事前学習モデルを活用した End-to-end 型動画キーフレーム物語生成法

仲村祐希 [1] * 工藤慧音 [1] * 鈴木潤 [1] 清水伸幸 [2]
[1] 東北大学 [2] ヤフー株式会社
{keito.kudo.q4, yuki.nakamura.r1}@dc.tohoku.ac.jp jun.suzuki@tohoku.ac.jp
nobushim@yahoo-corp.jp

概要

深層ニューラルネットワークや事前学習済みモデルの発展に伴って視覚および言語の横断タスクが多くの研究分野にて注目されている．映像と言語の融合領域タスクの一つとして「動画キーフレーム物語生成タスク」が提案されている．このタスク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，動画全体の要約を生成するタスクと言える．本研究では，既存のベースラインモデルの改善に加えて，新たに End-to-end での動画キーフレーム物語を生成する手法である解候補制約付き生成法を提案する．実験の結果，提案法は改善したベースラインモデルに比べて，キーフレームの選択精度は劣るものの，生成された説明文の質はベースラインモデルに比べて向上する結果となった．

1 はじめに

本研究では，映像と言語の融合タスクの一つである動画キーフレーム物語生成タスク [1] に着目し，更なる性能向上のため新たな方法論の確立を目指す．動画キーフレーム物語生成タスクは，動画要約タスクの一種で，簡潔に述べると動画を絵コンテのように数枚の画像と説明文で説明するタスクである．このタスクの難しさは，キーフレームの相対的な重要度を考慮しつつ，対応する説明文も生成しなくてはならない点である．キーフレームと説明文は相互依存関係にあり，どちらかが決まるともう一方も決まるといった関係である．よって，キーフレームと説明文は理想的には同時に予測する必要がある．本タスクのベースライン法を提案している文献 [2] では，同時予測が計算コスト面でも実装面でも困難であることから，問題を分解してキーフレームの抽出と説明文の生成を独立に実施するモデルを本タスクのベースラインとして提案している．本研究では，まずベースラインの改良として，キーフレームの抽出と説明文の生成の一方を固定し交互に反復して予測する方法を提案する．また，本タスクの性質に適した新しい方法論として，事前学習済みモデルを活用した End-to-end 方式のモデルも提案する．従来のベースラインの改良と End-to-end 方式の提案法を本タスクのデータを用いて評価し，その有効性を検証する．

2 動画キーフレーム物語生成タスク

概要． 最も有名な映像と言語の融合タスクは，動画説明文生成タスクと考えられる [3, 4]．また，その発展タスクとして，動画から一連のシーンを抽出し，そのシーン単位に説明文を付与するタスクを Dence Video Captioning (DVC) と呼ぶ [5, 6, 7, 8, 9]．本研究で取り上げる動画キーフレーム物語生成タスクは，DVC タスクをより現実の利用場面に合わせて改良したタスクに位置付けられる．より具体的には，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を抽出し，更に各キーフレームに対応する説明文を生成し，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提示するタスクである．この時，抽出すべきキーフレームと説明文のペアの数 N は，事前に与えられることを前提とする．

評価方法 動画キーフレーム物語生成タスクを提案した文献 [1] に従い，抽出したキーフレームの評価には，正解と予測結果のキーフレーム間の一致度合い計測する **aligned key-frame matching (AKM)** スコアを用いる．また，説明文の評価には，正例と予測結果の説明文間の一致度合いを計測する **METEOR** [10] を用いる．本研究では，説明文の評

