キャプション生成のためのアテンション機構を用いたデータ拡張

◯ 岩村 紀与彦 †，ルイ笠原 純ユネス †，アレッサンドロ モロ †,§，山下 淳 †，淺間 一 †
†: 東京大学，§: ライテックス
iwamura@robot.t.u-tokyo.ac.jp

概要： 画像から自動的にキャプション（内容の描写）を生成する研究が行われている．これは，視覚障害者のために内容の描写など様々な用途がある．従来研究の多くは，発展の目覚ましい深層学習を用いているが，依然として解決困難な問題がある．その 1 つに，訓練データセットの種類が限られているため，特定の映像に過度に適合してしまい，汎化性能が低下する点が挙げられる．そこで，本研究では，アテンションを活用し，過適合している領域を考慮するデータ拡張手法を提案する．実験の結果より，本研究は従来研究よりも優れた精度を示した．
< キーワード > キャプション生成, 深層学習, データ拡張

1. 序論

映像から自動的にキャプション（内容の描写）を生成することは重要な課題であり，動画の索引作成や視覚障害者のために動画内容の描写など多くの用途に用いられる．従来は，人間が映像に対して手作業でキャプションを生成していたため，労力がかかっている．そのため，コンピュータを用いたキャプション生成の自動化が求められている．近年，自動的にキャプションを生成する方法として，ラーニングベースド手法が提案されている [1][2]．この手法では，大量の訓練データセットを学習させることで，人間が画像から特徴量を設計する必要がなく，未知の映像に対しても精度の高いキャプションを生成することが可能である．しかし，課題の 1 つに特定の映像に過度に適合（過適合）してしまい，汎化性能が低下する点が挙げられる [3]．

画像分類の分野では，過適合に対処するための手法として，画像や動画にノイズを加えて（マスク処理），訓練データセットの種類を増やすデータ拡張手法があり，精度の向上が報告されている [4]．しかし，この手法は画像領域中から無作為に部分領域をマスク処理をほどこしてデータ拡張するため，画像全体の特徴を活用する画像分類での効果は期待できるが，画像中の特定領域が特に重要となるキャプション生成において，過適合に対処できない場合がある．

キャプション生成の分野では，アテンション画像を用いた研究が存在する [5]．アテンション画像とは，モデルが画像から単語を生成するときに，画像中のどの領域を重視しているかを可視化したものである．アテンション画像を使うと，例えば，画像から「車」という単語を生成するときに，画像中のどの領域に注意を払っているかを理解することができる．

Fig. 1: 提案手法のコンセプト．

本研究では，過適合している領域を考慮するデータ拡張を提案することを目的とする．提案手法のコンセプトは，モデルが重視している領域にマスク処理をほどこし，訓練データセット中で特にモデルが重視している領域の種類を増やすことである．Fig. 1 に示すように，画像中のアテンション値が高い領域が過適合しやすいと仮定し，その領域に優先的にマスク処理を行う．具体的には，入力映像に対してデータ拡張手法を適用しマスク処理がなされたデータをキャプション生成モデルに入力する．そのデータ拡張として，アテンション機構を利用してキャプション中の単語と画像の部分領域の関係性が強い部分にマスク処理を加えたデータセットを活用し，画像からのキャプション生成を行う．

2. 提案手法

2.1 提案手法概要

第 1 章で述べたように，本研究のコンセプトは，キャプション生成モデルが画像中の全ての画素を平等に扱うのではなく，キャプション内容に対応する画像の部分領域を重視していると仮定し，その領域に優先的にマスク処理を加えることである．モデルの訓練時に，画像

と対応するキャプションの組が用いられ，例えば，「人」というキャプションを生成するときに，画像中の人間が写っている領域に反応するように，モデルが訓練されている．一方で，その他の部分領域は，キャプションに対応しないため，モデルを訓練するときにはさほど重要ではないという点から，部分領域を重視しているという仮定は着想を得ている．Fig. 2 に示すように，提案するマスク処理の流れは，画像からアテンション画像を獲得し，アテンション画像中の値が高い箇所（ピーク）を取得し，ランダムにマスク領域を選択し，その領域中に特定の閾値 T を超えるピーク n_{peak} 個が含まれる場合に，そのマスクを採用する．既存手法も画像に対してマスク処理をするが，領域の選択方法が無作為であり，本研究では特定の領域を選択する点が異なる．

