피처레벨 비디오 분석과, 적응적 장면 선택을 이용한 비디오 캡셔닝 피처 생성

Ju-Hee Lee, *Je-Won Kang

Ewha Womans University * Ewha Womans University

juhee69@ewhain.net, * jewonk@ewha.ac.kr

요 약

본 논문에서는 비디오의 피처레벨 분석을 통해 비디오의 장면 구성 특징을 파악하고, 그에 적응적으로 대표 프레임을 선택하는 방법을 제안한다. 제안된 방법으로 생성된 캡셔닝 피처는 비디오를 잘 요약하고, 이를 통해 효과적인 캡셔닝을 수행할 수 있다. 기존 비디오 캡셔닝 연구에서는 비디오의 장면 구성을 고려하지 않고 단순 등간격으로 프레임 추출을 통하여 비디오 캡셔닝을 수행하였다. 이는 다양한 장면의 모임으로 이루어진 비디오의 특성을 고려하지 않은 방법으로, 경우에 따라 주요 장면을 놓치거나, 불필요하게 중복된 프레임을 선택하는 문제가 발생한다. 본 논문에서는 비디오의 피처레벨 분석을 통해 비디오의 구성 특징을 파악하고, 이를 고려해 적응적으로 주요 프레임을 추출하여 이와 같은 문제를 해결하여 비디오 캡셔닝 에서의 성능향상을 보인다. 제안 알고리즘을 이용하여 생성된 피처는 비디오를 잘 요약하여 비디오 캡셔닝 수행 시, MSVD 데이터 셋에서 4 개의 평가지표에 대해 약 0.78%의 성능향상을 보였고, MSR-VTT 데이터 셋에서 약 0.6%의 성능향상을 보였다.

1. 서론

비디오 캡셔닝은 비디오 프레임 시퀀스를 입력으로 하며 그에 대한 자연어 설명을 생성하는 기법이다. 이미지에서 비디오로 연구가 확장되면서, 시계열 상의 특징을 고려하는 문제가 주목되어 왔다[1].

기존 연구는 비디오에서 등간격으로 정해진 수의 프레임을 샘플링 하고 시계열 어텐션을 주는 방법을 사용하였다[2,3]. 그러나 단순 등간격 샘플링은 다양한 변화를 담고 있는 비디오의 특성을 고려하지 않아 한계가 있다. 많은 수의 이미지 데이터로 구성된 비디오는 길이가 가변적이고, 객체의 움직임이나 시점의 움직임 등의 시간적인 변화를 담고 있다. 또한 촬영기법, 편집기법에 따라 장면의 변화 또한 존재한다. 이러한 비디오의 다양성을 무시하고, 단순 등간격을 이용해 캡셔닝 입력 피처를 생성하면, 움직임이 드문 비디오에서는 다수의 중복 프레임을 선택해 불필요한 계산을 반복할 가능성이 크며, 반대로 객체나 시점의 움직임이 크고, 장면 변화가 존재하는 비디오의 경우 중요한 정보가 누락될 수 있다.

이를 해결하기 위해 비디오 요약 연구에서는 비디오 프레임간의 차이를 이용해 비디오의 장면 구성을 분석하고자 하였다[4]. 하지만 이러한 방법은 조명이나 움직임에 취약하며, 장면이 오버랩 되는 특정 편집기법에서 장면의 변화를 파악할 수 없다.

본 논문에서는 이를 해결하기 위해, 비디오를 조명이나 움직임의 영향을 받지 않도록 피처레벨에서 분석하여 장면 구성의 특성을 파악할 것이며, 이를 바탕으로 장면의 중복성을 최소화하며 비디오를 대표할 수 있는 적응적 프레임 선택방법을 제안한다. 제안 알고리즘은 언급한 기존의 방법의 단점을 개선함으로써 비디오 캡셔닝에 적합한 피처를 생성하여 비디오 캡셔닝의 성능을 향상시킬 것이다.

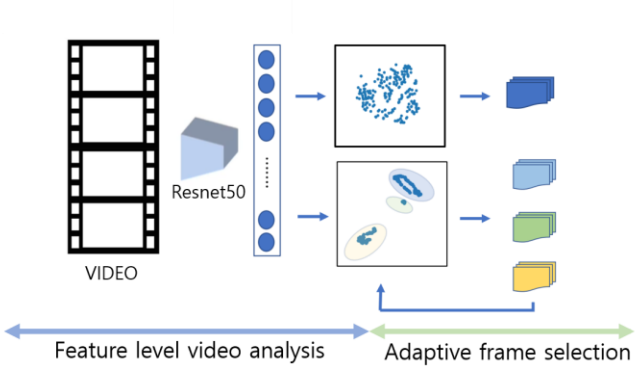

그림 1 제안하는 적응적 샘플링 방법 모델

2. 제안 알고리즘

2.1 피처레벨 비디오 특징 분석

비디오는 여러 장의 프레임으로 구성되어 있고, 장면의 변화가 있지 않다면 시간적으로 프레임사이의 상관관계가 높다. 그러나 조명이나 움직임의 변화가 있기 때문에 고차원 데이터로 장면의 유사도를 파악하는 것에는 어려움이 있다. 따라서 위해 고차원 데이터인 프레임을 resnet50[5]을 통과하여 주요 특징만으로 표현된 피처를 추출하였다. 추출된 피처는 조명과 움직임에 상관없이 인간의 관점과 유사하게 장면을 구분할 수 있다. 이후 장면의 유사한 정도를 보이도록 t-Stochastic Neighbor Embedding (t-SNE)[6]을 이용하여 매핑하였다.

