

アテンション機構を用いたクロップとマスクによる キャプション生成のためのデータ拡張*

岩村紀与彦** ルイ笠原 純ユネス*** モロ アレッサンドロ†,†† 山下 淳*** 浅間 一***

Image Captioning with Data Augmentation Using Cropping and Mask Based on Attention Image

Kiyohiko IWAMURA, Jun Younes LOUHI KASAHARA, Alessandro MORO, Atsushi YAMASHITA and Hajime ASAMA

Automatic image captioning has various important applications such as the depiction of contents for the visually impaired. Most approaches use Deep Learning and have achieved remarkable results. However there are still some unresolved issues. One of them is the overfitting of the trained model to specific images, usually caused by limited training dataset sizes. In order to augment the training dataset size in such scenarios, previous researches proposed data augmentation using random cropping or mask. However, those do not specifically target overfitted regions in images and, therefore, may remove areas in images that are needed to generate captions and lower performance. In this study, we propose a novel data augmentation method that targets specifically regions in images subject to overfitting by using attention. Experimental results show that the proposed method allows generation of better image captions.

Key words: data augmentation, deep learning, attention, mask, cropping

1. 序 論

画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る。利用例として、動画の索引作成や視覚障害者のために動画内容の描写や、SNS 上の大量の画像を自動索引するなど多くの用途が挙げられる。従来は、人間が手作業で画像に対してキャプションを生成していたため、労力がかかっていた。そこで、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。

キャプション生成の自動化に関しては、近年、深層学習を用いた手法が提案されている¹⁾²⁾。この手法は、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、問題点の 1 つに訓練データセットの数が限られているため、特定の画像に過度に適合（過適合）してしまい、汎化性能が低下する点がある³⁾。ここでの過適合とは、例えば、複数枚の車の画像を用いてモデルを訓練する際に、画像の枚数が限られているために車のナンバープレートに過適合し、ナンバープレートを車と誤検出することを指す。これにより、車側面などのナンバープレートが映らない画像から「車」というキャプションを生成することが困難になる場合がある。

訓練データセットの数の制限によって過適合が発生する問題に対処するために、訓練データセットの数を増やすデータ拡張手法が提案され、その有効性が報告されている。画像に対してデータ拡張を行う手法として、画像処理技術を用いて、訓練データとなる画像にノイズを加えてデータ数を増やす手法が

ある。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として四角形のマスクを加えて画像の種類を増やす、データ拡張を行う手法があり、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にマスク処理をほどこしてデータ拡張を行うため、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重要となることが多い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。また、事前に設定した画像の特定領域から無作為に切り出しを行うクロップ処理を組み合わせ、データ拡張を行う手法が報告されているが³⁾、この方法ではキャプション生成においてモデルが必要な領域（車や人間など）を削除する可能性がある。

これらの手法は、画像中から無作為にデータ拡張を行う領域を選択しているが、キャプション生成では車や人間などの画像中の特定領域が必要となるため、その領域に優先的にデータ拡張を行う手法が報告されている⁵⁾。その領域情報の取得には、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参照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活用して、画像中の特定領域にマスク処理をほどこして、データ拡張を行う手法が提案されている⁵⁾。この手法の利点として、モデルが参照している画像領域のデータ数を直接増やすことが可能となる。具体的に車の場合では、車の領域に優先的にマスクを付与してデータ数を増やすことで、例えば、ナンバープレート的一部分にマスクを付与する画像が作成され、ナンバープレート全体の画像特徴量と「車」という単語を一对一で結び付けることを困難にする。これにより、ナンバープレートなどの車の局所情報のみだけでなく、車全体の特徴を考慮することが期待できる。しかし、この方法ではマスクの大きさがアテンション領域の大きさにかかわらず一定であることから、マスクが車や人間などのキャプション生成に必要な領域を削除する問題が発生すると考えられる。

本研究では、キャプション生成においてモデルが重視する画

* 原稿受付 令和 2 年 5 月 6 日

掲載決定 令和 2 年 8 月 20 日

** 学生会員 東京大学大学院（東京都文京区本郷 7-3-1）

*** 正 会 員 東京大学大学院

† 東京大学大学院

†† ライテックス（東京都立川市柴崎町 3-5-11）

像中の特定領域の種類を増やすために、アテンション領域の大きさに応じて変化するマスク、クロップを用いたデータ拡張手法を提案することを目的とする。提案手法のコンセプトは、アテンション画像中の値が高い領域内に過適合しやすい領域が内包されていると仮定している。車の例では、アテンション画像中の値が高い領域が車、過適合しやすい領域が車のナンバープレートを含んでおり、車とナンバープレートは両方ともアテンション画像中の値が高い。

