VPAI_Lab at MedVidQA 2022: A Two-Stage Cross-modal Fusion Method for Medical Instructional Video Classification

Bin Li[1]*, Yixuan Weng[2]*, Fei Xia[2,3]*, Bin Sun[1], Shutao Li[1]
[1] College of Electrical and Information Engineering, Hunan University
[2] National Laboratory of Pattern Recognition, Institute of Automation, CAS
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences
{libincn, sunbin611, shutao_li}@hnu.edu.cn,
wengsyx@gmail.com, xiafei2020@ia.ac.cn

Abstract

This paper introduces the method of VPAI_Lab team's experiments on BioNLP 2022 shared task 1 Medical Video Classification (MedVidCL). Given an input video, the MedVidCL task aims to correctly classify it into one of three following categories: Medical Instructional, Medical Non-instructional, and Non-medical. Inspired by its dataset construction process, we divide the classification process into two stages. The first stage is to classify videos into medical videos and non-medical videos. In the second stage, for those samples classified as medical videos, we further classify them into instructional videos and non-instructional videos. In addition, we also propose the cross-modal fusion method to solve the video classification, such as fusing the text features (question and subtitles) from the pre-training language models and visual features from image frames. Specifically, we use textual information to concatenate and query the visual information for obtaining better feature representation. Extensive experiments show that the proposed method significantly outperforms the official baseline method by 15.4% in the F1 score, which shows its effectiveness. Finally, the official results show that our method ranks the Top-1 on the official unseen test set. All the experimental codes are open-sourced at https://github.com/Lireanstar/MedVidCL.

1 Introduction

One of the key goals of artificial intelligence (AI) is to develop a multimodal system that uses natural language queries to facilitate communication with the visual world (i.e., images, videos) (Cukurova et al., 2019). In recent years, the gap between language and visual understanding has narrowed (Guo et al., 2016; Lu et al., 2019) due to the development of pre-trained models (Devlin et al., 2018) and the introduction of large-scale language-vision datasets (Lei et al., 2018, 2020a,b). Improvements have been made in numerous vision-and-language tasks, such as visual classification (Servières et al., 2021), video question answering (Huang et al., 2020) and natural language video localization (Yuan et al., 2019; Chen et al., 2019; Zhang et al., 2020).

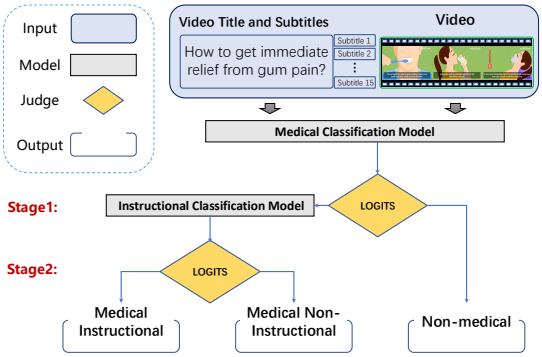

Figure 1: The overview of the proposed two-stage cross-modal fusion method.

The recent proliferation of online videos has changed the way people acquire information and knowledge. More and more people are accustomed to using instructional videos to teach or learn specific tasks. Medical instructional videos are more suitable and conducive to conveying key information through both visual and verbal communication in an effective and efficient manner (Gupta et al., 2022; Gupta and Demner-Fushman, 2022).

To better distinguish medical instructional videos from other videos, MedVidQA proposes Medical Video Classification (MedVidCL) task[1]. Given an input video, the MedVidCL task aims to correctly classify it into one of three following categories: Medical Instructional, Medical Non-instructional, and Non-medical.

Inspired by its dataset construction process (Gupta et al., 2022), we divide the classification process into two stages. As shown in Figure 1, given the question "How to get immediate relief

*These authors contribute equally to this work.

