Neural Baby Talk による注意喚起を目的とした運転シーンのキャプション自動生成

森 優樹† 福井 宏† 平川 翼† 西山 乘‡ 山下 隆義† 藤吉 弘亘†
†中部大学 ‡日産自動車
E-mail: yukiri@mprg.cs.chubu.ac.jp

1 はじめに

事故防止を目的とした運転支援システムの実現には，搭乗者への適切な注意喚起が必要である．また，自動車の走行シーンには複数の危険因子が存在する．歩行者や対向車などの危険因子を運転中でも安全に搭乗者へ注意喚起する方法として，画像キャプション生成が考えられる．

画像キャプション生成とは，1 枚の入力画像から対応したキャプションを生成する手法である．近年 Convolutional neural network (CNN) 及び Recurrent neural network (RNN) を活用したキャプション生成のアプローチが多数提案されている．このアプローチは，入力画像を特徴量へエンコードする CNN と，エンコードされた特徴量をキャプションへデコードする RNN で構成される Encoder-Decoder モデルを採用しており，より自然なキャプション生成を可能としている．

キャプション生成モデルの学習には，画像に対して人手により付与された正解キャプションのデータセットが必要である．データセットの作成には多大なコストがかかる上，一定の品質の確保が難しいという問題点がある．理由として，正解キャプションが作成者ごとの特性により異なり，付与されるキャプションの品質に差が生じることがあげられる．

また，従来のキャプション生成モデルの多くは，1 枚の画像に対して 1 つのキャプションのみを生成する．これは図 1 のような運転シーンなど，複数の注目したい物体があるシーンには不適であるという問題点がある．

そこで本研究では上記の 2 つの課題を解決し，キャプション生成による運転支援システムを実現する．提案手法では 2 つの取り組みを行う．1 つ目は，物体検出を用いて運転シーン上の物体からルールベースによる属性抽出を行い危険因子を選別し，注意喚起に適したデータセットの自動作成を行う．2 つ目は，キャプション生成の既存手法である Neural Baby Talk (NBT) に対して Attention マスクを適用し，運転シーンにおける各危険因子ごとにキャプションの生成を可能とする．

本論文の貢献は次の通りである．

図 1 従来手法の生成キャプション例

- ルールベースによる運転シーンに適した独自データセットの自動作成を提案する．Faster R-CNN による物体検出と，ルールベースによる属性抽出により，画像に対してキャプションの正解ラベルを自動で付与する．これにより従来と比べ少ないコストでデータセットを作成できる．
- 画像キャプション生成による運転シーンの注意喚起システムを実現する．NBT に対して Attention マスクを適用することで，運転シーンにおける各危険因子に注目したキャプションの生成が可能となる．これにより 1 枚の画像に対して複数のキャプションの生成が可能となり，既存モデルの問題点を解決できる．

2 関連手法

2.1 Show and Tell

深層学習による画像キャプション生成の代表的な手法に，LSTM [2] を用いたキャプション生成 [3] がある．LSTM を用いたキャプション生成は，画像を特徴ベクトル $\boldsymbol{x_{-1}}$ に変換する DCNN 部分，文章中の単語を特徴ベクトル W_e に変換する部分，そして，特徴ベクトル $\boldsymbol{x_t}$ を LSTM を入力し，次の単語の出現確率 p_t を求める部分から構成される．入力画像を I，キャプション開始記号を S_0，各ステップ t における LSTM の出力結果を $S = \{S_1, S_2, ..., S_{N-1}\}$ とすると，式 (1) より，画像を特徴ベクトル $\boldsymbol{x_{-1}}$ へ変換，時刻 t の LSTM の入力

は式 (2) により求められ，単語の出現確率は式 (3) で示される．W_e は単語埋め込みによる分散表現である．

$$\boldsymbol{x_{-1}} = P^{CNN}(I) \quad (1)$$

$$\boldsymbol{x_t} = W_e S_t, \quad t \in \{0, ..., N-1\} \quad (2)$$

$$p_{t+1} = P^{LSTM}(x_t), \quad t \in \{0, ..., N-1\} \quad (3)$$

2.2 Show, Attend and Tell

RNN [4] や LSTM を用いたキャプション生成では，扱う系列情報が長いほど情報の伝播がしづらくなり精度が低下する問題がある．この問題を解決するために，Attention 機構を取り入れたキャプション生成 [6] の手法が提案されている．Attention 機構は，ネットワークが抽出した特徴から重視する特徴を選択，学習し重み付けする手法であり，キャプション生成において高い精度を実現している．Attention 機構には，複数個の入力系列に由来するベクトルの重み付け平均を用いる Soft Attention 機構と複数の要素の中から 1 つを選択する Hard Attention 機構がある．

