複数動画に対する抽象的キャプション生成のための基本モデルの検討

高橋 力斗　　清丸 寛一　　Chu Chenhui　　黒橋 禎夫
京都大学情報学研究科
{r-takahashi, kiyomaru, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

概要

複数動画に対する抽象的キャプション生成は，複数動画に共通する内容を自然言語で説明する Vision and Language タスクである．本研究では複数動画に対する抽象的キャプション生成を行うための基本モデルとして，End-to-End モデル及び Cascade モデルを検討する．モデル構造の比較実験の結果，End-to-End モデルが複数の評価指標で Cascade モデルを凌駕していることを示す．加えて，モデルに対する複数動画の入力手法がモデルに与える影響を報告する．

1　はじめに

動画キャプション生成の研究は単一動画の具体的説明に焦点が当てられてきた [1, 2]．一般的な動画キャプション生成のタスク設定では，モデルは動画内で起こっているイベントを単文で説明する [3]．動画キャプション生成には動画と自然言語の深い理解が求められ，Vision and Language 領域における主要なタスクとして盛んに研究されている．

一方，動画理解で重要なもう一つの側面に抽象的動画理解がある．図 1 に例を示す．われわれは左右の動画をそれぞれ大人たちが鏡の前で踊っている動画，小学生の女の子たちが体育館で踊っている動画と認識できる．しかしそれと同時に，二つの動画はいずれも人々がジムで踊っている動画であると抽象的に理解することも可能である．このような抽象的動画理解は，複数動画に共通する内容を見つけ，動画間の関係性を理解する上で重要な役割を果たす．

抽象的な動画理解に焦点を当てたタスクとして，われわれは複数動画に対する抽象的キャプション生成を提案し、データセット（本論文では AbstrActs と称する）を構築した [4]．このタスクは，与えられた複数の動画に共通する情報をできる限り多く説明するタスクである．このタスクでは，各動画の内容を詳細に理解する能力に加えて，動画間に共通する内容を見つけるための抽象化能力が求められる．

図 1　複数動画に対する抽象的キャプション生成の例．入力された複数の動画に共通する内容を説明する．

本論文では，複数動画に対する抽象的キャプション生成を解くための種々のモデルを検討し，AbstrActs を用いてそれらの性能を評価する．具体的には，複数の動画特徴量を結合してモデルに入力する方法と，モデル全体の構造（End-to-End 及び Cascade）について検討・評価を行う．実験結果は，異なる時間における動画内容の類似度を考慮した入力手法（Soft Alignment）及び End-to-End モデルの有効性を示す．

2　関連研究

複数動画に対する抽象的キャプション生成は複数動画から抽象的キャプションを生成するタスクである [4]．入力は n 件の動画が含まれる動画グループ G である．出力は抽象的キャプション y である．抽象的キャプション y に求められるのは，動画グループ G 内の動画に共通する内容を可能な限り多く説明することである．タスクの最終目標は訓練データを学習した複数動画に対する抽象的キャプション生成を行うモデル $p_\theta(y|G)$ を得ることである．θ はモデルのパラメータ集合である．

動画キャプション生成では，与えられた単一動画

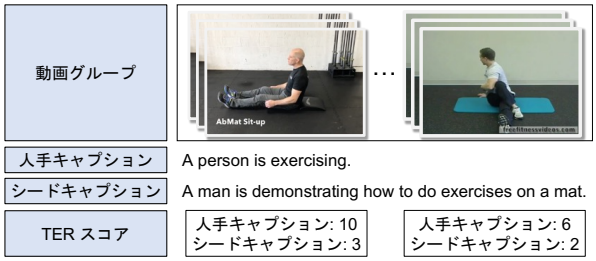


図 2　AbstrActs データセットの一例.

