動作推定を取り入れた1枚の画像からのキャプション生成

東京大学　◯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，淺間 一

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，視覚障害者のために動画内容の描写やインターネット上の大量の画像の索引付け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，手間がかかっている．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，深層学習手法が提案されている [1][2][3]．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質の良いキャプションを生成することが可能である．

これらの既存の深層学習の手法は，一般的に入力画像から畳み込みニューラルネットワーク（CNN）を用いて画像特徴を獲得し，その画像特徴に応じてリカレントニューラルネットワーク（RNN）を用いてキャプションを生成する．そして，これらの手法群は，画像特徴の活用方法で大きく次の 2 つに分類することができる．1），キャプション生成モデルが画像特徴を全て活用して，キャプションを生成する．2），キャプション生成モデルが画像特徴の一部を活用してキャプションを生成する．

画像特徴を全て活用して，キャプション生成をするモデルは，この分野の初期の頃に多く提案されている．代表的なものに，Vinyals らの手法 [1] がある．この手法は，主に，CNN と RNN から構成されており，入力画像から CNN を用いて色や形状などの画像特徴を獲得し，その特徴に応じて RNN を用いてキャプションの生成を行う．例えば，”車”という単語を生成するときに，画像特徴全体（人や車など多くの特徴が含れる）から，”車”という単語を生成する．

画像特徴の一部を活用して，キャプションを生成するモデルは，近年多く提案されている [2]．このモデルは，画像特徴から特定の一部を活用し，そこからキャプションを生成する．例えば，”車”という単語を生成するときには，画像特徴から車の特徴のみを活用して，”車”という単語を生成する．そのため，単語の生成時に不要な特徴を活用することを避けることができ，高いキャプション生成の精度が報告されている．しかし，上記の 2 つの手法群の課題の 1 つとして，動作に関する単語の生成が困難な場合が存在する事が挙げられる．

動作に関する単語の生成が困難な場合が存在することから，従来法の色や形状などの画像特徴を活用することに加えて，新たに動作特徴を考慮することを考える．この動作特徴は，画像中の物体の動きやすさを表す特徴である．例えば，立っている人間の足の開き具合といった形状の画像特徴に加えて，足の動きやすさを表す動作特徴を考慮することで，動詞の生成の精度向上を期待する．

本論文では，動作推定を取り入れるキャプション生成手法を新規に提案する．Fig. 1 に示すように，提案手法は，大きく 2 つの要素から構成されている．1 つ目は，従来の色や形状などの画像特徴を活用するためである．2 つ目は，動作特徴を活用するためである．

2. 提案手法

2.1 コンセプト

本研究のコンセプトは，画像から動作推定を行い動作特徴を活用することである．このカギとなるモチベーションは，人間が長年にわたる経験を基に画像から動作を推定することができる点にある．例えば，画像を与えられたときに，人間は”歩く’’や”立つ’’の動作を区別することが可能である．さらに，これは神経科学の研究にも支持されている．Kourtzi らは，人間が画像を見たときに，動画を処理する脳の領域が反応することを報告しており [4]，画像から動作の情報を活用していると考えることができる．

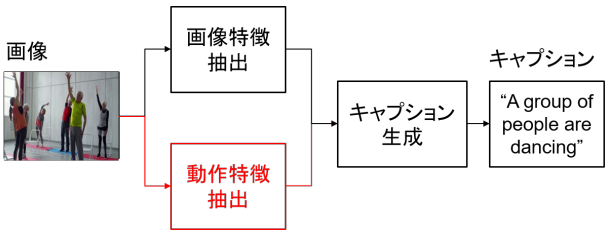

Fig. 1: 提案手法の概念図．

2.2 全体のネットワーク構造

コンセプトを実現するためには，画像特徴を抽出するネットワークに加えて，動作特徴を抽出するネットワークを備える必要がある．そこで，Fig. 1 に示すように，従来の色や形状を表す画像特徴を獲得するネットワーク，画像中の物体の動きやすさを表す動作特徴を獲得するネットワークと獲得する特徴に応じて単語を生成するキャプション生成のネットワークから提案手法は構成される．

(a) 画像特徴を抽出するネットワーク構造

画像特徴の獲得を行うために，本研究では CNN のモデルの一つである ResNet-101[5] を用いる．この ResNet-101 を用いることで，人間が画像から特徴を設計することなく，訓練データからネットワークが自動的に特徴を獲得でき，高い性能が示されている [1][2]．そのため，先行研究と同様のネットワークを活用し，画像の特徴抽出を行う．

(b) 動作特徴を抽出するネットワーク構造

本研究では，入力は画像であるため，複数枚の画像から差分を計算し，動きの特徴を獲得することができない．そのため，1 枚の画像から動作を推定する必要がある．そこで本研究では，学習ずみのニューラルネットワークを用いて，1 枚の画像から動作を推定することを行う．具体的には，Fig. 2 に示すように，入力される画像は，CNN によって動作推定が行われ．その後，CNN により特徴抽出が実行される処理の流れになっている．この動作推定を行う理由は，前述の通りに，入力が 1 枚の画像であるため，動作特徴を含まないため，動作推定が必要となる．次に，特徴抽出は，より複雑な特徴を獲得するために，動作推定の結果に対して実行される．この動作推定と特徴抽出を組み合わせることで 1 枚の画像から複雑な動作特徴を獲得可能となる．以上の点を踏まえて，本研究では，動作推定と特徴抽出を行うネットワーク構造を新たに提案する．その動作推定には，Im2Flow[6]，特徴抽出には，ResNet-101[5] を用いる．

