A Dual-level Detection Method for Video Copy Detection

Tianyi Wang[1][*], Feipeng Ma[1,2][*], Zhenhua Liu[1][*], Fengyun Rao[1][†]

[1]WeChat of Tencent, [2]University of Science and Technology of China

tyewang@tencent.com, mafp@mail.ustc.edu.cn, edinliu@tencent.com, fengyunrao@tencent.com

Abstract

With the development of multimedia technology, Video Copy Detection has been a crucial problem for social media platforms. Meta AI hold Video Similarity Challenge on CVPR 2023 to push the technology forward. In this paper, we share our winner solutions on both tracks to help progress in this area. For Descriptor Track, we propose a dual-level detection method with Video Editing Detection (VED) and Frame Scenes Detection (FSD) to tackle the core challenges on Video Copy Detection. Experimental results demonstrate the effectiveness and efficiency of our proposed method. Code is available at https://github.com/ FeipengMa6/VSC22-Submission.

1. Introduction

In the past decade, the development of information technology has led to a shift in the main carrier of information from text to images and then to videos. Moreover, with the rise of User-generated Content (UGC), the producer of information has shifted from Occupationally-generated Content (OGC) to UGC. As a result, a large number of videos have emerged on social media platforms and have been widely shared, leading to the increasingly important and challenging problems of video copyright protection. The core challenges of video copy detection are twofold: effective video descriptors and computational costs. Copied videos often involve edited portions, making it difficult for general visual models to differentiate between copied and original content. A powerful model must be capable of discriminating between videos, even when significant editing has taken place. Additionally, the cost of time and resources required to process each query video and identify the most similar reference video is a significant concern, necessitating the development of efficient and cost-effective methods.

In this paper, we summarize our proposed method for Meta AI Video Similarity Challenge, which tackle the core

---
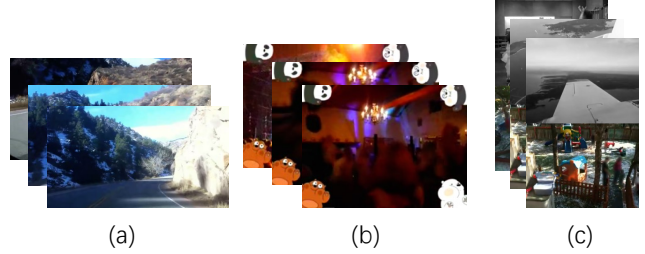
[*]Equal contribution.
[†]Corresponding author.

Figure 1. Three typical situations of query videos. (a) is an unedited video, which is the most of query videos. (b) is a copied video with general editing operation. (c) is a copied video with multiple scenes in each frame.

challenges of video copy detection through a dual-level detection method. In Fig. 1, there are three typical situations of query videos, including unedited video, copied video with general editing operation and copied video with multiple scenes in each frame. Our proposed dual-level detection method first identify if the video has been edited in video-level. For unedited videos, we use random vectors with small norm as their descriptors. What's more, we replace the bias term of these descriptors with a negative value during score normalization. For edited videos, we notice that it is necessary to deal with the situation that multiple scenes are concatenated along edge. We adopt traditional image processing method to detect and split the scenes in one frame. With our dual-level detection method, we can reduce the storage cost for unedited videos and improve the efficiency and accuracy of retrieval.

The main contributions are summarized as follows:

- We propose a dual-level detection method for Descriptor Track, which detects edited videos at video-level and multiple scenes at frame-level. With the dual-level detection, we can reduce the computational cost and improve the performance.

