글. 홍순기 SBS 미디어기술연구소 사원

딥러닝 기반 방송 클립 서비스

최근 시청자는 미디어 소비에 대한 주도권을 자신이 갖기를 원한다. 원하는 미디어를 원하는 시간에 소비하고자 하는 시청자의 욕구를 충족시키기 위해, 미디어 서비스 업체들은 호흡이 긴 동영상을 짧은 길이의 동영상으로 축약하여 제공하는 클립형 미디어 서비스를 제공하고 있다. 이러한 클립형 미디어 서비스를 위해서는 동영상 축약이 필수적으로 필요한데, 현재 대부분의 서비스 업체들은 동영상 축약을 수동으로 진행하고 있다. 따라서 수동으로 동영상을 축약하는데 소요되는 경제적/시간적 비용을 줄이기 위해, 예전부터 다양한 동영상 자동 축약 기술이 시도되었다[1][2]. 비교적 최근까지 연구된 자동 축약 기술은 이미지 인식/분석 기술을 활용하여 이미지의 품질을 다양한 관점으로 수치화하고, 선형 예측 기술을 통해 수치화된 이미지 품질들로부터 해당 이미지의 중요도를 추정한 후, 목표 시간에 맞추어 중요도가 낮은 프레임을 소거하는 방법을 사용하였다[3]. 이러한 방법을 통해 동영상 축약의 자동화 가능성을 확인할 수 있었으나, 이미지 인식/분석 기술의 한계에 의해 자동 생성된 축약 결과물을 바로 클립형 미디어 서비스에 적용할 수 있을 정도의 정확도를 확보하지는 못하였다. 하지만 최근에 이미지 인식/분석 능력을 비약적으로 발전시킨 딥러닝 기술을 동영상 자동 축약 기술에 적용하여 자동 생성된 축약 결과물의 정확성을 크게 향상시킨 기술들이 제안되었다.

이에 본 기고에서는 클립형 서비스를 위한 딥러닝 기반 동영상 자동 축약 기술의 발전 과정에 대해 살펴보고, 최근 딥러닝 기반 동영상 자동 축약 기술을 활용하여 서비스를 시작한 딥러닝 기반 UHD 방송 A-ESG 클립 서비스를 소개하려 한다.

딥러닝 기반 동영상 자동 축약 기술 현황

클립형 미디어 서비스를 제공하기 위해서는 동영상을 축약하는 과정이 필수적이다. 현재 동영상 축약 과정은 사람이 전체 영상을 확인하고 클립형 미디어로 생성할 짧은 길이의 구간을 발췌하여 파일로 생성하고 있다. 이러한 수동 축약 과정을 거칠 경우 제작 비용과 서비스 품질 측면에서 한계점이 존재한다. 현재 클립형 미디어 서비스가 선택하고 있는 수동 축약 과정은 클립형 미디어 서비스를 제공하는 채널과 프로그램이 많아질수록 제작 비용이 크게 증가할 수밖에 없다. 또한 수동으로 축약을 진행하는 시간에 의해 발생하는 서비스 지연은 서비스 품질 측면에서 시청자의 만족도를 저해하는 요소가 된다. 따라서 위의 문제를 보완하기 위해 동영상을 자동으로 축약하기 위한 기술의 개발이 필수적으로 필요하다.

자동 축약 기술의 핵심은 영상을 분석하여 사용자에게 유의미한 특징값들을 얼마나 정확하게 찾아내느냐로 정의할 수 있다. 특히 보통 사용자에게 유의미한 특징값들은 영상 내에 존재하는 객체(Object)에 의해 좌우되는 경우가 많으므로, 동영상 축약 기술은 영상 내에 존재하는 객체를 정확하게 인식하는 것이 매우 중요하다. 최근 딥러닝 기술의 등장과 함께 객체 인식 기술의 정확도가 딥러닝 이전의 기술과 비교하여 매우 가파르게 향상되고 있다. 실제로 MS社가 제공하는 머신러닝 데이터셋인 Common Objects in Context (COCO)[4]을 이용하여 객체 인식 성능을 테스트하였을 때, 2015년 딥러닝 기술에 의해 기존 알고리즘과 비교하여 3배의 성능이 향상(5% → 15%)되었으며, 이후 2017년까지 2.5년 동안 딥러닝 기술이 발전하면서 3배의 추가 성능 향상(15% → 46%)이 이루어졌다.