図1に示すように、本研究では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、その領域の種類を増やすためにマスク処理及びクロップ処理を行う。具体的には、アテンション値が高い画像領域の種類を複数に増やして、1種類の画像領域のみに適合させないことで、モデルの過適合を妨げることを行う。例えば、車の場合では、車の領域に優先的にマスクを付与してデータ数を増やす。これにより、ナンバープレートの一部にマスクを付与する画像を生成して、ナンバープレート全体の画像特徴量と「車」という単語を一对一で結び付けることを困難にする。これは、ナンバープレートという特にアテンション値が高い画像の種類を複数に増やし、1種類の特定領域のみに適合させないことに相当する。また、アテンション値の取得に関しては、車の写っている入力画像と同じ画像サイズのアテンション画像を用いる。このアテンション画像の各画素は、アテンション値を表しており、入力画像と画素毎に対応する。これにより、ナンバープレートのような画像の特定領域から「車」などの単語を生成することを防ぐ。つまり、ナンバープレートなどの車の局所情報のみだけではなく、車全体の特徴を考慮することが期待できると考えられる。そこで、入力画像に対してデータ拡張手法を適用しマスク処理がほどこされた画像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箇所にマスク処理及びクロップ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. 特定領域を考慮するデータ拡張手法

2.1 概要

本研究では、キャプション生成における過適合問題に対して、アテンション画像 (I_{att}) を用いた画像処理を行うデータ拡張に着目する。アテンション画像 (I_{att}) 中の値の高い領域を用いるキャプション生成モデルは優れた精度を報告していることから⁶⁾、値の高い領域はキャプション生成に良い影響を与える。そして、訓練データセットの数が限られている場合に、モデルの学習が進むにつれて、この領域内部は参照されるため、過適合が発生しやすいと考えられる。そこで、この特定領域のデータ数を増やすことを考え、その特定領域の取得には、アテンション画像 (I_{att}) を活用する。車の例では、車の映る領域は、アテンション画像 (I_{att}) 中の値が高く、データ拡張を行うべきである。

データを増やす提案手法の流れを図2に示す。まず、入力画像に対して、学習済みのモデルを活用してアテンション画像 (I_{att}) を生成する。本手法におけるアテンション画像 (I_{att}) とは、画像の各画素毎に値を持つ白黒の濃淡で表現される画像であり、値が高い (白い) ほどモデルがキャプション生成時に参照していることを意味し、本研究でデータ拡張を行いたい領域であることを表す。このアテンション画像 (I_{att}) は画像が

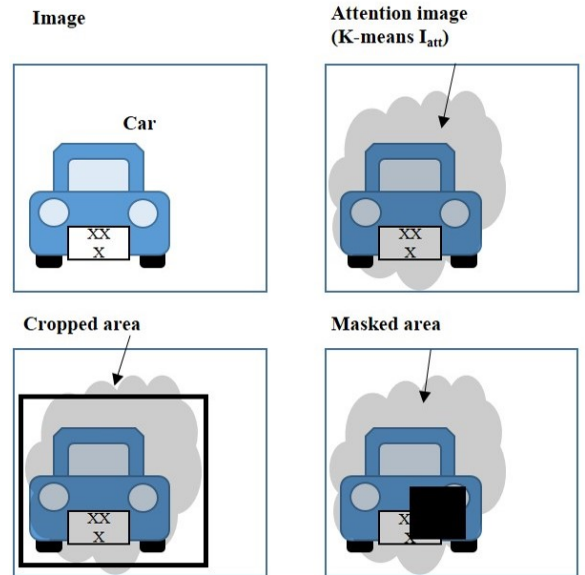


Fig. 1 Concept of the proposed method. When the model generates a caption, high value areas of the attention image ($K\text{-means } I_{att}$) are recognized as areas that are prone to overfitting, such as license plates. Therefore, data augmentation is performed in order to increase the number of data samples with variations in those high attention value areas

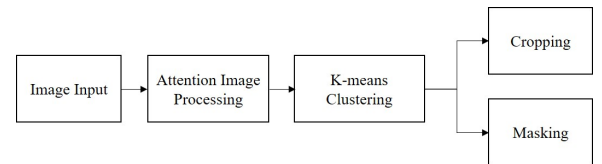


Fig. 2 Outline of the proposed method for performing data augmentation with mask and cropping

ら生成する単語毎に変化する。次に、獲得したアテンション画像 (I_{att}) に K-means クラスタリング処理を行い、アテンション画像中 (I_{att}) のモデルが強く参照する箇所の領域を推定する。その後、1) K-means 処理をほどこされたアテンション画像 ($K\text{-means } I_{att}$) を用いたクロップ処理、2) K-means 処理をほどこされたアテンション画像 ($K\text{-means } I_{att}$) を用いたマスク処理、3) データに変化を加えない、の中からランダムに1つを実行し、新たな訓練データを作成する。