2.3 Adaptive Attention

Attention 機構を取り入れたキャプション生成では，文章中の前置詞や接続詞などの画像の情報を必要としないと考えられる単語に対しても画像の特徴量を考慮してキャプションを生成している．そこで，単語を生成する際に，画像の特徴量を利用すべきか判断する Adaptive Attention 機構を取り入れたキャプション生成 [8] の手法が提案されている．Adaptive Attention 機構では，CNN により k 個のグリッドに分割された画像の特徴量 $V = \{v_1, v_2, ..., v_k\}$, $v_i \in \mathcal{R}^d$ と LSTM の中間層の出力 $h_t \in \mathcal{R}^d$ を用い，重み行列 $W_v, W_g \in \mathcal{R}^{kxd}$ 及び $w_h^T \in \mathcal{R}^k$ を用いて z_t を式 (4) で求め，式 (5) より，画像の特徴量に対する Attention の重み $\alpha \in \mathcal{R}^k$ が得られる．キャプションの生成に用いられる重み付き平均ベクトル c_t は式 (6) のように表される．

$$z_t = w_h^T \tanh(W_v V + (W_g h_t)\mathbb{1}^T) \quad (4)$$

$$\alpha = \mathrm{softmax}(z_t) \quad (5)$$

$$c_t = \sum_{i=1}^{k} \alpha_{ti} v_{ti} \quad (6)$$

画像の特徴量をキャプション生成に用いるか否かの判断には，visual sentinel ベクトル s_t を用いる．LSTM の入力 x_t と時刻 $t-1$ の LSTM の中間層の出力 h_{t-1}，LSTM のセル状態 m_t とすると，s_t は式 (7)，(8) となる．s_t は LSTM を拡張して求められるもので，LSTM のセル状態 m_t に対して g_t と要素ごとの積を取ることで，キャプション生成に画像の特徴量を考慮すべきか判断できる．

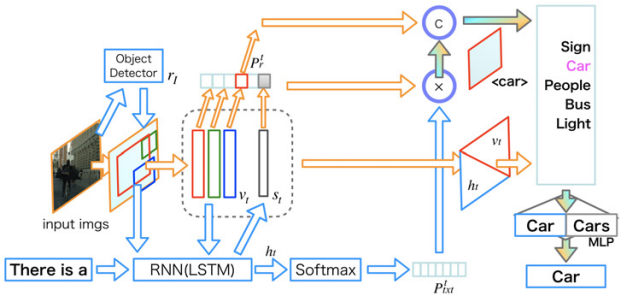

図 2　Neural Baby Talk のネットワーク図

$$g_t = \sigma(W_x x_t + W_h h_{t-1}) \quad (7)$$

$$s_t = g_t \odot \tanh(m_t) \quad (8)$$

visual sentinel を考慮した重み付き平均ベクトル $\hat{c}_t$ は，式 (9) で求めることができる．β_t は時刻 t のキャプション生成に画像の特徴量を考慮するかどうかを示す $[0, 1]$ の範囲を取るゲートとなっており，値が 0 ならば s_t，値が 1 ならば c_t がキャプションの生成に用いられる重み付き平均ベクトル $\hat{c}_t$ となる．式 (4) で求められる z_t と重み行列 W_s，式 (4) で用いた重み行列 $W_g \in \mathcal{R}^{k \times d}$ 及び LSTM の中間層の出力 $h_t \in \mathcal{R}^d$ から式 (10) により求められる visual sentinel ベクトル s_t を考慮した Attention の重み $\hat{\alpha}_t$ の最後の要素 $\hat{\alpha}_t[k+1]$ を β_t として用いる．

$$\hat{c}_t = \beta_t s_t + (1 - \beta_t) c_t \quad (9)$$

$$\hat{\alpha}_t = \mathrm{softmax}([z_t; w_h^T \tanh(W_s s_t + W_g h_t)]) \quad (10)$$

