車載カメラ映像と動き情報による近未来キャプション生成に関する研究

TP19052 森優樹　　指導教授：山下隆義

1. はじめに

画像キャプション生成は，入力画像に対する説明文を生成するタスクであり，ニュース文の自動生成や画像検索のタグ生成などに利用されている．また，自動運転においては，搭乗者の心理的負担を軽減するために，運転制御の判断根拠の言語的説明への応用も期待されている．一方で，これまでの画像キャプション生成は，入力された画像に対するキャプション生成に留まっており，近未来に起きうるイベントに対するキャプションを生成していない．自動運転においては，事故防止や搭乗者への注意喚起には，現時刻よりも今後起きうる近未来のイベントに対するキャプション生成が重要となる．本研究では，近未来キャプション生成という新たなタスクを提案するとともに，車載カメラ映像からの近未来キャプション生成に適したモデルの提案を行う．Berkeley Deep Drive eXplanation Dataset を用いた評価実験では，近未来キャプション生成が可能であることを示した．

2. 車載カメラを対象とした画像キャプション生成

画像キャプション生成とは，入力画像に適した説明文を生成するタスクであり，入力画像の要約や判断根拠の言語的説明が可能である．深層学習の登場以降，CNN と RNN を活用した手法が複数提案されている [1].

特定のシーンに特化する場合，車載カメラ映像が対象として挙げられる [2]．車載カメラ映像を対象とした画像キャプション生成では，自動運転制御の自然言語による判断根拠の証明や，搭乗者への周辺状況の注意喚起などが可能となる．

従来の車載カメラ映像を対象とした手法は，一般画像を対象とした手法を車載カメラ映像の環境に適したように発展させている．そのため，現時点における判断根拠の言語的説明を行うことが可能である．一方で，搭乗者への注意喚起など事故予防や危険因子に対する言語的説明は，現時点ではなく，未来の事象を対象としなければならない．従来手法では，このような近未来の事象を対象とした画像キャプション生成は行っていないという問題点がある．

3. 提案手法

本研究では，新たなタスクとして近未来キャプション生成を提案し，それに適したキャプション生成モデルの提案を行う．まず新たなタスクである近未来キャプション生成について説明し，その後，提案手法のネットワーク構造および学習方法について述べる．

3.1. 近未来キャプション生成

車載カメラ画像から事故防止や搭乗者への注意喚起を行う場合，数秒後の前方車の状況や歩行者の動きといった，近未来の情報が必要である．これまでのキャプション生成手法は与えられた時刻における画像からキャプションを生成しているため，このような要求に応えることができない．そこで，提案タスクでは，図 1 に示すように，複数の観測した画像を Captioning Near-Future Model に 入力し，近未来の動きを考慮してキャプション生成できるようにする．このとき，未観測の近未来の画像は利用しない．画像からその後の近未来に発生するイベントに対して注目すべき領域を捉えて，キャプション生成を行う．

3.2. Captioning Near-Feature Model

近未来のキャプションを生成するモデル構造を図 2 に示す．本モデルでは，複数の画像から特徴ベクトルを抽出する．そして，特徴ベクトルと合わせて自車の動き情報となるセンサデータをエンコーダに入力する．これにより，画像情報から捉えることのできない自車情報を考慮することを可能とする．Action Regressor は，エンコーダで獲得した中間表現から近未来の動きを予測する．そして，エンコーダで獲得した中間表現と Action Regressor で予測した動き情報をデコーダに入力して，近未来のキャプション生成を行う．

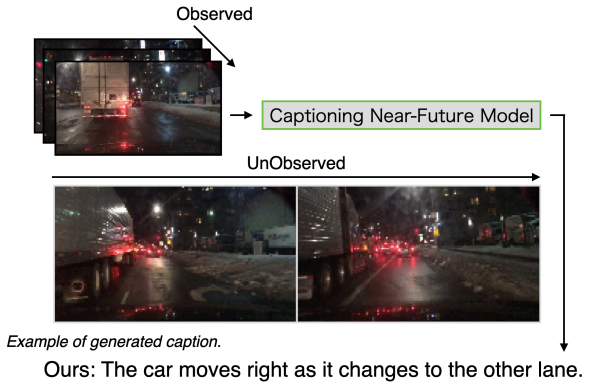

図 1：近未来キャプション生成

3.3. Action Regressor

Action Regressor は，近未来の自車の動き情報を推定するネットワークである．本ネットワークは 5 層の全結合層で構成されている．エンコーダで獲得した中間表現を入力し，動き情報として自車の速度とステアリング角度を推定する．本ネットワークの各層のユニット数は，1 層目が 1164 ユニット，2 層目が 100 ユニット，3 層目が 50 ユニット，4 層目が 10 ユニットとなっている．Action Regressor は，n フレーム分の速度とステアリング角度を出力する．出力層の出力は，エンコーダで獲得した中間表現と結合してデコーダに与える．Action Regressor はステアリング角度と速度で別々のネットワークである．