以上のように本研究では、アテンション画像 (I_{att}) 中の値が高い領域を残して、そのデータ数を増やすために、アテンション領域の大きさに応じて変化するマスク、クロップを用いたデータ拡張手法を新規に提案する。

2.2 アテンション画像

アテンション画像 (I_{att}) にはモデルが単語を生成するときに注視している画像領域情報が表れているため、その領域情報を活用することで、モデルの過適合を妨げることを目指す。具体的には、アテンション値が高い画像領域の種類を複数に増やして、1種類の画像領域のみに適合させないことで、モデルの過適合を妨げることを行う。例えば、バイクの場合では、バイクの領域に優先的にマスクを付与してデータ数を増やす。これにより、ナンバープレートの一部にマスクを付与する画像を生成して、ナンバープレート全体の画像特徴量と「バイク」と

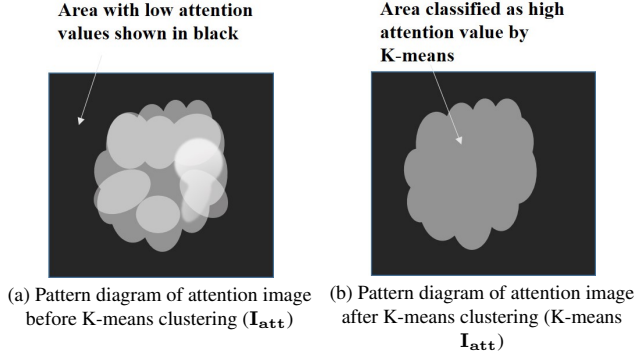


Fig. 3 Separation of high and low attention image areas using K-means clustering

いう単語を一对一で結び付けることを困難にする。これは、ナンバープレートという特にアテンション値が高い画像の種類を複数に増やし、1種類の特定領域にのみ適合させないことに相当する。また、アテンション値の取得に関しては、バイクの写っている入力画像と同じ画像サイズのアテンション画像 (I_k) を用いる。このアテンション画像 (I_k) の各画素は、アテンション値を表しており、入力画像と画素毎に対応する。これにより、ナンバープレートのような画像の特定領域から「バイク」などの単語を生成することを防ぐ。つまり、ナンバープレートなどのバイクの局所情報のみだけではなく、バイク全体の特徴を考慮することが期待できると考えられる。まず、入力する画像からアテンション画像 (I_{att}) を生成するために、事前にデータセットを用いて学習済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⁶⁾ のモデルを用いて事前に学習済みモデルを用意する。この学習済みモデルに画像を入力し、モデルが生成するキャプションとそのキャプション中の各単語に対応するアテンション画像 (I_k) が獲得できる。その後、それらを足し合わせ1つのアテンション画像 (I_{att}) を作成する。また、図3の(a)にアテンション画像 (I_{att}) の模式図の例を載せる。図3の(a)のようにグレーの領域がデータ拡張を行いたい箇所である。アテンション画像 (I_{att}) の計算には、式(1)を用いる。

$$I_{att}(i, j) = \sum_{k=0}^m I_k(i, j), \quad (1)$$

ここで、 $I_{att}(i, j)$ は足し合わせたアテンション画像、 i, j が画像の縦横の画素位置、 m は獲得するアテンション画像 (I_k) の総数、 $I_k(i, j)$ は k 番目の単語から獲得されたアテンション画像を意味する。また、獲得する単語毎のアテンション画像 (I_k) の各画素値を 0 から 1 の範囲に正規化する。

2.3 アテンション画像に対する K-means クラスタリング

アテンション画像 (I_{att}) には値の高い箇所と低い箇所が存在する。本研究では、高い箇所にデータ拡張処理を行うために、値を分類する必要がある。そこで、本研究では簡便な K-means を用いてクラス分類を行う。図3の(b)に K-means 処理をほどこしたアテンション画像 ($K\text{-means } I_{att}$) の模式図の例を示す。本研究では、2クラスでの K-means クラスタリングを行う。前述のアテンション画像 (I_{att}) と同様にグレーの領域はデータ拡張を行いたい箇所である。

図4にアテンション画像 (I_k) からアテンション画像 ($K\text{-means } I_{att}$)

