2022 - 3 - 4

異常状態を捉えた動画における キャプション自動生成手法の有効性について

対馬 陣 † 松原 雅文 ‡

岩手県立大学大学院ソフトウェア情報学研究科 † 岩手県立大学ソフトウェア情報学部 ‡

1. はじめに

近年，インターネットや SNS の普及により，動画などの情報メディアが増加しており，それに伴い，動画による情報伝達の機会も増加してきている．また，深層学習を用いた動画におけるキャプション生成技術が向上しており，人間の動作を日本語で説明することが可能になってきている．自然言語による説明文付与を用いることで，人が理解可能な形でデータへの解釈を与えることが可能となる．

キャプション生成技術を応用した研究は盛んに行われているが，そこで生成対象となっているデータの多くは画像であり，動画から生成されたキャプションについての応用はあまり行われていない．データに対して，システムがどういう解釈をしているかは，外部からでは分かりにくく，システムと外部で認識の齟齬が発生する可能性がある．そのため，キャプション生成技術を用いて，動画内容における説明文を自動的に生成することで，システムとユーザ間の適切な情報共有が可能になると考えられる．

人間の日常的な動作を捉えて活用する事例として，監視カメラや見守りカメラが考えられる．見守りカメラなどを用いて，日常的な動作を記録した動画から異常を検知する研究が行われている [1]．しかし，これらの研究では，異常検知というタスクに重点が置かれており，検知した異常をキャプションとして出力・活用するといったことはあまり行われていない．

そこで，本研究では，深層学習を用いてキャプション生成モデルを構築し，日常的な動作を記録した動画を用いて学習を行っている [2]．本稿では，この学習したモデルを用いて，見守りカメラ等からの異常動作を捉えた動画を用いてキャプション生成を行い，それについて評価を行うことで，異常状態を捉えた動画におけるキャプション自動生成が有効であるかどうかを検証する．

Effectiveness of Automatic Caption Generation Method for Videos Capturing Abnormal Conditions
Jin TSUSHIMA†, Masafumi MATSUHARA‡
†Graduate School of Software and Information Science, Iwate Prefectural University, ‡Faculty of Software and Information Science, Iwate Prefectural University

2. 関連研究

2.1. 人間の動作を日本語で説明するためのキャプションデータセット

人間の動作を認識し，日本語で説明するための研究が存在する [3]．この研究では，人間の動作を認識・説明するための日本語キャプションデータセットを構築し，構築されたデータセットが人間の動作を認識・説明するのに有効であることが示されている．

提案されているデータセットは，79,822 本の動画と 399,233 文の日本語キャプションで構成されている．構成されている動画には，家庭やオフィスで見られる人間の 100 種類の動作が使用されている．付与されているキャプションは，どこ・誰・動作の要素に分かれて記述されている．そのため，各要素の間に「で」と「が」を補完することで，文章として成立するようになっている．

本提案手法では，このデータセットを学習データとして，キャプション生成モデルに人間の動作を学習させることで，検知されうる異常動作をキャプションとして自動生成する．

2.2. 説明文生成を用いた動作行動予測

説明文生成を用いて，支援行動を想定した動作行動予測を行った研究が存在する [4]．この研究では，ユーザ支援システムにおける柔軟な状況理解と，システムとユーザの適切な情報共有を実現を目的とし，動作行動予測を行っている．そのために，動作行動予測のためのデータセットをクラウドソーシングを用いて構築し，シーングラフの予測を動作行動予測の補助タスクとして同時に学習させることで，キャプショニングにおける精度向上を図っている．

人間を対象とした支援において，システムがどのような状況理解を行っているかを言葉で表現することは重要である．そのため，本研究では，見守りカメラ等から異常動作を捉えた動画を用いてキャプション生成を行うことで，システムとユーザ間の適切な情報共有を支援する．

2.3. BERTScore

自然言語生成における自動評価指標として，BERTScore が提案されている [5]．この研究では，事

前学習された BERT から得られる，文脈上の埋め込みと生成文の各トークンと正解文の各トークンを利用した自動評価指標を提案している．それらのベクトル表現を利用して，トークン間の類似度を算出することにより，文章の意味を考慮した評価が可能になっていると考えられる．

また，363 種類の機械翻訳と画像キャプション生成システムから得られる出力を用いて，よく使用されている評価指標である，BLEU や CIDEr などの評価手法との比較実験を行っている．この実験から，BERTScore は他の評価指標と比べて，人手評価値との相関が高く，評価指標として有効であることが示されている．そのため，本研究では，この評価指標を用いて，生成されたキャプションにおける定量的評価を行う．

