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：2019 年 4 月～2021 年 3 月
研究者番号：50221
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。

この研究では、画像や動画と言った多様な形態のデータそれぞれを説明するキャプションを自動で生成する技術の確立をめざす。特に本研究では、キャプション生成に本質的に必要であり未解決である、(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と、加速フェーズではさらに、(iv)非言語的に埋め込まれた文法構造の獲得を実現し、多様なデータへのキャプション生成を可能にする大きなブレークスルーをもたらす。

まず、画像を説明するキャプションは、(i)何を説明するか、どう表現するか、ということがユーザの属性や好みに応じて変化するべきである（図 1）。

図1　個人の属性や好みに応じたキャプションの生成。

また、これらのキャプションは(ii)データ中の小さな領域をも詳細に表現できる必要がある。現在の画像キャプション生成は、「猫がこちらを見ている」など、画像中に大きく映った事物のみを、抽象的な語彙で表現する傾向がある。本研究では、注目されそうな部分に応じて図1のように「凶悪な目つき」という小さな領域と「猫」という大きな領域の対象を同時に扱える技術を確立する。

さらに、(iii)教師キャプションを持たないデータへの対応が必要である。現在のキャプション生成技術は、教師キャプションが付与された画像を数万枚の規模で作成したデータセットを用いる。しかし、実世界のマルチメディアデータには動画や音声と言った多様なモダリティが存在し、さらに画像でも撮影位置や時刻といったメタデータが付随する場合も多い。しかしながら、大規模データ収集の高いコストにより、上記のような様々なモダリティのマルチメディアデータについて、他の要求機能を満たすようなキャプション生成を実現するために都度大規模なデータセットを収集するのは困難である。

最後に加速フェーズでは、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、料理などの作業動画といった手順が埋め込まれたデータを理解する技術も確立する。そのためには、工程を抽出してその前後関係を学習する必要がある。この様な(iv)非言語的に埋め込まれた文法構造の獲得は、より一般的なコンテンツのストーリー理解にも必須の機能である。

2. 研究成果

(1)概要

本研究が画像キャプション生成において満たしたい要求機能は、(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、(iv)非言語的に埋め込まれた文法構造の獲得の 4 点であった。その成果として、各要求機能への取組とその統合的な評価を並列で進めてきた。

具体的には以下の通りである。(i)個人の属性や好みへの対応の解として**研究テーマ A「転移学習の研究」**に取り組み、既存研究より現実的かつ難しい問題設定におけるドメイン適応についての技術を確立し、特許を出願中である。(ii)詳細な表現への対応の解としては**研究テーマ B「ユーザーフレンドリーな参照表現生成」**を遂行し、詳細な差異に注目できるようなキャプション生成基盤技術としてトップ国際会議での発表に至っている。(iii)教師キャプションを持たないデータへの対応については**研究テーマ C「教師なしキャプション生成の研究」**として疑似教師データを活用しながらの教師なし学習によるキャプション生成を実現し、トップ国際会議での発表を予定している。(iv)非言語的に埋め込まれた文法構造の獲得については、**研究テーマ D「作業手順の構造的理解」**に取り組んだ。調理の様子をおさめた動画像に埋め込まれた作業の手順をグラフ構造として理解し、そのレシピを生成する技術を開発し、国際論文誌での発表に至っている。

またこれらを統合して評価するために**研究テーマ E「料理動画・レシピデータセットの構築と統合評価」**として新規データセットの収集と、上記技術を統合したキャプション生成技術の評価に取り組んだ。

(2)詳細

研究テーマ A「転移学習の研究」は、個人に依らない基本的なキャプション生成モデルを個人ごとの属性に応じた画像や自然言語に転移させることで、教師データが少ない状態(iii)での個人適応(i)を実現するための研究である。そもそも従来の転移学習は、その評価実験に明示的に/暗黙的に仮定されている条件が複数あり、実用性に疑問があった。本研究では、キャプション生成への適用を見据えて現実に即した評価条件設定に改めた深層転移学習手法を開発した。本成果は特許出願中かつ、コンピュータビジョン分野でのトップ国際会議である CVPR 2021 に採択され、発表予定である。

研究テーマ B「ユーザーフレンドリーな参照表現生成」は、「データ中の小さな領域をも詳細に表現できる」という(ii)の要求機能だけを切り出したものに対応する。参照表現生成では、画像内にある複数の物体の中から、どれに言及しているのかが一意に定まるようなキャプションを生成する。例えば図 1 に対して「物が写った写真の一部」というキャプションを生成しても、間違ってはいないもののどの領域についての記述かが全く分からない。そこで「猫の首輪についた魚」と言うと、画像中央下部にある金色の物体に定まる。単に一意に定まることだけを目指すと「魚」だけでも良いことになってしまうが、それだと画像のあらゆる領

