公開

（事後評価・研究者用）

研　究　報　告　書

「多様なデータへのキャプションを自動で生成する技術の創出」

研究期間：平成２９年１０月～平成３１年３月
研究者番号：５０１３４
研 究 者：牛久　祥孝

1．　研究のねらい

画像や動画のキャプション生成は、データ内の事物とそれぞれの関係性を理解して自然言語で表現する、メディア理解の究極の形態の一つである。本研究では、多様なマルチメディアデータ、特に画像や動画に対する自動キャプション生成技術の創出を目指す。特に、本質的に必要かつ未達の３つの要求機能、(i)個人の属性や好みへの対応、(ii)詳細な表現への対応、(iii)教師キャプションを持たないデータへの対応を実現し、多様なデータへのキャプションを自動で生成する技術を創出することが目的である。

キャプション生成は、入力データを説明する重要な語彙の推定、それらの語彙を文法モデルでつなぐことによるキャプション生成、という前後半のステップからなる。本研究では、上記 3 要求機能に共通した 2 つのボトルネック―語彙推定と文法モデルにおける、〈a〉訓練サンプルの偏りと〈b〉少数データしか存在しない状況―が存在すると考え、自然言語処理・画像認識・機械学習といった諸分野の先端的な知見を統合し、これらの解消を目指す。

2．　研究成果

（1）概要

本研究では、画像や動画を対象に、各要求機能を満たしたかどうかを報酬として算出する評価器を組み合わせるアプローチをとった。正解キャプションとの差を最小化しつつ報酬の和を最大化し、上記３要求機能を達成する。

(i)個人の属性や傾向への対応については、あるユーザのために生成されたキャプションかどうかの報酬を計算できる評価器を開発した。(ii)詳細な表現への対応については、キャプション内容の差分に注目した。詳細さに欠けたキャプションは、複数の入力データに対して同様の文になると考えられる。生成されたキャプションから逆にデータを検索したとき、入力したデータの順位がそのままキャプションの表現の詳細さとして評価できる。そこで、生成したキャプションからのデータ検索器を報酬として採用した。(iii)教師キャプションが無い状況への対応については、他のマルチメディアデータで教師キャプションが付与されているデータセットやテキストのみのデータからの知識転移を検討した。

（2）詳細

研究テーマ(ii)「データ中の小さな領域をも詳細に表現できるキャプション生成技術」

(ii)詳細な表現への対応のための生成したキャプションからのデータ検索器については、ユーザがより注目しやすいような差分についてのキャプションを生成するべきだとして、新たな問題定義を提案した。

まず、画像 1 枚 1 枚の差分ではなく、画像内の同種の物体についての差分に注目してキャプションが生成できるかという問題に取り組んだ。具体的には、画像中に複数の人や物などが写っているものを対象として、ある一つの人・物を対象としてキャプションを生成する。逆に、生成されたキャプションを手掛かりとして画像内の人・物を探した時に、入力として与えられた人・物が最もふさわしいものとして見つかるかどうかを評価器で計算させる。このような評価を最大化するキャプションが生成できれば、ある画像内で与えた人・物を表現するのに十分な詳細さのキャプションが生成できたことになる。この技術は直ちに拡張可能で、ある画像群内で与えた画像を表現するのに十分な詳細さのキャプションを生成することで、要求機能(ii)を達成できる。

既存の研究でも生成したキャプションから入力の人・物を検索できるかという評価器を用いるものが存在したが、それらはユーザにとって「探しやすい」―つまり、目立つような差分に注目しつつ、冗長な記述で読解の時間を長引かせないよう簡潔に記述できる―技術ではなかった。本研究では視覚的顕著性とアテンションを活用した差分キャプション生成手法を提案し、このように人に追って探しやすいような詳細さをもった差分キャプション生成手法を開発した。この研究に基づいた知的財産については国内特許を出願中である。また、レポジトリでのレポートおよびデータセットの公開も進めている。

アウトリーチ活動

本研究分野のアウトリーチ活動についても精力的に進めた旨を報告したい。2017 年 9 月～2019 年 3 月までの間に、キャプション生成の成果を含む講演を、国際会議ワークショップ基調講演を含めて 12 回つとめた。また画像・動画キャプション生成について書籍 1 件、解説記事 2 件を執筆したほか、現在もう 2 件の執筆を予定している。執筆済み解説記事 1 件は、映像情報メディア学会誌 2018 年 9 月号で Vision & Language 分野を俯瞰する特集記事としてゲストエディタを務めるとともに、1 章として画像・動画キャプション生成について執筆したものである。提案者が執筆した章は会員の投票による支持を受け、結果として提案者は当該号のベストオーサーに選出された。

3. 今後の展開

本研究者は、記事や動画といったコンテンツが計算機によって自動的に生成されるという未来ビジョンを持っている。現在はこれらのコンテンツ生成には人的コストや専門知識が必要だが、今後は膨大で多様なテキスト・マルチメディアデータを組み合わせたり、生成したりといった処理によって自動で提供されると考えている。ユーザが望む内容のコンテンツを提供するには、一般的なユーザにとって易しい方法で、コンテンツを編集・再検索できる技術が重要である。従って、ユーザがコンテンツを入手するための入力として自然言語が考えられ、そのためには言語によるコンテンツ理解や検索が必須だと考えている。本研究は、言語を通じて AI とインタラクションしながらコンテンツを入手し楽しむ未来へと繋がる、極めて重要な取り組みである。

ここまで、従来にはない 3 要求機能を満たす新たなキャプション生成技術を掲げ、基礎研究を進めてきた。本研究者は引き続き JST ACT-I の加速フェーズで同研究を継続する予定である。特に、本研究で掲げる問題設定に即したデータセットおよび評価基盤を構築して多くの研究者が挑戦するような研究領域を形成すること、およびこの 1 年半で得られた研究成果をその評価プロトコルに即して応用・発展させていくこと、この2点を進める予定である。