[1]https://codalab.lisn.upsaclay.fr/competitions/1058

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 212–219

from gum pain?" and subtitles and videos, we first classify it into medical and non-medical videos by the Medical Classification Model in stage one. If the input is classified as a medical video, in the second stage, we further classify it into medical instructional videos and non-instructional videos through the Instructional Classification Model.

For monomodal (Language) setting, it is feasible to classify the input video with the corresponding subtitle texts since the content of the video is directly related to its subtitles (Mahdisoltani et al., 2018; Perez-Martin et al., 2021). We choose various pre-trained models combined with our designed two-stage method to perform video classification. The experimental results show that pre-trained language models can achieve better semantic understanding.

Moreover, visual information is equally important for the MedVidCL task. To make full use of the information of visual and textual modality, we perform feature extraction on them separately and perform the query concatenation mechanism (Zhang et al., 2020) for better feature representation.

In this paper, we propose a two-stage cross-modal fusion method, by fusing the extracted visual features and textual features from the pre-trained language model. Compared with the official multimodal method, our multimodal method improves by 15.4% in F1, and the results show the effectiveness of our cross-modal method.

2 Proposed Approach

In this section, we will elaborate on the proposed approach for the medical video classification (MedVidCL) track. As the pre-training language method can enhance the performance of semantic representation queried by the text subtitles (Perez-Martin et al., 2022), we design the two-stages cross-modal fusion method, which is described in turn as follows.

2.1 Two-stage modeling for classification

Acquisition of the MedVidCL dataset mainly goes through (1) Extraction of medical and health-related tasks from WikiHow[2]; (2) Identification of relevant health-related tasks; (3) Expert label annotation for medical instructional videos. Therefore, we consider that the overall three-category (non-medical, medical instructional, and non-medical instructional) can be turned into a two-stage taxonomy. We first perform the binary classification of medical-related then process the binary classification of medical instructional-related.

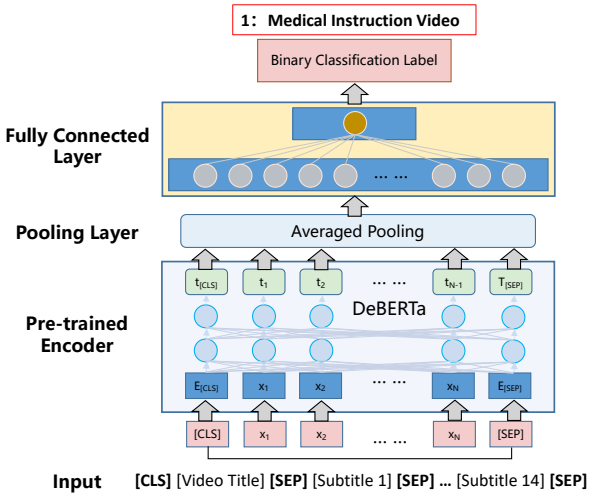

Figure 2: An example of the language-only classification model. Given the video title and its subtitles, it is required to perform the binary classification on the MedVidCL datasets.

2.2 Language-only video classification

Because the content of the video is directly related to its subtitles, it is feasible to use the corresponding subtitle texts to perform the classification of the input video (Miech et al., 2020). As shown in Figure 2, we concatenate the video title with the subtitles which are segmented into text spans for text encoding. Then the tokenized tokens are encoded through the DeBERTa model (He et al., 2020) for learning well-formed representations. An averaged pooling with the fully connected layer is designed to obtain the final features for the binary classification prediction.

