Prompting 기반 매개변수 효율적인 멀티 모달 영상 하이라이트 검출 연구

한동훈[1○], 남성욱[1], 박은환[1], 곽노준[2*]
버즈니AI 연구소, 서울대학교
{owen, zaid, jude}@buzzni.com
nojunk@snu.ac.kr

Parameter-Efficient Multi-Modal Highlight Detection via Prompting

DongHoon Han○, Eunhwan Park, Seong-Uk Nam, Nojun Kwak
Buzzni AI Lab
Seoul National University

요 약

본 연구에서는 비디오 하이라이트 검출 및 장면 추출을 위한 경량화된 모델인 Visual Context Learner (VCL)을 제안한다. 기존 연구에서는 매개변수가 고정된 CLIP을 비롯한 여러 피쳐 추출기에 학습 가능한 DETR과 같은 트랜스포머를 이어붙여서 학습을 한다. 하지만 본 연구는 경량화된 구조로 하이라이트 검출 성능을 개선시킬 수 있음을 보인다. 그리고 해당 형태로 장면 추출도 가능함을 보이며 장면 추출의 추가 연구 가능성을 시사한다. VCL은 매개변수가 고정된 CLIP에 학습가능한 프롬프트와 MLP로 하이라이트 검출과 장면 추출을 진행한다. 총 2,141개의 학습가능한 매개변수를 사용하여 하이라이트 검출의 HIT@1(>=Very Good) 성능을 기존 CLIP보다 2.71% 개선된 성능과 최소한의 장면 추출 성능을 보인다.

주제어: 하이라이트 검출, 장면 추출, 비디오 하이라이트, 멀티모달, 트랜스포머

1. 서론

사전학습 된 멀티모달 모델 [1, 2, 3, 4, 5, 6]들은 여러 모달리티 간의 학습 공간을 공유하는 것이 가능해지면서 이미지, 텍스트, 오디오, 비디오 모달리티를 동시에 사용하는 범용적인 문제 해결이 용이해졌다. 이에 따라, 비디오 분야도 멀티모달 기반 사전학습 모델을 활용하여 과거보다 다양한 과업를 우수하게 수행하는 것이 가능해졌으며 그에 따라 멀티모달 비디오-자연어 이해를 기반으로한 하이라이트 검출 (Highlight Detection)/장면 추출 (Moment Retrieval)에 대한 관심도 높아지고 있다.

하이라이트 검출(Highlight Detection)/장면 추출 (Moment Retrieval)은 비디오의 프레임들과 주어진 텍스트 간의 적합도 점수 (saliency score)를 정확하게 유추하고 해당 텍스트에 해당하는 장면 구간을 뽑아내는 과업이다. 하이라이트 검출/장면 추출 시에도 멀티모달 사전학습의 이점을 활용하기 위해 사전학습된 CLIP [1]으로 비디오 프레임과 텍스트의 피쳐를 뽑아낸 후, 입력값으로 사용한다. 하지만 기존의 하이라이트 검출/장면 추출을 수행하는 모델 [7, 8, 9]들이 CLIP 피쳐와 다른 피쳐 추출기의 피쳐 [10, 11]를 동시에 입력값으로 받아서 새로운 매개변수를 학습 하는 방식은 CLIP의 특징공간(feature space)를 온전히 활용하지 못하게 한다. 최근 멀티모달의 특징공간을 온전하게 활용하기 위해서 사전학습된 멀티모달 모델의 매개변수를 고정한 채, 멀티모달 피쳐 자체를 활용하는 비디오 분야의 과업를 진행하는 연구들을 찾아볼 수 있다 [12, 13].