(c) キャプションを生成するネットワーク構造

獲得される画像特徴と動作特徴は，単に連結されて，キャプション生成のネットワークに渡される．このキャプション生成には，高い性能を報告している先行研究 [2] と同様に LSTM を活用する．そのため，提案手法は，両特徴に基づいて，キャプション生成することが可能である．

3. 実験

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．

Fig. 2: 動作特徴を抽出するネットワークの処理の流れ．動作推定の画像中に存在する矢印が，各画素に推定される動きを表す．

Table 1: キャプション生成結果

	BLEU4	METEOR	ROUGE-l	CIDEr
先行研究 [7]	19.9	18.5	-	-
先行研究 [2]	24.0	20.3	46.4	49.6
提案手法	**24.9**	**20.4**	**46.5**	**50.8**

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 FLICKR30K データセットを用いる．このデータセットは，キャプション生成モデルが性能比較を行うために良く用いられ，一般的に公開されている大規模なデータセットである．その全ての画像はカラー画像である．このデータセットは，31,783 枚の画像で構成されており，その内，検証データとテストデータについては，それぞれ 1,000，1,000 枚で構成されている．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 3 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，8,510 語彙数となる．

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，METEOR，ROUGE，CIDEr を用いる．

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

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．これらの指標は基本的に生成されたキャプションと正解との一致度を評価する．そのため，高い値は良い結果を表す．

3.3 キャプション生成実験準備

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．また，バッチサイズは 16，画像特徴のエンコーダーの学習率は 1e-5，動作特徴のエンコーダーの学習率は 5e-4，デコーダーの学習率は 1e-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．そして，比較を行うキャプション生成のモデルは，先行研究 [2] と [7] を用いる．先行研究 [7] は，CNN と LSTM から構成されるモデルである．先行研究 [2] は，CNN と 2 つの LSTM から構成されるモデルである．両手法共に CNN には ResNet-101 を用いる．その理由は，ResNet-101 は画像処理の分野で最も使われるモデルの 1 つであり，優れた特徴抽出を実現できることと，画像特徴のモデルを統一することで，複数の手法で公平な比較が可能になるためである．動作特徴を考慮しない．特に，提案手法と先行研究 [2] の違いは，動作特徴の考慮の有無のみである．

4. 結果と考察

実験結果を Table 1 に示す．この Table 1 は，1,000 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．動作特徴を活用しない先行研究 [7] の値は最も低く，同様に動作特徴を活用しない先行研究 [2] が高い値を記録した．そして，提案手法は全ての評価指標で，先行研究よりも高い精度を記録した．提案手法と先行研究 [2] の違いは，動作特徴の有無であるため，動作特徴は本実験において有効に作用したと考えられる．この結果より，提案する手法は有効であった．特に，先行研究 [2] と比較すると，提案手法は CIDEr の増加点数が高く，CIDEr は頻度の低い単語を考慮する評価指標である．このことから，動作特徴を考慮することで，頻度の低い単語の生成に良い影響を付与した考えることができる．

提案手法は，動作特徴を活用することでキャプション生成の精度向上を達成したが，短所が存在する．今回の動作を予測するネットワークは，人間の動作など動きの速いデータセットで事前に学習を行った．そのため，動きの遅い物体に関しては，動作の予測に失敗する可能性が存在する．そのため，そのような場合に対処するためには，動きの遅い物体を含むデータセットでの学習が必要となる．

5. 結論

本研究では，動作推定を取り入れるキャプション生成手法の提案を行った．具体的には，従来の色や形状を表す画像特徴を獲得する処理に加え，画像中の物体の動きやすさを表す動作特徴を考慮する処理を加えた手法を提案した．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，提案手法は，画像特徴と動作特徴を単純に連結して用いているが，より良い特徴の活用方法を検討することが挙げられる．

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．

参考文献

[1] Vinyals, O.; Toshev, A.; Bengio, S.; Erhan. Show and Tell: A Neural Image Caption Generator. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, **2015**, 3156–3164.

[2] Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, **2018**, 6077–6086.

[3] Wang, C.; Yang, H.; Bartz C.; Meinel, C. Image Captioning with Deep Bidirectional LSTMs. *Proc. ACM International Conference on Multimedia*, **2016**, 988–997.

[4] Kourtzi, Z; Kanwisher, N. Activation in Human MT/MSTby Static Images with Implied Motion. *Journal of Cognitive Neuro-science*, **2000**, 12(1), 48-–55.

[5] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, **2016**, 770–778.

[6] Gao, R.; Xiong, B.; Grauman, K. Im2Flow: Motion Hallucination from Static Images for Action Recognition. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, **2018**, 5937–5947.

[7] Xu, K.; Ba, J.L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proc. International Conference on Machine Learning*, **2015**, 2048–2057.