3. 提案手法

3.1. キャプション生成モデルの構築

提案手法における構築モデルの概要を図 1 に示す．動画は画像が時系列に並んでいるデータであり，テキストは単語が時系列に並んでいるデータである．そのため，本手法では，系列データを別の系列データに変換することができる Seq2Seq[6] をベースとして，モデルを構築する．しかし，Seq2Seq の Encoder は，入力の長さに関係なく，特徴量を固定長ベクトルに変換してしまうという問題点がある．特徴量が入力の長さに関係なく，固定長ベクトルに変換されてしまうことで，動画の特徴が捉えきれない可能性がある．

そのため，本手法では，固定長ベクトルでは捉えることのできない特徴を学習させるため，Seq2Seq に Attention 機構 [7] を付与したモデルを用いて，キャプション生成モデルを構築する．また，Seq2Seq を構成する Recurrent Neural Network(RNN) には，RNN の一種である Long Short Term Memory(LSTM) ネットワークを用いる．Encoder で LSTM における各タイムステップごとのベクトルのスコアを計算し，Softmax を取ることで，各タイムステップの出力に影響を与える入力における情報の割合を計算する．その後，その割合に基づいて入力の隠れ層を足し合わせ，OutputStates として，Decoder における各タイムステップの出力と結合して変換し，Decoder における各タイムステップの最終的な単語予測確率を算出する．単語予測における次元数は，学習に用いたテキストの語彙数とする．

3.2. キャプション生成モデルの学習

学習データには，動画とその内容を説明するキャプションを用いる．また，学習前に，各データを学習に適した形に変換する．

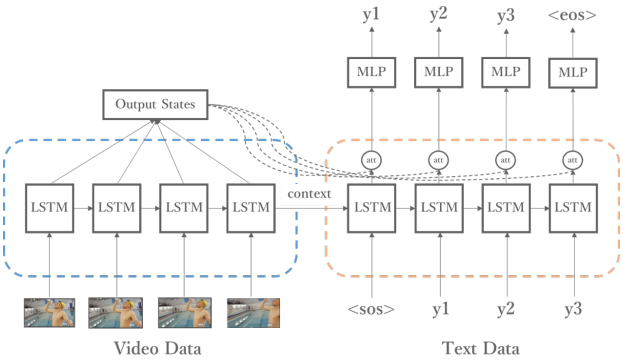

図 1: 構築モデルの概要図

動画データについては，学習に使用する動画における，各フレームの特徴量を学習済み CNN モデルを用いて抽出する．データセットに付与されているキャプションの例を表 1 に示す．データセットにおける，各動画のキャプションについては，付与されているキャプションが，どこ・誰・動作の要素に分かれている．そのため，まず，各要素の間に「で」と「が」を補完し，文章として成立させる．表 1 における例の場合,「部屋で赤いトレーナーの男の子がソファの上でジャンプしている」となる．

補完したキャプションに対して，分かち書きを行い，単語ごとに分割する．分かち書きに用いる形態素解析器には MeCab[2]，形態素解析用の辞書データには mecab-ipadic-NEologd[3] を使用する．分かち書きを行ったテキストに文頭と文末の情報をそれぞれ<sos>，<eos>として付与する．その後，学習に用いるキャプションの語彙数に基づき数値に変換する．

抽出された各フレームの特徴量と，数値に変換されたキャプションを入力としてモデルの学習を行う．

表 1: 付与されている正解キャプション例

どこ	誰	動作
部屋	赤いトレーナーの男の子	ソファの上でジャンプをしている
白い壁紙のレースのカーテンのあるリビング	赤い服を着た男児	ソファの上で飛び跳ねている
窓のある白い壁の部屋	赤い長袖を着た男の子	ソファの上で飛び跳ねている
白い壁の部屋	赤い服を着た少年	ソファの上で飛び跳ねている
白い壁の部屋	赤い服を着た男の子	ソファの上でジャンプしている

3.3. キャプション作成

学習後のモデルに対して，変換した動画と文頭情報<sos>を入力することで，動画の特徴量をもとに文の最初に出現する単語を予測する．その後，予測された単語の次に出現する単語の予測を，文末情報<eos>が

[2] https://taku910.github.io/mecab/
[3] https://github.com/neologd/mecab-ipadic-neologd