본 연구에서는 하이라이트 검출/장면 추출을 수행할 때에도 CLIP을 입력 데이터로 사용하기보단 멀티모달 피쳐 자체로 사용할 때, 더 효율적이고 효과적임을 보이고자 한다. 본 연구에서 제안하는 Visual Context Learner (VCL)은 별도의 학습가능한 매개변수를 최소화한 채 CLIP만을 비디오와 텍스트의 피쳐 추출기로 활용하여 하이라이트 검출 부문에서 기존 연구에 필적할 성능을 내고, 굉장히 간단한 구조로 트랜스포머 기반의 검출기 없이도 장면 추출이 가능함을 보인다. 본 연구는 CoOp [14]에서 영향을 받아, CLIP의 모든 매개변수를 고정하고 텍스트 입력 값에 학습가능한 프롬프트만을 삽입하여 하이라이트 검출을 수행한다. 그리고 이 과정에서 나온 프레임과 텍스트 간 적합도 점수(saliency score)를 가지고 MLP를 사용하여 준수한 장면 추출 성능을 보인다.

2. 관련 연구

2.1 프롬프트 학습

Prompting 은 추가적인 가상 토큰을 구성하는 매개변수를 추가한다. 그리고 전체 매개변수가 아닌 가상 토큰 매개변수만을 학습하는 경량화 미세 조정을 제안하였고 이는 도메인 이동 (Domain Shift) 을 해소하고 더 나아가 과적합 방지에 효과적임을 보였다 [15]. [16, 17] 은 이산 레이블 공간을 MLM Head 의 토큰으로 변환하는 함수 $\mathsf{V}(\cdot)$ ($\mathsf{V}(0)$ = "bad", $\mathsf{V}(1)$ = "amazing")와 입력 X를 사전 정의한 템플릿 함수 $\mathsf{T}(\cdot)$ ($\mathsf{T}(\mathsf{X})$ = "X It was [MASK] .")를 통해 변환한다. 자연어처리 분야에서 활발히 연구가 진행된 Prompting 은 CLIP과 같은 이미지 – 텍스트를 동시에 다루는 멀티 모달 모델에도 적용되어 좋은 성능을 보여주었다 [18, 19].

*교신저자 (Corresponding author)

2.2 하이라이트 검출 (Highlight Detection)/장면 추출 (Moment Retrieval)

하이라이트 검출은 비디오와 이에 상응하는 텍스트가 주어졌다고 가정했을 때, 비디오 프레임 중 텍스트와 가장 적합성이 높다고 판단되는 프레임을 유추하는 연구이다. 또한 장면 추출은 마찬가지고 비디오와 이에 상응하는 텍스트가 주어졌을 때, 해당 텍스트와 상응하는 영역을 구한는 연구이다.

해당 과업를 해결하기 위해 최근 연구 [7, 8, 9]들에서는 DETR [20]의 구조를 차용한 형태를 띈다. CLIP에서 추출한 피디오 프레임의 피쳐와 텍스트의 피쳐를 모델의 입력 값으로 사용하여 복수 층의 트랜스포머 [21] 인코더와 디코더를 학습하는 구조이다. 이때, 트랜스포머 인코더의 반환 값은 하이라이트 검출을 위해 사용되고, 트랜스포머 디코더는 N개의 학습 가능한 쿼리를 입력으로 넣어서 텍스트와 유관한 영역을 최대 N개까지 예측할 수 있도록 학습한다. 본 연구에서는 기존 연구들에서 사용하던 트랜스포머 및 학습 가능한 매개변수를 거의 사용하지 않고 경쟁적인 성능을 보인다.

3. 프롬프트 기반의 하이라이트 검출 성능 개선 모델

하이라이트 검출 시 총 N개의 비디오 프레임 ($[\mathsf{f}_1, \mathsf{f}_2, \cdots, \mathsf{f}_N]$)와 입력 토큰 $\mathsf{t} = [\mathsf{x}_1, \cdots, \mathsf{x}_\mathsf{m}]$가 주어진다. 하이라이트 검출 시에는 N개의 프레임 중 주어진 텍스트 t와 가장 연관성이 높은 프레임 f_k를 선택하는 것이 목표이다.

장면 검출 시에는 입력 토큰 t와 부합하는 장면 구간 (t_{start}, t_{end})를 최대한 가깝게 유추하는 것이 목표이다.