表 1　AbstrActs の動画及びキャプションの統計．キャプション数は人手キャプションとシードキャプションの数の合計を表す．動画数は各動画グループに含まれる動画数の合計を表す．平均含有動画数は動画グループに含まれる平均的な動画数を表す．

	訓練データ	検証データ	テストデータ
動画グループ数	10,983	840	1,674
キャプション数	21,966	1,660	3,348
動画数	38,514	2,452	5,157
動画種類数	16,732	1,475	2,848
平均含有動画数	3.5	3.0	3.1

に対してキャプションを生成するモデルを学習する．動画キャプション生成モデルは，まずはじめに事前学習済みの動画特徴量抽出器を用いて動画特徴量を得る．一般に，動画特徴量抽出器には CNN ベースのモデル [5, 6, 7] あるいは Transformer ベースのモデル [8, 9, 10] が用いられる．動画特徴量からキャプションを生成するモデルとしては，多くの先行研究が LSTM ベースのモデルを採用しているが，最近は Transformer ベースのモデルを採用する研究が増加している [11, 12].

既存の動画キャプション生成の研究の大部分は，単一動画の内容を詳細に理解して正確なキャプションを生成することに焦点が当てられてきた．本研究で注目するのは抽象的な動画理解であり，既存のキャプション研究とは焦点が異なる．

3　AbstrActs データセット

AbstrActs は複数動画に対する抽象的キャプション生成のために構築されたデータセットである [4]. AbstrActs は VATEX [13] を元データとしている．図 2 に AbstrActs の例を示す．データは動画グループ，2 種類の抽象的キャプション（人手キャプションとシードキャプション），TER スコアで構成される．本研究では人手キャプションを複数動画に対する抽象的キャプション生成の正解キャプションに用いる．

表 1 に AbstrActs の動画及びキャプションの統計を示す．データの分割は VATEX の分割に対応している．動画グループには平均 3 件以上の動画が含まれる．AbstrActs に含まれる動画には，Kinetics-600 [14, 15] に記載の 600 種類の人間の動作が含まれる．

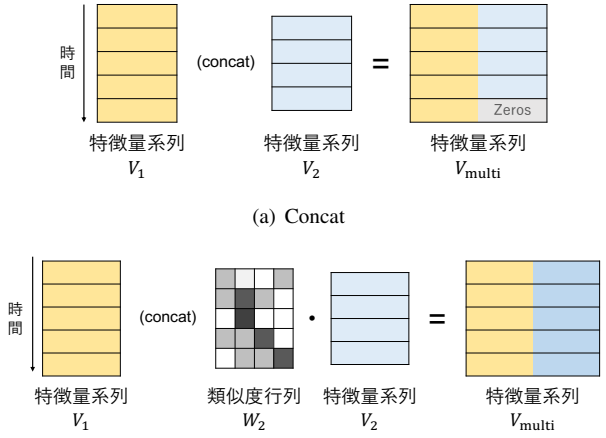


図 3　二つの動画特徴量を結合する方法.

4　基本モデルの検討

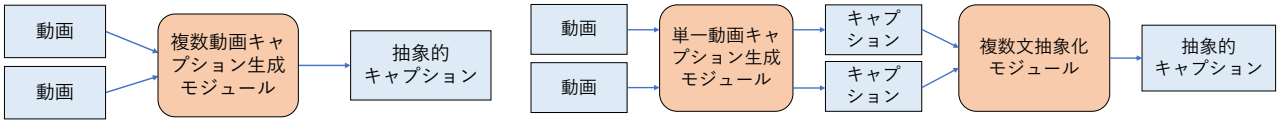
本研究では提案タスクを解くモデルとして Transformer ベースのモデルを考える．モデルに対するどのような改善が複数動画に対する抽象的キャプション生成に効果的かを確かめるために，複数の動画特徴量の入力手法及びモデル構造について検討する．

4.1　複数の動画特徴量の結合方法

Transformer ベースのモデルは一つの特徴量系列を入力として受け取る．動画特徴量は時間的な特徴量系列で表され，各時間ステップが動画の数フレームに対応している．複数の動画をモデルに入力するには，何らかの手法で複数の動画特徴量を処理して一つの特徴量にする必要がある．