2.3 Cross-modal video classification

When people watch videos, they may not always judge the video contents through the subtitle texts. For the non-audio parts, the visual information counts a lot (Gabeur et al., 2020). Therefore, for each subtitle span, we can add the visual feature to predict the video content. As shown in Figure 3, we design the cross-modal video classification model. Specifically, we focus on the feature joint alignment of video frames and subtitle text. The binary classification is performed after mapping the subtitle spans with their corresponding video frame into the same vector space. For the text modality, we input the subtitle texts into the pre-trained model

[2]https://www.wikihow.com/Main-Page

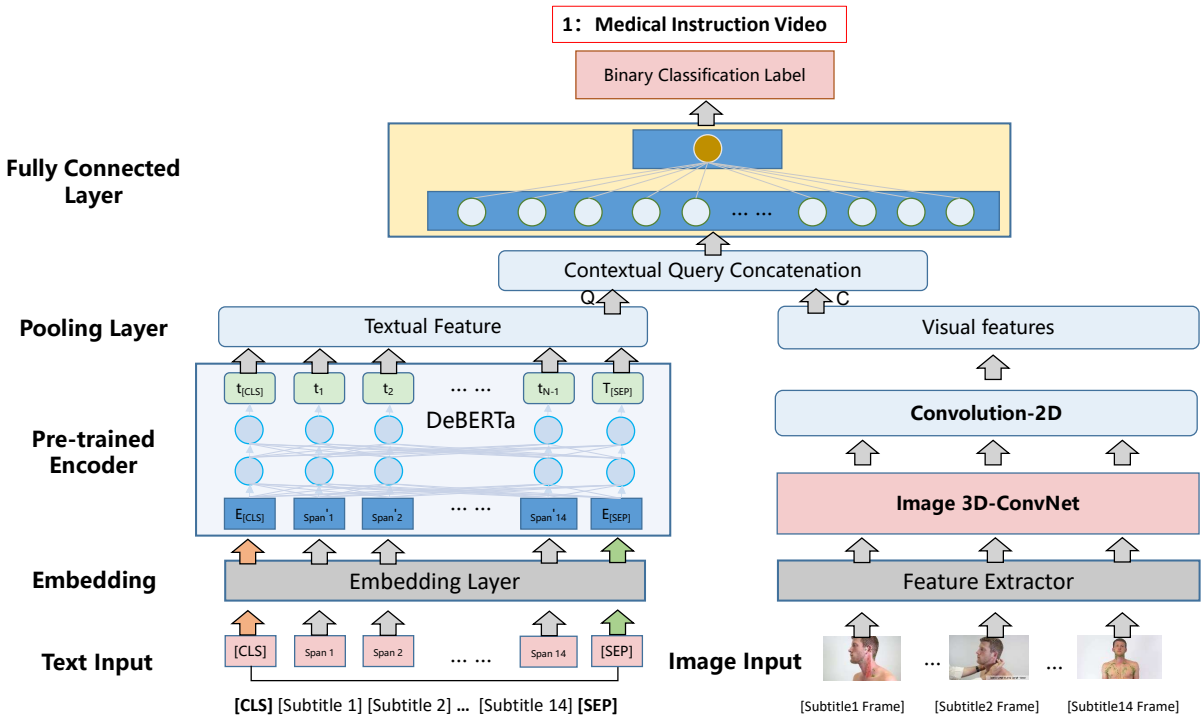

Figure 3: An example of the cross-modal classification model. The cross-modal features encoded separately are sent to the context query concatenation module for joint alignment. The binary classification is performed through the fully connected layer.

Video Category	Train	Validation	Test	Total
Medical Instructional	789	100	600	1489
Medical Non-instructional	2394	100	500	2994
Non-Medical	1034	100	500	1634
Total	4217	300	1600	6117

Table 1: Statistics of the medical video classification task dataset for the experiments.

for obtaining the textual feature. For the visual modality, we extract the raw frames with down-sampling, where 20 frames are derived from each video at a uniform time interval. Then we utilized a 3D ConvNet (I3D) module (Balaguer and Gobbetti, 1995) with the Convolution-2D for obtaining the visual features, which was pre-trained on the Kinetics dataset (Kay et al., 2017). We perform the Context Query Concatenation (Cq_Concat) (Zhang et al., 2020) for joint alignment of the textual features (Q) and the visual features (C) for the final binary classification prediction.

