社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

森　優樹† 平川　翼† 山下　隆義† 藤吉　弘亘†

† 中部大学　〒487–8501 愛知県春日井市松本町 1200
E-mail: †{yukiri,hirakawa}@mprg.cs.chubu.ac.jp, ††{takayoshi,fujiyoshi}@isc.chubu.ac.jp

あらまし　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に留まっており，近未来に起きう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事象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．

キーワード　画像キャプション生成，image captioning，LSTM，CNN，車載カメラ映像，in-vehicle camera，near future

Image Captioning in Near-Future from Vehicle Camera images and motion information

Yuki MORI†, Tsubasa HIRAKAWA†, Takayoshi YAMASHITA†, and Hironobu FUJIYOSHI†

† Chubu University,　1200 Matsumoto, Kasugai, Aichi, 487–8501 Japan
E-mail: †{yukiri,hirakawa}@mprg.cs.chubu.ac.jp, ††{takayoshi,fujiyoshi}@isc.chubu.ac.jp

Abstract　Image caption generation is a task to generate explanatory text for input images, which is used for automatic generation of news sentences and tags for image search. In addition, in order to reduce the psychological burden on passengers, automonous driving is also expected to be applied to provide verbal explanations of the decision-making basis for driving control. On the other hand, the image caption generation so far has been limited to captions for the input images and has not generated captions that can occur in the near future. In automonous driving, it is more important to generate captions for events in the near future than for the current time to prevent accidents and alert passengers. In this paper, we propose a new task of near-future caption generation and a model suitable for generating near-future captions from images captured by in-vehicle cameras. We showed that near-future caption generation is possible in evaluation experiments using the Berkeley Deep Drive eXplanation Dataset.

Key words　image captioning, LSTM, CNN, in-vehicle camera, near future

1. はじめに

コンピュータビジョンの分野において，畳み込みニューラルネットワーク (CNN) を中心とした深層学習が画像分類問題において高い認識精度を達成して以降，物体検出や自動運転制御などの様々なタスクで高い精度を達成している．そして，どのような領域に着目して識別しているか，というその説明可能性に対する注目が集まっている．これまで，物体認識では，ネットワークの判断根拠を視覚的説明として可視化できる手法が提案されている [11]〜[13]．一方で，ネットワークの判断根拠を可視化するだけでなく，人間が解釈しやすい自然言語による行動理由の言語的説明も重要である．判断根拠の言語的説明には，画像キャプション生成を用いることで実現が可能である．

画像キャプション生成として，CNN や RNN を活用した手法が多数提案されている [1]〜[5]．これらの手法は，エンコーダ・デコーダモデルを採用しており，画像を特徴ベクトルにエンコードする CNN と，エンコードされた特徴ベクトルを自然言語のキャプションにデコードする RNN で構成される．これにより，入力画像に対応したより自然なキャプション生成を可能としている．近年では，画像キャプション生成は，自動運転

車の説明可能性を向上させる目的 [6] や，事故防止を目的とした運転支援システムの実現 [7] などへの応用も提案されている．

これらは，車載カメラ画像から入力画像をもとに，既に起きた出来事に関するキャプション生成を行う．一方で，自動運転モデルにおける説明可能性の向上や事故防止や乗客への注意喚起を行う場合，数秒後の前方車両や歩行者の動きといった，近未来の情報が必要になる場合がある．したがって，近未来に起こる可能性のある出来事を考慮可能なキャプション生成が行えるモデルが求められている．従来の画像キャプション生成の多くは，近未来を考慮したキャプション生成は行っていない．

そこで，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する．さらに，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を提案する．本研究の貢献を以下に示す．

- 新たなタスクとして，近未来を対象とした画像キャプション生成を提案
- 近未来を考慮した画像キャプション生成を行うことで，自動運転における深層学習モデルの説明可能性の向上，及び事故防止や注意喚起を目的とした運転支援システムへ応用が可能

2. 関 連 研 究

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手法であり，入力画像の要約や判断根拠の言語的説明が可能である．画像キャプション生成は，深層学習が注目される以前から研究されている [14],[15] が，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]〜[5].