出力されるまで再帰的に行い，生成された単語列を入力動画に対するキャプションとして出力する．ここで，予測時に尤度が一番高い単語を選択したからといって，文全体で尤もらしい文が生成されるとは限らない．そのため，本手法では，キャプション生成に Beam Search を使用する．

Beam Search を用いたキャプション生成における，単語予測の流れを図 2 に示す．Beam Search を用いたキャプション生成では，単語予測の各タイムステップにおいて，そこまでの対数尤度が高い単語列を K 個保持しながら単語を選択する．長い系列を見渡して尤度を評価することで，より適切な文を生成することが可能となる．

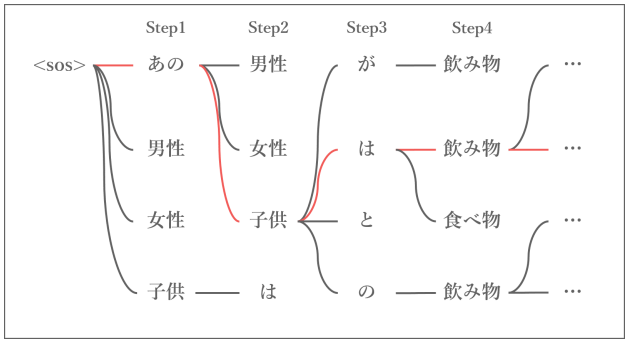

図 2: BeamSearch を用いたキャプション生成 (K=4 の場合)

4. キャプション生成実験

4.1. 概要

提案手法において，学習したモデルが検知されうる異常動作をキャプションとして生成可能か，また，システムとユーザ間の情報共有に適した精度であるかを確認する．

検知されうる異常動作としては，見守りカメラ等で捉えられる日常的な動作における,「倒れている」や「泣いている」等の動作を対象とする．異常と考えられる動作を捉えた動画を用いてキャプション生成を行い，生成されたキャプションについて比較・評価を行う．

4.2. 実験条件

学習データとして，STAIR Actions キャプションデータセット [3] を用いて学習を行う．データセット内の 100 種類ある各動作から，動画を 100 本ずつ，合計 10,000 本使用した．データセットの各動画には 5 つのキャプションが付与されている．そのため，今回は学習テキストとして，合計 50,000 文のテキストを使用した．また，学習の際に用いる CNN モデルには，VGG16 モデル [8] を使用し，Beam Search における K は 5 とした．

評価用データとして，データセットから学習に使用していない動画を任意で選択し，使用する．

4.3. 実験結果と考察

評価用データにおける，正解キャプション，各動画において生成されたキャプションを表 1 に示す．

表 2: 正解キャプションと生成キャプション

正解キャプション	生成キャプション
白い壁の室内で青色の服を着ている男性が床に寝転がっている	白い壁の部屋で黒い服を着た男の子が寝そべっている
室内でボーダーの服の赤ちゃんが泣いている	屋内で白い服を着た赤ちゃんが寝転んで泣いている
屋外でエンジの洋服のひとがあおむけで寝ている	屋外で赤い服を着た女性が寝転がっている
白い壁の部屋で白いパーカを着た子供が泣いている	屋外で白い服を着た男の子が泣いている

表 1 より，生成されたキャプションが，自然な日本語文章で生成されていることから，提案モデルは，システムとユーザ間の情報共有に適した精度で生成可能である可能性が示唆された．また，正解キャプションと同様の行動が出力されていることから，検知されうる異常動作をキャプションとして生成可能である可能性が示唆された．

行動については正解と同様のものであると考えられるが，服の色や背景情報を多く誤認識していた．これは，学習データが少なく，服の色や背景情報についての学習が足りていないことが原因であると考えられる．また，場所を「屋外」と誤認識しているものについては，動画における背景の壁が白く，学習データにおける屋外には明るくて白い場所が多かった．そのため,「屋外」であると誤認識したのは，背景が白い以外の細かい特徴を捉えることができなかったことが原因であると考えられる．

5. 生成キャプション評価

5.1. 概要

提案手法によって生成されたキャプション (以下，生成キャプション) において，正解キャプションと同様の内容が出力されているかどうかを定量的に評価する．キャプション生成モデルには，STAIR Actions キャプションデータセット [3] を用いて学習を行ったものを用いる．