- The proposed method achieve outstanding performance on Meta AI Video Similarity Challenge and we got second prize on Descriptor Track. Our ablation study shows the effective of each module.
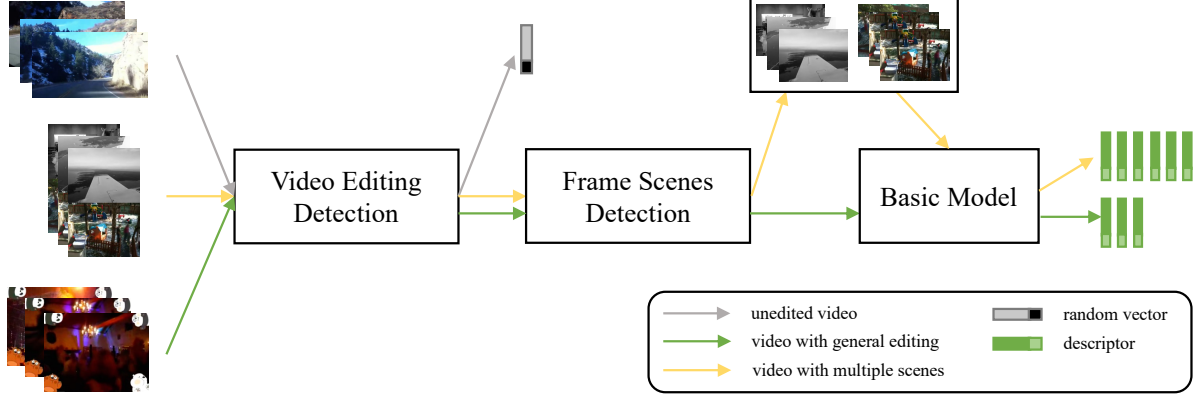
Figure 2. Overview of our proposed pipeline. For unedited video, we directly use a random vector with samll norm and negative bias term. For video with general editing, we extract features of each frame by our pre-trained basic model as descriptors. For video with multiple scenes, we detect and split the scenes in each frame, then use basic model to generate descriptors.

2. Method

In this section, we first introduce the design of our basic model. Then we explain the details of our proposed dual-detection method including video editing detection, frame scenes detection and score filter normalization.

2.1. Basic Model

We train a basic model to extract descriptors for video copy detection. There are two potential types of descriptors that can be employed in video copy detection: video-level features or frame-level features. Given that our objective is not only to discriminate copied videos but also to identify copied portions between query and reference videos, we selected frame-level features as the video descriptor. Therefore, we adopt image transformer [4, 7] as backbone. We follow SSCD [8] to train our basic model in a self-supervised manner. As SSCD uses, we combine SimCLR [1] method with entropy loss [10].

The InfoNCE Loss. We use the InfoNCE loss in SimCLR, which is softmax cross-entropy loss with temprature. The loss function is formulated as follow:

$$L_{\text{InfoNCE}} = -\frac{1}{|P|} \sum_{i,j \in P} \log \frac{\exp(\cos(z_i, z_j)/\tau)}{\Sigma_{k \neq i} \exp(\cos(z_i, z_k)/\tau)}$$

(1)

Where P is the set of positive pairs, z_i represents the descriptor, τ is the temperature.

The Entropy Loss. We follow SSCD using the entropy loss proposed in [10]. The loss function is formulated as:

$$L_{\text{KoLeo}} = -\frac{1}{N} \sum_{i=1}^{N} \log(\min_{j \neq i} ||z_i - z_j||)$$

(2)

Where N is the size of training set.

The Final Loss. The final loss function is:

$$L = L_{\text{InfoNCE}} + \lambda L_{\text{KoLeo}}$$

(3)

Where λ is the weights of Entropy Loss term. Cause our training process of basic model is based on SSCD [8], more details can be found in this paper.

2.2. Video Editing Detection

To address the issue of high computational costs in video copy detection and provide an efficient solution, we propose a straightforward method to identify edited videos before generating frame-level descriptors. We observe that videos with copies are often edited, incorporating techniques such as blending, blurring, rotations, and other manipulations. This is due to the fact that copied videos must frequently merge multiple clips, necessitating additional editing operations. By filtering out videos that have not been edited among the query videos, we can reduce computational costs. To achieve this goal, we have developed a model capable of discriminating between edited and unedited videos. Since editing operations can be viewed as strong augmentations, we aim to identify videos with such augmentations using a binary classification approach. We utilize CLIP [9] to extract frame features without any post-processing and feed these features into RoBERTa [6]. And we employ the class token to calculate cross entropy. We find that the edited video detection can achieve high accuracy, and using a small value α as threshold can filter most of unedited videos. For the unedited video, we use a random vector with very small value as descriptor. This processing can reduce the storage cost for query videos and speed up searching.

