지식의 증류기법을 이용한 샷 경계 검출 모델

박성민, 윤의녕, 조근식
인하대학교 컴퓨터공학과
zzererer@naver.com, entymos@hotmail.com, gsjo@inha.ac.kr

Shot Boundary Detection Model using Knowledge Distillation

Sung Min Park, Ui Nyoung Yoon, Geun-Sik Jo
Department of Computer Engineering, Inha University

요 약

샷 경계 검출(Shot Boundary Detection)은 영상 콘텐츠 분석을 위한 필수적인 기술이며, 다양한 방식으로 편집된 영상의 샷 경계를 정확하게 검출하기 위한 연구가 지속되어 왔다. 그러나 기존에 연구들은 고정된 샷 경계 검출 알고리즘이나 매뉴얼한 작업과 같이 학습이 불가능한 과정이 포함되어 있어 성능 개선에 한계가 있었다. 본 논문에서는 이러한 과정을 제거한 End-to-End 모델을 제안한다. 제안하는 모델은 시공간 정보 추출성능을 높이기 위해 행동 인식 데이터셋을 이용한 전이학습을 사용하고, 샷 경계 검출 성능을 높이기 위해 개선된 지식의 증류기법(Knowledge Distillation)을 결합한다. 제안하는 모델은 ClipShots 데이터셋에서 DeepSBD 에 비해 cut transition 과 gradual transition 이 각각 5.4%, 41.29% 높은 성능을 보였고, DSM 과의 비교에서 cut transition 의 정확도가 1.3% 더 높은 결과를 보였다.

1. 서론

샷 경계 검출은 영상 콘텐츠를 분석하기 위해 필요한 과정으로 현재까지 지속적으로 연구되고 있다. 최근에는 시공간 정보를 분석하는데 있어서 기존에 사용하던 2D 합성곱 신경망보다 더 좋은 성능을 보이는 3D 합성곱 신경망[1]으로 개선하여 검출 성능을 높이는데 성공하였다. 하지만 모델의 성능 개선을 위해 학습이 불가능한 방법들을 추가하여 오히려 성능 향상의 방해요소로 작용하고 있다. 3D AlexNet 을 사용하는 DeepSBD[2]의 경우 매뉴얼한 방법으로 데이터셋을 생성하는 과정과 프레임 간의 유사도 비교 방법을 사용하고 있으며, DSM[3]의 경우 프레임 간의 유사도를 이용한 샷 경계 탐지 모델을 전면부에 사용하고 있다. 이러한 방법들은 학습으로 인한 성능개선이 불가능해 모델의 한계를 결정짓는 요소이다. 본 논문에서는 고정된 방법론들을 제외한 End-to-End 모델을 제안한다. 행동 인식 데이터셋을 이용한 전이학습으로 시공간 정보 추출 능력을 향상시키고, 지식의 증류기법을 통해 샷 경계 검출 성능을 개선한다.

2. 배경지식 및 관련연구

2.1 배경지식

지식의 증류기법은 이미 학습된 모델(교사모델)을 이용하여 학습이 부족한 모델(학생모델)을 학습시키는 방법이다. 이 방법[5]은 학생 모델의 예측 확률 분포가 교사모델을 따라가므로 원하는 성능까지 빠르게 도달한다는 장점이 있다. 하지만 이미 학습된 모델의 성능을 넘을 수 없다는 단점이 있다. 이를 보완하기 위해 지도학습과 결합하는 방법[6]을 사용하여 성능을 개선할 수 있다.

2.2 관련연구

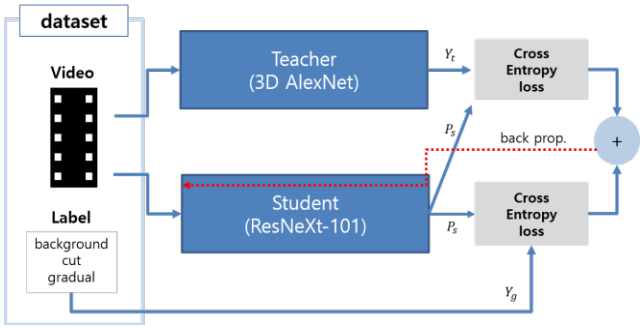
DeepSBD[2]는 기존의 AlexNet 을 3D 합성곱 신경망 형태로 바꾼 뒤, 매뉴얼한 데이터셋 생성과정과 프레임 간의 유사도 비교 방법을 결합한 샷 경계 검출 방법이다. 이 방법은 여러 단계에 걸쳐 학습을 진행한다. 먼저 데이터셋 생성을 위해 수집한 학습 데이터를 A, B 로 나눈 후, A 로 AlexNet 을 학습시킨다. 그 다음 B 를 이용해 해당 모델을 테스트한 뒤 분류된 결과를 매뉴얼하게 cut transition, gradual transition, no transition 클래스로 분류한다. 마지막으로, 분류를 마친 데이터셋 B 과 기존에 학습 데이터로 사용되었던 A 를 이용해 AlexNet 을 학습시켜 완성된다. 이렇게 학습시킨 결과는 TRECVID 데이터셋에서 최고의 성능을 내게 된다. 본 논문에서는 End-to-End 모델 간의 비교를 위해, DeepSBD 에서 기존에 결합된 매뉴얼한 방법을 제거한 AlexNet 을 비교대상으로 사용한다.