日本語文章には，異音同義表現が多く存在しており，同じ光景に対する説明文においても，異なる表現が用いられる可能性がある．生成キャプションの一般的な指標として用いられている BLEU などは，表面的な類似性に依存しており，文章の意味は考慮されていない．BERTScore は他の評価指標と比べて，人手評価値と

の相関が高く，評価指標として有効であることが示されているが，日本語においては評価されておらず，日本語文章評価における有効性は示されていない．そこで，我々は日本語文章評価における BERTScore の有効性について検証を行い [9]，その結果，日本語文章における有効性が示唆された．そのため，本実験においては，評価指標として，意味を考慮した評価が可能である BERTScore を用いて定量的評価を行う．

使用するデータセットの動画には，1 つの動画に対して正解キャプションが 5 つ付与されている．そのため，BERTScore を用いて，正解キャプション同士のスコアを総当たりに算出する．対象となるキャプションが 5 つであるため，10 通りのキャプションの組み合わせに対してスコアを算出し，正解キャプションにおけるスコアリングのぶれを確認する．スコアリングのぶれについては，10 通りのスコアの四分位数における，四分位範囲で算出する．その後，5 つの正解キャプションに対する生成キャプションのスコアをそれぞれ算出し，平均値を求め，その結果について比較・評価を行う．

正解キャプションにおけるスコアリングのぶれの範囲内に，正解キャプションに対する生成キャプションのスコアの平均値が収まっていれば，正解キャプションと同等の意味を持つキャプションが生成されていると考えられる．

5.2. 実験条件

4.2. 節と同様に，学習データとして，データセット内の 100 種類ある各動作から，動画を 100 本ずつ，合計 10,000 本使用した．データセットの各動画には 5 つのキャプションが付与されている．そのため，今回は学習テキストとして，合計 50,000 文のテキストを使用した．また，学習の際に用いる CNN モデルには，VGG16 モデル [8] を使用し，Beam Search における K は 5 とした．

評価用データとして，データセットから学習に使用していない動画を任意で選択した．

BERTScore を算出する際に用いる事前学習済みモデルについては，東北大学の乾・鈴木研究室が開発した日本語学習済みモデル[1]を使用する．

5.3. 実験結果と考察

正解キャプションにおける各スコアリング結果のぶれを図 3 に，使用したキャプションを表 3 に記す．また，図 3 における，箱ひげ図と重なっている点線は，正解キャプションに対する生成キャプションのスコアリング結果の平均値を表している．

[1]https://huggingface.co/cl-tohoku/bert-base-japanese

表 3: 使用キャプションとそのカテゴリ

動画カテゴリ：lying_on_floor, 床に横たわっている
正解キャプション
部屋でチェックのシャツを着た男性が床にあおむけで寝ている
白い壁の部屋でチェック柄のシャツを着た女性が仰向けになっている
カーペットの敷かれた白い壁の部屋で
青いチェックのシャツを着た男性が床に仰向けに寝ている
金属のラックがある部屋で黒い長ズボンをはいた
男性が仰向けになっている
白い壁の室内で青色の服を着ている男性が床に寝転がっている
生成キャプション
白い壁の部屋で黒い服を着た男の子が寝そべっている
動画カテゴリ：crying, 泣いている
正解キャプション
茶色の床の部屋で縞模様の服を着た赤ちゃんが
マットの上に寝転がって泣いている
室内でボーダーの服の赤ちゃんが泣いている
部屋で赤ちゃんが泣いている
屋内でボーダー柄の服を着た赤ちゃんが寝転んで泣いている
室内で赤ちゃんが泣いている
生成キャプション
屋内で白い服を着た赤ちゃんが寝転んで泣いている
動画カテゴリ：baby_crying, 赤ちゃんが泣いている
正解キャプション
部屋で男の子がテーブルに手を置きながら泣いている
白い壁に茶色のドアがある部屋で白い洋服をきた幼児が
椅子に座り机に両手を置いて泣いている
白い壁の部屋で白いパーカを着た子供が泣いている
白い壁紙のリビングで白い服を着た子供が泣いている
白い壁の部屋で白い長袖パーカを着た黒髪の子供が
テーブルに手をついて泣いている
生成キャプション
屋内で白い服を着た黒髪の男性が泣いている