* 第一，第二著者の本論文への貢献は同等である

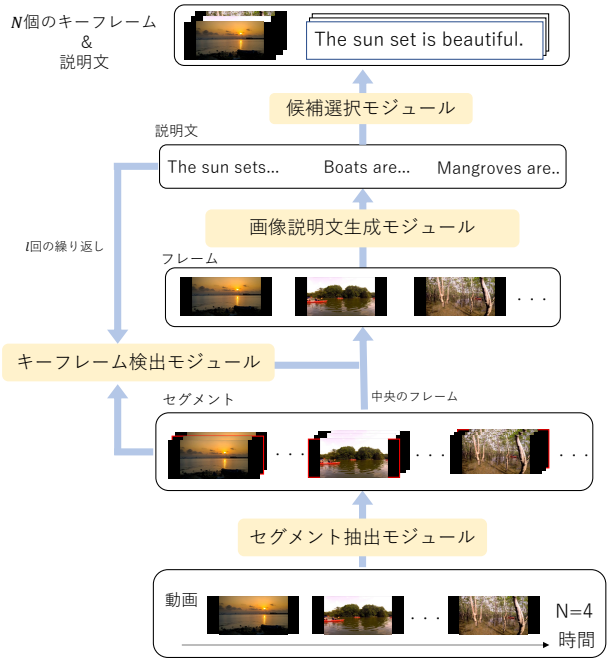

図 1　ベースラインモデルの概要

価指標として，更にニューラル言語モデルをベースとした評価指標である **BLEURT** [11] を追加する．これは，説明文間の意味的な類似性をより評価するためである．各動画ごとに，正例となるキーフレームと説明文のペアは N 個以上存在することから，評価時には，モデルが生成した N 個のキーフレームとの $\mathrm{AKM}_{\mathrm{cos}}$ が最大となる N 個の評価データのキーフレームと説明文に対して，各種評価指標を用いた評価を行う．詳細な定義は，これまでの研究に関する文献 [2] を参照されたい．

3　評価データの改善

文献 [1] にて構築されたデータは，ActivityNet Captioning [5] のデータを拡張して作成されたものである．基本的にクラウドワーカーが作成したデータのため，評価データ中に低品質な説明文が散見される．そのため，潜在的にシステムの性能を正しく評価できない可能性がある．そこで，本研究では評価データの改善も合わせて実施した．具体的には，英語母語話者のいる信頼性の高いデータ作成業者に，元データの部分集合に対して正解説明文の再付与を依頼した．最終的に，442 動画に対して高品質な説明文を付与した評価データを構築した．

4　ベースラインモデル概要

本研究では，本タスクにおけるベースラインとして提案されたモデル [2] をキーフレームの抽出と説明文の生成を一方を固定し交互に反復して予測するように改良を加えたものを新たなベースラインとして用いる．図 1 に本研究で用いるベースラインモデルの概要を示す．

4.1　システムの概要

改良したベースラインモデルは次の 4 つの要素から構成される．

セグメント抽出モジュール　動画を入力として動画をシーンごとに複数のフレームをまとめた**セグメント**に分割する．セグメント抽出モジュールとして，既存研究のベースラインモデル [2] で利用した Masked Transformer に加えて [9]，動画処理のツールである PySceneDetect[1] を用いた．このモジュールによって N 個以上のセグメントを生成する．

画像説明文生成モジュール　候補となるフレームに対して，事前学習済みの画像説明文生成モデルを使って説明文を生成する．事前学習済みの画像説明文生成モデルとして，fairseq-image-captioning[2]，clipcap [12]，clip-reward [13] を利用する．それぞれのモデルは，本研究で用いるデータセット [1] に含まれれるキーフレームと説明文を用いて微調整する[3]．

キーフレーム検出モジュール　セグメントと先に選択したフレームに対する説明文を入力として，セグメントの中から説明文にマッチするフレームを予測する．既存研究 [2] と同様に双方向 LSTM により構築したモデルを利用する．セグメント内の全てのフレームを ResNet200 [14] と BN-inception [15] を用いて特徴量としたものと，説明文を BERT [16] を用いて特徴量に変換したものを連結した系列を入力として，各フレームがキーフレームか否かの 2 値分類を行うモデルである．

候補選択モジュール　既存研究 [2] と同様，セグメントと説明文のペアを受け取り，N 個のペアを選択する．選択の指標として，セグメント抽出モ

1）www.github.com/Breakthrough/PySceneDetect
2）www.github.com/krasserm/fairseq-image-captioning
3）clip-reward はキャプションニングモデルに加えて，学習に用いる CLIP の微調整も行なう．

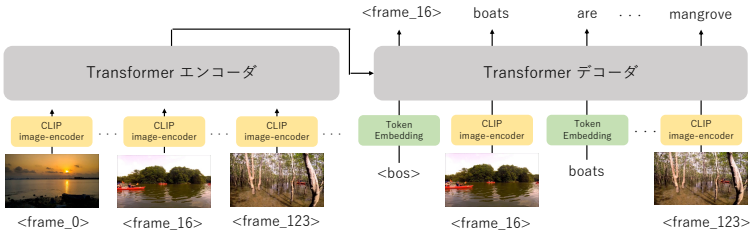

図 2　解候補制約付き生成法のモデルの概要.