3.4. 損失関数

自車の速度とステアリング角度の推定する損失関数には，平均二乗誤差を用いる．また，Action Regressor とエンコーダ・デコーダ部分は End-to-End で学習する．自車の速度 a とステアリング角度 c，エンコーダ LSTM の出力を x_k，デコーダ LSTM の中間層の出力を h_k とすると Action Regressor の損失関数は，平均二乗誤差を用いた L_{action} であり，式 (1)，エンコーダ・デコーダ部分の損失関数には交差エントロピーを用いた $L_{caption}$ である，式 (2) で表せる．損失関数は，式 (1) と式 (2) より，式 (3) で定義できる．

$$L_{action} = \sum_t ((a_t - a'_t)^2 + (c_t - c'_t)^2) \tag{1}$$

$$L_{caption} = \sum_k \log p(y_k \,|\, y_{k-1}, h_k, x_k) \tag{2}$$

$$L = L_{action} + L_{caption} \tag{3}$$

4. 評価実験

提案手法の有効性を確認するために，評価実験を行う．評価実験では，Berkeley Deep Drive eXplanation Dataset[2] を用いて，提案手法の近未来キャプション生成における性能を評価する．

本実験で使用する Berkleley Deep Drive eXplanation Dataset は，6,984 本の車載カメラ映像から構成されているデータセットであり，自動車の速度，加速度，進行角度，速度，及び 26,228 個の自動車の制御イベントのアノテーションが付与されている．

イベントの平均時間はデータ全体で 7.26 秒である．動画は 30fps で撮影されており，本実験では計算量削減のために 1fps に調整して利用する．学習サンプルは，4,356 本，14,933 文，評価サンプルは 536 本，1742 文を利用する．本実験では，学習の更新回数は 30 エポック，バッチサイズは 50 とする．画像サイズは 160x90 であり，入力フレーム n は 5 である．また，デコーダの出力ユニット数 M は 1,290 である．学習時，パラメータの初期化には Xavier，各構成

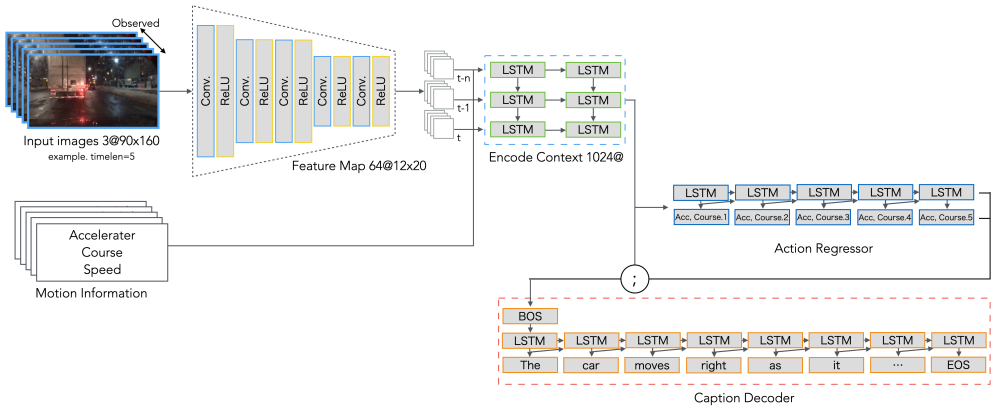

図 2：Captioning Near-Future Model

表 1：キャプション生成時刻による精度比較

生成時刻	BLEU@4	METEOR	CIDEr
現在	15.97	28.20	74.96
近未来	**17.02**	**29.26**	**83.73**
未来	11.97	27.70	47.11

ネットワークの学習最適化には Adam を用いる．生成キャプションを評価する指標として，BLEU，METEOR および CIDEr を用いる．

4.1. キャプション生成時刻の定義

本実験で利用するデータセットには，イベントが発生している区間がアノテーションされている．そこで，以下のようにキャプション生成区間の時間的な定義を行う．

- 現在：イベント発生区間全体を入力し，キャプション生成
- 近未来：イベント発生区間の前半を入力し，後半部分のキャプションを生成
- 未来：イベント発生区間以前を入力し，イベント発生区間中のキャプションを生成