研究領域の形成としては、学習用キャプションが少なく、かつ個人適合と詳細な表現が求められるデータセットが必要である。キャプションが限られるのは、そのマルチメディアデータ自体の作成からコストがかかる場合である。また、データの内容が複雑になるほど個人適合と詳細表現の必要性が増す。そこで本研究では、画像としてレシピや組み立て書のような図とテキストが混在したマニュアルを、動画として料理などの作業動画を検討している。これらが言語化されれば、一般的なユーザが新たなスキルや手順を身につける際の検索が容易になり、ユーザが理解する際の大きな手助けとなる。

4. 自己評価

・研究目的の達成状況

本研究の3つの要求機能のうち、(i)についてはまず、個人の好みに適合したキャプションかどうかを判定する評価器自体が新規な研究となるため、これを進めた。(ii)については、詳細なキャプションかどうかを判定する評価器の実現のみならず、それを用いて実際にキャプション生成を詳細になるよう改善させる技術を確立し、国内特許の申請とレポジトリでの公開を進めた。(iii)の少数データへの対応、およびこれら(i)〜(iii)の評価器を統合した評価実験については、今後の加速フェーズで引き続き精力的に取り組む。

・研究の進め方（研究実施体制及び研究費執行状況）

【実施体制】本研究のうち、(ii)データ中の小さな領域をも詳細に表現するという要求機能については、株式会社デンソーとの共同研究として進めた。また、研究遂行に当たっては東京大学の学生 1 名も実験データ収集と解析に従事した。またその他の要求機能について、第 2 年次にはほかの学生 1 名も実験データ収集と解析に従事した。

【研究費執行状況】第 1 年次は概ね当初の計画通りに執行された。第 2 年次は研究者が東京大学からオムロンサイニックエックス株式会社へ異動し、所属部署自体の研究予算が増加した。そちらを優先して執行した結果、本研究自体の進行は円滑に進みながらも、研究費については 75%程度の執行となった。

・研究成果の科学技術及び学術・産業・社会・文化への波及効果

上記のとおり、本研究は既に産業界との共同研究を含めている。本研究は画像や動画などの視覚データを自然言語へと変換するものであり、人間の視覚機能を補助する技術としてすぐに波

及が期待できるものである。

更に本研究は、視覚データと自然言語を融合して理解するための基本的な要素技術を対象としている。最近の周辺学術分野では、人間と計算機とで自然言語のみならず関連する視覚データなどの別モーダルデータを含めて対話行為を実現する研究に注目が集まり始めている。本研究での成果は、このような対話行為に(i)個人への適合、(ii)より細部へ注目した対話行為、そして(iii)少数の対話データで学習可能なシステムを提供し、学術のみでの取り組みを産業や社会へ波及させるために必要不可欠な知見を提供すると考えられる。

本研究者は、このような着想の元、研究期間中も精力的にアウトリーチ活動を進め、このような波及効果に共感する研究者および企業とのつながりを広げている。

・研究課題の独創性・挑戦性

本研究者は画像キャプション生成については第一人者であり、世界に先駆けて取り組みを始めている。現在の類似研究では、入力画像領域についての最大公約数的なキャプション生成の精緻化に注目した研究がほとんどであり、提案者と同一の要求機能を掲げる研究はまだ存在しない。

また提案者は、3つの要求機能を実現するための各要素技術についても既に知見を有している。画像認識から自然言語処理、機械学習に至るまでの広範な知識と実績を有する研究者は世界的にも少ない状態であり、本提案を遂行する上での大きなアドバンテージである。

5. 主な研究成果リスト

(1) 論文(原著論文)発表

特になし

(2) 特許出願

研究期間累積件数:1 件

発　明　者: 板持　貴之、牛久　祥孝、田中　幹大、佐藤　育郎
発明の名称: 説明文章生成装置、対象情報表現システム、及び説明文章生成方法
出　願　人: 株式会社デンソー、国立大学法人東京大学
出 願 番 号: 特願 2018-136333

(3) その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

- [基調講演] Y. Ushiku, Frontiers of Vision and Language: Bridging Images and Texts by Deep Learning. Workshop of Machine Learning under International Conference on Document Analysis and Recognition, International Conference on Document Analysis and Recognition, Kyoto, 2017/11/11.
- [招待講演] 牛久祥孝, Deep Learning による視覚・言語融合の最前線. 画像符号化シンポジウム(PCSJ) / 映像メディア処理シンポジウム(IMPS), 伊豆, 2017/11/22.
- [招待講演] 牛久祥孝, 視覚と言葉をつなげる技術. 情報処理学会 IPSJ-ONE, 東京, 2018/3/15.

- 米谷竜 (著, 編集), 斎藤英雄 (著, 編集), 池畑諭 (著), 牛久祥孝 (著), 内山英昭 (著), 内海ゆづ子 (著), 小野峻佑 (著), 片岡裕雄 (著), 金崎朝子 (著), 川西康友 (著), 齋藤真樹 (著), 櫻田健 (著), 高橋康輔 (著), 松井勇佑 (著), 画像キャプションの自動生成/コンピュータビジョン―広がる要素技術と応用. 未来へつなぐデジタルシリーズ, 37 巻, 共立出版, 2018/6/28.
- 牛久祥孝 (著, 編集), 山口正隆 (著), 福井啓 (著), 中山英樹 (著), 齋藤真樹 (著), 吉川友也 (著), 重藤優太郎 (著), 竹内彰一 (著), 視覚･言語融合の最前線　. 映像情報メディア学会誌, Vol.72, No.5, 2018.