ジュールによって計算されたセグメントのスコア [4] と，画像説明文生成モジュールによって計算された説明文の尤度の和を用いる．動的計画法を用いて，この指標の和が最大となるように，できるだけセグメントの時間的重複が発生しないように N 個のセグメントと説明文のペアを選択した．選ばれたセグメントのフレーム（キーフレーム）および説明文がベースラインモデルの最終的な出力となる．

ベースラインモデルはこれら 4 つのモジュールを利用し，以下の手順で N 個のキーフレームと説明文を選択・生成する．

1. セグメント抽出モジュールを用いて，入力動画をセグメントに分割する．
2. 分割された全てのセグメントの中央のフレームに対して，画像説明文生成モジュールを用いて説明文を生成する．
3. 生成された説明文とセグメントを入力として，キーフレーム検出モジュールにより説明文に最も合うフレームを選択する．
4. 各セグメントから選ばれたフレームを入力として再び，画像説明文生成モジュールにより説明文を生成する．
5. 3.，4. を l 回繰り返す（本研究では全ての設定において $l = 4$ で固定した）．
6. 候補選択モジュールにより選択されたキーフレームと説明文をモデルの出力とする．

5　解候補制約付き生成法

本研究では，新たに事前学習モデルを活用した End-to-end 型の手法である**解候補制約付き生成法**を提案する．この手法ではフレームに分割した動画を入力として N 個のキーフレームの選択と説明文のペアを生成を行うモデルを学習する．キーフレームと説明文は異なるモーダルの情報であるが，提案法のモデルは 1 つの系列として処理を行う．解候補制約付き生成法の詳細は付録 A を参照されたい．

5.1　アーキテクチャと入出力

モデルのアーキテクチャとして Transformer [17] のエンコーダ，デコーダをベースとし，入力系列長などを変更したモデルを利用する．

モデルの概要を図 2 に示す．動画から 0.5 秒ごとにサンプリングした全てのフレームを CLIP [18] の画像エンコーダを用いて特徴量ベクトルとし，解候補制約付き生成モデルのエンコーダに入力する．出力はその動画を要約した N 個のキーフレームとなるフレームのインデックスとそれに対する説明文のペアである．デコーダは一度の推論で，キーフレームとそれに対応する説明文のペア N 個を 1 つずつ生成（選択）するように学習させる（図 2）．デコーダへは入力がテキストトークンの場合，そのトークンに対応する埋め込みを入力とする．一方，入力がフレームの場合，フレームを CLIP の画像エンコーダによって特徴量に変換したもの入力とする．デコーダは通常の系列変換モデルと同様に，次のテキストトークンを予測するか，どのフレームを選択すべきかを表すフレームのインデックスを予測する．

5.2　学習方法

事前学習　MS COCO キャプションデータセット [19] を用いて，疑似的な動画キーフレーム物語生成データセットを作成し事前学習を行なう．これにより，モデルの生成文の質の向上を目指す．

微調整　これまでの研究 [1] において付与されたアノテーションを元に，各動画から N 個のキーフレームと説明文のペアを 8 パターンサンプリングし学習データとして用いた．

5.3　推論時

初めに，動画からサンプリングした全てフレームに対し，4 節で用いたものと同様の，事前学習済み画像説明文生成モデルを用いて説明文の生成を行な

4）Masked Transformer のセグメントのスコアはモデルから出力されるスコア，PySceneDetect のスコアは 1 で固定とした．

表 1 ベースラインモデルおよび解候補制約付き生成法の各評価スコア．$\mathrm{AKM}_{\mathrm{ex}}$ は $\mathrm{AKM}_{\mathrm{cos}}$ が最大となるようにマッチングした正解フレームに対して，一致しているキーフレームの割合を表す．

手法	キャプションモデル	$\mathrm{AKM}_{\mathrm{ex}}$	$\mathrm{AKM}_{\mathrm{cos}}$	BLEURT	METEOR
ベースラインモデル	clip-reward	**.420**	**.811**	.364	.111
	fairseq-image-captioning	.397	.796	.403	.118
	clipcap	.384	.801	.361	**.147**
解候補制約付き生成法	clip-reward	.258	.712	.443	.134
	fairseq-image-captioning	.268	.721	**.454**	.130
	clipcap	.258	.726	.364	.123