本研究では複数の動画特徴量を入力する手法として，Concat と Soft Alignment の二つを考える．Concat は複数の動画特徴量を各時間ステップごとに結合する．図 3(a) に概要を示す．この手法は動画間の異なるフレーム同士の内容の違いを考慮していない．

Soft Alignment では，複数の動画特徴量のうち一つに着目し，その動画特徴量に似ている部分を他の動画特徴量から集めて結合する．この手法は注意機構 [16] に着想を得ている．図 3(b) に概要を示



(a) End-to-End モデル　　(b) Cascade モデル

図 4　複数動画に対する抽象的キャプション生成を解く二つのモデル．End-to-End モデルは複数の動画から抽象的キャプションを直接生成する．Cascade モデルでは，まず単一動画キャプション生成モジュールが各動画のキャプションを生成し，その後得られた複数のキャプションを複数文抽象化モジュールに入力して抽象的キャプションを生成する．

す．動画特徴量系列同士で時間ごとに類似度を計算し，その類似度で動画特徴量を重み付けして結合する．n 件の動画特徴量系列 $V_1, V_2, ..., V_n$ に対して Soft Alignment を適用することで，一つの特徴量系列 $V_{\text{align}} = \text{concat}(V_1, V_2', ..., V_n')$ を得る．ただし，$V_i' = W_i \cdot V_i$ である．

ここで W_i は二つの動画特徴量系列 V_1, V_i に対応する類似度行列である．また，$V_1 \in \mathbb{R}^{T_1 \cdot M}$ かつ $V_i \in \mathbb{R}^{T_i \cdot M}$ である．T_1, T_i はそれぞれ V_1, V_i の系列長を表す．M は各時間の特徴量の次元である．このとき，$W_i \in \mathbb{R}^{T_1 \cdot T_i}$ である．二つの動画の各時間同士に対応する類似度は次の式で計算される：

$$W_i(t_1, t_i) = \frac{V_1(t_1) \cdot V_i^{\mathrm{T}}(t_i)}{|V_1(t_1)||V_i(t_i)|} \tag{1}$$

時間 t における特徴量 $V_{\text{align}}(t)$ は，特徴量 $V_1(t)$ に類似する特徴量を特徴量系列 V_i から集め，類似度で重み付けして結合したものである．特徴量系列 V_{align} の系列長 l は特徴量系列 V_1 の系列長に等しい．

4.2　モデル構造

4.2.1　End-to-End モデル

End-to-End モデルは入力された動画グループから直接抽象的キャプションを生成するよう学習するモデルである．図 4(a) に End-to-End モデルの概要を示す．まず学習済みの動画特徴量抽出器を用いて，入力された n 件の動画の特徴量を得る．次に，4.1 節で説明した Concat あるいは Soft Alignment を用いて，一つの特徴量系列 V_{multi} を得る．結合された特徴量系列 V_{multi} は Transformer ベースのエンコーダモデルに入力され，特徴量の系列 $\mathbf{z} = f_{\text{enc}}(V_{\text{multi}}) = (z_1, z_2, ..., z_l)$ を得る．l は結合された特徴量系列 V_{multi} の系列長である．

最後に，Transformer ベースのデコーダモデルで抽象的キャプション y を得る．デコードのステップ t における生成単語 $y_t = f_{\text{dec}}(\mathbf{y}, \mathbf{z})$ は，過去のステップで生成された単語列 $\mathbf{y} = (y_1, y_2, ..., y_{t-1})$ と特徴量系列 $\mathbf{z}$ から生成される．