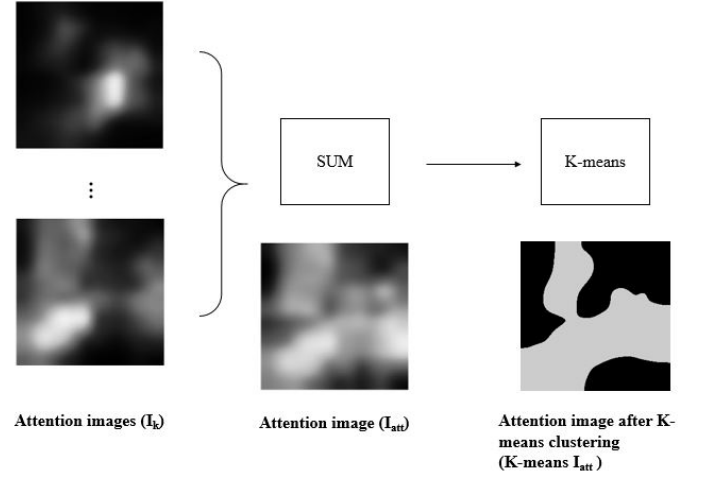


Fig. 4 Outline of attention image processing

means I_{att}) を獲得する流れを示す。図4のアテンション画像 (I_k , I_{att} , $K\text{-means } I_{att}$) は、実計測画像である。まず、正規化された複数枚のアテンション画像 (I_k) を足し合わせ、1つのアテンション画像 (I_{att}) を生成する。そして、K-means クラスタリングを行い、特定の値を持つ画素値を Class1 に、それ以外の画素値を Class0 に分類することで、K-means 処理後のアテンション画像 ($K\text{-means } I_{att}$) の生成を行う。

ここで図4のアテンション画像 (I_k) からアテンション画像 ($K\text{-means } I_{att}$) を生成する流れにおける K-means 処理による分割過程を図5の(a), (b)を用いて説明する。この図5の(a)はアテンション画像 (I_{att}) を、図5の(b)はアテンション画像 ($K\text{-means } I_{att}$) を簡単化しており、図5の(a)に示すように、図5の(a)は 3×3 の大きさの濃淡画像 (I_{att}) とする。この図5(a)に K-means 処理を行うことで、例えば、図5の(a)の点線で囲まれた領域の attention 値が 0.9, 0.4, 0.4, 0.3 の画素を class1 に、それ以外の領域に存在する画素を class0 に分類することができるため、図5の(b)に示すような画素毎にクラス分類された画像 ($K\text{-means } I_{att}$) を獲得することができる。実際には、実計測画像である図4のアテンション画像 (I_{att}) から実計測画像である図4のアテンション画像 ($K\text{-means } I_{att}$) を獲得する。ここで、提案手法は、アテンション画像 (I_k) を足し合わせるため、獲得するアテンション画像 (I_{att}) の画素値は 1 以上になる可能性がある。しかし、前述の通り、提案手法は K-means 処理をほどこすため、1以上の値を含む全ての画素値がクラス分類され、クロップ処理、マスク処理に活用される。そのため、提案手法では 1 以上の値を扱うことが可能である。

2.4 アテンション画像を用いたクロップ

従来手法³⁾の通りに、入力画像にクロップを適用すると、キャプション生成時にモデルが参照する画像領域を全て削除する可能性がある。そこで、獲得した K-means クラスタリング済みのアテンション画像 ($K\text{-means } I_{att}$) を利用する。

図6を例に、データ拡張を行いたい領域を四角形で検出し、その領域を中心にクロップ処理を行う。四角形を検出する理由としては、K-means クラスタリングの結果分類された領域は、必ずしも頂点を持たず、そのままではクロップ処理をほどこすのが困難なためである。その四角形の検出には OpenCV⁷⁾ の