3.1 프롬프트 기반 CLIP 인코더 학습

$$\boldsymbol{E}_i = \left\{ [\mathsf{e}_{\mathsf{i},\mathsf{k}_1}], \cdots, [\mathsf{e}_{\mathsf{i},\mathsf{k}_\mathsf{n}}], [\mathsf{e}_{\mathsf{i},\mathsf{x}_1}], \cdots [\mathsf{e}_{\mathsf{i},\mathsf{x}_\mathsf{m}}] \right\} \quad (1)$$

그림 1 중 하단은 하이라이트 검출을 하는 부분에 대한 설명이다. 멀티모달 사전학습 모델의 파괴적 망각(catastrophic forgetting)을 막는 동시에 학습의 효율을 최대화 하기 위해 CLIP의 이미지 인코더와 텍스트 인코더의 매개변수는 모두 고정을 하였다. 그리고 비디오의 텍스트 앞에 수식 (1) 학습 가능한 매개변수(k차원) N개를 붙여서 해당 부분만 학습을 진행하였다. 이 때 N개의 학습 가능한 프롬프트는 기존의 비디오에 대한 텍스트와 연결되어 CLIP의 텍스트 인코더를 통과하게 된다.

$$\mathsf{s}_{\mathsf{saliency}} = \{[\mathsf{ImgEnc}(\mathsf{f}_1) \cdot \mathsf{TxtEnc}(\mathsf{t})], \cdots, [\mathsf{ImgEnc}(\mathsf{f}_\mathsf{N}) \cdot \mathsf{TxtEnc}(\mathsf{t})]\} \quad (2)$$

그리고 비디오의 각 프레임들은 CLIP의 이미지 인코더를 통과하게 된다. 수식 (2)에서와 같이 k차원으로 된 각각의 프레임의 피쳐는 텍스트의 피쳐와 순차적으로 유사도 계산을 진행하게 되며, 이는 프레임별 적합도 정답 점수에 근사하도록 학습이 된다. 이 과정에서 N개의 학습 가능한 프롬프트는 수식 (3)와 같이 최대의 성능을 낼 수 있는 방향으로 학습이 되게 된다. 여기서 나온 프레임 별 텍스트와의 유사도로 하이라이트 검출을 진행하게 된다.

$$\mathsf{highlight} = \arg\max(\mathsf{s}_{\mathsf{saliency}}) \quad (3)$$

표 1. QVHighlights 데이터셋의 *val split*으로 Highlight Detection (>=Very Good) 성능을 비교한 결과.

Method	mAP	HIT@1
Moment-DETR	36.52	56.45
QD-DETR	39.13	63.03
UMT	**39.85**	64.19
CLIP	37.01	63.03
OURS (7218 shot)	38.84	**65.74**
OURS (3000 shot)	38.47	65.35
OURS (1000 shot)	38.14	64.12
OURS (500 shot)	37.79	62.65

3.2 적합도 점수 기반 장면 추출 모듈

그림 1 중 상단은 장면 추출을 진행하는 부분에 대한 설명이다. 이 전 단계에서 구한 비디오 프레임 별 텍스트와의 유사도 정보를 MLP의 입력값으로 사용하여 각 프레임이 텍스트에 부합하는 장면인지를 판단한다. 그리고 부합한 연속된 프레임들을 연결하여 장면 추출을 위한 예측 값으로 사용한다.

3.3 학습

VCL은 QVHighlights 데이터셋을 통해 학습이 되었다. VCL을 학습하기 위해 총 3개의 손실함수가 사용되었다. 첫 번째 손실함수는 프레임 별 텍스트 피쳐 유사도 점수와 프레임 별 적합도 점수(saliency score) 라벨의 평균제곱오차함수이다. 두 번째로 사용된 손실함수는 장면 추출을 위한 Generalized IoU [22] 손실함수를 사용하였다. 세 번째로 사용된 손실함수는 DETR [20]에서 사용된 것과 같은 방식으로, 모델의 장면 추출 예측 값과 정답 값으로 이분 매칭(bipartite matching) 알고리즘을 사용하였다.