한편, Tensorflow[5], Darknet[6] 등과 같이 딥러닝 신경망을 쉽게 구성할 수 있는 플랫폼들이 오픈소스로 공개되고, 이러한 플랫폼 위에서 구동할 수 있는 Faster R-CNN[7], Single Shot Multibox Detection(SSD)[8], You Only Look Once(YOLO)[9]와 같은 객체 검출 신경망 모델들도 오픈소스로 공개되면서 딥러닝 기술을 활용하기 위한 문턱이 많이 낮아졌다. 따라서 자연히 딥러닝 기반의 객체 인식 기술을 자동 축약 기술에 적용하려고 하는 움직임도 생겨났다.

MS社의 Yao[10]는 [그림 1]과 같이 각각의 이미지(Frame)를 공간적으로 분석하기 위한 딥러닝 신경망(AlexNet)[11]과 순차적인 이미지들(Clip)을 시간적으로 분석하기 위한 딥러닝 신경망(C3D)[12]을 사용하여 각 세그먼트의 하이라이트 지수를 구한 후, 하이라이트 지수를 활용하여 동영상 축약을 수행하는 알고리즘을 제안하였다. 딥러닝 신경망을 통해서 기존의 알고리즘에 비해 객체 검출 성능을 향상시켰다는 점과 순차적인 이미지들을 사용하여 시간적 분석을 하는 신경망을 추가했다는 점에서 제안된 알고리즘은 이후 딥러닝을 사용한 동영상 축약 알고리즘에 많은 영향을 주었다. 또한 동영상 축약의 형태를 Timelapse와 Skimming의 2가지 타입으로 정의한 것도 의미가 있다. 두 가지 형태 모두 하이라이트 지수를 활용한다는 공통점이 있지만, Timelapse의 경우는 하이라이트 지수가 낮을수록 빠른 속도로 프레임을 재생하고 Skimming의 경우는 하이라이트 지수가 낮은 프레임을 삭제하는 방식으로 동영상을 축약한다는 차이점이 있다. 위의 장점에도 불구하고 Yao의 알고리즘은 프레임 레벨의 학습데이터가 필요하다는 점과 공간적/시간적 분석을 위한 딥러닝 신경망이 분리되어 있어 신경망 학습이 힘들고 알고리즘 수행 시간이 오래 걸린다는 단점이 있다. 따라서 이러한 문제를 극복하기 위해 세그멘트 레벨의 학습데이터와 공간적/시간적 분석이 하나의 딥러닝 신경망에서 수행되도록 하기 위한 연구가 진행되었다.

Panda[13]는 영상의 상황에 따라 딥러닝 네트워크상에서 공간적으로 활성화되는 특정 영역이 존재하는 것에 영감을 얻어 DeSumNet이라는 영상 축약 솔루션을 제안하였다. DeSumNet은 입력 영상을 시공간적으로 분석할 수 있는 3D 합성곱 신경망(Convolution Neural Network, CNN) 구조를 활용하여 특정 상황에서 시공간적으로 활성화되어 있는 영역을 찾는 '합성곱 신경망'과 사용자의 의도를 반영하여 활성화된 영역을 중요도 점수(Importance score)로 변환해 주는 '완전히 연결된 망'으로 구성되어 있다.
예를 들어, [그림 1]과 같이, 서핑(Surfing) 영상이 입력 영상인 경우, 영상을 일정한 크기의 클립으로 나눈 후 나뉜 클립 중에서 사용자가 중요하다고 생각한 클립을 선택하여 DeSumNet을 학습시키면, 차후에 테스트 영상을 DeSumNet에 공급하였을 때 사용자가 선택한 영상과 유사한 클립의 중요도 점수가 높게 계산되어 출력된다. 이후, 중요도 점수가 높은 클립들을 우선적으로 추려서 축약 동영상을 생성하게 된다.