Adaptive Attention 機構を取り入れたキャプション生成の最終的な単語の生成確率は，重み行列 W_p 及び式 (9) で求めた重み付き平均ベクトル $\hat{c}_t$，LSTM の中間層の出力 $h_t \in \mathcal{R}^d$ を用いて，式 (11) で表すことができる．

$$p_t = \mathrm{softmax}(W_p(\hat{c}_t + h_t)) \quad (11)$$

2.4 Neural Baby Talk

Adaptive Attention 機構を取り入れたキャプション生成では，画像の特徴量をキャプション生成に利用するか判断することにより，高い精度を実現している．これに加え，キャプション生成に物体検出を利用したキャプション生成の手法である Neural Baby Talk (NBT) [1] が提案されている．NBT は，入力画像に対して，Region Proposal Network (RPN) による物体検出を行い，RoI Pooling を施すことで，検出された物体領域の特徴量及びラベルをキャプション生成に用いる手法である．NBT のネットワーク構造を図 2 に示す．

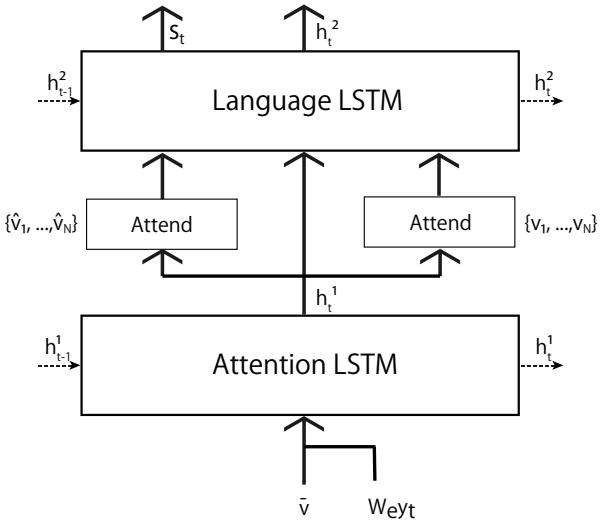

図 3　NBT における言語モデル

NBT は，検出した物体候補領域の特徴量及びラベルをキャプション生成に取り入れる機構を持ち，物体候補領域のラベルと言語モデルの出力 y^{txt} の生成確率を求めることで，この機構を実現している．物体候補領域ごとのキャプションに加える確率分布 P_{rI}^t は，Pointer Networks [9] を用いて算出される．Pointer Networks で用いる入力要素へのポインタ u_i^t は，重み行列 $W_v, W_g \in \mathcal{R}^{kxd}$ 及び $w_h^T \in \mathcal{R}^k$ 及び物体検出により得られた候補領域の特徴量 v_t，LSTM の中間層の出力 $h_t \in \mathcal{R}^d$ により式 (12) で示され，式 (12) で得られた u_i^t から式 (13) により，確率分布 P_{rI}^t を算出する．

$$u_i^t = w_h^T \tanh(W_v v_t + (W_g h_t)\mathbb{1}^T) \quad (12)$$

$$P_{rI}^t = \mathrm{softmax}(u_i^t) \quad (13)$$

また，言語モデルの出力 y^{txt} を採用する確率は，式 (14) で求める．言語モデルの出力を採用する確率は，Adaptive Attention 機構を応用し，式 (7)，(8) と同様に求められる visual sentinel ベクトル s_t を用いて算出する．visual sentinel ベクトルを考慮した物体候補領域ごとのキャプションに加える確率分布 P_r^t は式 (15) となり，式 (15) の最後の要素が visual sentinel ベクトル $\tilde{r}$ である．

$$p(y_t^{txt}|y_{1:t-1}) = p(y_t^{txt}|\tilde{r}, y_{1:t-1})p(\tilde{r}|y_{1:t-1}) \quad (14)$$

$$P_r^t = \mathrm{softmax}([u^t; w_h^T \tanh(W_s s_t + W_g h_t)]) \quad (15)$$

visual sentinel ベクトル $\tilde{r}$ を，式 (14) の $p(\tilde{r}|y_{1:t-1})$ に適用し，式 (16) により，言語モデルの出力 y^{txt} の条件付き確率を求める．式 (15) と式 (16) を式 (14) に適用することで，visual sentinel を考慮した言語モデルの出力 y^{txt} の条件付き確率が求められる．P_r^t と P_{txt}^t のうち，NBT はより高い確率を選択することで，キャプションを生成する．$W_q \in \mathcal{R}^{V \times d}$ は重み行列である．