域を探さないといけないような参照表現生成になってしまうため、人に対して生成するキャプションとしては利便性が少ない。そこで本テーマでは、「猫の首輪」のように画像内で視覚的に顕著性の高い領域を紐づけながら、冗長性の少ないキャプションを生成する技術をアテンション機構とランキング学習を組み合わせた深層学習モデルで実現した。この成果は代表者が東京大学に在籍している際にデンソー、デンソーアイティーラボラトリとの共同研究として進めたものであり、コンピュータビジョン最高峰の国際会議である ICCV 2019 でデータセットの公開と併せて発表済みである。

研究テーマ C「教師なしキャプション生成の研究」は、転移学習が難しい場合でも(iii)を解決するための手段として進めている。教師データとなる画像/動画とキャプションの対が無い状態でも、そうした画像/動画といったデータやキャプションのような文のデータといった、ペアになっていないデータであれば容易に収集可能である。そこで本テーマでは、そうした画像と文が独立に存在するデータについて、一部の画像から尤もそうな画像・キャプションの疑似ペアを少しずつ増やして教師あり学習する手法を開発した。この成果は NAIST および理研 AIP と共同で取り組み、自然言語処理分野においてメジャーである国際会議 EACL 2021 で発表済みである。

研究テーマ D「作業手順の構造的理解」は、動画のような動的なデータの中での非言語的に埋め込まれた文法構造(iv)を獲得するためのテーマである。長い動画のようなデータの中では、時々刻々を言語で記述するキャプションの上位に、そのストーリーのような遷移が存在する。例えば調理動画であれば、各調理ステップに対応する記述と、それらの系列を同時に理解する必要がある。そこで本テーマでは、調理動画の理解を各調理ステップのキー画像からなる画像列の理解と捉え、途中経過の食材や初出の材料がどのステップで組み合わされているのかを示す木構造を定義し、そうした木構造を用いたレシピ生成技術を開発した。画像列と材料のリストから木構造を生成し、その木構造からレシピを生成する。また生成されたレシピからも木構造を推定し、前述の木構造と一致するかどうかを評価させてレシピと動画列の一貫性を担保している。この成果は京都大学との共同研究として進め、速報誌である IEEE Access に掲載されている。

研究テーマ E「料理動画・レシピデータセットの構築と統合評価」は、テーマ D のような調理動画を対象として、その多様性への対応を正しく評価できるようなベンチマークを作成し、実際にそれぞれの研究テーマから得られた技術を統合したキャプション生成技術を適用して評価を行うことを目的としている。料理レシピのデータベースとそれらに基づいたサービスで定評のあるクックパッドと、京都大学や東京大学の研究者との合同でデータセットの構築にあたった。

本研究テーマでは 2 段階でデータセット収集とその公開を進めている。まず、クックパッドが既に所持している画像列データとレシピのデータについて、その作業手順構造をグラフとして付与したデータセットを構築、こうした自然言語を含めたリソースについて公開して共有する国際会議としてはもっとも著名な

LREC 2020 で発表済みである。

次の段階として、既存のレシピに基づいて調理する様子を収録した動画像を新規に収集し、その上で研究テーマ A~D の成果を統合したキャプション生成技術を実行して評価する予定である。残念ながら新型コロナウィルスの影響によってデータ収集プロセスの策定の再検討を余儀なくされ、収集スピードも低下したことで 1 年弱の遅れを来しているが、データ収集自体は着手済みであり、2021 年度の前半でデータセット構築を完了して公開・キャプション生成結果の報告へとつなげる予定である。

3. 今後の展開

この 2 年間によって、画像のような静的なデータから動画のような動的なデータまで、個人に応じて詳細なキャプションを生成する一連の技術が確立された。これらは画像や動画と言った非言語情報を自然言語で検索・閲覧したり、更に近年盛んに研究されているデータ生成ネットワークに応用したりと言った形で、自然言語によってマルチモーダルなデータを活用する基盤技術となる。例えば、本研究で最終的に実現した手順映像の言語理解は、作業記録の自動生成や非熟練者への言語サポートに活用できる。

最近では自己注意機構を活用した Transformer と呼ばれるネットワーク構造が自然言語処理のみならずマルチモーダルなデータの理解についても有効性を認められつつある。本研究でもこのネットワーク構造の活用について試行中であるが、Transformer を用いる場合はより大量な学習データが必要となる。そのため、より少ないデータでもそのバイアスを活用してキャプション生成を可能にする畳み込みニューラルネットワークと再帰ニューラルネットワークの組合せや、そうしたバイアスをデータセット間で吸収する転移学習など、本研究で取り組んできた諸技術については引き続ききわめて有効な技術であり続けると考えている。