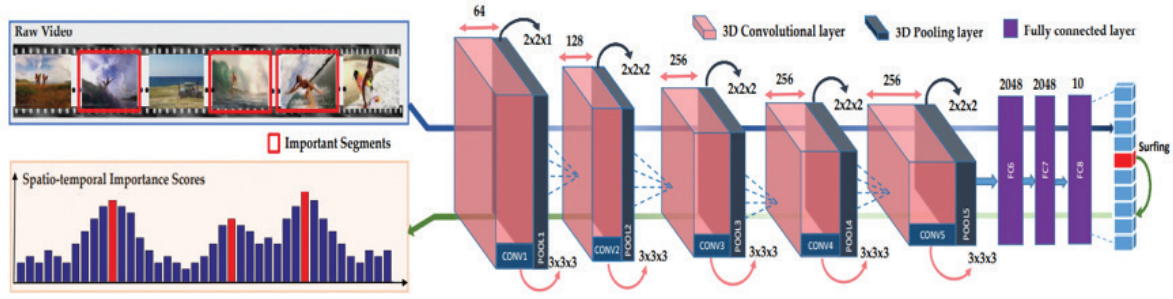

그림 1. DeSumNet 구조, 3D CNN을 사용하여 시공간 중요도 점수를 계산함

Panda의 알고리즘은 하나의 통합된 딥러닝 신경망인 DeSumNet을 세그멘트 레벨의 학습데이터를 활용하여 학습시키고, 이를 이용하여 동영상 축약을 수행할 수 있다는 점에서 딥러닝 신경망을 활용한 최근 알고리즘 중에서도 매우 의미 있는 결과를 보여줬다고 할 수 있다.

홍순기
SBS 미디어기술연구소
사원

딥러닝 기반 UHD 방송 A-ESG 클립 서비스

그림 2. 지상파 UHDTV A-ESG 서비스

UHD 방송 A-ESG는 방송 일정 안내와 함께 '하이라이트 클립'을 시청자에게 제공하는 양방향 서비스로, 시청자가 손쉽게 방송콘텐츠의 하이라이트 부분을 [그림 2]와 같이 시청할 수 있다. 하지만 현재는 하이라이트 클립을 제작하기 위해 발생하는 비용과 시간 문제로 선별된 몇몇 방송콘텐츠에 대해서만 클립 서비스가 제공되고 있는 실정이다.

따라서 '딥러닝 기반 UHD 방송 클립 서비스'는 하이라이트 클립을 제작하는 과정을 딥러닝 기반 인공지능 클립 생성 시스템이 수행하게 함으로써, 클립 제작 비용과 시간을 획기적으로 낮추면서도 안정적인 품질을 획득할 수 있다. 결과적으로 [그림 3]과 같이 '딥러닝 기반 UHD 방송 클립 서비스'를 통해 방송사는 모든 방송콘텐츠에 대하여 UHD 방송 A-ESG 서비스를 어려움 없이 제공할 수 있다.