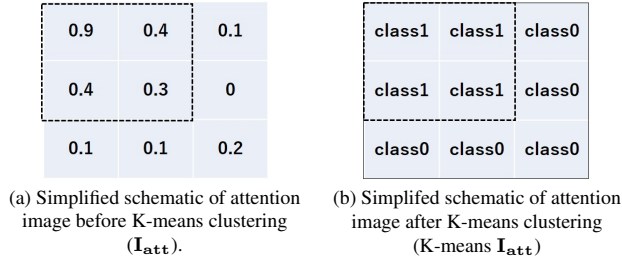


Fig. 5 Simplified schematic of attention image

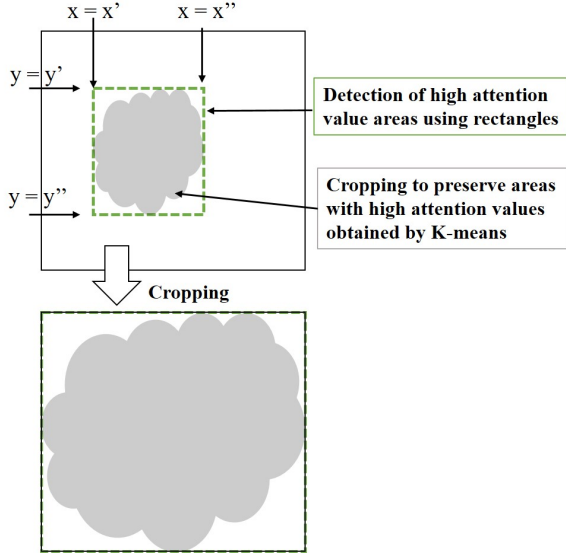


Fig. 6 Schematic of cropping

ライブラリーを活用する．具体的には，アテンション値が高い領域の輪郭を表す点の座標を取得して，その点群に外接する短径の座標を導出することで，アテンション値が高い領域に外接する四角形の検出を行う．図 6 に K-means 処理されたアテンション画像 ($\mathbf{I}_{att}$) がある．図 6 のグレーの領域は，データ拡張を行いたい領域であり，残してクロップ処理を行うため，図 6 の $x' \sim x''$ と $y' \sim y''$ を含むようにクロップする領域を選択していく．具体的には，0 から x' ，0 から y' ， x'' から画像の横幅， y'' から画像の縦幅までの間からそれぞれ無作為にサンプルを行う．入力画像からサンプルした 4 点を頂点とする四角形の画像領域をクロップし，クロップされた領域をにリサイズし，新たな訓練データとする．

ここまでは，K-means によって正しいアテンション画像 ($\mathbf{I}_{att}$) のクラスタリングが行われる場合で考えているが，実際には，K-means クラスタリングは推定を誤る．そのため，アテンション値が高いとクラスタリングされた $x' \sim x''$ と $y' \sim y''$ をある程度含めた，0 から $x' + \alpha x'$ ，0 から $y' + \alpha y'$ ， $x'' - \alpha$ (画像横幅 $- x''$) から画像横幅， $y' - \alpha$ (画像縦幅 $- y''$) から画像縦幅までを設定する．ここで， x', y' が 0， x'', y'' が画像幅と同じ場合には，先行研究³⁾と同様の事前設定値を用いて，クロップ処理を行う．

Table 1 MSCOCO'14 dataset

MSCOCO'14	Training data	Validation data	Test data
Number of images	113,000	5,000	5,000

2.5 アテンション画像を用いたマスク

従来手法⁴⁾の通りに，入力画像にマスクを付与すると，画像中の特定領域の種類を増やすことが難しい場合がある．そこで，獲得した K-means クラスタリング済みのアテンション画像 ($\mathbf{K}\text{-means } \mathbf{I}_{att}$) を利用する．具体的には，アテンションが高い領域を四角形で検出し，その領域を中心にマスク処理を行う．四角形の検出理由は，クロップ処理と同様である．アテンション画像 ($\mathbf{I}_{att}$) 中のアテンション値は，入力画像から獲得されるアテンション画像 ($\mathbf{I}_k$) の枚数によって大きく変化するため，付与するマスクのサイズも一定ではなく，アテンション画像 ($\mathbf{I}_{att}$) のアテンション値に応じて変化することが望ましい．

そこで，本研究では，従来研究⁴⁾のマスクサイズを参考に，OpenCV⁷⁾により検出した四角形の大きさ $\times \beta$ をマスクのサイズとする．これにより，画像中の特定領域の大きさに応じてマスクのサイズを変更することが可能となる．また，クロップの場合と同様に，K-means クラスタリングは推定を誤ることを考慮し，アテンション値の高い領域と低い領域の境界となる領域にマスクを付与する．具体的には，マスク内に含まれるアテンション値の高いクラスの画素数が， $\gamma \times$ マスクの縦幅 $\times$ マスクの横幅よりも大きくなるようにする．ここで，付与するマスク領域内の画素値はランダムな値で埋める．そして，入力画像にマスクを付与し，新たな訓練データとする．

3. 特定領域を考慮するデータ拡張手法の性能評価実験

3.1 実験概要

提案手法の有効性を確かめるために 2 種類の実験を行った．1 つ目の実験は，Web 上の著作権フリーの画像に対して，各手法を適用して，キャプション生成を行った．この画像には，正解となるキャプションは付与されていない．

2 つ目の実験では，一般的に公開されている画像とキャプションが付与された大規模データセットを用いた．そして，提案するデータ拡張手法の有効性を評価するために，大規模データセットを Karpathy 分割法で学習，検証とテスト用に分割しキャプション生成の実験を行った．実験で利用する評価指標は，キャプション生成の分野で代表的に活用されている BLEU 評価指標⁸⁾，METEOR 評価指標⁹⁾，ROUGE-L 評価指標¹⁰⁾，CIDEr 評価指標¹¹⁾を用いた．全ての指標は値が高いほど，精度が高いことを意味する．

3.2 データセット

本実験 2 で用いるデータセットは，キャプション生成の分野で最も活用されるデータセットの 1 つであり，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¹²⁾を用いる．このデータセットは 123,000 枚の画像を含んでいる．表 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．データの预处理として，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，8,791 語彙数となる．

3.3 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る BLEU-N (N=1,2,3,4) 評価指標⁸⁾, METEOR 評価指標⁹⁾, ROUGE-L 評価指標¹⁰⁾, CIDEr 評価指標¹¹⁾を用いる。まず、BLEU 評価指標⁸⁾は式(2)で表される。

$$BLEU_N = \min(1, e^{1-\frac{r}{c}}) \cdot e^{\frac{1}{N} \sum_{n=1}^N \log p_n}, \quad (2)$$

