複数映像の抽象化を要するキャプション生成

高橋 力斗　Chu Chenhui　黒橋 禎夫
京都大学情報学研究科
{r-takahashi, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

概要

本研究では，新たな映像キャプション生成タスクとして抽象的映像キャプション生成を提案する．抽象的キャプション生成は，複数の映像を入力として，それらの映像全てに共通する内容を説明する文を生成するタスクである．このタスクでは，モデルは複数の映像に含まれる情報を抽象化し表現することが求められる．提案タスクのための新たなデータセットを構築し，モデルによる実験を行った．実験の結果，抽象的キャプション生成タスクは挑戦的なタスクであることが分かった．

1　はじめに

映像キャプション生成は，単一の映像を入力として，映像の説明文を生成するタスクである [1][2]．このタスクでは，モデルは与えられた映像に対して詳細かつ正確な説明を出力することが求められる．

本研究では新たな映像キャプション生成タスクとして，**抽象的映像キャプション生成**を提案する．抽象的映像キャプション生成は，複数の映像を入力として，それらの映像全てに共通する内容を説明する文を生成するタスクである．

抽象的映像キャプション生成タスクでは，モデルは複数の映像に含まれる情報を抽象化し表現することが求められる．したがって，抽象的映像キャプション生成タスクはモデルの情報抽象化能力を測る指標となることが期待される．また，抽象的映像キャプション生成タスクによってモデルが学習した抽象的な表現は，クラスタリングされた動画群に対する自動ラベル付与や，テキスト・動画をクエリとした動画検索などに応用できる．

本研究では抽象的映像キャプション生成データセットの構築及びそのデータセットを用いた実験を行った．データセットの構築では，既存の映像キャプションデータセットをもとにして，動画検索・テキスト含意認識・クラウドソーシングを用いて構築を行った．図 1 に構築したデータセットの一例を示す．実験では，構築したデータセットを用いてモデルを訓練し，抽象的映像キャプション生成を行った．モデルに複数の動画を入力するための工夫として，動画特徴量を類似度で重み付けして結合し，モデルに入力した．実験の結果，抽象的映像キャプション生成は挑戦的なタスクであることが分かった．

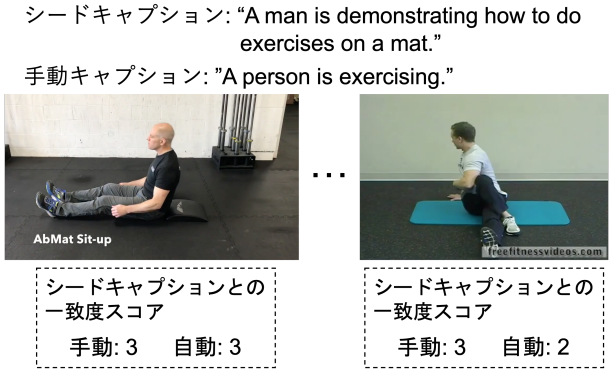

図 1　構築したデータセットの一例．

2　関連研究

ここ数年で，映像キャプション生成に利用できるデータセットが数多く提案されている．分割されていない長い動画に対して複数のキャプションを付与しているデータセットとして，ActivityNet Captions [3]，YouCook2 [4] などがある．10 秒程度に分割された短い動画とキャプションを収集した大規模データセットとして，HowTo100M [5]，VATEX [6]，MSR-VTT [7], MSVD [8] が有名である．本研究では，抽象的映像キャプション生成のデータセットを構築する際に VATEX データセットを利用した．

動画と言語を繋ぐ埋め込みモデルとして CLIP4Clip [9] がある．CLIP4Clip は動画埋め込みとテキスト埋め込みを同じ表現空間で学習するモデルである．動画の各フレームの画像と動画キャプションを同時に入力し，出力する動画特徴量とテキスト

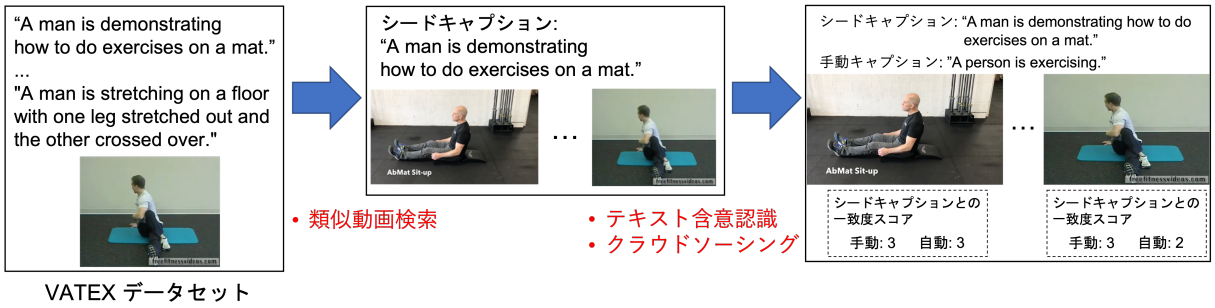

図 2 データセットの構築手順.

