# 動画キーフレーム物語生成手法の提案

佐藤俊 [1] * 佐藤汰亮 [1] *　鈴木潤 [1]　清水伸幸 [2]
[1] 東北大学 [2] ヤフー株式会社
{shun.sato.p8,tasuku.sato.p6}@dc.tohoku.ac.jp　jun.suzuki@tohoku.ac.jp
nobushim@yahoo-corp.jp

## 概要

本稿では言語画像横断型タスクの一つとして提案した動画キーフレーム物語生成タスクに対してベースラインとなる手法を提案する．このタスクでは，指定した数の「キーフレーム」とそれに対応する「説明文」を用いて，絵コンテのように動画を中身を瞬時に把握可能な要約を生成することが目的である．我々は既存の動画要約データセットである ActivityNet Captions を拡張したデータセットを用いて，教師あり学習に基づくベースラインを構築し，性能評価及び分析を行った．

## 1　はじめに

近年スマートフォンのようなカメラ搭載型デバイスの普及により，あらゆる人が容易に動画を投稿できるようになったことで，動画コンテンツの総量は爆発的に増えつつある[1)]．こうした背景から，人間が即座に動画のおおよその内容を把握し，特定の目的に適/不適の判断をしたい場面が増えている．例えば，大量の動画の中から (1) 視聴したい好みの動画を見つける場面，(2) 特定のシーン/人物/事象を含む動画のみを抽出したい場合，などである．この状況に対処するために動画の中身を適切に要約する技術への期待が高まっている．過去の動画要約の研究としては，動画の中から重要と思われるフレームを抽出する**キーフレーム検出タスク** [1, 2, 3] や動画全体や動画をシーンごとに分割したセグメントに対して説明文を付与する**動画説明文生成タスク** [4, 5, 6, 7] などがあげられる．

これらの先行研究に対して，動画の内容を瞬時に理解できる要約を第一要件においた動画要約タスクとして，我々はこれまでに**動画キーフレーム物語生成タスク**を提案した [8]．このタスクでは１つの動画に対し，重要と思われるフレーム (**キーフレーム**) を抽出した上でそれぞれのフレームを説明する文を付与し，動画を絵コンテのような形で要約することで，動画の中身を瞬時に理解可能な要約を行う．文献 [8] では，動画説明文生成タスクのデータセットである ActivityNet Captions に対してキーフレームを決定するためのアノテーションを付与し，動画キーフレーム物語生成タスクのためのデータセットを構築した．本研究ではこのデータセットを用いた教師あり学習に基づいたアプローチで動画キーフレーム物語生成タスクのベースラインモデルを構築し，その性能評価と分析を行う．

## 2　関連研究

動画説明文生成の手法としては動画全体に対して説明文を付与する手法 [9, 10] やセグメントと呼ばれる動画をシーンごとに分割したものに説明文を付与する Dence Video Captioning と呼ばれる手法がある [6, 11, 12, 13, 7]．本研究の提案法は Dence Video Captioning の方法論をベースにし，動画キーフレーム物語生成タスクを解くために発展させた方法と位置づけることができる．Chen ら [14] は強化学習を用いて説明文とともに動画を表現する上で重要なフレームを数枚提示する手法を提案している．

文献 [8] では，１つの動画に対してその内容を表現する上で重要と思われるキーフレームのアノテーションとその説明文を付与したデータを作成し，動画を数枚のキーフレームと説明文で絵コンテのように瞬時に理解可能な形で要約する動画キーフレーム物語生成タスクの枠組みを提案した．このデータセットにより，Chen らの手法と類似のシステムを教師あり学習のアプローチで構築できるようになったとも言える．本研究では，このデータセットを用いて動画キーフレーム物語生成タスクを解くベースラインモデルを構築する．