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 の precision である。次に METEOR 評価指標⁹⁾は式(3)~(5)で表される。

$$Score = Fmean * (1 - Penalty), \quad (3)$$

$$Fmean = \frac{10PR}{R + 9P}, \quad (4)$$

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams_matched} \right)^3, \quad (5)$$

式(3)~(5)において、 $Score$, P , R , $\#chunks$, $\#unigrams_matched$ は、それぞれ、METEOR 評価指標⁹⁾, uni-gram 一致に基づく precision, uni-gram 一致に基づく recall, 順番が同じ部分文字列の数, 正解と生成キャプション間で一致した uni-gram の数を表す。次に precision と recall の調和平均である ROUGE-L 評価指標¹⁰⁾は式(6), (7)を用いて算出される。

$$Recall = \frac{LSC(output_words, reference_words)}{reference_words}, \quad (6)$$

$$Precision = \frac{LSC(output_words, reference_words)}{output_words}, \quad (7)$$

ここで、 $LSC(\cdot)$, $output_words$, $reference_words$ は、それぞれ、生成キャプションと正解とで一致する最大の文字列の単語数, 生成キャプション中の単語数, 正解キャプション中の単語数を表す。そして、CIDEr 評価指標¹¹⁾は式(8)~(10)で表される。

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log \left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))} \right), \quad (8)$$

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|}, \quad (9)$$

$$CIDEr(c_i, S_i) = \sum_{n=1}^N w_n CIDEr_n(c_i, S_i), \quad (10)$$

式(8)~(10)において、 $g_k(s_{ij})$, S_i , s_{ij} , c_i , ω , Ω , $h_k(s_{ij})$, I , $g^n(c_i)$, $\|g^n(c_i)\|$, w_n はそれぞれ、 s_{ij} の ω に対する TFIDF ベクトル, 正解キャプションの集合 $S_i = \{s_{i1}, \dots, s_{im}\}$, 正解キャプション s_i の j 番目の文章, 生成された文章, N-gram の集合, 全ての n-gram によって構成される語彙, 正解キャプション s_{ij} に現れる ω の数, データセット内の全ての画像の数, 全ての n-gram の長さ n に対するベクトル ($g^n(s_{ij})$ に関しても同様である), $g^n(c_i)$ の大きさ ($g^n(s_{ij})$ に関しても同様である), 重み, を表す。

3.4 実験環境

実験に使用される全ての画像は、256×256 画素の大きさ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



Fig. 7 Image used in experiment 1 and captions generated by each method.

Previous method⁶⁾: A woman sitting on a motorcycle next to a green building.

Previous method³⁾: A woman sitting on a motorcycle in front of a building.

Proposed method: A woman in a red dress sitting on a motorcycle.

and Tell⁶⁾を用いる。パラメータの更新方法は、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 4×10^{-4} , エポックは 120, 交差エントロピー誤差を用いて学習を行う。学習・検証・テストにおいて、GPU は Intel 製の Core i9-7900X を使用する。また、学習済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、 α, β, γ は、それぞれ 0.2, 0.4, 0.6 と設定し、本実験では、各モデルの学習において、約 1,130,000 枚のデータ拡張を行った。

3.5 実験結果および考察

実験 1 で用いた画像の例を図 7 と図 8 に示す。

図 7 には、バイクの上に座る女性の画像と、各手法が生成したキャプションを表している。従来手法（データ拡張なし）⁶⁾は、「A woman sitting on a motorcycle next to a green building」、従来手法³⁾は「A woman sitting on a motorcycle in front of a building」、提案手法は「A woman in a red dress sitting on a motorcycle」とそれぞれキャプションを生成した。ここで、従来手法⁶⁾は、CNN に Resnet101 を用いたものであり、データ拡張を行っていない。また、従来手法³⁾は、データ拡張としてクロップ処理を組み合わせた手法であり、従来手法⁶⁾に適用したものである。全ての手法で、「motorcycle」という単語が生成できていることがわかる。

次に図 8 の (a) には、森林をバイクで走行中の画像と、各手法が生成したキャプションを載せている。従来手法⁶⁾は「A man riding a bike through the woods」、従来手法³⁾は「A man riding a motorcycle down a dirt road」、提案手法は「A man riding a motorcycle through a forest」とそれぞれキャプションを生成した。

図 8 の (b) には、図 8 の (a) の画像からバイクのナンバープレートが写っている領域を取り除いた画像と、各手法が生成したキャプションを載せている。従来手法⁶⁾は「A man riding a skateboard through the woods」、従来手法³⁾は「A man riding a horse through a forest」、提案手法は「A man riding a dirt bike through a forest」とそれぞれキャプションを生成した。結果から、データ拡張を行わない従来手法⁶⁾は、「skateboard」、従来手法³⁾は「horse」、提案手法は「bike」とそれぞれ単語を生成したことから、提案手法の効果が確認できる。そして、従来手法⁶⁾³⁾は、ナンバープレートを取り除いたことで、「skateboard」「horse」とそれぞれキャプションを生成していることから、ナ



(a) Image with numberplate.