$$P_{txt}^t = \mathrm{softmax}(W_q h_t) \quad (16)$$

NBT では，従来の Attention モデルとは違い，2 層の LSTM 層からなる Attention モデル [10] を採用しているため，グリッドごとでは無く物体候補領域ごとに Attention を適用することが可能となっている．図 3 に言語モデルの構成を示す．物体候補領域の特徴量は $V = \{v_1, v_2, ..., v_k\}$, $v_i \in \mathcal{R}^d$，CNN により k 個のグリッドに分割された画像の特徴量は $\hat{V} = \{\hat{v_1}, \hat{v_2}, ..., \hat{v_k}\}$, $\hat{v_i} \in \mathcal{R}^d$ とすると，Attention の重みは式 (4)，(5) で算出される．

NBT では，物体検出をキャプション生成に取り入れることで，より精度の高いキャプション生成を可能とする．

3　提案手法

提案手法では，NBT をベースとした注意喚起に適したキャプション生成法を実現するために，2 つの取り組みを行う．1 つ目は，学習データに対してルールベースの自動アノテーションを施し，注意喚起に適したデータセットを作成する．2 つ目は，NBT に Attention マスクを適用する．これにより，従来法の問題点を解決でき，RPN で検出した危険因子に注目したキャプションを生成できる．

3.1　注意喚起に適したデータセットの作成

従来のキャプションのデータセット作成は，人手により行われている．そのためデータセットの作成には多大なコストがかかるうえ，正解キャプションのアノテータごとの特性により，付与されるキャプションの品質に差が出るため，一定の品質の確保が難しいという問題点がある．従来のデータセットは運転シーンに適しておらず，適切な注意喚起のためには独自データセットを作成する必要がある．そこで，本研究ではコストの削減を目的としたデータセットの自動作成を提案する．

表 1　アノテーションルール

優先度	状況
1	人が道路を横断中
2	人が歩道上に存在
3	信号検出
4	看板検出
5	駐車多数
6	距離が近い検出クラス

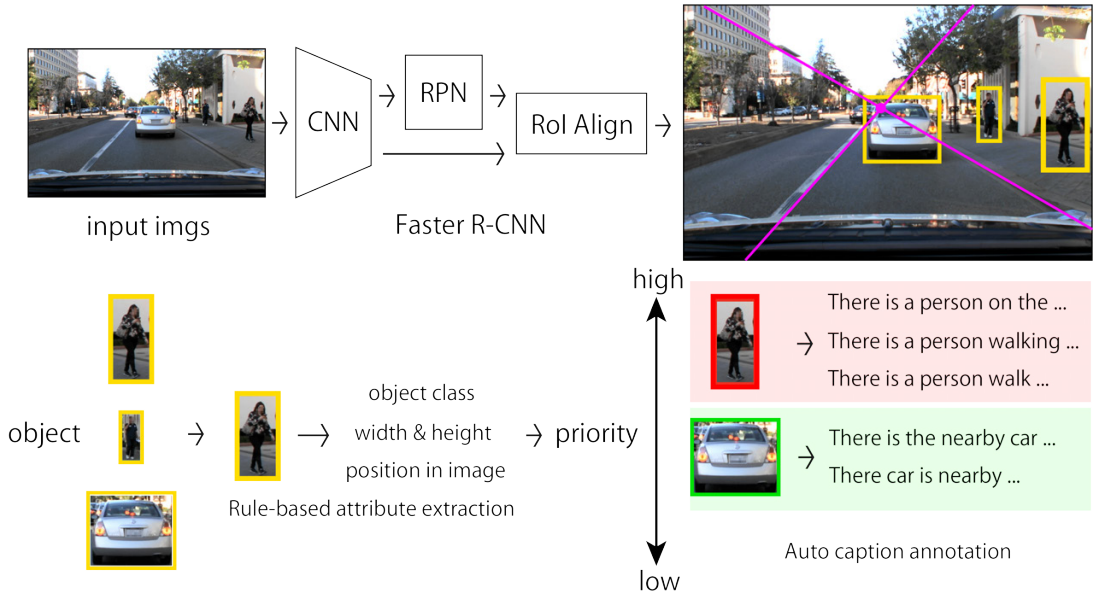

図 4　自動アノテーションによるデータセットの作成

図 4 にデータセットの自動作成の概要を示す．まず，運転シーン画像に対して物体検出を行う．検出物体の中から抽出した危険因子をもとに，一般道の運転シーン画像 30,320 枚からなる独自データセットに対してキャプションの正解ラベルを自動で付与する．運転シーンにおける危険因子の抽出には，物体の種別，位置，距離が重要であり，これらの要素を考慮して検出物体から危険因子を抽出する．物体検出器には Faster R-CNN を用いる．Faster R-CNN の学習は COCO Dataset [11] により行う．Faster R-CNN の検出閾値は 0.9 とする．