2.3. Frame Scenes Detection

We notice that stacking multiple scenes in one frames is an obvious augmentation of copied videos, and simple traditional image processing method can deal with this situation

well. Due to the continuity of the video, the combination of multiple scenes in one frames are limited. As shown in Fig. 3, scenes are usually concatenated along one side and one frame usually has an even number of scenes, most of them have two or four scenes.
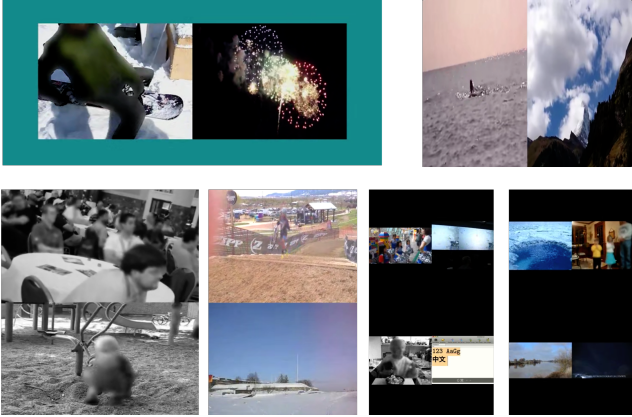

Figure 3. Multiple scenes in one frame.

2.4. Score Filter Normalization

We follow [5, 8] using similarity normalization in our evaluation. It introduce a background image dataset and only queries whose similarity score with reference is much higher than images in background dataset will have high scores. Based on it, we modify the integrated bias to suppress the score of unedited videos. In Sec. 2.2, we use a random vector with very small value as descriptor for unedited video, but scores of these videos are clustered around 0. Because scores of hard positive pairs are clustered around 0 too, we should further suppress the score of unedited videos. Inspired by the integrated bias term of similarity normalization, we can replace it by a negative value and the similarity score with any reference videos will reduce to a negative value.

3. Experiments

3.1. Implementation Details

In video editing detection, we adopt ViT-L-16 of CLIP to extract frame features. The initial weights for RoBERTa is chinese-roberta-wwm-ext [2, 3] in huggingface.

3.2. Results

Our proposed method achieve outstanding performance on Meta AI Video Similarity Challenge. The results of Phase 1 and Phase 2 on Descriptor Tracks are presented in Tab. 1. On Descriptor Track, our method got second place Phases 1, just 0.021 away from the first place. Although we got the first place on Phase 2, we notice that the performance drop a lot when transfer the model to Phase 2. The

reason is that we only ensemble 4 models and our ensemble results are not much better than single model. Without a strong ensemble method, the transfer ability is limited.

User or teams	Phase 1 μAP	Phase 2 μAP
do something(Ours)	<u>0.9176</u>	**0.8717**
FriendshipFirst	**0.9197**	0.8514
cvl-descriptor	0.8534	0.8362
Zihao	0.7841	0.7729

Table 1. Leaderboard results on Descriptor Track. **Bold** indicates the best result and <u>underline</u> indicates the second best result.

4. Ablation Study

To validate the effective of our proposed method, we split validation set by ourselves and analyze each module on validation set. At the beginning of the competition, we randomly divide queries in training set into 8:2 as offline training set and validation set. And the trend of performance on validation set can reflect the trend on test set. We use single basic model in ablation study because our ensemble method do not improve much. The results are shown in Tab. 2, our basic model can achieve 0.8580 on μAP, it shows that our basic model is a very strong baseline for video copy detection. Then combining frame scenes detection with basic model, the performance increased by 5%. And with the video editing detection and frame scenes detection, the performance achieve 0.9492.

Method	μAP
Basic model	0.8580
+ FSD	0.9075
+ VED	0.9492

Table 2. Ablation study.

5. Conclusion

We introduce a dual-detection method for Video Copy Detection in this paper. The video editing detection in video-level can identify unedited videos and use random vectors with small norm and negative bias term as descriptors. The frame scenes detection in frame-level can detect scenes and split them into multiple videos, where video only have one scenes in each frame. Thought the dual-detection method, we got second place on the Descriptor Track of Meta AI Video Similarity Challenge 2022. And our descriptor give strong support to our first-place solution on Matching Track.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[2] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online, Nov. 2020. Association for Computational Linguistics. 3

[3] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*, 2019. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Matthijs Douze, Giorgos Tolias, Ed Pizzi, Zoë Papakipos, Lowik Chanussot, Filip Radenovic, Tomas Jenicek, Maxim Maximov, Laura Leal-Taixé, Ismail Elezi, et al. The 2021 image similarity dataset and challenge. *arXiv preprint arXiv:2106.09672*, 2021. 3

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[7] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[8] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022. 2, 3

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[10] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018. 2