4.2.2　Cascade モデル

Cascade モデルは単一動画キャプション生成モジュールと複数文抽象化モジュールを組み合わせたモデルである．図 4(b) に Cascade モデルの概要を示す．はじめに，単一動画キャプション生成モジュールでは，Transformer モデルを用いて各動画に対応するキャプションを生成する．得られたキャプションは，学習済み単語埋め込みモデルによって単語埋め込みの系列に変換される．これらのキャプション特徴量系列は，4.1 節で説明した入力手法を適用することで，一つの特徴量系列に変換される．最後に，特徴量系列を Transformer モデルに入力し，抽象的キャプションを得る．

4.2.3　Cascade (Gold) モデル

Cascade (Gold) モデルは，Cascade モデルの単一動画キャプション生成モジュールが十分高い性能を持っている状態を想定したモデルである．このモデルでは単一動画キャプション生成モジュールを使用せず，代わりに各動画の正解キャプションを複数文抽象化モジュールに入力することで抽象的キャプションを生成する．正解キャプションは動画の内容を十分に説明しているはずであるため，正解キャプションは完璧な性能を持つ単一動画キャプション生成モジュールが生成するキャプションとみなせる．

5　モデルの評価実験

本節では 4 節で説明したモデルを用いて複数動画に対する抽象的キャプション生成の実験を行う．

5.1　実験設定

データセットには AbstrActs 及び VATEX [13] を用いた．AbstrActs の動画グループには最大 6 件の動画が含まれているが，実験設定の簡略化のた

表 2 複数の動画特徴量の結合方法に関する End-to-End モデルの性能比較.

入力手法	BLEU-4	CIDEr	METEOR	ROUGE-L
Concat	16.0	84.5	18.2	43.6
Soft Alignment	**18.6**	**130.1**	**21.5**	**47.3**

表 3 異なるモデル構造に関する性能比較．特徴量の入力手法には Soft Alignment を用いた.

	BLEU-4	CIDEr	METEOR	ROUGE-L
End-to-End	**18.6**	**130.1**	**21.5**	**47.3**
Cascade	14.8	65.4	17.4	40.9
Cascade (Gold)	17.5	103.5	19.8	44.1

めにモデルに入力する動画数を 2 件に固定した．VATEX データセットは Cascade モデルの訓練及び Cascade (Gold) モデルの推論に用いた．End-to-End モデルに入力する動画特徴量の抽出には CLIP4Clip を用いた．Cascade モデルに入力するキャプション特徴量の抽出には，fastText [17] が提供する学習済み CBOW モデルを用いた．推論結果の評価には，動画キャプション生成タスクで広く用いられている自動評価指標である BLEU-4 [18], CIDEr [19], METEOR [20], ROUGE-L [21] を用いた.

5.2 複数の動画特徴量の結合方法の比較

4.1 節で説明した複数の動画特徴量の結合方法である Concat 及び Soft Alignment の二つを比較した．この実験では，4.2.1 節で述べた End-to-End モデルを使用した．複数の動画特徴量を結合する 2 種類の手法の性能を比較した．実験結果を表 2 に示す．Soft Alingment がいずれの評価指標においても Concat を上回った．Soft Alignment は，動画同士の異なる時間における特徴量の類似度を考慮しているという点で，Concat と異なる．この違いが複数動画の共通内容を発見することに役立っていると考えられる.

5.3 モデル構造の比較

モデル構造に関する性能を比較した．図 3 に結果を示す．実験の結果，End-to-End モデルが Cascade モデルを凌駕していることが分かった．End-to-End モデルは全ての評価指標において Cascade モデルと Cascade (Gold) モデルの両方を上回った.