---

* 第一，第二著者の本論文への貢献は同等である
1)　付録 A に詳細を示す．

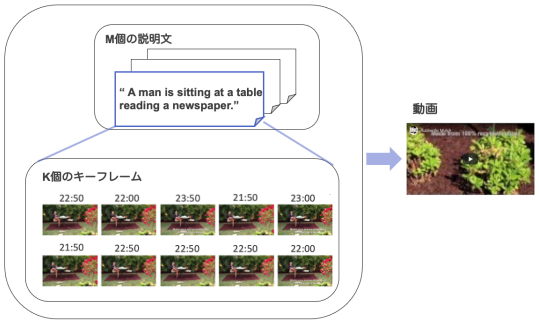



**図 1** 動画キーフレーム物語生成タスクに用いるデータセットの概要．各動画は $M$ 個の説明文とそれぞれの説明文に対応する $K$ 個のキーフレームについてのアノテーションを持つ．キーフレームの最小単位は 0.5s となっている．

# 3 動画キーフレーム物語生成タスク

本章では動画キーフレーム物語生成タスクのタスク全体の定義とその評価方法について述べる．

## 3.1 タスクの定義

$\mathbf{x}$ を 1 つの動画とし $\mathbf{y} = (y_1, \ldots, y_N)$ を動画 $\mathbf{x}$ 内のキーフレームとする．ここで $y_i$ は動画 $\mathbf{x}$ の冒頭から数えて i 番目のキーフレームを指し $N$ は事前にシステムの外で決められたシステムが出力すべきキーフレーム/説明文の数を表している，$\mathbf{z} = (z_1, \ldots, z_N)$ はシステムが出力する $N$ 個のキーフレームに対する説明文を表す．すなわち $z_i$ は $i$ 番目のキーフレーム $y_i$ を説明する文となっている．

要約として出力すべきキーフレーム/説明文の数 $N$ についてはビデオの長さ，要約の用途など状況に応じて変化することが想定される．またそれをシステムが自動的に決定することは困難であるため，$N$ は動画ごとに事前に人間が決定することを前提としている．出力する要約の数については文書要約においても同様の議論がなされている [15]．したがって，動画キーフレーム物語生成タスクは動画 $\mathbf{x}$ と出力すべき要約の数 $N$ を受け取り，$N$ 個のキーフレームと説明文を出力する $\mathcal{T}: (\mathbf{x}, N) \to (\mathbf{y}, \mathbf{z})$ のように記述される．

## 3.2 評価方法

動画キーフレーム物語生成タスクではシステムが $N$ 個のキーフレームと説明文のペア $\{(y_i, z_i)\}_{i=1}^{N}$ を出力する．本節ではキーフレームと説明文のそれぞれについて評価する方法を記述する．

### 3.2.1 キーフレームの評価方法

図 1 に今回用いるデータセットに付与されているアノテーションの概要を示した．それぞれの動画は $M$ 個のセグメントとそれを説明する説明文が付与されている．またそれぞれのセグメントごとに対応する $K$ 個のキーフレームに関するアノテーションがつけられている．

付録 B の図 3 に本タスクの評価全体の概要を示す．ここで $\mathbf{A} = (A_1, \ldots, A_N)$ は時系列順のセグメントを表す．上述の通り 1 つのセグメント $A_i$ は複数のキーフレームに関するアノテーションをもち，1 つのキーフレームを $a \in A_i$ として表記する．同様に $\mathbf{p} = (p_1, \ldots, p_N)$ はシステムが予測したキーフレームの位置を表す．システムは事前に指定した $N$ 個のキーフレームを出力するが，正解の数 $M$ が $N$ よりも大きい場合には $M$ 個の正解から $N$ 個選んでできる ${}_M C_N$ 通りの組み合わせから最適なものを探索する．

次に文献 [8] に従い式 (1) で記述される評価のための **aligned key-frame matching (AKM)** スコアを導入する．

$$\mathsf{AKM}_{\mathsf{ex}} = \max_{\mathbf{p}' \in \Gamma(\mathbf{p})} \left\{ \frac{1}{N} \sum_{i=1}^{N} \max_{a \in A_i} \left\{ \mathsf{m}_{\mathsf{ex}}(p'_i, a) \right\} \right\}, \qquad (1)$$

ここでマッチング関数 $\mathsf{m}_{\mathsf{ex}}(\cdot,\cdot)$ は 2 つのフレームを受け取り，それらが完全に同じときに 1，それ以外の場合には 0 を返す関数である．$\Gamma(\cdot)$ は入力系列の順番を入れ替えた順列集合を表す．$p'_i$ は $\mathbf{p}'$ の中の $i$ 番目の要素であり，式 (1) は解答候補 $A_i$ 中の 1 つの $a$ と予測 $p_i$ とのマッチングの中で最大の値を取るという操作を $N$ 回繰り返し，その平均値がもっとも大きくなるときの値をその予測に対するスコアとする．