Show and Tell [1] は，エンコーダである CNN と，デコーダである RNN から構成される画像キャプション生成の手法である．エンコーダでは，入力画像から画像特徴量を抽出する．そして，画像特徴量をデコーダに入力し，適した単語を出力する．出力された単語は次の時刻の入力となり，次に適した単語を出力する．このとき，画像特徴量と言語特徴量が同じ埋め込み空間で得られるように学習されている．ここで，デコーダの RNN には，LSTM を用いている．

LSTM を用いたキャプション生成では，入力系列が長いほど情報の伝播がしづらくなり，精度が低下する．一方，Show, Attend and Tell [2] で提案されたアテンション機構を用いた学習を行うことで，精度の低下を抑制し，高精度なキャプション生成を可能としている．

これらの手法は，一般画像を対象としている．特定のシーンに特化したキャプション生成としては，車載カメラ映像が挙げられる．車載カメラ映像を対象とした場合，自動運転制御の自然言語による説明可能性の確保や，搭乗者への周辺状況の注意喚起なとが可能となる．

Kim ら [6] は，車載カメラ映像からキャプション生成する手法を提案している．本手法は，入力画像から，自車のセンサ情報を推論および学習する Vehicle Controler モデル，自車の行動を説明する Textual Explanation Generator モデルから構築されている．Vehicle Controler で獲得したアテンションを Textual Explanation Generator で利用することで，自車の行動に基づいたキャプション生成を可能としている．

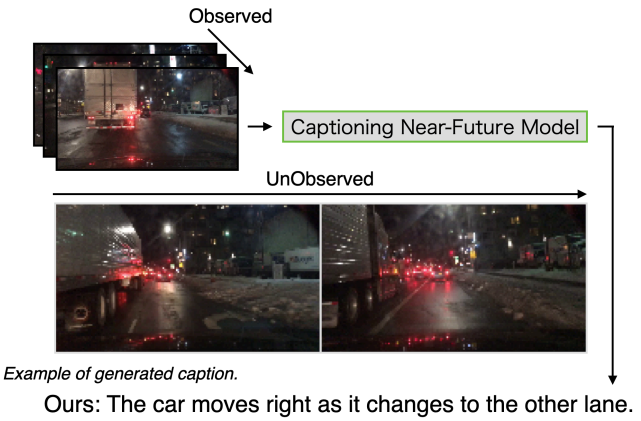


図 1　近未来画像キャプション生成

森らは，自動車の周囲の環境から危険となる因子を物体検出器により発見し，画像キャプション生成に組み込み，乗客に伝えることを目的とした Attention Neural Baby Talk を提案している．本手法では，ルールベースと物体検出器により危険因子を選択し，入力特徴量に Attention マスクを適用することで，特定の危険因子に関する注意喚起のための画像キャプション生成を可能としている．

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，それに適した近未来を考慮したキャプション生成モデルの提案を行う．本章では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，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．