2.4 Late fusion method

Since there is a huge gap between visual features (Zhang et al., 2021) and language features, we design the late fusion method to use the Bagging algorithm (Breiman, 1996) to obtain the results of the above two models. Specifically, we use the logits from the different models for this ensemble method, where these logits are summed together before the softmax. We adopt the softmax to perform the final prediction. The Bagging algorithm is used during the prediction, which can effectively reduce the variance of the final prediction by bridging the prediction bias of different models, enhancing the overall generalization ability of the system.

3 Experimental setup

3.1 Data Description

Recently, with the rapid development of the video field, the informative nature of video has changed the way human beings obtain information (Lin et al., 2019). Medical Video CLassification (MedVidCL) (Gupta et al., 2022) is a data set about medical instructional video classification, which has been validated by human annotators. The medical classification datasets contain a collection of 6,617 videos, and it is required to classify the video into "Medical Instructional", "Medical Non-Instructional", and "Non-Medical" classes.

The statistics of the medical video classification datasets (seen datasets for experiments) are shown in Table 1. To construct the MedVidCL dataset, the

Experimental Items	Medical-related			Instructional-related			Overall
Method	F1	Precision	Recall	F1	Precision	Recall	F1(macro)
SVM	/	/	/	0.802	1.000	0.670	0.874
BERT-Base-Uncased	/	/	/	0.915	0.960	0.875	0.929
RoBERTa-Base	/	/	/	0.934	0.980	0.893	0.947
BigBird-Base	/	/	/	0.942	0.982	0.907	0.957
DeBERTa One-Stage	0.996	0.996	0.996	0.992	0.984	1.000	0.963
DeBERTa Two-Stage	0.980	0.980	0.980	0.936	**1.000**	0.880	0.934
BigBird One-Stage	0.996	0.996	0.996	0.994	0.996	0.990	0.983
BigBird Two-Stage	0.996	0.996	0.996	0.998	0.996	0.990	0.985
Ensemble	**0.996**	**0.996**	**0.996**	**0.998**	0.999	**1.000**	**0.988**

Table 2: Results of the monomodal with language on the seen test set.

Experimental Items	Medical-related			Instructional-related			Overall
Method	F1	Precision	Recall	F1	Precision	Recall	F1(macro)
L + V (I3D) + LSTM	/	/	/	0.726	0.797	0.667	0.757
L + V (ViT) + LSTM	/	/	/	0.773	0.902	0.677	0.814
L + V (I3D) + Transformer	/	/	/	0.727	0.762	0.695	0.748
L + V (ViT) + Transformer	/	/	/	0.791	0.922	0.692	0.824
Ours (One-Stage) + DeBERTa + I3D	0.990	0.988	0.992	0.984	0.969	1.000	0.967
Ours (Two-Stage) + DeBERTa + I3D	**0.998**	**1.000**	**0.996**	0.986	0.973	1.000	0.971
Ours (One-Stage) + BigBird + I3D	0.992	0.992	0.992	0.986	0.981	0.992	0.975
Ours (Two-Stage) + BigBird + I3D	0.992	0.988	0.995	0.973	0.947	1.000	0.977
Ensemble	0.994	0.994	0.994	**0.992**	**0.984**	**1.000**	**0.978**

Table 3: Results of the Multimodal with Language (L) and Vision (V) on the seen test set.

organizers first train the machine learning model based on the data marked by medical experts from HowTo100M and YouTube8M datasets (Abu-El-Haija et al., 2016). After that, the videos with high confidence are selected and sorted out with the machine learning method (Gupta et al., 2022).

3.2 Evaluation metrics

We follow the standard evaluation metrics of answer prediction in MedVidQA. The performance of the system is evaluated through two evaluation indicators (Gupta and Demner-Fushman, 2022). Each experiment was conducted for 10 rounds with different random seeds for eliminating the random bias, and we select the model with the highest F1 score on the valid set and then report its score on the test set. The metrics are introduced as follows.