特徴量が類似するようにモデルを訓練する．本研究では CLIP4Clip をデータセット構築の際に用いる動画埋め込みモデルとして利用した．

3 データセットの構築

抽象的映像キャプション生成は既存の映像キャプションデータセットで訓練・評価することができないため，新たにデータセットを構築した．図 1 に構築したデータセットの一例を示す．構築したデータセットは次のもので構成される．

- 動画グループ: 2-6 件の動画が含まれる動画群
- 手動キャプション: 人手によって付与したキャプション
- シードキャプション: 自動的に付与したキャプション
- シードキャプションとの一致度スコア
  - 自動スコア
  - 手動スコア

キャプションは動画グループ内の全ての動画に共通する内容を説明する英文である．シードキャプションとの一致度スコアは動画グループ内の各動画に対して付与されるスコアであり，シードキャプションと動画の内容一致度を表すスコアである．シードキャプションとの一致度スコアには，自動スコアと手動スコアがある．

手動キャプションはデータセットの正解ラベルとしての利用を想定している．シードキャプション及び各動画に対応するスコアは，モデルの訓練及びテスト時に利用することを想定している．例えば，正解ラベルである手動キャプションに加えて，シードキャプションの内容をスコアで重み付けして学習させることで，モデルに抽象的キャプション生成をより効率的に学習させることが期待できる．また，手動スコアが高いシードキャプションについては，テスト時の正解ラベルとしても利用できる．

データセットの構築手順を図 2 に示す．既存の映像キャプションデータセットである VATEX データセットを使用し，いくつかの処理を行なってデータセットを構築した．最初に，VATEX データセット内の動画に対して類似動画検索を行うことで，複数動画とシードキャプションからなるデータセットを構築した．その後，テキスト含意認識とクラウドソーシングを用いてデータセットにシードキャプションとの一致度スコア及び手動キャプションを付与した．また，VATEX データセットにおける訓練データ・検証データ・テストデータに対してそれぞれ処理を行ったものを，構築したデータセットの訓練データ・検証データ・テストデータとした．以下でデータセット構築のために行った処理の詳細を述べる．

3.1 類似動画検索によるデータセット構築

既存の映像キャプションデータセットである VATEX データセットに対して類似動画検索を行い，動画グループに対してシードキャプションを付与したデータセットを構築した．類似動画検索を用いてデータセットを作る手順は次のとおりである．

1. 動画埋め込みモデルを用いて，VATEX データセット内の動画の特徴量を抽出する．
2. k 近傍法を用いて，検索元となる動画に類似する動画を VATEX データセット内から上位 6 件集めて一つの動画グループとする．
3. 検索元の動画に付与されていた英語キャプション 10 件の中で最も語数が少ないものをシードキャプションとし，動画グループに付与する．

動画埋め込みモデルには CLIP4Clip [9] を用いた．CLIP4Clip による埋め込みでは，動画全体を一つの特徴ベクトルに変換する．したがって，動画の特徴ベクトルの集合に対して k 近傍法を適用することで類似動画を検索することが可能である．k 近傍法を

実行するツールには Faiss[1)] [10] を用いた．

3.2 シードキャプションとの一致度スコアの自動付与

テキスト含意認識を用いた処理によって，動画グループ内の各動画に対してシードキャプションとの一致度スコアを自動的に付与した．シードキャプションは，元々は動画グループ内の動画とは異なる動画に対して付与されていた説明文である．そのため，シードキャプションは動画グループ内の動画の内容と食い違っている可能性がある．動画に対してスコア付けを行うことで，シードキャプションに合わない動画を動画グループから除外することができる．また，モデルの訓練時にシードキャプションの内容を用いる場合，自動スコアに応じて重み付けをした学習を行える．

テキスト含意認識は，二つのテキスト間の含意関係を判定するタスクである．あるテキストが正しいと仮定した場合に，もう一つのテキストが正しいと言えるかを判定する．動画グループ内のある動画に VATEX データセット内で元々付与されていたキャプションが正しいと仮定した場合にシードキャプションが正しいと言えるならば，シードキャプションはその動画の内容を正しく説明していると考えられる．

テキスト含意認識を用いて動画に対して自動スコアを付与する手順は次の通りである．

1. シードキャプション c' と，動画グループ内の動画に VATEX データセットで元々付与されていたキャプション $c_1, c_2, ..., c_{10}$ を準備する．
2. テキスト含意認識モデルを用いてキャプション c' とキャプション c_1 の含意関係を判定する．キャプション $c_2, ..., c_{10}$ それぞれについても同様の操作を行う．
3. テキスト含意認識で含意関係にあると判定されたキャプションの数を動画のスコアとする．