本研究では，キャプション中の単語と画像の部分領域の関係性を活用するためにアテンション機構を用いる．まず，Step 1 として，入力する画像は学習ずみのモデルに渡されて，アテンション画像を出力する．この学習ずみのモデルは学習目標のモデルと同じ構造をもつニューラルネットワークであり，事前にデータセットを用いて学習されたモデルである．次に，Step 2 において，そのアテンション画像と画像を用いてデータ拡張を行い，学習目標のモデルを訓練する．このアテンション画像にはモデルが単語を生成するときに注意している画像領域情報が表れているため，その領域にマスク処理を施すことで，学習目標のモデルの過適合を妨げることを目指す．

2.2 アテンション機構を用いたデータ拡張

2.1 節で述べたように，アテンション機構を活用することでキャプション中の単語と画像の部分領域の関係性をとらえることができる．そして，関係性の高い領域にマスク処理を加えることでキャプション生成の精度向上を目指すのが提案手法のコンセプトである．提案手法は，各 epoch 毎に次の手順を実行する．まず，特定の画像に対して，キャプション中の単語数分のアテンション画像を獲得する．次に，単語数分のアテンション画像を全て足し合わせる．そして，値の最も高いピクセル群を選び，マスク処理を加えるピクセルを選択する．最後に，アテンション画像に対応する元の画像のピクセルにマスク処理を加えて訓練データとする．

(a) **学習ずみのアテンション機構の準備**

前述の通り，アテンション機構を活用するために，事前にデータセットを用いて学習ずみのキャプション生成モデルを準備する必要がある．本研究では Xu ら [5] のモデルを用いて事前に学習ずみモデルを用意する．

(b) **単語毎のアテンションの処理**

学習ずみのアテンション機構を用いて，画像から対応するキャプション中の単語数分のアテンション画像を取得する．その後，それらを単に足し合わせ，1 つのアテンション画像を作成する．例えば，画像に対応するキャプション中の単語が 3 単語とすると，3 つのアテンション画像を獲得し，それらを足し合わせ 1 つのアテンション画像とする．

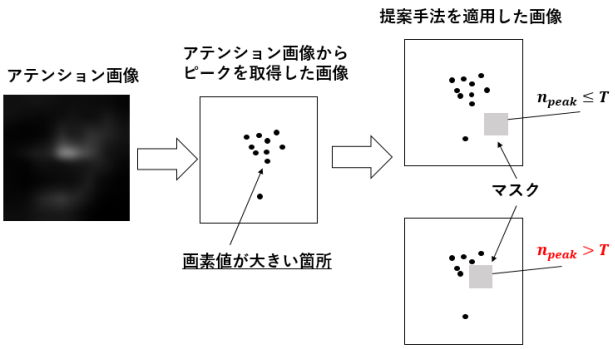

Fig. 2: 提案手法のマスク処理．

(c) **アテンション画像中の最も高いピクセル群の取得**

本研究では，単純に単語数分を加算したアテンション画像から値が大きいピクセルを上から 100 ピクセル数を選択する．これにより，値の低い領域（モデルがキャプションを生成するときにあまり注意しない部分領域）を除外することができる．

(d) **付与するノイズ及びノイズを加えるピクセルの選択**

先行研究 [4] と同様に，ノイズとして四角形のマスクを付与する．ここでマスク領域内のピクセルの画素値はランダムな値で埋める．また，ノイズを加えるピクセルの選択は，ノイズを加えるマスク領域が選択した 100 ピクセル中から 10 ピクセルより多く含むようにする．

3. 実験

画像分類の分野で用いられているデータ拡張手法及び提案手法の有効性を検証するために画像を用いたキャプション生成の実験を行う．

3.1 データセット

本実験で用いるデータセットには，画像 1 枚当たり 5 個程度のキャプションが付与された MSCOCO'14 データセット [6] を用いる．Fig. 3 に画像とキャプションの概念図を載せる．画像と対応するキャプション（1 つの文章）を表している．このデータセットは 123,000 枚の画像を含んでいる．Table. 1 に示すように，検証データ，テストデータについては，それぞれ 5,000 枚を使う．データの前処理として，全ての文章を小文字に，”や-などの英数字以外を無視．訓練データ中で 5 回未満の出現頻度の単語はフィルタリングする．その結果として，8791 語彙数となる．

Table.1: データセットの種類

MSCOCO14	訓練データ	検証データ	テストデータ
枚数 [枚]	113,000	5,000	5,000

Table.2: キャプション生成結果

	BLEU1	BLEU2	BLEU3	BLEU4
baseline[5]	72.6	55.7	41.4	30.5
ref[4]	72.7	55.8	41.5	30.6
提案手法	**73.1**	**56.1**	**41.7**	**30.8**