Previous method ⁶⁾: A man riding a bike through the woods.
 Previous method ³⁾: A man riding a motorcycle down a dirt road.
 Proposed method: A man riding a motorcycle through a forest.



(b) Image without numberplate.

Previous method ⁶⁾: A man riding a skateboard through the woods.
 Previous method ³⁾: A man riding a horse through a forest.
 Proposed method: A man riding a dirt bike through a forest.

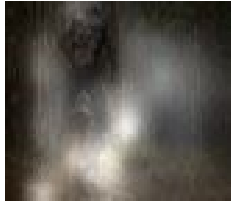
Fig. 8 Image used in experiment 1 and captions generated by each method



(a) Previous method ⁶⁾
 attention image (I_k) related to
 keyword [skateboard]



(b) Previous method ³⁾
 attention image (I_k) related to
 keyword [horse]



(c) Proposed method
 attention image (I_k) related to
 keyword [bike]

Fig. 9 Attention image (I_k) for Fig. 8 (b)

ナンバープレートの情報から「bike」「motorcycle」を生成していたと考えられる。さらに、図 9 の (a), (b), (c) にそれぞれ、図 8 の (b) から「skateboard」, 「horse」, 「bike」のキャプションを生成するときのアテンション画像 (I_k) を載せる。図 9 から、提案手法のアテンション値は、他の手法に比べ、広範囲の領域にアテンション値が高いことが確認できることから、ナンバープレートのような画像中の特定領域のみだけでなく、その周囲の情報を活用できていると考えることができる。以上のことから、提案手法の「過適合」に対する効果が確認できる。

実験 2 の結果を表 2 に示す。この表 2 は、5000 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。実験は 3 回行い、獲得した値の平均値を表 2 に載せている。全ての評価指標は値が高いほど良く、太字は最も値が高かった手法を示している。CNN に Resnet101 を用いた従来手法（データ拡張なし）⁶⁾ はデータ拡張は行っていない。また、従来手法 ⁴⁾ は、画像分類の分野で行われているデータ拡張手法を従来手法 ⁶⁾ に適用したものである。従来手法 ³⁾ は、データ拡張としてクロップ処理を組み合わせた手法であり、従来手法 ⁶⁾ に適用したものである。そして、従来手法 ⁵⁾ は、アテンション画像とマスク処理を組み合わせたものであり、従来手法 ⁶⁾ に適用した。

提案手法も同様に、従来手法 ⁶⁾ を活用した。実験結果より、全ての評価指標で提案手法が最も高い数値を示した。この結果より、提案するデータ拡張は既存手法よりも有効であった。

また、図 10 の (a), (b) は、順に、実験 2 での従来手法、提案手法のモデルの学習時における誤差を示している。横軸が学習ステップを意味する Epoch、縦軸が各 Epoch での誤差を表す。これらの結果から Epoch が増えるにともない、従来手法、提案手法は、訓練誤差が小さくなるのがわかる。しかし、検証誤差は、従来手法が Epoch 6 の時に最も小さい値をとり、以降増加していくのに対して、提案手法は Epoch 10 の時に最も小さい値を記録するまで減少し続けた。一般的に、過適合が発生している場合には、訓練誤差の減少にともない検証誤差が増加する現象が確認される。従来手法は Epoch 7-10 間で訓練誤差の減少と検証誤差の増加が確認できることから過学習が発生していると考えられる。しかし、提案手法は、Epoch 0-10 において、訓練誤差及び検証誤差が減少していることから、従来手法に比べ、過学習に効果があることが確認できる。

最後に、本研究で明らかにならなかったことの 1 つに、提案手法を適用した際に性能が低下する場合がある点が挙げられる。これは、提案手法はアテンション値が高い画像領域の種類を複数に増やして、1 種類の画像領域のみに適合させないことで、モデルの過適合を妨げることを行うが、アテンション値が高い画像領域は認識にとって重要な領域であるため、種類の増やし方によっては、性能が低下する場合があることが考えられるからである。例えば、モデルの訓練中に継続して、画像中の同一の領域にマスクを付与して、アテンション値が高い画像領域の種類を効果的に増やすことができない可能性がある。しかし、この問題に対処するために、提案手法では、データ拡張なし、データ拡張としてクロップ、データ拡張としてマスク、の 3 つの方法を組み合わせている。そのため、本提案法を適用すると性能が低下する場合は生じにくいと考えられる。以上のことから、提案手法による性能低下の検証には、更なる調査が必要である。