4. 실험

표 1는 *val split*에서 VCL을 전체 데이터셋과 데이터셋의 일부만을 가지고 프롬프트 튜닝을 진행한 결과이다. 전체 학습 데이터 (7,218개 샘플)를 사용했을 때는 하이라이트 검출에서 mAP가 38.84, HIT@1이 65.74가 나왔으며, 절반 이하의 학습 데이터인 3,000개의 샘플을 사용하여 학습을 하였을 때는 mAP

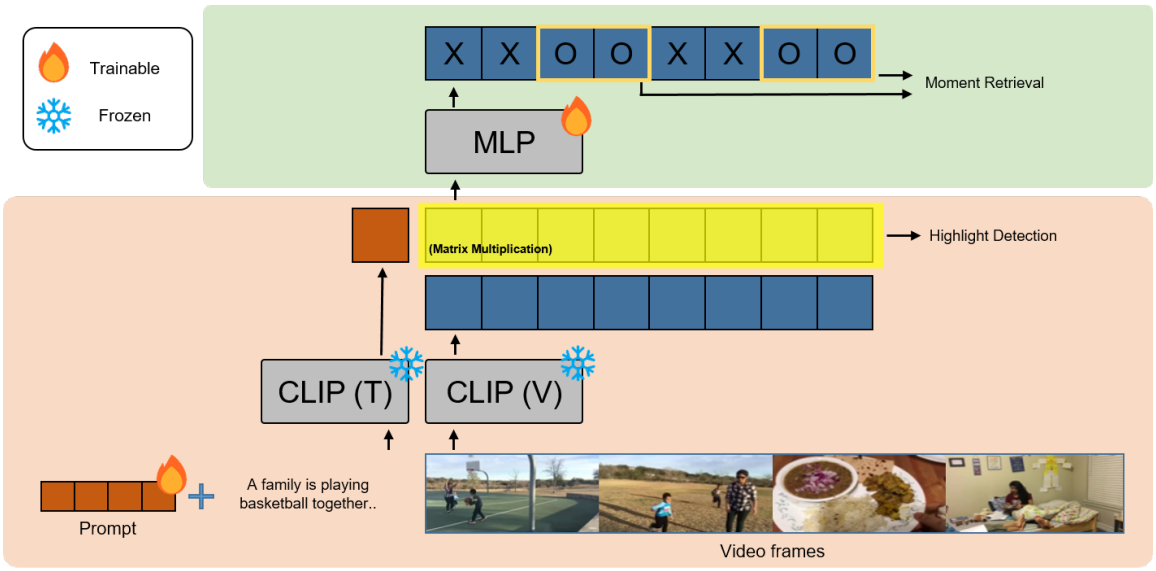

그림 1. VCL의 구조. 일부 매개변수만 학습가능(Trainable)한 상태이다.

표 2. QVHighlights 데이터셋의 *val split*으로 성능을 비교한 결과.

Method	Moment Retrieval					Highlight Detection	
	R1		mAP			>= Very Good	
	@0.5	@0.7	@0.5	@0.75	avg	mAP	HIT@1
CLIP	-	-	-	-	-	37.01	63.03
Moment-DETR	53.94	34.84	-	-	32.20	35.65	55.55
Moment-DETR PT	59.68	40.84	-	-	36.30	37.70	60.32
QD-DETR	**62.68**	**46.66**	**62.23**	**41.82**	**41.22**	39.13	63.03
UMT	60.26	44.26	-	-	38.59	**39.85**	64.19
VCL	43.33	25.75	39.23	24.95	21.13	38.845	**65.74**

가 38.47, HIT@1이 65.35가 나왔다. 추가적으로 학습 데이터를 줄여가며 학습을 진행했을 때, 점진적으로 성능 저하가 일어나는 것을 확인하였다. 표 2는 VCL과 기존의 연구의 성능을 비교한 표이다. QVHighlights [7] 데이터셋의 *val split*에서의 하이라이트 검출/장면 추출 성능을 사용하여 비교하였다.

