公開

（事後評価・研究者用）

研　究　報　告　書

「映像とテキストを組み合わせたストーリー理解の実現」

研究期間：平成 29 年 10 月～平成 31 年 3 月
研究者番号：５０１３５
研 究 者：大谷　まゆ

1．　研究のねらい

映像の内容を理解することはコンピュータビジョン分野における大きな目標であり，現在に至るまで，映像中のオブジェクトやアクション認識など多くの取り組みがなされてきた．近年，映像理解タスクとして Video captioning や Video question answering が提案されている．Video captioning は映像の内容を説明する文の生成を目的としており，Video question answering は映像に関する質問に対する回答文を生成するタスクである．これは本来，静止画像理解のために考案されたものであり，この指標で評価可能な範囲は映像理解の一部に過ぎない．このようなベンチマークのために開発された手法の多くは，映像のフレームの並びを無視したものが多く，一方でそのような手法であっても既存のタスクではある程度の性能を達成することが可能となっている．つまり、従来の映像理解タスクは時間方向の変化を扱うことを要求していないといえる．

そこで本研究は，さらなる映像理解手法の開発を推し進めるため，映像の時間に関する構造に着目した新たな映像理解タスクを設計し、現在の映像理解技術の限界を調査する．具体的には，映像中の複数のイベントとその関係性に着目した映像キャプションデータセットを構築する．そしてこのデータセットを使った複数のタスクを現在主流となっているベースライン手法で評価することで、既存の映像理解のための手法における課題を明らかにし今後の映像理解研究の方向性を示す．

2．　研究成果

（1）概要

本研究では、これまでの映像理解タスクでは考慮されてこなかった「イベント間の関係性の理解」に着目したキャプションデータセットを開発した．これまでのデータセットの多くは映像を単一のイベントやアクションを撮影した単純なものに限定している．提案するデータセットは映画から時間的に近いクリップを2つ抽出して組み合わせることでストーリー上関係のある複数のイベントを含む映像を構成した．またクラウドソーシングサービスを活用し、映像の内容を説明するキャプションを収集した．

最後に構築したデータセットを使った簡単なタスクを設計し、現在主流となっている映像理解手法のベースラインで評価した．

（2）詳細

研究テーマ A「映像理解データセットの構築」

本研究ではまず2つのデータセットを検討した．

1．映像の順序データ

自然言語で記述されたストーリーを手がかりに，映像をそのストーリーに沿った順序に並び替えるタスクを考案し、ストーリーテキスト、映像クリップの組、およびその並び順をクラウドソーシングを使って収集した．このタスクはアノテーターの間でも回答のバリエーションが大きく、映像の順序には高い曖昧性があることが明らかとなった．また既存の映画のキャプションデータセットを用いて予備実験を実施した．実験結果から、簡単な並び替え問題に限定すれば、既存の手法である程度解くことができることを確認した．

2．イベントの関係性に着目したキャプションデータ

映像の並び替えのためのデータは曖昧性が高く、映像理解タスクとしての定式化が困難であると判断し、新たなデータとして、イベントの関係性に着目したキャプションデータ収集に取り組んだ．具体的には映画から抽出した2つのクリップを組み合わせることで複数のイベントを含んだ短い映像を作成し、イベントとその関係性を説明したキャプションを収集した．キャプションの収集にはクラウドソーシングサービスを利用した．この時、キャプションのフォーマットを「__ because/but __」と指定した．これにより、アノテーターが複数のイベントの説明を記述し、イベント間の関係性を「because」または「but」から選択することでラベル付するようにした．このデータ収集により図1のような映像とキャプションのペアデータが得られた．

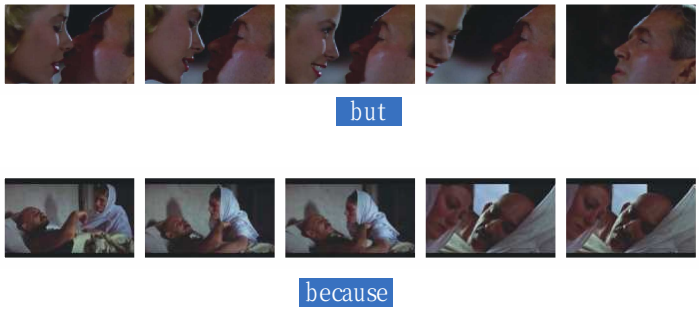

図 1 収集した映像キャプションの例

研究テーマ B「ベースライン手法による映像理解タスク評価」

研究テーマ A により得られたデータセットを用いた簡単なタスクを設計し、ベースライン手法で評価することにより提案するデータセットと既存の映像キャプションデータセットの違いを実験的に調査した．まず映像キャプションデータを用いたシンプルなタスクとして、映像に対して正しいキャプションを選択するタスクを設計し、評価した．例を図2に示す．例ではクエリ映像に対して上段のキャプションを選択することで正解となる．

図 2 キャプション選択問題の例

ベースライン手法として各キャプションがクエリ映像と対応づいたものである確率を出力するニューラルネットワークベースのモデルを構築した．この実験では映像のフレーム順を考慮するモデルとフレーム順を無視したモデルを作成し、その性能を比較した．また既存の映像キャプションデータセットを用いて同様の実験をした．

この実験の結果、従来の映像キャプションデータセットを用いた実験ではフレーム順を無視した手法がフレーム順を考慮した手法より高いスコアとなった．一方で提案するデータセットではフレーム順を考慮した手法がもう一方よりも高い性能を示した．これは本データセットにおいては時間方向の変化が重要な要素であることを示唆している．また従来のデータセットを使った実験に比べ、提案データセットを使った実験では性能が限られていることを確認した．これは既存の映像理解のための技術では時間方向の変化を扱うことが困難であるためであると推測される．

次に、映像の前半とキャプションから、映像の後半を予測するタスクで同様のベースライン手法を評価した．この実験でも前述のキャプション選択タスクと同様の傾向が観察され、提案するデータセットにおいて映像の時間方向の変化を扱うことの重要性が確認された．

3. 今後の展開

今回の成果は従来の研究では考慮されてこなかった映像中のイベント間の関係性に着目した映像理解への研究を進めるために必要不可欠であるデータの収集と、基礎的な分析である．今後はこの成果に基づき、高度な映像理解に向けて時間方向の変化を考慮した手法の発展が期待される．

4. 自己評価

当初に想定した映像の並び替えタスクは調査により、合理性のあるタスクとして定式化することが困難であることが確認され、そこで得られた知見に基づきイベントの関係性に着目した映像キャプションの収集に取り組んだ．研究のアプローチに関しては方向性の転換はあったが、当初の目標としていた映像の時間方向の変化を考慮したタスクの設計とベースライン手法を用いた基礎的な分析は達成できた．

クラウドサービスに詳しいエンジニアが研究に加わり、クラウドサービスを利用したデータ収集環境の構築を担当した．これにより、当初想定したものより大幅にデータのホスティングおよびデータ収集に関わる費用が削減された．これまでの成果は今後国内の学会などで発表し、フィードバックを得ていく．また今回得られた成果を元に研究を進め、国際会議などへの投稿を目指す．

5. 主な研究成果リスト

(1) 論文 (原著論文) 発表

該当なし

(2) 特許出願

該当なし

(2) 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)

該当なし