DSM[3]은 세 가지 모델을 결합한 샷 경계 검출 프레임워크다. 이 프레임워크의 샷 경계 검출 과정을 보면, 먼저 Initial Filtering 으로 transition 후보군을 추출한 뒤 Cut Model 로 cut transition 을 분류한다. 마지막으로 Gradual Model 에서 64 프레임 길이의 영상에서 객체 탐지 방식을 적용해 3 프레임에서 40 프레임까지 다양한 길이의 gradual transition 을 찾는다. 이렇게 구성된 DSM 은 ClipShots, TRECVID 2007, RAI 데이터셋에서 최고의 성능을 보여주게 된다. 본 논문에서 제안하는 모델의 경우 입력하는 영상의 길이가 16 프레임으로 Gradual Detector 에 입력되는 길이와 서로 다르다. 따라서 실험 분석 단계에서는 gradual

transition 의 성능은 비교하지 않고, cut transition 의 성능 비교를 통해 제안한 방법의 효과를 증명한다.

DSM[3]의 마지막 모델인 Gradual Model 에서 사용한 3D ResNet-18 은 Kinetics 데이터셋으로 학습된 모델을 사용했다. Kinetics 는 행동 인식 문제에 관한 데이터셋으로, 3D 합성곱 신경망으로 변형시킨 ResNet, ResNeXt, DenseNet 등의 성능평가[4]를 위해 사용되었다. 그 결과, 3D ResNet-18 의 정확도가 제일 낮았고, 3D ResNeXt-101 은 단일 신경망 모델 중 가장 좋은 성능을 보였다.

3. 제안하는 End-to-End 모델



<그림 1> 제안하는 모델의 구조

제안하는 모델의 구조는 <그림 1>과 같다. 지식을 전달할 교사(Teacher)모델은 ClipShots 데이터셋으로 이미 학습된 3D AlexNet 을 사용하고, 학습할 학생(Student)모델은 ResNeXt-101 을 사용한다.

3.1 전이학습

학습할 학생모델은 시공간 변화 정보 추출의 성능을 높이기 위해 Kinetics 데이터셋으로 ResNeXt-101 을 학습시킨다. 행동 인식과 샷 경계 검출은 시공간 정보를 추출하여야 한다는 점에서 유사하다. 그러므로 행동 인식 데이터셋으로 학습시킨 뒤 전이학습을 할 경우 샷 경계 검출에 있어서 더 좋은 성능을 기대할 수 있다.

3.2 지식의 증류기법을 적용한 모델의 학습과정

먼저 입력할 데이터 영상에서 샷 경계에 해당하는 영역을 추출한다. 추출된 영상을 <그림 1>의 교사모델과 학생모델에 각각 입력시킨다. 각 모델에서는 각 클래스 별 분류 점수인 $\text{Score} = (S_n, S_c, S_g)$를 출력한다. S_n은 no transition 클래스 점수, S_c 는 cut transition 클래스 점수, S_g 는 gradual transition 클래스 점수이다. 그 후 교사모델이 예측한 클래스 정보 Y_t 와 입력한 영상에 대한 Ground Truth 정보 Y_g 를 기준으로 학생모델의 클래스 별 확률 $P_s = \text{Softmax}(\text{Score}_s)$ 과의 오차를 계산한다. 여기서 Score_s 는 학생모델이 계산한 클래스 점수를 의미한다. 분류를 위한 오차 함수는 Cross Entropy(CE) loss 를 사용하며 전체 오차 함수는 다음과 같다.

$$\text{Loss}_{\text{sbd}} = \text{CE}(P_s, Y_t) + \text{CE}(P_s, Y_g)$$

교사모델과 학생모델 간의 확률분포 차이를 학습시키는 기존의 지식 증류 기법과는 다르게 교사의 예측도 하나의 Ground Truth 로써 지도학습을 진행한다. 이러한 차이는 교사모델의 성능에 수렴하는 것을 피하면서, 경사 하강법으로 학습이 진행될 때 교사모델의 예측이 하나의 모멘텀으로 작용하여 local minima 문제를 개선할 수 있다.