4. 結 論

本研究では、キャプション生成においてモデルが重視する画像中の特定領域の種類を増やすために、アテンション領域の大きさに応じて変化するマスク、クロップを用いたデータ拡張手法の提案を行った。具体的には、アテンション機構を用い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

Table 2 Experimental results

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Previous method ⁶⁾ (without data augmentation)	72.6	55.7	41.4	30.5	25.63	53.56	99.64
Previous method ⁴⁾	72.7	55.8	41.5	30.6	25.47	53.58	100.04
Previous method ³⁾	73.0	56.3	42.0	31.2	25.92	54.02	102.24
Previous method ⁵⁾	73.1	56.1	41.7	30.8	25.43	53.57	99.41
Proposed method	73.6	57.0	42.8	31.8	26.09	54.37	103.73

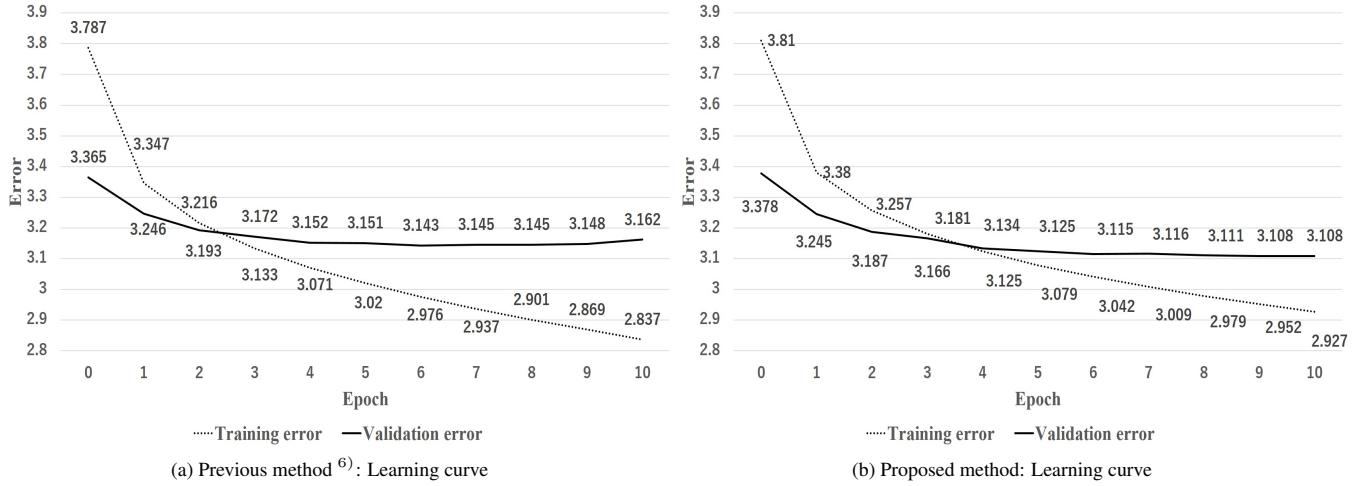


Fig. 10 Learning curve in experiment 2

クロップとマスク処理を加えてデータを増やす手法を提案した。キャプション生成の評価実験では、提案手法が既存手法よりも精度向上を達成した。

今後の予定として、K-means クラスタリングのクラス数を変化させた場合のキャプション生成の精度や、各種パラメータを変更した際の影響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。ここに謝意を表する。

参 考 文 献

- Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille: Explain Images with Multimodal Recurrent Neural Networks. *arXiv preprint arXiv:1410.1090*, (2014).
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Erhan: Show and Tell: A Neural Image Caption Generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 3156.
- Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel: Image Captioning with Deep Bidirectional LSTMs. *Proceedings of the ACM International Conference on Multimedia*, (2016), 988.
- Terrance DeVries and Graham W. Taylor: Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*, (2017).

- 岩村 紀与彦, ルイ笠原 純ユネス, モロ アレッサンドロ, 山下 淳, 浅間 一: 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. ビジョン技術の実利用ワークショップ講演論文集, (2019), IS2-A3.
- Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the International Conference on Machine Learning*, (2015), 2048.
- <http://opencv.org/>
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu: BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the Annual Meeting on Association for Computational Linguistics*, (2002), 311.
- Michael Denkowski and Alon Lavie: Meteor Universal: Language Specific Translation Evaluation for Any Target Language. *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, (2014), 376.
- Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie: An information-theoretic approach to automatic evaluation of summaries. *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter Association for Computational Linguistics*, (2006), 463.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh: Consensus-based image description evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 4566.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick: Microsoft COCO: Common Objects in Context. *Proceedings of the European Conference on Computer Vision*, (2014), 740.