上記の ‘現在’ は，イベント発生区間中の画像を入力してキャプション生成する．キャプション生成できた時点ではイベントが終了している．これは従来のキャプション生成に相当する．‘近未来’ は，イベント発生区間中に，その後に生じるイベントに対するキャプションを生成する．‘未来’ は，イベント発生区間より以前の画像を入力して将来に起きうるイベントに対するキャプションを生成する．‘近未来’ の場合，ブレーキをかけて停車するなどのイベント時に速度が最初に低下するとその根拠と合わせて将来停車するかどうかをキャプション生成できることが期待される．

4.2. 評価結果

表 1 にキャプション生成時刻によるキャプション生成精度を示す．BLEU@4 の評価指標において，イベント発生区間の前半を入力とした ‘近未来’ が 17.02 と最も良い精度となっており，‘近未来’ に適したキャプション生成ができていることがわかる．また，METEOR や CIDEr の評価指標でも，それぞれ 29.26，83.73 とイベント全体を入力とする ‘現在’ よりも精度向上していることがわかる．表 2 に，動き情報および Action Regressor の有無によるキャプション生成精度の比較結果を示す．これより，動き情報を追加することで近未来の精度が BLEU@4 において 17.02 と動き情報を用いない場合に比べて向上していることがわかる．METEOR および CIDEr でも同様に向上している．特に，CIDEr は，近未来のキャプション生成において，49.04 から 83.73 と大幅に向上していることがわかる．

4.3. 動き情報および Action Regressor の有用性

各比較手法によって生成したキャプションおよび入力特徴量の可視化を行う．ここでは，近未来を生成時刻とし，可視化結果を図 3 に示す．灰色画像はキャプション生成を行った範囲の画像であり，本提案手法では入力に利用していない画像である．

表 2：動き情報および Action Regressor の有用性評価

条件	生成時刻	BLEU@4	METEOR	CIDEr
なし	現在	12.10	26.92	45.47
	近未来	12.83	26.56	49.04
Action Regressor のみ	現在	13.75	27.91	59.35
	近未来	12.85	27.06	48.50
動き情報のみ	現在	16.16	28.67	75.61
	近未来	16.74	28.77	77.33
動き情報・	現在	15.97	28.20	74.96
Action Regressor あり	近未来	**17.02**	**29.26**	**83.73**

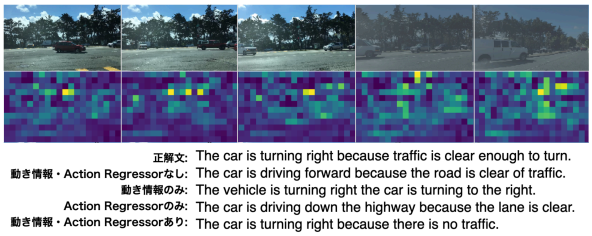

図 3：生成キャプション例

図 3 は，右に曲がるイベントが発生しているシーンである．動き情報を用いる場合，および動き情報と Action Regressor を利用する場合ともに，右に曲がるキャプションを生成できている．特に，traffic is clear というような前方が安全であるという表現に近いキャプションが生成できている．画像特徴は，前方や右折先に強く反応していることがわかる．

このように，画像および動き情報を利用することで，近未来に発生するイベントに適したキャプションを生成できていることがわかった．図 3 から，自車の動き情報を入力に用いた場合のキャプションが，アノテーションとして人が作成したキャプションに近いことが分かる．

5. おわりに

本研究では新たなタスクとして，近未来を考慮したキャプション生成を提案した．また，車載カメラ映像を対象とした，近未来のイベントに対する画像キャプション生成に適したモデルの提案を行った．評価実験により，近未来の画像キャプション生成が可能であることを示した．今後の課題としては，車載カメラを対象とした画像キャプション生成以外への応用が考えられる．

参考文献

[1] O. Vinyals, *et al.*, ”Show and tell: A neural image caption generator.” CVPR，2015.

[2] J. Kim, *et al.*, ”Textual explanations for self driving vehicles.” CVPR，2018.

研究業績

Y. Mori, *et al.*, ”Image Captioning for Near-Future Events from Vehicle Camera Images and Motion Information.” IV，2021.（他 3 件）