4. 실험 및 분석

4.1 데이터셋 및 평가 기준

실험을 위한 데이터셋은 YouTube, Weibo 에서 추출한 ClipShots 데이터셋[3]을 사용했다. 이 데이터 셋은 총 4039 개의 영상, 128,636 개의 cut transition, 38,120 개의 gradual transition 으로 구성되어 있다. 학습은 3539 개의 영상으로 진행했으며 테스트는 500 개의 영상으로 실험을 진행했다.

평가기준은 TRECVID 표준 평가방식을 사용했으며, 이 방식은 예측한 경계가 Ground Truth 와 최소한 1 개 프레임이라도 겹치면 예측한 것으로 간주한다. 정확도 성능 평가는 정밀도(Precision)와 재현율(Recall)의 조합인 F1 score 를 사용했다. 정밀도는 예측한 결과 중 Ground Truth 의 비율이고, 재현율은 Ground Truth 중 예측한 결과의 비율을 의미하며, F1 score 는 정밀도와 재현율의 조화평균을 의미한다. F1 score 에 대한 식은 다음과 같다.

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

4.2 학습 및 테스트 데이터의 전처리

학습을 위해 입력하는 영상의 모든 프레임에 대해 crop 과정을 거친다. 5 개의 임의의 위치(top-left, top-right, bottom-left, bottom-right, center)에서 128×128 크기의 이미지를 crop 하여 사용한다. 모델에 입력되는 영상 데이터의 크기는 128×128×3×16 으로 해상도는 128×128, 채널은 RGB, 프레임 길이는 Ground Truth 인 샷 경계를 포함하는 영역을 16 프레임 단위로 잘라 사용하였다. 학습 데이터의 Ground Truth 는 cut transition, gradual transition 이 각각 122,760 개, 35,698 개로 구성되어 있다. 그 결과 두 데이터 수가 균형을 이루고 있지 않아 상대적으로 데이터 수가 적은 gradual transition 이 학습이 잘 안되는 문제점이 있다. 그래서 두 transition 을 1:1 비율로 입력시켜 학습되도록 하였다.

테스트를 위한 영상 데이터의 크기는 입력 데이터와 마찬가지로 128×128×3×16 이며 영상의 시작 프레임부터 8 프레임 간격으로 이동하며 16 개 프레임만큼 입력했다.

4.3 전이학습 및 지식의 증류기법에 관한 실험

<표 1>은 행동 인식 문제에서 사용되는 Kinetics 데이터셋으로 인한 전이학습과 지식의 증류기법이 샷 경계 검출의 성능에 미치는 효과를 실험한 결과다. 실험은 ClipShots 데이터셋의 100%를 사용하여 ResNeXt-101 을 5 epoch 학습시켰다.

<표 1> 전이학습 및 지식의 증류기법에 따른 성능비교

Kinetics	지식 증류	F1 score	
		cut	gradual
X	X	0.770	0.504
	O	0.772	0.516
O	X	0.825	0.722
	O	**0.859**	**0.746**

그 결과 전이학습과 지식의 증류기법으로 인해 모든 샷 경계 검출 성능이 상승하는 현상을 보였다. 이는 행동 인식 문제에서 학습한 시공간 변화 정보를 추출하는 능력이 샷 경계 검출 능력향상에 큰 영향을 끼쳤으며, 지식의 증류기법으로 local minima 문제를 개선할 수 있었음을 의미한다.

4.4 ClipShots 데이터셋에서의 성능평가

<표 2>와 <표 3>은 ClipShots 데이터셋으로 샷 경계 검출을 실험한 결과다. 실험은 데이터셋의 100%를 사용하여 5 가지 실험 대상들을 5 epoch 학습시켰다. 실험 대상인 5 가지 모델들에 대한 설명은 다음과 같다. <표 2>에서 (1)과 (2)는 End-to-End 모델 간의 비교를 위한 모델이다. (1)은 기존의 DeepSBD 에서 학습이 불가능한 요소를 제거한 3D AlexNet 이며, (2)는 DSM 의 Gradual Detector 에서 객체 탐지 방법을 제거한 3D ResNet-18 로 입력하는 영상 길이를 16 프레임으로 수정하여 제안하는 모델과 조건을 맞추었다. (3)은 기반 모델의 검출 성능을 평가하기 위해 지식의 증류기법을 적용하지 않은 ResNeXt-101 이며, (4)는 본 논문에서 제안하는 지식의 증류기법을 적용한 ResNeXt-101 이다. 마지막으로 <표 3>에서 DSM 은 본 논문에서 제안하는 모델의 성능평가를 위한 비교대상으로 ClipShots 데이터셋에서 가장 좋은 성능을 보여준 모델이다.