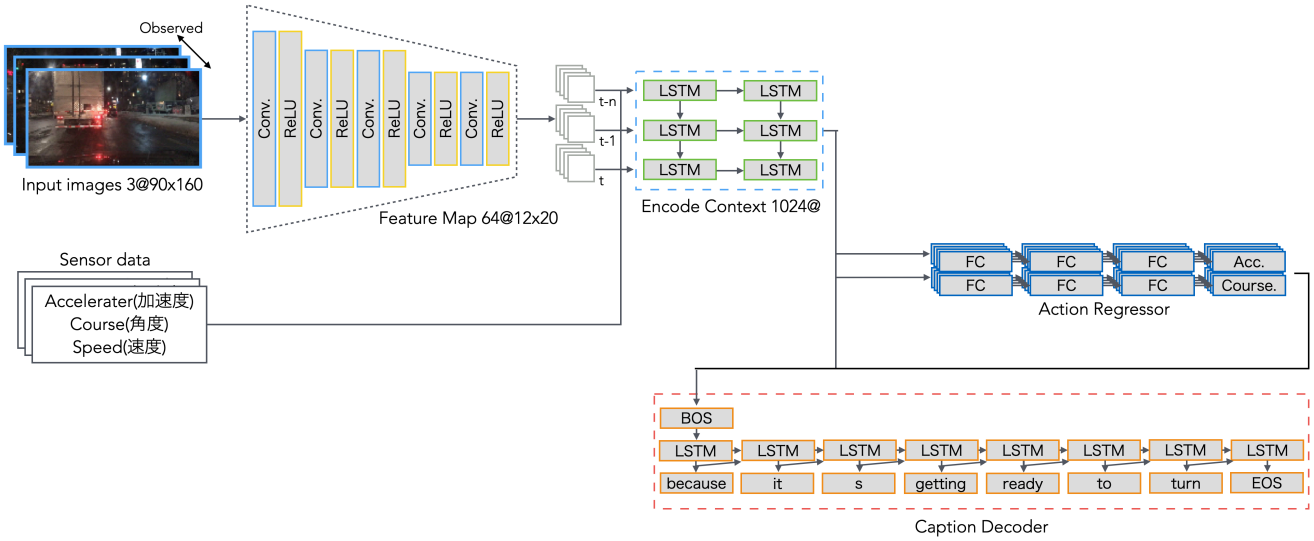


図 2　ネットワーク構造

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．

3.3　特徴抽出ネットワーク

特徴抽出ネットワークは，5 層の CNN で構成された畳み込みニューラルネットワークである．各活性化関数には ReLU を使用する．各時刻の画像をそれぞれ特徴抽出ネットワークに入力し，特徴ベクトルを獲得する．ここでは，各時刻の特徴ベクトルを個別に抽出する．そして，各時刻の観測情報としてエンコーダに入力する．特徴抽出ネットワークは，Berkeley Deep Drive eXplanation Dataset を用いて事前学習を行う．本データは車載カメラ映像から自動車の加速度や角度などのセンサ情報を推定することができる．これにより，車載カメラ画像を対象としたキャプション生成に適した特徴を獲得することができる．事前学習したネットワークのうち，最終層の手前までを特徴抽出ネットワークとして利用する．

3.4　エンコーダ

エンコーダは 2 層の LSTM から構成されている．特徴抽出ネットワークで獲得した各時刻の特徴ベクトルを入力し，時系列変化を考慮した中間表現を獲得する．LSTM は，1024 ユニットあり，中間表現のベクトルは 1024 次元である．　ここで，特徴抽出ネットワークから 5 フレーム分の特徴ベクトルを入力するため，時系列変化を考慮した中間表現ベクトルとなっている．

3.5　動 き 情 報

時系列での変化を捉えるために，画像情報だけでなくセンサから獲得した動き情報もエンコーダに入力する．動き情報としては，速度，加速度，ステアリング角度を利用する．速度と加速度の両方を利用することでどの程度の速度において加減速したかどうかを考慮した特徴となる．また，ステアリング角度を入力することで直進中なのかカーブを曲がっているかなどを考慮できる．

3.6　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，5 フレーム分の速度とステリ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．

3.7　デ コ ー ダ

デコーダは 1 層の LSTM から構成されている．エンコーダで獲得した中間表現 1024 次元と Action Regressor の出力 10 次元をデコーダに入力し，単語列を出力する．入力次元数は 1034 次元であり，LSTM のユニット数は，1024 ユニットある．

3.8　提案手法の学習

提案手法は，画像から特徴ベクトルを獲得する特徴抽出ネットワーク，画像特徴ベクトルを中間表現に変換するエンコーダ，自車の近未来の動き情報を推定する Action Regressor，近未来のキャプションを生成するデコーダから構成されている．特徴抽出器ネットワークとエンコーダ・デコーダ部分は別々に学習する．特徴抽出ネットワークは Berkeley Deep Drive eXplanation Dataset を用いて，自車の速度とステアリング角度を教師データとして事前学習した 5 層の CNN を用いる．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

表 1　Berkeley Deep Drive eXplanation のイベント例

Event	Description	Explanation
加減速	The car is accelerates	because the street is clear of traffic.
右左折	The car turns left	since the are no cars blocking the way.
車線変更	The car moves right	as it changes to the other lane.
合流	The car merges left	as it enters the highway.
後退	The car backs up	to start parallel parking.