次にアノテーションの方法について説明する．まず，注意喚起に必要なクラスを COCO Dataset の 80 カテゴリの中から 5 クラス (Person，Car，Bicycle，Stop sign，Traffic light) を選択する．そして，Faster R-CNN により検出した物体領域のうち，該当するクラスの物体領域を選択する．また，各物体の矩形の下辺中央を基準点 (図 5 の左参照) として定める．運転シーン画像に対してハフ変換を行い，検出した直線が交差する部分を消失点とする．消失点と基準点から，図 5 に定めるように，”left”，”right”，”center”の 3 種類の方向属性を，各検出物体に対し付与する．また，各検出物体の矩形の面積から，検出物体の距離を推定する．クラスごとに近距離，遠距離の閾値を設定し，”normal”，”nearby”，”far” の 3 種類の距離属性を，各検出物体に対して付与する．

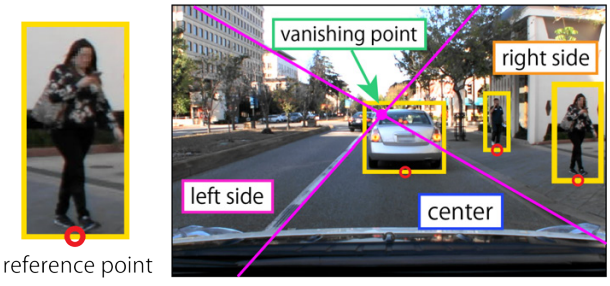

図 5　アノテーションルールの例

付与された方向，距離属性の情報をもとに，危険因子属性を付与する．危険因子属性とは，運転シーンにおいての危険な状況を考慮した属性のことであり，人が道路を横断している状況などが対象である．本研究では，危険因子属性に表 1 に示す 6 状況を設定する．危険因子属性には優先度を設定し，優先度の高い 2 つの属性を運転シーン画像に対して付与する．

データセットの作成方法について説明する．独自データセットとして用意した一般道の運転シーン画像 30,320 枚を学習用に 25,987 枚，評価用に 4,333 枚に分けて利用する．アノテーションで付与した方向属性，距離属性，危険因子属性をもとに，対応する正解キャプションのテンプレートを付与する．正解キャプションの例を図 6 に示す．運転シーン画像に付与された危険因子属性から，優先度の最も高い危険因子属性に対応するキャプションを 3 文，2 番目に優先度の高い危険因子属性に対応するキャプションを 2 文，計 5 文を正解キャプションのラベルとして付与する．しかし，危険因子が優先度 6 のみの場合は，優先度 6 に関するキャプションを 3 文，COCO Dataset で学習した NBT により生成されたキャプションを 2 文，計 5 文を正解キャプションのラベルとして付与する．また，危険因子属性を持つ検出物体の優先度上位 3 つを評価時に用いる検出物体として用いる．

画像 30,320 枚に対して計 151,600 文のキャプションを付与して独自データセットの作成を行う．

3.2　Neural Baby Talk による学習

前節で作成した独自データセットを用いて，NBT の学習を行う．NBT の物体検出部分には COCO Dataset で学習した RPN を用い，NBT の言語モデルにおける Attention LSTM と Language LSTM のユニット数は 512 とする．COCO Dataset で事前学習を行ったモデ

ルをもとに，独自データセットを再学習する形で学習を行った．CNN 及び LSTM の最適化手法には Adam を使用し，CNN の学習率は 0.00001，LSTM の学習率は 0.0005 とする．学習には独自データセット 30,320 枚のうち 25,987 枚，129,935 文を利用した．また，学習回数は 15 エポック，バッチサイズを 10 とする．

3.3 Attention マスクを適用したキャプション生成

NBT において，RPN で検出した各物体候補領域に対する Attention の重み a_t をマスクを用いて制御することで検出物体群のうち特定の物体候補領域にのみ注目したキャプション生成を可能とする．提案手法の概要を図 7 に示す．