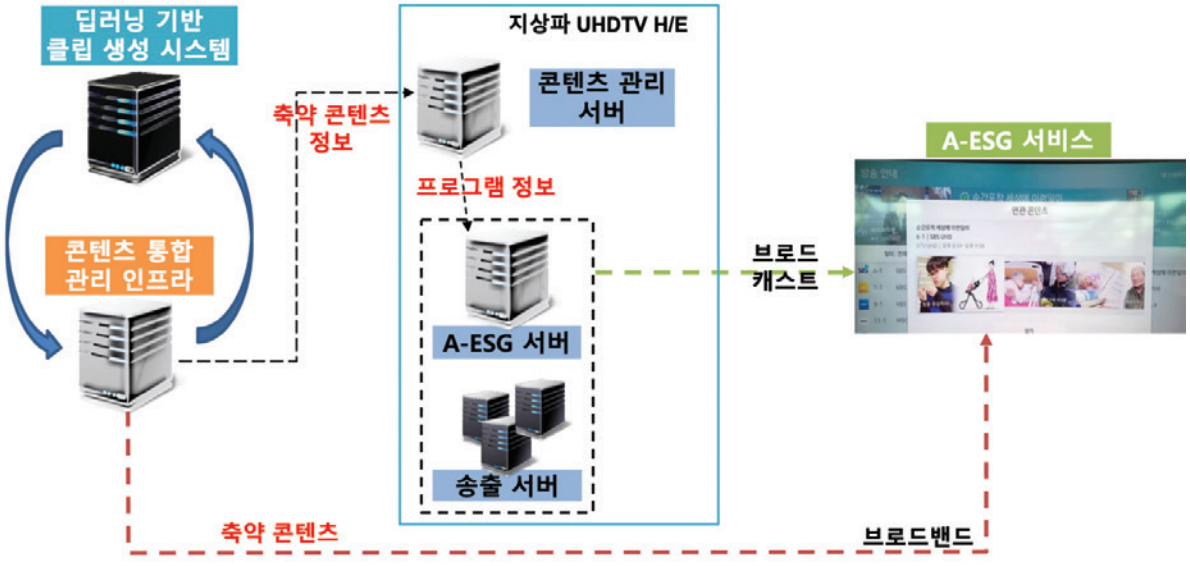

그림 3. 딥러닝 기반 UHD 방송 A-ESG 서비스 워크플로우

인공지능 클립 생성 시스템은 수많은 방송본 영상과 해당 영상의 하이라이트 클립들과의 관계를 시공간적으로 분석하여 학습하는 딥러닝 알고리즘에 기반하고 있다. 구체적으로 딥러닝 알고리즘은 [그림 1]에서 소개한 3D CNN 구조를 사용하여 구현할 수도 있지만 해당 기술은 프레임 단위의 중요도 점수를 계산하지 못하는 문제점이 있다. 따라서 최근에는 2D CNN을 통해 매 프레임의 특징값을 생성하고, 생성된 특징값을 1D 장단기 메모리망(Long-Short Term Memory, LSTM)을 통해 시계열 분석하여 중요도 점수를 프레임 단위로 예측하는 방법을 사용하고 있다.

딥러닝 기반 클립 생성 시스템은 기존의 인식 기술(인물, 객체 등) 기반 시스템의 한계를 극복한 높은 예측 정확도를 보여주지만, 근본적으로 방대한 양의 학습데이터가 필요하기 때문에 효율적으로 학습데이터를 생성하기 위한 방안이 꼭 필요하다. 따라서 방송사에서 딥러닝 기반 기술을 도입하기 위해서는 방송한 콘텐츠를 실시간으로 편집/관리/유통할 수 있는 콘텐츠 통합 관리 인프라가 필수적으로 필요하다. 이러한 콘텐츠 통합 관리 인프라로부터 학습 데이터를 받아 학습에 활용하고, 학습에 의해 진화한 알고리즘에 의해 생성된 클립들을 다시 콘텐츠 통합 관리 인프라를 통해서 유통함으로써, 시간이 지날수록 더 재미있는 하이라이트 영상을 찾아서 서비스가 이루어지는 딥러닝 기반 클립 생성 시스템 구축이 가능하다.

마지막으로 딥러닝 기반 클립 생성 시스템은 '딥러닝 기반 UHD 방송 클립 서비스'를 통해 풍부한 A-ESG 클립 서비스 제공이 가능해져, 시청자에게 더 나은 품질의 UHD 방송 서비스를 제공하는 것 자체에도 의미가 있지만, 딥러닝 기반 클립 생성 시스템을 구축/운용하는 과정에서 획득한 영상 분석 및 예측 기술은 향후 방송 콘텐츠 제작 및 관리 효율화 업무에 활용할 수 있는 중요한 자산이 될 수 있다.

참고 문헌

[1] Gygli, M., Grabner, H., Van Gool, L.: Video summarization by learning submodular mixtures of objectives. In: CVPR (2015)

[2] Zhang, K., Chao, W.l., Sha, F., Grauman, K.: Summary transfer: exemplar-based subset selection for video summarization. In: CVPR (2016)

[3] M. Gygil, etc., Creating Summaries from User Video, ECCV2014, pp 505~520

[4] COCO:Common Objects in Context (2016). http://mscoco.org/dataset/#detections-leaderboard. Accessed 25 July 2016.

[5] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng., TensorFlow:Large-scale machine learning on heterogeneous distributed systems, arXiv preprint, 1603.04467, 2016. arxiv.org/abs/1603.04467. Software available from tensorflow.org.

[6] Joseph Redmon, "Darknet: Open Source Neural Networks in C," Software available from http://pjreddie.com/darknet/, 2013-2016.

[7] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. E. Reed. SSD: single shot multibox detector. CoRR, abs/1512.02325, 2015.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640, 2015.

[10] T. Yao, T. Mei, and Y. Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 982-990, 2016.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.

[12] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In ICCV, 2015.

[13] R Panda, A Das, Z Wu, J Ernst, AK Roy-Chowdhury, Weakly supervised summarization of web videos, 2017 IEEE International Conference on Computer Vision (ICCV), 3677-3686