また上記の式 (1) のマッチングの条件を緩和した以下の式 (2) で表されるマッチング関数 $\mathsf{m}_{\mathsf{cos}}(\cdot,\cdot)$ を用いた $\mathsf{AKM}_{\mathsf{cos}}$ も評価指標として用いる．

$$\mathsf{m}_{\mathsf{cos}}(\mathbf{v}_p, \mathbf{v}_a) = \max\left(0, \mathsf{Cos}(\mathbf{v}_p - \bar{\mathbf{v}}, \mathbf{v}_a - \bar{\mathbf{v}})\right) \qquad (2)$$

ここで $\mathbf{v}_p$ と $\mathbf{v}_a$ はそれぞれ予測と正解のキーフレームの画像の特徴ベクトルを表す．$\bar{\mathbf{v}}$ は評価対象動画内の特徴ベクトルの平均とする．式 (2) の詳細については付録 B に記載した

### 3.2.2 説明文の評価方法

説明文の評価には METEOR [16] を用いる．METEOR は動画説明文生成タスクにおける説明文

の評価指標として一般的に用いられるものの一つであり，本研究でもそれに則って評価を行う．

# 4 実験

本章では実験に用いるデータおよびモデルの詳細，実験結果について述べる．

## 4.1 データセット

今回実験には，1 章で述べたように文献 [8] で提案したデータセットを用いる．データセットの訓練/開発データの分割は ActivitytyNet Captions の分割の仕方に従い，訓練データの数が 7,727 個， 開発データの数が 4,282 個となっている．ActivitytyNet Captions では評価データは公開されていないため本章では開発データ上での性能を報告する．また各動画ごとの生成する説明の個数 $N$ については，4 とした．ただしデータについているセグメント/説明文のアノテーションが 4 個未満の場合には，その数を $N$ とした．

## 4.2 ベースラインモデル

動画キーフレーム物語生成はキーフレームの選択とその説明文の生成を同時に行う必要がある．しかし，これらを同時に実行することはモダンなニューラルモデルを用いても容易ではない．そこで我々はこのベースラインとなるシステムの推論/生成の過程を多段階に分割して構築した．図 2(a) にモデル全体の概要を示した．本システムは次の 4 つの要素から構成される．

**(Step1: セグメント抽出モジュール)** 動画を入力として動画を内容ごとにいくつかのフレームをまとめたセグメントに分割する．このモジュールでは $N$ 個以上の候補を生成する．

**(Step2: 説明文生成モジュール)** Step1 のモジュールから出力されたセグメントを受け取り，それぞれのセグメントに対して説明文を生成し，そのペアを出力とする．

**(Step3: 候補選択モジュール)** セグメント/説明文のペアを受け取り，それらの中からできるだけ時間的重複が発生しないように事前に決めた数 $N$ 個を選択する．したがって出力は $N$ 個のセグメント/説明文のペアとなる．

**(Step4: キーフレーム検出モジュール)** 与えられた $N$ 個のセグメント/説明文の情報を用いてセグメントの中から説明文が示す内容にマッチするキーフレームを予測する．このモジュールの出力がシステム全体のキーフレームと説明文の予測結果となる．

**表 1** ベースラインモデルの性能値.

| | $AKM_{ex}$ | $AKM_{cos}$ | METEOR |
|---|---|---|---|
| ベースライン | 0.493 | 0.844 | 9.61 |
| ランダム | 0.308 | 0.781 | 8.57 |

### 4.2.1 各モジュールの実装

本節では上記のシステムを構成する各モジュール群の実装について詳細を述べる．まず初めに，本研究では動画を連続したフレームの集合としてとらえる．Zhou ら [7] の設定に従い，動画はフレームごとに ResNet200 [17] と BN-Inception [18] を用いて画像特徴量に変換する．またフレームはデータセットのアノテーションに合わせて 0.5s ごとに離散的にサンプリングを行っている．