さらに、本研究は本質的に、画像/動画といった画素数列データと自然言語列というグラフのような構造を持つデータの変換に取り組むものであった。今後もこうしたデータ構造を超えるモダリティ変換の研究を別の系に適用し、効率的なデータ活用基盤の研究を続行する予定である。

4. 自己評価

研究目的の達成状況および研究費執行状況については、要求仕様(i)~(iv)それぞれを満たす技術の開発に成功し、それぞれ特許出願や国際論文誌/会議での発表に至っている。ただし新たなデータセットを更に構築してその上で統合的に評価する部分については、新型コロナウィルスの影響で遅滞している。

研究実施体制については、各研究テーマの遂行にあたって研究代表者と相乗効果の見込める研究者と産学問わず連携して進めたことにより、複数の研究成果を並行して創出できた。

研究成果の波及効果については、企業との共同研究によって各企業が持つデータに対する社会実装が期待される。また、各研究テーマは世界的にも広く注目されているような論文誌/国際会議で発表しており、積極的に国内外での招待講演を通じてその更なる周知に努めている。更に、研究テーマ A や E で述べたようなデータセットの公開を通じて広く世界の研究者に本テーマでの更なる研究を促しており、今後より本研究領域が盛んに研究されるようになることを期待している。

研究課題の独創性・挑戦性としては、画像/動画キャプション生成という研究領域を真に社会実装するために必要となる機能を網羅的に押さえた研究課題であると自負している。現在もこうした画像/動画と自然言語を含めたメディア理解は Vision and Language というテーマでコンピュータビジョン、自然言語処理、機械学習といった諸分野で取り組まれているが、今も同一のデータセットに対する精度競争を注意機構や Transformer などの新たな関連技術によって更新する取り組みが多い。本研究では一貫して、基本的に既存の研究課題をより現実的な設定に改善しながらその解となる技術を提案し続けており、新規かつ有用な課題に挑戦し続けている。

5. 主な研究成果リスト

(1)論文(原著論文)発表

1. M. Tanaka, T. Itamochi, K. Narioka, I. Sato, Y. Ushiku, T. Harada. Generating easy-to-understand referring expressions for target identifications. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 5794-5803.
2. T. Nishimura, S. Tomori, H. Hashimoto, A. Hashimoto, Y. Yamakata, J. Harashima, Y. Ushiku, S. Mori. Visual grounding annotation of recipe flow graph. Proceedings of the Language Resources and Evaluation Conference. 2020, pp. 4275-4284.
3. T. Nishimura, A. Hashimoto, Y. Ushiku, H. Kameko, Y. Yamakata, S. Mori. Structure-Aware Procedural Text Generation from an Image Sequence. IEEE Access. 2020, vol. 9, pp. 2125-2141.
4. U. Honda, Y. Ushiku, A. Hashimoto, T. Watanabe, Y. Matsumoto. Removing Word-Level Spurious Alignment between Images and Pseudo-Captions in Unsupervised Image Captioning. Proceedings of the European Chapter of the Association for Computational Linguistics. 2021, pp. 3692-3702.
5. Q. Yu, A. Hashimoto, Y. Ushiku. Divergence Optimization for Noisy Universal Domain Adaptation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, accepted.

(2)特許出願

研究期間累積件数:2 件

1.
発　明　者: 郁　青、牛久　祥孝、橋本　敦史
発明の名称: 推定システム、推定装置および推定方法
出　願　人: オムロン
出 願 番 号 : 2019-206384

2.
発　明　者: 橋本　敦史、牛久　祥孝、森　信介、西村　太一
発明の名称: 構造化データ表現の獲得方法
出　願　人: オムロン
出 願 番 号: 2020-038785

（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）

1. 【基調講演】牛久祥孝. 深層学習によるビジョン＆ランゲ―ジの世界. ビジョン技術の実利用ワークショップ(ViEW), 2019/12/5.
2. 【招待講演】Yoshitaka Ushiku. Deep Learning for Natural Language Processing and Computer Vision. Tutorial on Asian Conference on Machine Learning, Nagoya, Japan, 2019/11/17.
3. 【招待講演】牛久祥孝. 画像・映像理解と自然言語への架け橋. 情報処理学会連続セミナー第 3 回, 2019/9/26.
4. 【招待講演】牛久祥孝. 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. 情報科学技術フォーラム(FIT), 2019/9/3.
5. 【招待講演】牛久祥孝. ビジョン＆ランゲージ～「意図」をどのようにモデリングするか？ 画像センシングシンポジウム(SSII), 2019/6/12.