Attention マスク $\mathbf{A_t}$ は検出した危険因子の個数だけ要素を持っており，キャプションに使用する要素を 1 にした one hot vector である．Attention マスクと物体候補領域の特徴量 v_i の重み付き和で求めたベクトル c_t は式 (6) を拡張した式 (17) となる．

$$c_t = \sum_{i=1}^{N} v_i \cdot a_{ti} \cdot \mathbf{A}_{\mathrm{ti}} \tag{17}$$

また NBT では，従来の Attention モデルとは違い，2 層の LSTM 層からなる Attention モデルを採用しているため，グリッドごとでは無く物体候補領域ごとに Attention を適用することが可能となっている．そこで，注目させたい危険因子の物体候補領域 v_t を Attention として用いることで，危険因子に注目したキャプション生成を可能としている．

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，作成したデータセットによる注意喚起キャプション生成の性能を評価する．そして，NBT に Attention マスクを適用した提案手法による各危険因子に注目した複数のキャプション生成の性能を評価する．本実験ではアンケート及び自動評価指標 [12, 13] を用いて生成キャプションを評価する．

作成したデータセットの有用性を示すため，アンケートを行った．アンケートには，COCO Dataset で学習

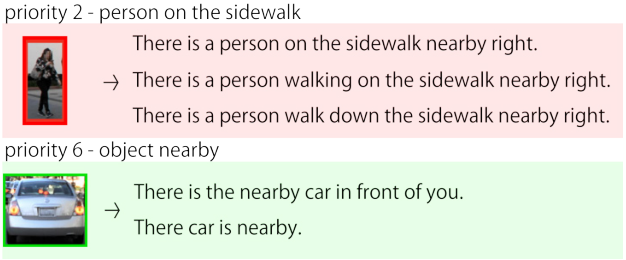

図 6　付与した正解キャプション例

した NBT と提案手法のデータセットで学習した NBT で優先度が最も高い物体に注目して生成したキャプションを用いた．アンケートは評価者 59 名を対象に行った．評価者は 5 グループに分かれ，各グループに画像 20 枚，計 100 枚に対して生成されたキャプションの評価を行った．アンケートの形式は画像 1 枚に対して，各手法で生成したキャプションを 1 文ずつ A，B として手法を隠した状態で表示し，画像のシーンの注意喚起に適したキャプションを選択する形式で行った．選択肢は，“A のキャプションが適している”，“B のキャプションが適している”，“両方適している”の 3 種類を提示し，生成キャプションが適している割合を適合率とし，比較評価した．

提案手法による各危険因子に注目した複数キャプション生成の有用性を示すため，自動評価指標による評価を行った．自動評価指標を用いた評価は，従来の NBT と，提案手法の NBT で生成したキャプションを用いた．それぞれの学習は独自データセットで行った．提案手法の NBT では，提案手法のルールベースで求めた危険因子の優先度をもとに Attention マスクを用い，優先順位順にキャプションを計 3 文用意した．優先度順に生成したキャプションの精度を確かめるため，独自データセットの正解キャプション 5 文のうち 3 文，優先順位第 1 位に関するキャプションを参照文として，優先度第 1 位の物体に注目した生成キャプションとの各自動評価指標を算出した．

4.1 アンケートによる評価

アンケートによる評価結果を表 2 に示す．評価結果では，従来手法に比べ，提案手法の適合率が 20 ポイント優れていることが分かる．提案手法では，危険因子の種別，位置，距離に関する単語を含んだキャプションが生成されたことから，適合率が上昇したと考えられる．各手法の適合率より，COCO Dataset に比べて独自データセットが注意喚起に適していることが分かる．

表 2　適合率

Method	Questionnaires Precision(適合率)
NBT	43.1%
Our NBT (priority1 only)	**63.2%**

4.2 自動評価指標による評価

自動評価指標による評価結果を表 3 に示す．危険因子の優先順位を考慮した評価において，提案手法が従来手法よりも高い精度を示した．これは，提案手法の NBT が，従来手法に比べて優先順位を考慮したキャプションを生成したことを示している．以上の結果より，提案手法は各危険因子に注目したキャプション生成を実現していることが分かる．