テキスト含意認識を行うモデルには，テキスト含意認識で高い性能が報告されている SemBERT [11] を用いた．全ての動画に対して自動スコアを付与した後に，自動スコアが 0 である動画は動画グループから排除した．また，動画の排除によって動画グループ内の動画数が 1 件以下となったデータはデータセットから排除した．

3.3 シードキャプションとの一致度スコアの手動付与

クラウドソーシングを用いて，動画グループ内の各動画に対してシードキャプションとの一致度スコアを手動で付与した．手動でのスコア付けの目的は，モデルの訓練・テスト時に利用できる情報の付与である．モデルの訓練時にシードキャプションの内容を用いる場合，人手によって付与したスコアに応じて重み付けした学習を行える．また，手動スコアが高い動画が多く含まれるシードキャプションについては，テストデータの正解ラベルとしての利用も期待できる．

シードキャプションと動画の組をクラウドワーカーに見せ，シードキャプションが動画を正しく説明しているかどうかを Yes,No で回答してもらった．シードキャプションと動画の組一つにつき 3 名のクラウドワーカーに回答してもらい，Yes と回答したクラウドワーカーの人数を手動スコアとした．

シードキャプションは複数の動画の共通内容を説明する抽象的なキャプションであると想定している．この想定を考慮し，シードキャプションが動画の一部分のみを説明している場合でも Yes と答え，シードキャプションが動画内容と食い違う場合のみ No と回答するようクラウドワーカーに指示した．

3.4 手動キャプション付与

クラウドソーシングを利用して，3.1 節で準備した動画グループに対して手動でキャプション付与を行った．手動で動画にキャプションを付与する目的は，シードキャプションよりもより信頼性の高いキャプションを集め，データセットの正解ラベルとして利用することである．

クラウドワーカーにデータセット内の動画グループを提示し，全ての動画に共通する内容を説明する抽象的なキャプションを記述してもらった．キャプションの信頼度や統一性を高めるため，次のようなルールを定めた．

- 動画に現れていない内容を補完したキャプションを書かない．
- 複数の動画の内容を列挙しただけのキャプションを書かない．
- 動画に現れる人や物の数を合計したキャプションを書かない．

1） https://github.com/facebookresearch/faiss

表 1　構築したデータセットの統計．

	訓練データ	検証データ	テストデータ
データサイズ	10,983	830	1,674
キャプション数	21,966	1,660	3,348
動画数	38,514	2,452	5,157

3.5　データセットの統計

構築したデータセットの統計を表 1 に示す．キャプション数は，シードキャプション及び手動キャプションの数の合計である．動画数は，データセットの各動画グループに含まれる動画数を重複を含めて合計したものである．構築したデータセットの詳しい統計及び分析については付録 A に記す．

4　実験

4.1　実験設定

構築したデータセットを用いて，抽象的映像キャプション生成の実験を行った．本実験では，2 件の動画のみを入力としてキャプションを生成した．データセットの正解ラベルは手動キャプションとした．

4.2　提案モデル

抽象的キャプション生成を行うモデルには，双方向 LSTM をエンコーダ，単方向 LSTM をデコーダとする Encoder-Decoder モデルを用いた．また，動画の埋め込みには行動認識モデルとして有名な I3D [12] を使用した．

モデルに対して複数の動画を同時に入力するために，各動画の各フレームの特徴ベクトルをコサイン類似度で重み付けし，各フレームごとに結合したものをモデルの入力とした．図 3 に入力する動画が 2 件の場合の前処理の説明図を示す．前処理の具体的な手順は次のとおりである．

1. 一つ目の動画特徴量 $R^{(1)}$ と，二つ目以降の動画特徴量 $R^{(2)}, R^{(3)}, \ldots$ の各フレームについて特徴ベクトルの類似度を計算し，類似度行列 $A_{12}, A_{13}, \ldots$ を作成する．類似度の計算にはコサイン類似度を用いる．
2. 類似度行列と動画特徴量の積 $R'^{(2)} = A_{12} \cdot R^{(2)}$ を計算する． $R'^{(3)}$ 以降についても同様に計算する．
3. $R^{(1)}, R'^{(2)}, R'^{(3)}, \ldots$ を同じフレームについて結合し，新たな動画特徴量 V とする．