う．その後，特徴量に変換した全てのフレームをモデルのエンコーダへの入力とし，N 個のキーフレームと説明文のペアの候補を生成した際の尤度を元に，モデルは最終的な出力を行う．

サンプリングされた全てのフレームとそれらに付与された説明文の中から N 個の説明文とキーフレームを選択する組み合わせは膨大に存在するため，全ての組み合わせについて尤度の計算を行うのは計算量的に困難である．そこで，1 つの説明文とフレームのペアに対する尤度の計算を 1 単位としたビームサーチを行いながら，最終的に N 個のキーフレームと説明文に対する尤度の計算を行なう．[5)]

6 動画キーフレーム物語生成実験

データセット 訓練データは文献 [1] で提案されたものを用いた．ただし，評価データに関しては 3 節で説明した，説明文を改善したデータを用いた．

実験設定 基本的な実験設定は既存研究 [2] に従った．$\mathrm{AKM}_{\mathrm{cos}}$ の計算に用いる画像の特徴量ベクトルは ResNet200 [14] を用いた．説明文の評価の前処理として Moses[20] の lowercase.perl と tokenizer.perl [6)] を用いてトークナイズを行った．なお，説明文の評価は出力された N 個の説明文をビデオ単位で一つに連結してから各評価スコアを算出した．最終的に一つの動画から出力するキーフレームとキャプションのペアの数 N は 4 とした．

6.1 実験結果

表 1 にベースラインモデルおよび解候補制約付き生成法における評価結果を示した．

まず，$\mathrm{AKM}_{\mathrm{cos}}$ および $\mathrm{AKM}_{\mathrm{ex}}$ についてはベースラインモデルの方が総じてスコアが高く，より良いキーフレームを選択できていることが分かった．キーフレームの選択において，ベースラインモデルの方がスコアが高かった理由として，動画をセグメント単位に区切ってから処理していることが挙げられる．解候補制約付き生成法は，動画全体のフレームを入力としているため，キーフレームの候補が多い．そのため，キーフレームを当てるのが困難であり，解制約付き生成法のスコアが低かったと考えられる．

次に，生成した説明文の評価については概ね解候補制約付き生成法の方がスコアが高いという結果となった．これは解候補制約付き生成法がキーフレームと説明文を同じ系列として処理していることによる影響ではないかと考えられる．

7 おわりに

本研究では，既存の動画キーフレーム物語生成タスクに対して，反復法によるベースラインの改善と，事前学習モデルを活用した End-to-end 型の方法論を提案した．提案法では，事前学習済みの画像説明文生成モデルを活用し，各フレームに対して網羅的に説明文を生成し，その候補を制約として説明文の生成とキーフレームを抽出を同時に行う．また，付加的な要素として，品質の低い評価データを改善するために，従来の 1/10 程度の部分集合に対してではあるが，人手による高品質な説明文を再付与し，再構築した評価データを用いて提案法の有効性を検証する実験を行った．実験の結果，提案法は改善したベースラインモデルに比べて，キーフレームの選択精度は劣るものの，生成された説明文の質はベースラインモデルに比べて向上する結果となった．

まずは本研究にて End-to-end 型の動画要約法が完成したので，今後は今回の実験で顕著に良い結果が得られなかったキーフレーム抽出の性能向上を実現することを目指す．

5) ビームサーチアルゴリズムの詳細は付録 A に示す．
6) www.github.com/moses-smt/mosesdecoder

謝辞

本研究は，Yahoo 研究所と東北大学との共同研究として実施されたものである．また，本研究の一部（データ作成）は，JST ムーンショット型研究開発事業 JPMJMS2011 の助成を受けて実施されたものである。

参考文献

[1] 北山晃太郎, 鈴木潤, 清水伸幸. 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.

[2] 佐藤俊, 佐藤汰亮, 鈴木潤, 清水伸幸. 動画キーフレーム物語生成手法の提案. 言語処理学会 第 28 回年次大会 発表論文集, Vol. NLP2022, No. 0, 2022.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[12] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. **CoRR**, Vol. abs/2111.09734, , 2021.

[13] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with CLIP reward. In **Findings of the Association for Computational Linguistics: NAACL 2022**, pp. 517–527, Seattle, United States, July 2022. Association for Computational Linguistics.

[14] 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.

[15] 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.

[16] 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.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L 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.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **ICML**, 2021.