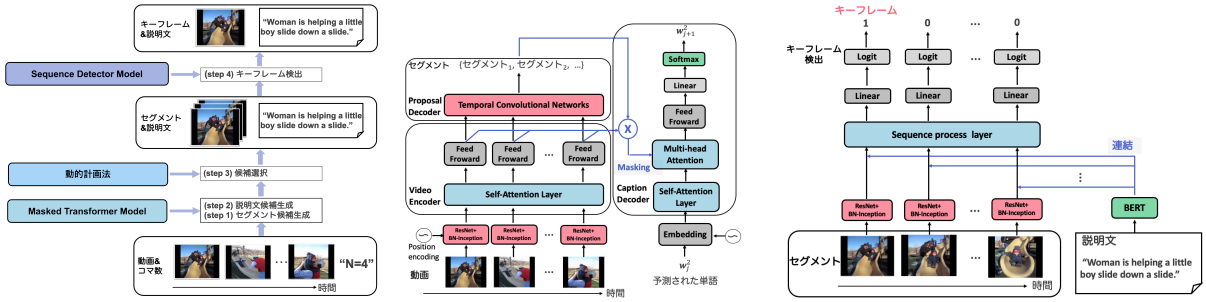
Step1 と Step2 のモジュールは Zhou らが動画説明文生成タスクのために提案している Masked Transformer Model (**MTM**) [7] を用いる．図 2(b) に MTM の概要を示している．MTM は動画内部からセグメント抽出を行いそれと同時にセグメントの内容に対応する説明文を生成する．本研究では MTM が出力するセグメントを Step1, 説明文を Step2 の出力として扱う．

MTM はありえそうなセグメント/キーワードの候補をできる限り多く出力するため，Step3 では動的計画法を用いてこれらの中から $N$ 個を選択する．また動的計画法のパス決定の決定にはセグメント/説明文の尤度を用いる．

Step4 のキーフレーム検出には Neural Sequence Detector Model (**SeqM**) を用いる．図 2(C) に SeqM の概要を示している．このモジュールの中ではまず入力されたセグメント内部の各フレームを画像特徴量に変換する．同様に入力された説明文を BERT [19] で特徴量に変換し，最終的に画像特徴量と連結して，双方向 LSTM(**Bi-LSTM**) で実装された系列処理層に入力する．系列処理層では各フレームに対してそのフレームがキーフレームか否かの 2 値分類を行う．このとき予測確率がもっとも高いフレームをシステムの最終的なキーフレームの予測とする．各モジュールのパラメータは付録 C に記載した．

## 4.3 実験結果

表 1 にベースラインモデルの性能値を示す．比較対象として，動画内部のフレームの中からランダム




(a) (b) (c)

**図 2** 動画キーフレーム物語生成モデルの概要; (a) 動画キーフレーム物語生成モデルの全体の流れ, (b) Masked Transformer Model [7] (c) Neural Sequence Detector Model.

**表 2** Step3 および Step4 において用いるモジュールに関する分析．**オラクル**は正解のセグメントを入力した場合，**予測**.DP は動的計画法, **予測**.Greedy は貪欲法で上位 $N$ 個の候補を選択した場合 のシステムの出力結果を表す．

| 候補選択モジュール | | キーフレーム検出モジュール | 評価指標 $AKM_{ex}$ | $AKM_{cos}$ | METEOR |
|---|---|---|---|---|---|
| オラクル | | Bi-LSTM | 0.574 | 0.881 | 28.29 |
| | | Tranformer | 0.398 | 0.820 | 25.87 |
| 予測. | DP | Bi-LSTM | 0.493 | 0.844 | 9.61 |
| | | Transformer | 0.366 | 0.802 | 9.32 |
| | Greedy | Bi-LSTM | 0.478 | 0.837 | 9.56 |
| | | Transformer | 0.360 | 0.795 | 8.93 |

に 1 秒分のセグメントを選択し，そのセグメントに対して説明文を生成した場合の METEOR，そのセグメントの中からランダムに選択したフレームをキーフレームの予測とした時の $AKM_{ex}$ と $AKM_{cos}$ をランダムの性能値として記載した．この結果から提案したベースラインモデルがランダムの結果よりも性能値が高く，キーフレーム検出およびその説明文生成についてデータセットを教師信号として活用できていることがわかった.[2] またモデルの出力例については付録 E に記載した．

# 5 分析

本章では提案した動画キーフレーム物語生成モデルのシステム構成について検討する．

## 5.1 候補選択モジュール (Step3)

表 2 に Step3 の候補選択モジュールとしてセグメントと説明文の候補の中から動的計画法と貪欲法を用いて上位 $N$ 個を選択した場合の結果を示す．オラクルは正解の説明文を入力した場合のキーフレーム検出と説明文生成の評価値である．動的計画法と貪欲法の結果を比べると動的計画法を用いた場合の予測が優れていることがわかった．これは一般的に動的計画法が貪欲法と異なり，厳密な最適解に到達できることから直感的な結果と言える．