2.2 적응적 주요 장면 추출

데이터 분포를 바탕으로 변화가 없는 비디오는 등간격 샘플링을 통해 캡셔닝에 필요한 피처를 추출하고, 장면의 변화가 있는 비디오의 경우 비디오를 하위 장면 그룹으로 재구성한다. 유사한 장면의 피처는 가까운 곳에 밀집되며, 반대의 경우 상대적으로 멀리 위치하기 때문에 밀도 기반 클러스터링을 이용하면, 유사 장면 그룹으로 분해할 수 있다. 본 논문에서는 Hierarchical density-based spatial clustering of applications with noise(HDBSCAN)[7]을 이용하여 비디오를 하위 장면 그룹으로 재구성하였다. 이때, 밀도 기반 클러스터링에서 중요한 파라미터인 mcs(mean-cluster-size)는 최소 몇 개의 데이터를 하나의 클러스터로 정할 것인가에 대한 기준이다. 파라메터가 너무 크면, 유사도가 낮은 장면도 포함하는 커다란 클러스터를 형성할 것이며, 너무 작으면 유사한 장면도 다른 클러스터에 속하게 되어 비디오를 하위 장면그룹으로 적절히 재구성할 수 없다. 비디오의 길이와 장면의 길이에 따라 그룹의 크기가 달라지기 때문에, 일차적인 클러스터링 이후의 결과와 비디오의 총 길이, 장면의 길이 등을 고려하여 적응적으로 파라메터를 설정하여 재그룹화를 진행하였다. 성공적으로 그룹화된 비디오에서의 그룹별 프레임 샘플링을 통해 캡셔닝의 입력 피처를 생성한다. 제안된 방법으로 생성된 피처는 기존 방법에서의 유사 프레임의 중복 또는 대표 프레임을 선택하지 못하는 문제가 해결된 피처이므로 캡셔닝 수행 시 성능향상을 보인다.

3. 실험결과

우리는 실험을 위해 비디오 캡셔닝에서 주로 사용되는 Microsoft Video Description (MSVD) [8]데이터 셋 과 MSR-VTT[9]를 이용하였다. MSVD 에는 1,970 개의 비디오 클립으로 구성되어 있어 트레이닝에 1520 개, 테스트에 450 개의 데이터를 사용하였다. MSR-VTT 는 10000 개의 데이터 셋으로 구성되어 있어 트레이닝에 7010 개, 테스트에 2990 개를 사용하였다. 우리는 기존 등간격 프래임 추출방식과 제안하는 적응적 추출방식을 비교하기 위해 엔코더-디코더 구조에 어텐션을 추가한 모델을 가본모델로 사용하였다 성능평가의 방법으로 4 개의 평가지표 BLEU[10], METOR[11], ROUGEL[12], CIDEr[13]을 사용하였다.

	Bleu_4	CIDEr	METEOR	ROUGE_L
기존 방법	24.8	28.3	24.0	58.4
제안 방법	25.9	29.3	24.6	68.9

표1 MSVD[8] **데이터에 대한 성능비교**

	Bleu_4	CIDEr	METEOR	ROUGE_L
기존 방법	31.0	38.1	25.4	55.8
제안 방법	32.5	37.8	26.0	56.5

표 2 MSR-VTT[9] **데이터에 대한 성능비교**

기존의 등간격 샘플링 방법과 제안하는 적응적 샘플링 방법을 비교하였을 때, MSVD 데이터 셋에서, 4 개의 평가지표에 대해 약 0.78%의 성능향상을 보였고, MSR-VTT 데이터 셋에서 약 0.6%의 성능향상을 보였다. 이를 통해 제안하는 적응적 피처 생성 방법이 캡셔닝의 피처 생성방법으로 적합하다는 것을 보여준다.

4. 결론

캡셔닝을 위한 피처 추출 과정에서 등간격 프레임 샘플링은 비디오의 장면 구성 특성을 고려하지 않기 때문에, 중복된 프레임과 정보의 손실을 야기하여 비디오를 잘 대표하지 못한다. 이를 해결하기 위해 피처레벨 비디오 분석과, 적응적 비디오 장면 선택 알고리즘을 제안하였다. 비디오를 피처레벨에서 분석하여 특성을 파악하면 비디오를 하위 장면 그룹으로 재구성 가능하다. 비디오의 특성을 고려하여 적응적으로 프레임을 샘플링하여 캡셔닝에 적합한 피처를 생성함으로 중복된 프레임을 최소화하고, 정보의 손실을 막아 캡셔닝의 성능을 향상시킬 수 있다.

Acknowledgement

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음" (IITP-2020-0-01460)

참고문헌

[1] Subhashini Venugopalan et al., “Sequence to sequence-video to text,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4534-4542

[2] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, “Video paragraph captioning using hierarchical recurrent neural networks,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016

[3] Yan, Chenggang, et al. "Stat: spatial-temporal attention mechanism for video captioning." IEEE transactions on multimedia 22.1 (2019): 229-241.

[4] Zhang, Ke, et al. "Video summarization with long short-term memory." European conference on computer vision. Springer, Cham, 2016.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778

[6] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.Nov (2008): 2579-2605.

[7] McInnes, Leland, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." Journal of Open Source Software 2.11 (2017): 205.

[8] D. Chen and W. B. Dolan, “Collecting highly parallel data for paraphrase evaluation,” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 190-200.

[9] Xu, Jun, et al. "Msr-vtt: A large video description dataset for bridging video and language." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[10] Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40thannualmeetingoftheAssociationforComputationalLinguistics, 2002, pp. 311-318.

[11] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65-72.

[12] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out, 2004, pp. 74-81.

[13] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus based image description evaluation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566-4575.