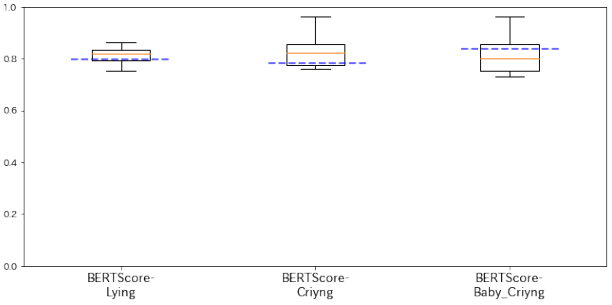

図 3: 各スコアリング結果のぶれ

表 3 より，正解キャプション同士は意味的に類似しているが，表現には，ばらつきが見られることが分かる．図 3 より，各スコアリング結果における，四分位範囲が小さいことから，正解キャプション同士は意味が類似している表現であることが分かる．

図 3 より，正解キャプションに対する生成キャプションのスコアリング結果の平均値が四分位範囲内に収まっていることが分かる．このように，BERTScore は，文章の意味を考慮した評価が可能であるため，正解キャプションに対する生成キャプションのスコアリング結果の平均値が四分位範囲内に収まっていたことから，生

成キャプションは正解キャプションと同様の意味を持っていると考えられる．

これらのことから，提案手法では，検知されうる異常動作と同様の意味を持っているキャプションが生成可能であると考えられる．そのため，提案モデルを用いて生成されたキャプションは，異常状態を捉えた動画において，有効である可能性が示唆された．

6. おわりに

本稿では，見守りカメラ等で撮影された動画から検知されうる異常動作をキャプションとして自動生成する手法と，その有効性について述べた．

Seq2Seq に Attention 機構を付与したモデルを用いて，キャプション生成モデルを構築し，キャプション生成時に BeamSearch を用いる手法を提案し，STAIR Actions キャプションデータセットを用いてモデルの学習を行い，異常と考えられる動作を捉えた動画を想定したキャプション生成実験を行った．その結果，自然な日本語文章で生成されていることから，提案モデルは，システムとユーザ間の情報共有に適した精度で生成可能である可能性が示唆された．また，正解キャプションと同様の行動が出力されていることから，検知されうる異常動作をキャプションとして生成可能である可能性が示唆された．

また，生成キャプションについて，BERTScore を用いて定量的評価を行った．生成キャプションは正解キャプションと同様の意味を持っていると考えられることから，提案手法では，検知されうる異常動作と同様の意味を持っているキャプションが生成可能であると考えられる．そのため，提案モデルを用いて生成されたキャプションは，異常状態を捉えた動画において，有効である可能性が示唆された．

今後は，服の色や背景情報の誤認識を防ぐために，学習データを増やして学習を行う予定である．また，今回の実験は，入力した動画に付与されている 5 つの正解キャプションのみを用いた評価であるため，同一カテゴリにおけるすべての動画の正解キャプションに対して，評価を行う予定である．

本稿では，異常動作に着目して，生成キャプションの有効性について検証したが，今後は，動画内容の理解促進や実況文生成など，他の分野に対しても生成キャプションが活用できるかどうかの検討・考察も行う予定である．

謝辞

本研究の一部は JSPS 科研費 21K12611 の助成を受けたものである．

参考文献

1）Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, Xiao Bai. 2020. Self-trained Deep Ordinal Regression for End-to-End Video Anomaly Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12173-12182.

2）対馬　陣，松原雅文, 見守りカメラを用いた異常状態に対するキャプション自動生成手法の提案, 情報処理学会 第 84 回全国大会, 4W-08，pp. 819-820, 2022.

3）重藤優太郎, 吉川友也, 藺佳慶, 竹内彰一, 人間の動作を日本語で説明するためのキャプションデータセット, 言語処理学会第 25 回年次大会, pp. 1173-1176, 2019.

4）中村 泰貴, 河野 誠也, 湯口 彰重, 川西 康友, 吉野 幸一郎, 説明文生成を用いた動作行動予測”, 研究報告自然言語処理（NL）,2022-NL-253(6),1-7 (2022-09-22) , 2188-8779

5）Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi“BERTScore: Evaluating Text Generation with BERT”International Conference on Learning Representations 2020 (ICLR 2020)

6）I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In NIPS, 2014.

7）Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin：”Attention Is All You Need”, Proc. of 31st Conference on Neural Information Processing Systems (NIPS 2017), pp.5998-6008, Long Beach, CA, USA.

8）Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations.

9）対馬 陣, 松原 雅文, キャプション自動生成における BERTScore の有効性について, 第 21 回情報科学技術フォーラム（FIT2022）