表 2　Description 部の頻出単語

BDD-X action descriptions	
Word	Count
stop	6879
slow	6122
forward	4322
drive	3994
move	3273

表 3　Explanation 部の頻出単語

BDD-X action explanations	
Word	Count
traffic	7486
light	6116
red	3979
move	3915
clear	3660

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．Action Regressor の損失関数には，平均二乗誤差，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，損失関数は，式 (1) のように定義できる．

$$L=\sum_t((a_t-a'_t)^2+(c_t-c'_t)^2+\sum_k \log p(y_k\,|\,y_{k-1},h_k,x_k) \quad (1)$$

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 [6] 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．

4.1 実 験 概 要

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．制御イベントのアノテーションには，自動車の制御イベントの所要時間，自動車の行動理由説明のキャプションアノテーションが含まれる．キャプションアノテーションは，自車の行動を言語化した Description 部分，行動理由を言語化した Explanation 部分からなる．主なイベントを表 1 に示す．代表的なイベントに，加減速や右左折，車線変更，合流，後退がある．また，それぞれのイベントが生じたときの理由として，前方が空いている，高速の合流，縦列駐車するためなどがある．

データセットの辞書単語数は 1,290 語である．各部分の頻出単語 5 単語を表 2 および表 3 に示す．これより，Description 部分は stop，forward といった加減速にかかわる単語が多い．Explanation 部分は，light，red などの信号機や move，clear などの前方の状況に対する単語が多い．

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 160 × 90 である．学習時，パラメータの初期化には Xavier，各構成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．

表 4　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	**16.58**	**28.71**	**78.94**
未来	11.97	27.70	47.11

4.2 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．

4.3 キャプション生成時刻の評価

本実験では，前節で定義した各キャプション生成時刻における精度を比較する．さらに提案手法で導入した自車のセンサ情報および Action Regressor の有用性を各生成時刻でも確認する．表 4 にキャプション生成時刻による生成精度を示す．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 16.58 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 28.71，78.94 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．一方で，イベント発生区間以前を入力とする ‘未来’ は，BLEU@4 の評価指標で 11.97 と大幅に精度低下している．これは，イベントが起こるまでの時間が長く，画像やセンサ情報の変化がイベントと一致しないためであると考える．これより，長時間先の未来を予測することに限界があることがわかる．

4.4 センサ情報および Action Regressor の有用性の評価

表 5 に，センサ情報および 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，センサ情報を追加することで近未来の精度が BLEU@4 において 16.74 とセ

表 5　センサ情報および 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
センサのみ	現在	16.16	28.67	75.61
	近未来	**16.74**	**28.77**	77.33
Action Regressor	現在	13.75	27.91	59.35
のみ	近未来	12.85	27.06	48.50
センサ情報・	現在	15.97	28.20	74.96
Action Regressor あり	近未来	16.58	28.71	**78.94**

ンサ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 77.33 と大幅に向上していることがわかる．Action Regressor を追加した場合も同様に各評価指標で精度向上しているが，センサ情報のみの場合よりも低下している．これは，Action Regressor の速度およびステアリング角度の予測が正しくできていない時，キャプション生成精度が低下していることが考えられる．一方，センサ情報と Action Regressor の両方を追加すると，BLEU@4 や METEOR の評価指標では，センサ情報のみ追加した場合よりも精度が若干低いが，CIDEr の評価指標では，78.94 とセンサ情報のみの場合よりも精度が向上している．CIDEr は，同じ単語の曖昧な表現に対する評価を考慮できる指標である．センサ情報と Action Regressor を導入することで，キャプション生成の表現力が向上していると考えられる．

4.5　生成キャプションの可視化

前節での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．1 つ目は，信号機が赤色となり前方車に合わせて速度を落としているイベントが発生しているシーンである．センサ情報がない場合，停止したというキャプションとなっているが，センサ情報を用いることでブレーキをかけるという速度を落とす表現になっている．また，センサ情報および Action Regressor を利用した場合は slow down とより速度を落とす表現になっている．また，画像特徴としては，信号機付近に強く反応していることがわかる．次に，２つ目は，右に曲がるイベントが発生しているシーンである．センサ情報を用いる場合，およびセンサ情報と 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．３つ目は，前方車が速度を落としたため，自車も速度を落とすイベントが発生しているシーンである．このシーンにおいてもセンサ情報があると slowed down や slows down のように速度を落とす表現を正しくキャプション生成できている．画像特徴も前方車に強く反応している．４つ目は，前方車が道路端に停車するため，自車が停車するイベントが発生しているシーンである．このシーンでは，slow down と速度を落とす表現が生成されている．センサ情報を確認すると速度が 0 ではなく，徐行している値であるため，アノテーションされたキャプションのように stop という表現になっていない．しかしながら前方車がいるため速度を落としたとキャプションできており，他車が影響していることを的確に捉えている．画像情報も前方車に強く反応していることがわかる．