1. F1 Score on Medical Instructional class.
2. Average macro-level F1 score across all the classes.

The calculation equation of each metric is shown as follows.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * PrecisionRecall}{Precision + Recall}$$

$$MacroF1 = \frac{\sum_{i=1}^{n} F1_i}{n}$$

3.3 Experimental Details

The Method provided by the organizer[3] provides four baseline methods based on different features. In order to obtain language features, the organizer extracted the subtitle information in the video using the Pytube library[4]. In addition, the organizer uses 3D convolution (I3D) (Carreira and Zisserman, 2017) to extract the visual features of the video in units of every second. And then they train statistical classifiers such SVM (Cortes and Vapnik, 1995) method, Transformer model (Vaswani et al., 2017), VIT model (Dosovitskiy et al., 2021) and LSTM model (Hochreiter and Schmidhuber, 1997). The details of each baseline are introduced below.

[3]Specific implementations can refer to https://github.com/deepaknlp/MedVidQACL/tree/master/MedVidCL

[4]https://github.com/pytube/pytube

Method	Med-Inst Precision	Med-Inst Recall	Med-Inst F1	Macro F1
BigBird Two-Stage (Monomodal)	0.9949	0.9775	0.9861	0.9893
Ours (Two-Stage) + DeBERTa + I3D	0.9948	0.9750	0.9848	0.9884
Ensemble	**0.9974**	**0.9775**	**0.9873**	**0.9901**

Table 4: Submitted official results of the unseen test set.

1. **Monomodal (Language)** They utilize the pre-trained Transformer models from Hugging Face(Wolf et al., 2020) such as BERT-Base-Uncase(Devlin et al., 2019) , RoBERTa-Base(Liu et al., 2019) and BigBird-Base(Zaheer et al., 2020).

2. **Monomodal (Vision)** After extracting features from I3D or VIT, the organizer uses the LSTM network and transformers network to build classifiers.

3. **Multimodal (Language + Vision)** After the text features and visual features are obtained, they are concatenated and then connected to a full connection layer for classification.

Our method use the DeBERTa-large-v3 (He et al., 2021) model. The DeBERTa improves the BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models using disentangled attention and enhanced mask decoder. It shares the base model with 24 layers and 1024 hidden size. We formulate the original three-classification task (one-stage) into two two-classification tasks (two-stage). And we trained two models separately to support our video classification under the two-stage setting. In the experimental table, we will report and compare the testing effect in one-stage and two-stage settings respectively.

3.4 Implementation details

We train the model based on the Pytorch framework (Paszke et al., 2019) and use the huggingface[5] (Wolf et al., 2020) framework. When training the model, we employ the AdamW optimizer (Loshchilov and Hutter, 2017). The default learning rate is set to 1e-5 with the warm-up (He et al., 2016). Four RTX3090 GPUs with 24G memory are implemented for all experiments.

We use the [SEP] token to concatenate the title and subtitles of the video. Experiments are carried out in maximum lengths 512. When it is necessary to distinguish whether it is a medical video or not at the first stage, we set the "Medical Non-instructional" video and the "Medical Instructional" video as the same category. When turning into the second stage, we exclude "Non-medical" video samples. All the experimental codes are open-source at https://github.com/Lireanstar/MedVidCL.

[5]https://github.com/huggingface/transformers

4 Results and discussions

In this section, we introduce the experimental results of the monomodal in language and the multimodal in language-version where the further discussions and official results are also presented.