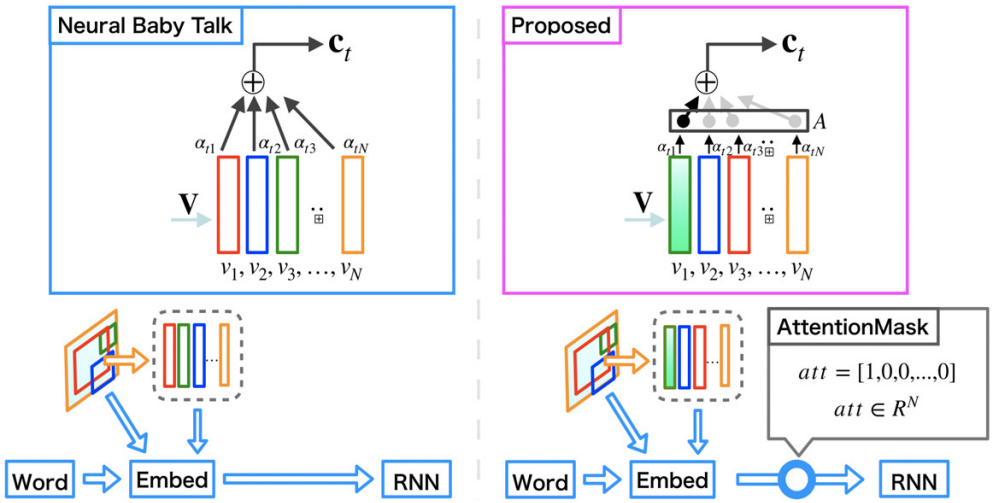

図 7　Attention 機構に対するマスクの適用

表 3　優先度第 1 位の危険因子に注目した自動評価指標による評価

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR
NBT (only pripority1)	62.9	56.6	50.4	46.7	35.1
Our NBT (only priority1)	**71.0**	**66.3**	**62.4**	**59.4**	**40.0**

4.3　生成キャプションの比較

COCO Dataset で学習した従来手法と独自データセットで学習した提案手法の生成キャプションの例を図 8 に示す．図 8 では，従来手法を N，提案手法を優先度順に 1，2，3 の順で示す．画像はキャプション生成に用いた画像であり，矩形は危険因子である．矩形の色は文字色と対応している．図 8 左上より，優先度第 1 位の生成キャプション例に注目する．提案手法において，道路を横断しようとしている女性に対し，“ There is a person on the sidewalk nearby right.”のキャプションが得られた．これは，クラス，距離，位置に関する正しい単語が含まれており，注意喚起に適したキャプションが生成されていると言える．

5　おわりに

本研究では，注意喚起に適したデータセットの自動作成及び各危険因子に注目した複数キャプションの生成を行った．提案手法は，画像に対して物体検出を行い，ルールベースによりアノテーションすることで，データセットの自動作成を可能にした．そして，従来モデルの NBT に対し，Attention マスクを導入することで，各危険因子に注目した複数キャプション生成を可能にした．今後は，提案したアプローチをベースに，より注意喚起に適したデータセットの自動作成手法の構築，及び運転シーンにおけるより危険な状況に対し，適切な注意喚起キャプションを生成できる手法を構築する．

参考文献

[1] J. Lu, J. Yang, D. Batra, *et al.* ”Neural baby talk”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7219–7228, 2018.

[2] S. Hochreiter, ”LONG SHORT-TERM MEMORY”, Neural Computation 9(8) pp. 1735–1780, 1997.

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

[4] A. Graves, A. Mohamed, G. Hinton, ”Speech Recognition With Deep Recurrent Neural Networks”, arXiv preprint arXiv:1303.5778v1, 2013.

[5] Y. Lecun, B. Boser, J. S. Denker, *et al.* ”Backpropagation applied to handwritten zip code recognition”, Neural Computation，vol.1，pp.541–551，1989.

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

[7] S. Ren, K. He, R. Girshick, *et al.* ”Faster R-CNN: Towards real-time object detection with region proposal networks”, In Neural Information Processing Systems, 2015.

[8] J. Lu, C. Xiong, D. Parikh, *et al.* ”Knowing when

図 8　注意喚起のためのキャプション生成例

to look: Adaptive attention via a visual sentinel for image captioning", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[9] O. Vinyals, M. Fortunato, N. Jaitly, *et al.* "Pointer Networks", Advances in Neural Information Processing Systems, pp. 2692–2700, 2015.

[10] P. Anderson, X. He, C. Buehler, *et al.* "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[11] L. Tsung-Yi, M. Michael, B. Serge, Hays, *et al.* "Microsoft coco: Common objects in context", The European Conference on Computer Vision, pp. 740–755. 2014.

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

[13] 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.