このように，画像およびセンサ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，自車のセンサ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，近未来を考慮した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，近未来を考慮したキャプション生成が可能であることを示した．また，自動車の加速度や速度などのセンサーデータを，車載カメラを対象としたキャプション生成の入力に用いる有効性を確認した．今後の課題としては，車載カメラを対象としたキャプション生成以外への応用が考えられる．また，より遠い未来を考慮するキャプション生成に必要な要素の探索，改善も検討する．

文　　献

[1] O. Vinyals, A. Toshev, S. Bengio, "Show and tell: A neural image caption generator." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2015.

[2] K. Xu, J. Ba, R. Kiros, "Show, attend and tell: Neural image caption generation with visual attention." In International Conference on Machine Learning，2015.

[3] J. Lu, C. Xiong, D. Parikh, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2017.

[4] J. Lu, J. Yang, D. Batra, "Neural baby talk." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018.

[5] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2019.

[6] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for selfdriving vehicles." In Proceedings of the European Conference on Computer Vision，2018.

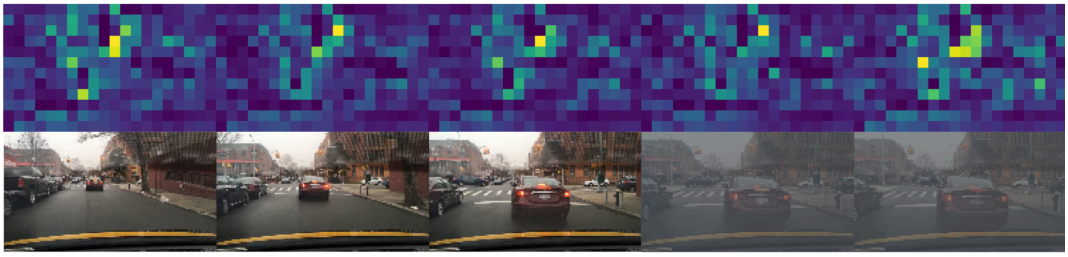
[7] Y. Mori, H. Fukui, T. Hirakawa, N. Jo, T. Yamashita, H. Fujiyoshi, "Attention neural baby talk: captioning of risk factors while driving." In Proceedings of the IEEE International Conference on Intelligent Transportation Systems，2019.

[8] K. Papineni, S. Roukos, T. Ward, "Bleu: a method for automatic evaluation of machine translation." In Annual Meeting of the Association for Computational Linguistics, 2002.

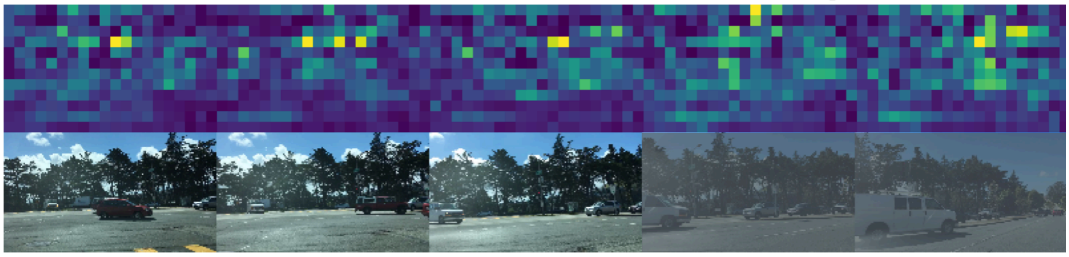
[9] M. Denkowski, A. Lavie, "Meteor universal: Language specific translation evaluation for any target language." In European Chapter of the Association for Computational Linguistics Valencia, 2014.

[10] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation. " In arXiv:1411.5726, 2015.