4.1 Experimental results

The experimental results of the monomodal with language on the seen test set are shown in Table 2, and the multimodal with language and vision can be found in Table 3. For the one-stage setting, we implement the classification for three categories of prediction. For the proposed two-stage setting, we exclude the non-medical category for the first stage classification, then perform the two-categories classification to differentiate the medical instructional and medical non-instructional videos for the final result. Specifically, for the monomodal results, our method outperforms all the baselines in the overall scores. What excites us is that the SVM method achieves the same recall score (1.000) as the deep learning DeBERTa model on medical instructional-related classification, indicating that the subtitle information of the video has strong semantics. As for the multimodal settings, the proposed two-stage cross-modal fusion method outperforms the one-stage cross-modal fusion method, which demonstrates its effectiveness. The proposed method is significantly ahead of the baseline methods. We believe that is because our model can recognize visual features more efficiently combined with the strong pre-trained language model. Moreover, the ensemble method can be a wise choice to enhance the final score compared with other single models. In the end, we find that the proposed method with cross-modal fusion can achieve similar perfor-

mance to the monomodal methods, which demonstrates the superiority of our proposed method.

4.2 Official results

As shown in Table 4, we present the results of official submissions on the unseen test set. Further conclusions can be found that the monomodal modality (language) can indeed effectively identify the semantic information from the video, which outperforms the cross-modal setting. It is in line with the experimental results under the test set. We perform the ensemble method by adding the logits generated from the two single model, and adopt the Softmax for the final prediction. Finally, by adopting an ensemble method, we achieve the Top-1 score in the final official stage.

5 Conclusion

This paper introduces our approach to solving the medical video classification (MedVidCL) task in BioNLP of the ACL2022. Specifically, we propose the two-stage method with cross-modal fusion using the pre-trained language model. We report the performance of our model compared on the test set in monomodal and multimodal settings. The experimental results show that our method obtains the best performance on the seen test set and unseen official test set, which proves that our method is effective. Also, experimental results show that language understanding is better than multimodal video understanding. In the future, we will further study how to design a more efficient structure to jointly learn the representation in visual language for better multimodal video understanding.

Acknowledgement

This work is supported by the National Key Research and Development Project of China (2018YFB1305200) and the National Natural Science Fund of China (62171183, 61801178).

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Jean-Francis Balaguer and Enrico Gobbetti. 1995. i3d: a high-speed 3d web browser. In *Proceedings of the first symposium on Virtual reality modeling language*, pages 69–76.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition*.

Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8175–8182.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20:273–297.

Mutlu Cukurova, Carmel Kent, and Rosemary Luckin. 2019. Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring. *British Journal of Educational Technology*, 50(6):3032–3046.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *ECCV 2020-European Conference on Computer Vision*, volume 12349, pages 214–229. Springer.

Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. 2016. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2022. A dataset for medical instructional video classification and question answering. *arXiv preprint arXiv:2201.12888*.

Deepak Gupta and Dina Demner-Fushman. 2022. Overview of the MedVidQA 2022 Shared Task on Medical Video Question Answering. In *Proceedings of the 21st SIGBioMed Workshop on Biomedical Language Processing, ACL-BioNLP 2022*. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11021–11028.

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

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020a. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020b. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer.

Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. 2018. Fine-grained video classification and captioning. *arXiv preprint arXiv:1804.09235*, 5(6).

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jesus Perez-Martin, Benjamin Bustos, Silvio Jamil F Guimarães, Ivan Sipiran, Jorge Pérez, and Grethel Coello Said. 2021. Bridging vision and language from the video-to-text perspective: A comprehensive review. *arXiv preprint arXiv:2103.14785*.

Jesus Perez-Martin, Benjamin Bustos, Silvio Jamil F Guimarães, Ivan Sipiran, Jorge Pérez, and Grethel Coello Said. 2022. A comprehensive review of the video-to-text problem. *Artificial Intelligence Review*, pages 1–75.

Myriam Servières, Valérie Renaudin, Alexis Dupuis, and Nicolas Antigny. 2021. Visual and visual-inertial slam: State of the art, classification, and experimental benchmarking. *Journal of Sensors*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le

Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *arXiv: Learning*.

Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence*.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554.