文からなる Microsoft COCO[*1]を使用する. 学習に関するハイパーパラメータは, 学習率を 0.0001 とした他, 学習アルゴリズムは Adam を用い, Encoder は事前学習した VGGNet を用いる.

3.2　実験 1: 勾配降下法の比較

勾配法毎の BLEU 値を表 1に, 勾配法 Adam と AdaGrad の LOSS 値を比較したものを図 2に示す.

一般にニューラルネットワークには予測精度を向上させるため, 回帰の損失を最小化するために, ネットワークで用いる重みの更新を行う必要がある. 先行研究では勾配降下法として Adam を用いたが, 出力結果に与える影響を調査するため, AdaGrad と比較した. Adagrad は各パラメータ方向の勾配の二乗和を保存して, 学習率をその平方根で割ることで稀な特徴に対して学習率を高めにする. 一方, エポックを重ねるごとに過去の学習率が累積していき, やがて 0 に漸近してしまうため, すぐに更新されなくなってしまうという問題も抱えている. Adam は勾配の平均と分散のモーメントを推定することで, パラメータ毎に適切なスケールで重みが更新されることを可能にする.

表 1, 図 2の結果から, 今回のデータに対しては勾配降下法として Adam を使用するのが適切であると言える.

図2　Adam と AdaGrad の LOSS 値

[*1] http://mscoco.org/

表1　勾配法毎の BLEU 値

勾配法	BLEU1	BLEU2	BLEU3	BLEU4
Adam	0.715	0.512	0.364	0.253
AdaGrad	0.540	0.258	0.073	0.026

3.3　実験 2: train データ数の比較

test 用画像に対する説明文および注意機構付き画像を図 3, 図 4, 図 5に，train 画像 20,696/41,392/82,783 枚毎の BLEU 値を表 2に示す．出力結果を比較することで, 今回のネットワークの大きさに対しては, データ量が多い方が適切な文章を生成できることが分かった.

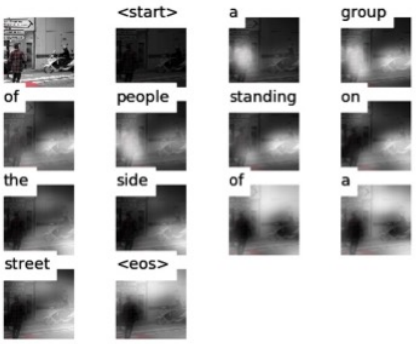

図3　train 画像 20,696 枚の結果

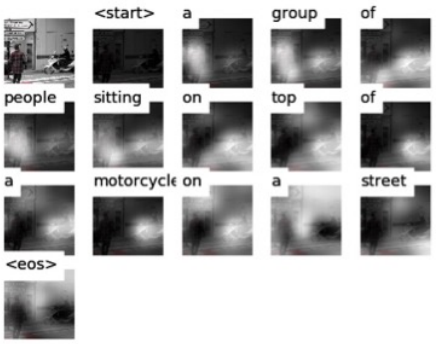

図4　train 画像 41,392 枚の結果

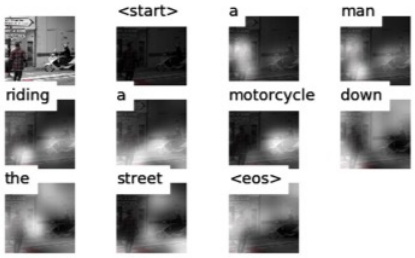

図5　train 画像 82,783 枚の結果

表2　train 画像枚数毎の BLEU 値

データ数	BLEU1	BLEU2	BLEU3	BLEU4
20,696	0.687	0.480	0.332	0.221
41,392	0.703	0.496	0.347	0.237
82,783	0.715	0.512	0.364	0.253

3.4　実験 3: 使用するネットワークの比較

ResNet を用いて学習させた際の, test 用画像に対する説明文および注意機構付き画像を図 6に，ネットワーク毎の BLEU 値を表 3に示す.

先行研究で用いられている VGGNet は, 100 万枚を超える画像を 1000 個のオブジェクトカテゴリに分類するタスクで学習したパラメータをそのまま使用している．今回比較した ResNet は, ある層で求める最適な出力を学習するのではなく, 層の入力を参照した残差関数を学習することで最適化しやすくしている. それにより, 高い表現力と小さい誤差の両立を実現させた.

今回の実験では ResNet の方が BLEU 値がよく, 生成される文章も画像の特徴を上手く捉えられている.

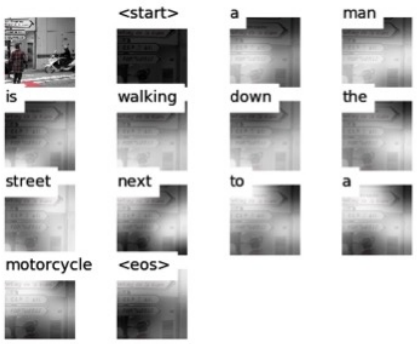

図6　ResNet

表3　ネットワーク毎の BLEU 値

NETWORK	BLEU1	BLEU2	BLEU3	BLEU4
VGGNet	0.715	0.512	0.364	0.253
ResNet	0.732	0.535	0.386	0.273

4　おわりに

本稿では，注意機構における基本性能についての検証を行った．その結果, 勾配法は Adam, ネットワークは ResNet を使用し, train データ量は多い方が良いと分かった.

今後の課題として, ピクセル毎に注意機構をつけることによる精度向上，分散表現を用いた実験結果の評価および考察などが挙げられる．また, Attention Branch Network [3] などについても調査を行い，注意機構に対するより詳細な考察を深めると共に説明可能な AI を実現する方法として開発を行いたい.

参考文献

[1] O.Vinyals, A.Toshev, S.Bengio, D.Erhan, ”Show and tell: a neural image caption generator,” in CVPR’ 2015, 2015.

[2] 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.

[3] , Hiroshi Fukui and Tsubasa Hirakawa and Takayoshi Yamashita and Hironobu Fujiyoshi, Attention Branch Network: Learning of Attention Mechanism for Visual Explanation, Computer Vision and Pattern Recognition, 2019.