3.2 評価指標

本研究では，評価のために，キャプション生成の分野で一般的に用いられている BLEU-N（N=1,2,3,4）評価指標 [7] を用いる:

$$BLEU_N = min(1, e^{1-\frac{r}{c}}) \cdot e^{\frac{1}{N}\sum_{n=1}^{N} \log p_n} \quad (1)$$

ここで，r，c は参照する文章と生成された文章の長さ，p_n は修正された n-gram 精度である．

3.3 キャプション生成実験

実験に使用される全ての画像は，256×256 ピクセルのサイズのカラー画像であり，チャンネル毎に平均と標準偏差を用いて正規化されている．キャプション生成のモデルは，Show，Attend and Tell[5] を用いる．パラメータの更新方法には Adam オプティマイザーを活用した．また，バッチサイズは 32，デコーダーの学習率は 4e-4，エポックは 120 で学習を行う．学習・検証・テストにおいて，GPU は NVIDIA 製の GeForce RTX 2080Ti を，CPU は Intel 製の Core i9-7900X を使用する．また，学習ずみのアテンション機構を準備するために，前述の条件でキャプション生成モデルの学習を行っている．そして，本実験で加えるマスクは四角形であり，その 1 辺の大きさは，24 と設定する．

実験結果を Table. 2 に示す．この Table. 2 は，5,000 枚のテストデータにおけるキャプション生成の精度を示している．一般的にキャプション生成の分野で用いられている BLEU-4 評価指標は，0-100 の範囲をとり，100 が最も良い値を意味している．baseline はデータ拡張を適用していない手法である．また，ref は画像分類の分野で行われているデータ拡張手法 [4] を baseline に適用したものであり，baseline より高い精度であった．そして，提案手法の結果が最も高い数値を示していることが確認できる．この結果より，提案するデータ拡張は有効であった．

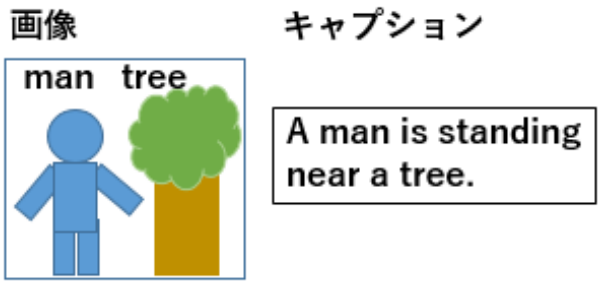

Fig. 3: 実験で用いるデータセット（キャプションと画像）の概念図．

4. 結論

本研究では，過適合している領域を考慮するデータ拡張の提案を行った．具体的には，アテンション機構を利用してキャプション中の単語と画像の部分領域の関係性が強い部分にマスク処理を加える手法を提案した．キャプション生成実験では，提案手法による精度向上を達成した．

今後の予定として，画像に付与するマスクのサイズや，マスク領域内のピーク値の個数に関する閾値 T を変化させた場合のキャプション生成の精度の調査を行うことが挙げられる．

謝辞

本研究の一部は，東京大学情報基盤センターの SGI Rackable C2112-4GP3/C1102-GP8 (Reedbush-U/H/L) を用いて遂行された．

References

[1] Mao Junhua, Xu Wei, Yang Yi, Wang Jiang and Yuille Alan L. Explain Images with Multimodal Recurrent Neural Networks. *arXiv preprint arXiv:1410.1090*, 2014.

[2] Vinyals Oriol, Toshev Alexander, Bengio Samy and Erhan. Show and Tell: A Neural Image Caption Generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164. 2015.

[3] Cheng Wang, Haojin Yang, Christian Bartz and Christoph Meinel. Image Captioning with Deep Bidirectional LSTMs. *Proceedings of the ACM International Conference on Multimedia*, pages 988–997, 2016.

[4] DeVries Terrance and W.Taylor Graham. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[5] Xu Kelvin, Ba Jimmy Lei, Kiros Ryan, Cho Kyunghyun, Courville Aaron, dinov Ruslan Salakhut, Zemel Richard S. and Bengio Yoshua. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the International Conference on Machine Learning*, pages 2048–2057, 2015.

[6] Lin Tsung-Yi, Maire Michael, Belongie Serge, Hays James, Perona Pietro, Ramanan Deva, Dollár Piotr and Zitnick C.Lanwrence. Microsoft COCO: Common

Objects in Context. *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014.

[7] Papineni Kishore, Roukos Salim, Ward Todd and Zhu Wei-Jing. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 311–318, 2002.