해당 실험을 통해 학습에 필요한 매개변수를 대폭 줄였음에도 불구하고 하이라이트 검출에서 기존 연구들과 비교했을 때 경쟁력이 있는 결과를 얻었음을 확인할 수 있다. 또한 소수의 데이터 집합으로 학습을 했을 때도 학습 성능이 어느정도 유지가 되는 것을 통해서 적은 자원에서도 강건하게 작동함을 확인할 수 있었다. 이러한 결과는 추후에 확장 연구를 통해 개선의 여지가 깊음을 의미한다.

5. 한계

장면 추출을 할 때, 간단한 MLP를 사용하여 각 프레임이 텍스트와 관련이 있는 장면인지를 이진 분류한다. 하지만 해당 모듈은 수용 영역(receptive field)이 좁고, 여러 상황에서 일반화 성능이 탁월하지 않아 성능의 편차가 생기는 모습을 보인다. 해당 지점을 개선하여 하아라이트 검출 성능의 이점을 최대한으로 활용할 수 있는 장면 추출 모듈이 필요하다.

6. 결론

하이라이트 검출과 장면 추출은 서로 연관이 있는 과업이다. 하이라이트 검출은 각 프레임과 텍스트에 대한 연관 정도를 잘 표현할수록 좋은 성능을 보인다. 그리고 하이라이트 검출 성능이 좋다면 이를 기반으로 장면 추출에서 좋은 성능을 만들 수 있는 여지가 생긴다. VCL은 적은 매개변수를 가지고 하이라이트 검출에서 경쟁력있는 결과를 얻었으며, 소수의 데이터 집합에서도 성능 보전이 이루어진다. 이러한 장점을 살리고 이에 맞는 장면 추출 모듈을 고안하는 방향으로 추가적인 개선이 가능할 것으로 보인다.

참고문헌

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natu-

ral language supervision," *International conference on machine learning*, pp. 8748–8763, 2021.

[2] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *International Conference on Machine Learning*, pp. 12 888–12 900, 2022.

[3] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[4] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, Vol. 34, pp. 9694–9705, 2021.

[5] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *Advances in Neural Information Processing Systems*, Vol. 34, pp. 24 206–24 221, 2021.

[6] J. Jang, C. Kong, D. Jeon, S. Kim, and N. Kwak, "Unifying vision-language representation space with single-tower transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 1, pp. 980–988, 2023.

[7] J. Lei, T. L. Berg, and M. Bansal, "Detecting moments and highlights in videos via natural language queries," *Advances in Neural Information Processing Systems*, Vol. 34, pp. 11 846–11 858, 2021.

[8] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query-dependent video representation for moment retrieval and highlight detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23 023–23 033, 2023.

[9] Y. Liu, S. Li, Y. Wu, C.-W. Chen, Y. Shan, and X. Qie, "Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3042–3051, 2022.

[10] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] Y. Jo, S. Lee, A. S. Lee, H. Lee, H. Oh, and M. Seo, "Zero-shot dense video captioning by jointly optimizing text and moment," *arXiv preprint arXiv:2307.02682*, 2023.

[13] Z. Lin, S. Geng, R. Zhang, P. Gao, G. de Melo, X. Wang, J. Dai, Y. Qiao, and H. Li, "Frozen clip models are efficient video learners," *European Conference on Computer Vision*, pp. 388–404, 2022.

[14] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, Vol. 130, No. 9, pp. 2337–2348, 2022.

[15] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.emnlp-main.243

[16] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 255–269, 2021. [Online]. Available: https://aclanthology.org/2021.eacl-main.20

[17] T. Schick and H. Schütze, "It's not just size that matters: Small language models are also few-shot learners," pp. 2339–2352, 6 2021. [Online]. Available: https://aclanthology.org/2021.naacl-main.185

[18] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision (IJCV)*, 2022.

[19] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *European conference on computer vision*, pp. 213–229, 2020.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, Vol. 30, 2017.

[22] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.