End-to-End モデルで高い性能が確認された理由の一つに，Cascade モデルで起こりうる誤り伝播問題が存在しないことがある．著者が人手で 50 件の推論結果を分析したところ，50 件中 7 件で，単一動画キャプション生成モジュールの生成キャプションに誤りがあることで Cascade モデルが抽象的キャプションの生成に失敗していることを確認した．図 5 に例を示す．この例では，End-to-End モデルは二つの動画に映る子供を具体的に説明できている一方，Cascade モデルは子供を “person” と説明した．図 6 に同じ例における Cascade モデルの中間出力である各動画に対するキャプションを示す．単一動画キャプション生成モジュールは図中右側の動画に映っている子供を “man” という抽象的な単語で説明した．抽象的キャプションには “child” や “kid” などの単語が含まれることが望ましいが，単一動画キャプション生成モジュールの誤りが複数文抽象化モジュールに伝播して，過度に抽象的な単語である “person” が生成された．End-to-End モデルが “kid” という望ましい単語を生成したのは，動画特徴量を直接使うために誤り伝播が起こらないからである.

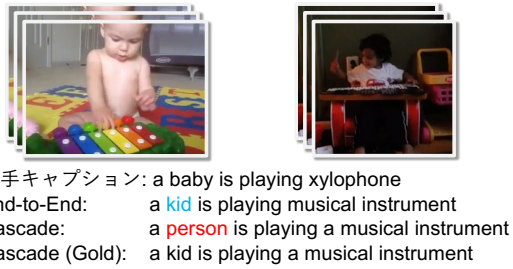


図 5 異なる構造を持つ各モデルによる推論結果の例．青色または赤色で強調されている単語は，それぞれ望ましい生成または望ましくない生成であることを表す.

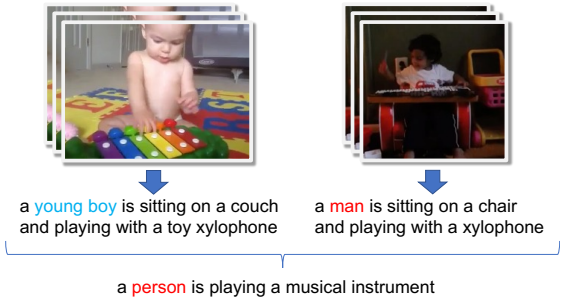


図 6 Cascade モデルで観測された，問題のあるキャプション生成の例.

6 おわりに

本研究では複数動画に対する抽象的キャプション生成タスクにおいて，基本的なモデルの検討を行った．また，モデルの抽象的キャプション生成の性能向上に効果的な要素を調べるために，AbstrActs データセットを用いて評価実験を行った．本研究で述べた実験結果及び考察により，今後の複数動画に対する抽象的キャプション生成に関する研究が促進されることを期待する.

謝辞

本研究はサムスン SDS 株式会社の助成を受けたものである.

参考文献

[1] Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. Visual to text: Survey of image and video captioning. **IEEE Transactions on Emerging Topics in Computational Intelligence**, Vol. 3, No. 4, pp. 297–312, 2019.

[2] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. **ACM Computing Surveys (CSUR)**, Vol. 52, No. 6, pp. 1–37, 2019.

[3] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. **arXiv preprint arXiv:1412.4729**, 2014.

[4] 高橋力斗, Chu Chenhui, 黒橋禎夫. 複数映像の抽象化を要するキャプション生成. 言語処理学会 第 28 回年次大会, pp. 1181–1186, 浜松, 2022.3.14.

[5] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In **Proceedings of the IEEE international conference on computer vision**, pp. 4489–4497, 2015.

[6] 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.

[7] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In **Proceedings of the European conference on computer vision (ECCV)**, pp. 305–321, 2018.

[8] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 6836–6846, 2021.

[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] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 3202–3211, 2022.

[11] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 7622–7631, 2018.

[12] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 8739–8748, 2018.

[13] 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.

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. **arXiv preprint arXiv:1705.06950**, 2017.

[15] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. **arXiv preprint arXiv:1808.01340**, 2018.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. **arXiv preprint arXiv:1409.0473**, 2014.

[17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. **Transactions of the association for computational linguistics**, Vol. 5, pp. 135–146, 2017.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[19] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 4566–4575, 2015.

[20] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In **Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization**, pp. 65–72, 2005.

[21] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In **Text summarization branches out**, pp. 74–81, 2004.