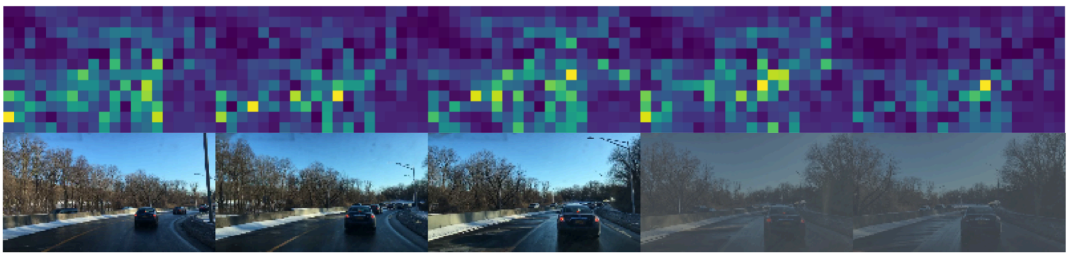
[11] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation." In Proceedings of the IEEE Conference on Computer Vision



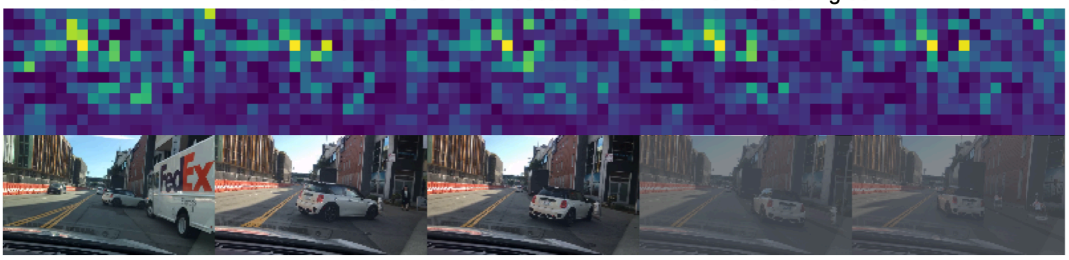
正解文: The car slows to a stop behind another car because the light ahead is red.
センサ情報・Action Regressorなし: The car is stopped because the light is red.
センサ情報のみ: The car brakes to a stop because the light is red.
Action Regressorのみ: The car is stopped because the light is red.
センサ情報・Action Regressorあり: The car slows down because the car in front is slowing.



正解文: The car is turning right because traffic is clear enough to turn.
センサ情報・Action Regressorなし: The car is driving forward because the road is clear of traffic.
センサ情報のみ: The vehicle is turning right the car is turning to the right.
Action Regressorのみ: The car is driving down the highway because the lane is clear.
センサ情報・Action Regressorあり: The car is turning right because there is no traffic.



正解文: The car slows down because the car is front is slowing down.
センサ情報・Action Regressorなし: The car is stopped because the light is red.
センサ情報のみ: The car slowed down because it s approaching a stop sign.
Action Regressorのみ: The car is stopped because the light is red.
センサ情報・Action Regressorあり: The car slows down because the car in front is slowing down.



正解文: The car stops because a car has backed up out of a driveway into the lane.
センサ情報・Action Regressorなし: The car is driving forward because the road is clear.
センサ情報のみ: The car slows down to a stop because it is approaching another light.
Action Regressorのみ: The car is driving forward because the road is clear.
センサ情報・Action Regressorあり: The car slows down because the car in front is slowing down.

図 3　特徴量の可視化および生成されたキャプション

and Pattern Recognition, 2019.

[12] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, "A.: Learning deep features for discriminative localization." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[13] S. Ramprasaath, R., C. Michael, D. Abhishek, V. Ramakrishna, P. Devi, and B. Dhruv, "Grad-cam: Visual explanations from deep networks via gradient-based localization." In International Conference on Computer Vision, pages 618–626, 2017.

[14] Y. Ushiku, T. Harada, and Y. Kuniyoshi, "Efficient image annotation for automatic sentence generation." In ACM International Conference on Multimedia, 2012.

[15] Y. Ushiku, M. Yamaguchi, Y. Mukuta, and T. Harada, “Common subspace for model and similarity: Phrase learning for caption generation from images.” In International Conference on Computer Vision, 2015, pp. 2668–2676.