[19] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. **CoRR**, Vol. abs/1504.00325, , 2015.

[20] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In **Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions**, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[22] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, pp. 961–970, 2015.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.

表 2　解候補制約付きモデルの学習設定

事前学習	
学習データサイズ	100,000 事例
最適化アルゴリズム	AdamW [23]
	($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$)
学習率のスケジューラ	Cosine decay
Warmup ステップ	4,000 ステップ
学習率 (最大値)	0.000005
ドロップアウト	0.1
バッチサイズ	168
学習エポック数	30 エポック
微調整	
学習データサイズ	23,216 事例
最適化アルゴリズム	AdamW [23]
	($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$)
学習率のスケジューラ	Cosine decay
Warmup ステップ	2,000 ステップ
学習率 (最大値)	0.000005
ドロップアウト	0.1
バッチサイズ	168
学習エポック数	200 エポック

A　解候補制約付き生成モデルの詳細

アーキテクチャ　モデルのアーキテクチャとして，Transformer [17] のエンコーダ，デコーダを用いる．実装は T5-large [21] をベースとし，入力系列長を 2048 とする．これは，0.5 秒おきにフレームをサンプリングした場合，約 17 分の動画まで全てのフレームを入力することができる長さである．データセットを構成する動画は最長 20 分で多くの動画が 5〜10 分の長さであり [22]，大半の動画を切り捨てることなく利用できる入力系列長の大きさとした．また，キーフレームの選択を行うためのフレームのインデックスを予測するため，最終層の出力次元数をエンコーダの語彙数+入力系列長 (2048) に増加させる．加えて，入力の画像特徴量の次元数を変換するための線形層も追加する．アーキテクチャの変更がなく利用可能な部分については，事前学習済みの T5[7] のパラメータの初期値として学習を行なう．

事前学習　MS COCO キャプションデータセット [19] を用いて，疑似的な動画データセットを作成し事前学習を行なう．以下の手順で学習に利用する各事例を作成する．

1. MS COCO データセットから N 個の画像と説明文のペアをサンプリングする．
2. エンコーダの入力系列長を長さ 1 以上の N 個の区間に無作為分割する．
3. それぞれの区間から 1 つインデックスを無作為に選択する．このインデックスが N 個の正解のフレームのインデックスとなる．

分割した各区間の位置のエンコーダの入力には，全てその区間の中の正解フレームと同じ特徴量を入力する．ただし，正解フレームのインデックスに入力する特徴量を除いて，以下の式で摂動を付与する．

$$Noise = v_{\text{average}} \beta N(0, 1) \tag{1}$$

ここで，v_{average} はミニバッチ内の画像特徴量の全ての数値を平均したスカラー値である．また，β は摂動の大きさを制御するパラメータであり，本実験では $\beta = 0.05$ に統一した．モデルは N 個の (ノイズの加えられていない) 正解フレームのインデックスとそれに付与された説明文を予測するように学習する．

学習設定　表 2 に学習時のハイパーパラメータを示す．

7)　Huggingface Transformers にて公開されているものを利用した https://huggingface.co/t5-large

ビームサーチアルゴリズム　アルゴリズム 1 に解候補制約付き生成モデルのビームサーチを用いた生成アルゴリズムを示す．また，実験では解候補制約付き生成法で用いるビーム幅は 3 とした．

Algorithm 1 解候補制約付き生成モデルで用いるビームサーチアルゴリズム

fnput $NumKeyFrame \in \mathbb{N}$: Number of key frame.
Input $BeamWidth \in \mathbb{N}$: Beam width.
Input $Video$: List of frames in a video.

```
Captions = []
for each frame ∈ Videos do
    Captions.append(ImageCaptioningModel(frame))
end for
Candidates = []
Beams = []
for i = 1 to NumKeyFrame do
    for each BeamFrames, BeamCaptions ∈ Beams do
        LastTime = BeamFrames[−1].time
        for each frame, caption ∈ Videos, Captions do
            if frame.time <= LastTime then
                continue
            end if
            InputFrames = BeamFrames + [frame]
            InputCaptions = BeamCaptions + [caption]
            score = Transformer(
              InputFrames, InputCaptions
            )
            Candidates.append(
              (score, InputFrames, InputCaptions)
            )
        end for
    end for
    Beams ← ScoreTopK(Candidates, BeamWidth)
end for
return Beams[0]
```