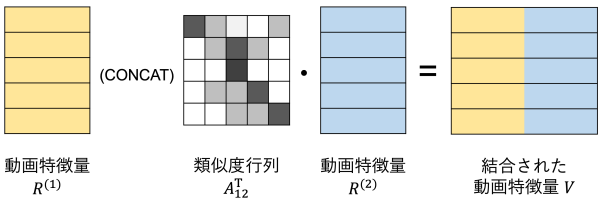

図 3　動画特徴量の前処理（二つの動画の場合）

表 2　実験で得られた各種スコア．参考として VATEX による映像キャプション生成のスコアを記載している．

	BLEU-4	Meteor	Rouge-L	CIDEr
Abstractive	9.0	13.1	38.2	12.6
VATEX [6]	28.1	21.6	46.9	44.3

各動画の特徴量は 1,024 次元の特徴ベクトルの時系列データで表現される．入力する動画が 2 件の場合，前処理によって得られる動画特徴量 V は動画特徴量 $R^{(1)}$ と同様のフレーム数を持ち，各フレームの特徴ベクトルの次元は 2,048 である．

4.3　実験結果

実験結果を表 2 に示す．Abstractive と書かれた行には，抽象的映像キャプション生成を行わせた結果得られた各種スコアを記している．VATEX と書かれた行には，VATEX の論文中で示されている，今回の実験で用いたモデルと同様の構造を持つモデルでの VATEX データセットを用いた映像キャプション生成の各種スコアを記している．

VATEX データセットを用いた映像キャプション生成と比べて，抽象的映像キャプション生成のスコアが顕著に低い．この結果から，抽象的映像キャプション生成が挑戦的なタスクであることが分かった．

5　おわりに

本稿では抽象的映像キャプション生成のデータセット構築及び実験について述べた．データセットの構築では，VATEX データセットをもとに動画検索，テキスト含意認識及びクラウドソーシングを用いて構築を行った．実験では，モデルに対して複数の動画を入力し，抽象的キャプション生成を行った．実験結果から，抽象的キャプション生成が挑戦的なタスクであることが分かった．

今後は抽象的映像キャプション生成タスクを解くより良いモデルを提案する予定である．具体的には，モデルに複数のエンコーダを用いたり，シードキャプションや自動スコア・手動スコアを利用してモデルを訓練することで精度向上が期待できる．

謝辞

本研究は富士通株式会社の助成を受けたものである.

参考文献

[1] Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. Nits-vc system for vatex video captioning challenge 2020. **arXiv preprint arXiv:2006.04058**, 2020.

[2] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 13278–13288, 2020.

[3] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In **Proceedings of the IEEE international conference on computer vision**, pp. 706–715, 2017.

[4] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In **Thirty-Second AAAI Conference on Artificial Intelligence**, 2018.

[5] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 2630–2640, 2019.

[6] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 4581–4591, 2019.

[7] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 5288–5296, 2016.

[8] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In **Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies**, pp. 190–200, 2011.

[9] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. **arXiv preprint arXiv:2104.08860**, 2021.

[10] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. **arXiv preprint arXiv:1702.08734**, 2017.

[11] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware bert for language understanding. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 9628–9635, 2020.

[12] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In **proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, pp. 6299–6308, 2017.

A　データセットの統計及び分析

本節には，抽象的映像キャプション生成のために構築したデータセットの統計・分析を記す．構築したデータセットの動画件数ごとのデータサイズを表 3 に示す．訓練データ・検証データ・テストデータのいずれにおいても，動画を 2 件含むデータが最も多い．また，動画の件数が多いデータほど，データサイズが少ない傾向にある．

自動付与・手動付与されたスコアに対する動画数の内訳をそれぞれ表 4，表 5 に示す．自動スコアに関しては，いずれのデータについてもスコアが 1 の動画が最も多く，スコアが高い動画ほど数が少ない傾向にある．手動スコアに関しては，いずれのデータについてもスコアが 2 以上の動画がデータ全体の 95%以上を占めている．手動スコアの内訳から，シードキャプションは総じて動画の内容を正しく説明していると考えられる．

表 3　データサイズの内訳（動画グループ内の動画数）．

動画数	訓練データ	検証データ	テストデータ
2 件	3,882	407	797
3 件	2,333	206	367
4 件	1,720	109	223
5 件	1,417	64	152
6 件	1,631	44	135
計	10,983	830	1,674

表 4　動画数の内訳（自動スコア）．

スコア	訓練データ	検証データ	テストデータ
1	11,819	792	1,615
2	6,863	461	880
3	4,879	271	678
4	3,712	216	509
5	2,897	195	427
6	2,414	129	308
7	1,998	110	243
8	1,629	97	197
9	1,253	93	180
10	1,050	88	120
計	38,514	2,452	5,157

表 5　動画数の内訳（手動スコア）．

スコア	訓練データ	検証データ	テストデータ
0	2	13	0
1	166	106	29
2	2,713	390	385
3	35,633	1,943	4,743
計	38,514	2,452	5,157