<표 2> End-to-End 모델 간의 성능비교

Model	F1 score	
	cut	gradual
(1)	0.815	0.528
(2)	0.831	0.688
(3)	0.825	0.722
(4)	**0.859**	**0.746**

<표 2>는 End-to-End 모델 간의 성능비교를 위한 실험 결과다. 먼저 지식의 증류기법을 제외한 기반 모델 간의 비교를 위해 (1), (2), (3)의 정확도를 비교해 보면 cut transition 은 성능차이가 크지 않지만 긴 시공간 변화 정보를 필요로 하는 gradual transition 의 검출 정확도는 큰 차이를 보이고 있다. 이는 모델의 시공간 정보 추출 능력은 인접한 프레임 간의 순간적인 변화보다는 여러 프레임에 걸친 변화를 찾는데 더 효과적임을 의미한다. 또한 지식 증류기법의 효과로 인해 제안하는 모델이 End-to-End 모델 중 가장 좋은 성능을 보였다.

<표 3> 제안하는 모델과 DSM 의 성능비교

Model	F1 score	
	cut	gradual
DSM	0.848	**0.870**
(4)	**0.859**	0.746

<표 3>은 제안한 모델의 성능 평가를 위해 ClipShots 데이터셋에서 최고의 성능을 보여준 DSM 과 비교한 실험결과다. 그 결과 제안한 모델의 cut transition 검출 성능이 DSM 보다 더 좋은 결과를 보였다. 이는 DSM 에 포함되어 있는 학습이 불가능한 모델이 샷 경계 검출 성능을 방해하는 요인이었음을 의미하며, 지식 증류 기법으로 인한 모델 최적화가 샷 경계 검출 문제에서 더 효과적이었음을 증명한다.

5. 결론 및 향후연구

본 논문에서는 행동 인식 데이터셋을 이용한 전이학습과 지식 증류기법을 적용한 End-to-End 모델을 제안했다. 본 논문에서 제안한 모델은 DeepSBD 에서 사용한 3D AlexNet 과 비교하여 cut transition 및 gradual transition 의 검출 성능이 각각 5.4%, 41.29% 높았으며, DSM 의 Gradual Detector 에 사용된 ResNet-18 과 비교하여 각각 3.37%, 8.43% 높은 성능을 보였다. 또한 ClipShots 데이터셋에서 최고의 성능을 보인 DSM 과 비교하여 cut transition 의 성능이 1.3% 더 높은 결과를 보였다. 이는 본 논문에서 제안한 전이학습과 지식 증류기법을 결합한 방법이 샷 경계 검출에 있어서 효과적임을 증명하였다.

향후 연구로는 End-to-End 모델의 특성을 유지하면서 gradual transition 의 검출 성능을 올리기 위해 입력할 프레임 길이를 늘리고 모델 후반부의 분류기를 개선할 계획이다.

6. 감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음 (IITP-2017-0-01642)

7. 참고문헌

[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks”, The IEEE International Conference on Computer Vision (ICCV), pp. 4489-4497, 2015

[2] A. Hassanien, M. Elgharib, A. Selim, S.-H. Bae, M. Hefeeda and W. Matusik, “Large-scale, Fast and Accurate Shot Boundary Detection through Spatio-temporal Convolutional Neural Networks”, arXiv:1705.03281, 2017

[3] S. Tang, L. Feng, Z. Kuang, Y. Chen and W. Zhang, “Fast Video Shot Transition Localization with Deep Structured Models”, arXiv:1808.04234, 2018

[4] K. Hara, H. Kataoka and Y. Satoh, “Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?”, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6546-6555, 2018

[5] G. Hinton, O. Vinyals and J. Dean, “Distilling the Knowledge in a Neural Network”, Deep Learning and Representation Learning Workshop: NIPS, 2014

[6] G. Chen, W. Choi, X. Yu, T. Han and M. Chandraker, “Learning Efficient Object Detection Models with Knowledge Distillation”, Advances in Neural Information Processing Systems 30 (NIPS), 2017