また結果からキーフレーム検出および説明文生成の両方でオラクルの結果と予測値の間に大きな差があり，モデル全体の性能には改善の余地があると言える．

## 5.2 キーフレーム検出モジュール (Step4)

表 2 にキーフレーム検出モジュールの系列処理層として Transformer [20] と Bi-LSTM を用いた場合の結果を記載した．$AKM_{ex}$ および $AKM_{cos}$ のいずれの評価指標でも Bi-LSTM を用いた場合の方がキーフレーム検出の性能値が高いことがわかる．近年多くのタスクにおいて優れた性能を発揮している Transformer であるが，それらの多くは大規模データセットを用いた訓練による部分が大きい．今回扱ったデータセットは相対的に小規模であり，直接的な系列処理アーキテクチャである Bi-LSTM の方がキーフレーム検出に適していたと考えられる．

これらの結果を踏まえて表 1 では動的計画法と Bi-LSTM を用いたモデルを本タスクのベースラインとして選出した．

# 6 結論

本研究では，動画キーフレーム物語生成タスクのために必要な複数の機能を組み合わせたベースラインとなるモデルを提案し，その構成要素について分析を行った．提案したモデルは一定の性能を示したものの，生成されるキャプションの質やキーフレームの検出についていずれもオラクルの性能値と大きな差があった．各機能を構成する個別のモジュールの性能向上や全体のアーキテクチャの見直しを行いさらなる性能改善を今後の課題としたい．

2) METEOR の値は先行研究と比較しても妥当な値である．詳細については付録 D に記載した

# 参考文献


[1] W. Wolf. Key frame selection by motion analysis. In **1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings**, Vol. 2, pp. 1228–1231 vol. 2, 1996.

[2] Sourabh Kulhare, S. Sah, Suhas Pillai, and R. Ptucha. Key frame extraction for salient activity recognition. In **23rd International Conference on Pattern Recognition (ICPR)**, pp. 835–840, 2016.

[3] X. Yan, Syed Zulqarnain Gilani, Hanlin Qin, Mingtao Feng, L. Zhang, and A. Mian. Deep keyframe detection in human action videos. **ArXiv**, Vol. abs/1804.10021, , 2018.

[4] Oriol Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 3156–3164, 2015.

[5] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, **Proceedings of the 32nd International Conference on Machine Learning**, Vol. 37 of **Proceedings of Machine Learning Research**, pp. 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

[6] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In **International Conference on Computer Vision (ICCV)**, 2017.

[7] 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 (CVPR)**, June 2018.

[8] 晃太郎北山, 潤鈴木, 伸幸清水. 動画キーフレーム物語生成タスクの提案とデータセットの構築. 人工知能学会全国大会論文集, Vol. JSAI2021, No. 0, pp. 4I4GS7e03–4I4GS7e03, 2021.

[9] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence – video to text. In **2015 IEEE International Conference on Computer Vision (ICCV)**, pp. 4534–4542, 2015.

[10] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. In **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1494–1504, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[11] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 7190–7198, 2018.

[12] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 7492–7500, 2018.

[13] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, **Computer Vision – ECCV 2020**, pp. 121–137, Cham, 2020. Springer International Publishing.

[14] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, **Computer Vision – ECCV 2018**, pp. 367–384, Cham, 2018. Springer International Publishing.

[15] P. Over and J. Yen. An introduction to DUC 2003: Intrinsic evaluation of generic news text summarization systems, 2003.

[16] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In **Proceedings of the Second Workshop on Statistical Machine Translation**, pp. 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **2016 IEEE Conference on Computer Vision and Pattern Recognition**, pp. 770–778, 2016.

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, **Proceedings of the 32nd International Conference on Machine Learning**, Vol. 37 of **Proceedings of Machine Learning Research**, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

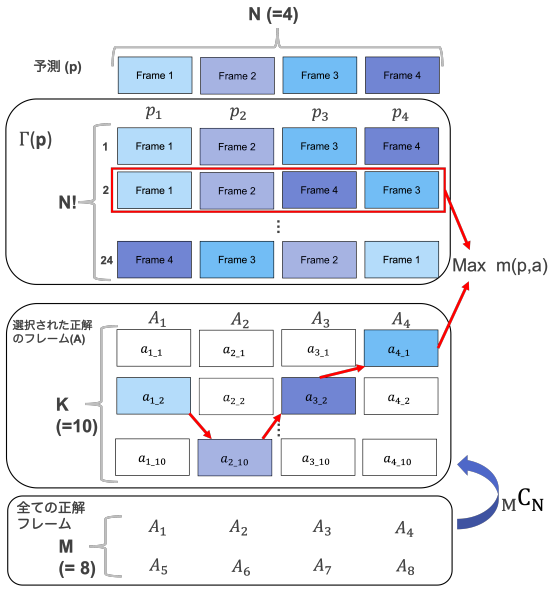



**図 3** キーフレームに関する評価の概要. $\mathbf{A} = (A_1, \ldots, A_N)$ は正例のキーフレーム集合を表し，それぞれの $A_i$ は K 個の正解のキーフレーム $(a_{i1}, \ldots, a_{iK})$ を持つ. $\mathbf{p} = (p_1, \ldots, p_N)$ はシステムが予測した $N$ 個のキーフレームを表す. $\Gamma(\mathbf{p})$ は予測されたキーフレームの順番を入れ替えて作られる順列を表す. キーフレームの評価においては

## A 動画データに関する統計情報

動画投稿サイト Youtube では毎分合計 500 時間以上の動画アップロードされていると報告されている.[3)]

## B 評価方法の詳細

式 (1) で表されるマッチング関数 $\mathtt{m}_{\mathtt{ex}}(\cdot,\cdot)$ は予測と正解のキーフレームの間に完全一致を要求しているが，セグメント内部では類似したフレームが複数あることが想定され，正解の候補が複数あったとしても完全に同じフレームを当てるのは容易でないことが考えられる．そこで我々は AKM スコアのマッチング関数 $\mathtt{m}_{\mathtt{ex}}(\cdot,\cdot)$ の代わりに制限を緩和した式 (2) で表されるマッチング関数 $\mathtt{m}_{\mathtt{cos}}(\cdot,\cdot)$ を用いた $\mathtt{AKM}_{cos}$ も評価指標の 1 つとして導入する.

$\mathtt{m}_{\mathtt{cos}}(\cdot,\cdot)$ は $\mathtt{m}_{\mathtt{ex}}(\cdot,\cdot)$ と異なり，異なる位置のフレームであっても正解のキーフレームと類似したフレームを選択できていれば 1 に近い値をとる.

また図 3 で示す評価全体では全ての組み合わせの探索を行うと膨大な計算コストがかかるため，AKM スコアを用いた動的計画法によって効率的な探索を行う.

**表 3** SeqM のハイパーパラメータ

| 入力 | 系列 | | 動画 | 説明文 |
|---|---|---|---|---|
| モデル | Bi-LSTM | Transformer | ResNet/BN | BERT |
| 入力次元 | 1792 | 1792 | - | - |
| 出力次元 | 512 | 512 | 512/512 | 768 |
| 学習率 | 0.0001 | 0.00005 | - | - |

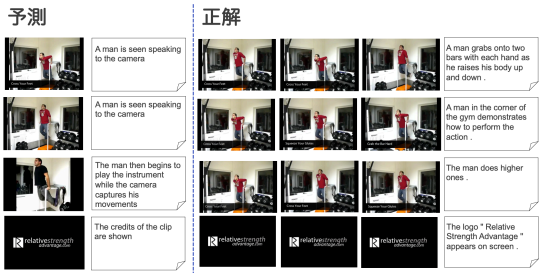



**図 4** システムの予測と正解のキーフレームと説明文と実例

## C モデルのパラメータ

MTM の実装およびハイパーパラメータについては MTM を提案している Zhou ら [7] が公開しているもの[4)]に従った.

表 3 に SeqM のパラメータを記載した．SeqM で用いる ResNet200 と BN-inception の次元数はそれぞれ 512 次元とし，説明文のテキストの特徴量として用いる BERT の次元数は 768 次元とした．最終的には，この 3 つの特徴量を連結した 1792 (=512+512+768) 次元のベクトルを SeqM の入力とした.

## D METEOR の値の妥当性

本研究で説明文の生成を担っている MTM の提案論文 [7] では ActivityNet Captions を Dence Video Captioning のデータセットとして扱い，MTM を用いて説明文を生成している．その論文における説明文の METEOR の値は 9.56 であり，本研究での METEOR が妥当な値であることを示している.

## E モデルの出力例

図 4 にシステムの出力と対応するキーフレームおよび説明文の正解の実例を示す.

3) https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

4) https://github.com/salesforce/densecap