실감영상 분류를 위한 VGG 수정 모델

배재현 박상효

경북대학교 컴퓨터학부

Modified VGG Model for Immersive Image Classification

Jaehyun Bae Sang-hyo Park

School of Computer Science and Engineering

Kyungpook National University

jayangie@naver.com, s.park@knu.ac.kr

요 약

본문은 실감미디어로 주목받는 MPEG-I 표준을 위한 영상을 전통적인 자연 영상과 분류할 수 있는 CNN 모델을 제안한다. 실감미디어의 효율적인 데이터 압축을 위해 view synthesis 가 이뤄지는 경우가 많은데, 이러한 영상의 경우 일반적인 자연 영상과 신호 특성이 현저히 다른 영역이 다수 존재할 수 있다. 그러나 일반적인 동영상 압축 기법들은 자연 영상에 최적화된 경향이 많아서 이러한 실감미디어 영상에 최적화되기 어려운 문제가 있다. 이를 해결하기 위해 본문에서는 영상 분류 문제에서 탁월한 효과를 보여왔던 CNN 모델을 활용하여 실감미디어 영상 분류 문제를 해결하고자 한다.

1. 서론

실감미디어(Immersive Media)의 수요가 증대됨에 따라, 동영상 압축 표준을 이끌어온 Moving Picture Experts Group(MPEG)에서도 이에 대한 표준화 프로젝트를 진행해왔다. 특히 전방위 몰입형 비디오 압축 표준을 위해 ISO/IEC 23090 Part 12(MPEG Immersive Video) 프로젝트를 진행하고 있다[1]. 이 프로젝트는 기존의 동영상 압축 표준과는 다르게, 코덱 내에 코덱이 존재하는 구조로 되어 있으며, 대표적으로 2013 년에 표준이 완료된 ITU-T H.265 | ISO/IEC 23008-2(High Efficiency Video Coding, 이하 HEVC)[2] 뿐 아니라 차세대 동영상 압축 표준으로 예상되는 ITU-T H.266 | ISO/IEC 23090-3(Versatile Video Coding, 이하 VVC)[3]역시 사용할 수 있도록 표준화를 진행하고 있다. 그러나 이러한 동영상 압축 표준들은 일반적으로 자연의 영상들이나 컴퓨터 그래픽스 영상들을 목표로 개발되어 온 기법들이 대다수이다 보니, MPEG Immersive Video 와 같은 다양한 시점의 합성 영상들(이하 아틀라스)의 특성을 살리기 어렵다. 그래서 이러한 아틀라스 영상을 압축하려 할 경우, 효율성이 떨어지기 쉽다.

아틀라스 영상들에 대한 효율을 증대하기 위해 최근에 연구가 발표된 바 있다. 다시점의 합성 과정에서의 중복성 제거 기법과 잔차 신호의 표현 방식을 활용해서 실감미디어 압축의 효율을 끌어올리는 기법이 제안된 바 있다[4]. 또한, 서로 다른 동영상 압축 코덱의 활용을 통해 전체 아틀라스 영상 압축 효율을 평가한 연구도 최근에 발표되었다[5]. 특히 VVC 의 주요 기법 중 하나인 디블로킹 필터의 적용 여부를 비교함으로서, 아틀라스 영상에 필터의 효용성이 평가되었다. 이에 따르면 어떤 영상에서 적절히 디블로킹 필터를 적용하지 않을지 판단하는 것이 중요함을 알 수 있다.

본문에서는 어떠한 영상이 아틀라스 중심의 영상인지, 그리고 어떠한 영상이 전통적인 동영상 압축에 적합했던 영상인지를 구분하는 기법을 제안한다. 이를 위해, 아틀라스 중심과 그렇지 않은 영상들의 데이터셋을 만들고, 이를 CNN 으로 학습하여 실감미디어 영상을 분류하는 문제를 해결한다. 특히, 널리 알려진 VGG 신경망[6], ResNet50[7]의 효용성을 평가하고, 이 모델의 성능을 극대화하기 위하여 수정한 딥러닝 아키텍쳐를 제안한다. 실험을 통해 우리의 제안하는 VGG 수정 모델의 정확도가 95%에 이르는 것을 밝힌다.

2. 실감영상 분류 데이터셋

본 논문에서 사용한 데이터셋의 경우, 영상 이미지는 총 71 장으로 이 중 True 로 사용된 데이터는 33

*이 논문은 2020 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2020R1I1A3072227).

장, False 의 경우 38 장이 사용되었다. 영상의 해상도는 4096 x 4096, 2048 x 1536, 2048 x 1024, 1920 x 1024, 1536 x 1024, 1536 x 768 으로 여러 종류로 구성되어 있다. 본 연구에서는 한 장의 영상 이미지를 128 x 128 로 쪼개어 여러 장으로 분리한 후, 훈련과 검증에 대한 데이터셋을 구성한다. 이렇게 분할된 영상 데이터는 총 14520 장이고, 이에 대해 모두 True/False 로 레이블을 추가한 후, 사이킷런을 활용하여 훈련 데이터와 검증 데이터의 비율을 각각 70:30 으로 구성하였다.

3. 제안하는 CNN 모델

ILSVRC 와 같은 영상 분류 문제에서 뛰어난 성능을 보여왔던 VGG[6]와 ResNet50[7]를 초기 모델로 선정하여 실감영상 분류 문제를 풀 수 있다. 그러나 실감영상 데이터셋의 특수성 때문에 기존 모델의 구조를 그대로 가져오는 경우 정확도의 한계나 훈련이 제대로 이루어지지 않을 수 있다. 특히 VGG16 의 경우, 본문에서 사용한 실감영상 데이터셋의 이진분류문제에 대한 예측 값과 실제 값의 오차에 기반한 gradient 를 일반적인 back-propagation 으로 학습하는 경우, 자칫 기울기 소실 문제가 발생하여 학습이 진행되지 않을 수 있다. 따라서 본문에서는 기존 VGG16 모델에서 상위의 전결합층 뿐 아니라 기존 CNN 내부의 구조를 변경하는 모델을 제시한다. 또한 제안하는 수정된 VGG16 모델(modified VGG16, 이하 mVGG16)의 성능 비교를 위하여, 또 다른 state-of-the-art 인 ResNet50 모델 역시 수정한 결과를 제시한다.

mVGG16 의 경우 상기 서술한 기울기 소실 문제를 고려하여 다음과 같은 구조 수정을 제시한다. 미니배치의 평균과 분산을 이용하여 정규화 작업을 수행하는 작업인 배치 정규화(Batch Normalization)[8]를 기존 모델에 추가해주었다. 본 연구에 사용된 mVGG16 에서는 모든 컨볼루션층 이후에 배치 정규화층을 추가하였다. 또한, 본 연구의 목적에 맞게 이진 분류를 위해 마지막 출력층의 노드 수를 2 개로, 활성화 함수를 소프트맥스로 구성하였다. 손실함수는 categorical cross-entropy 로 설정하였다. 다음으로 ResNet 의 경우, mVGG16 과 마찬가지로 이진분류를 위하여 마지막 출력층의 노드 수를 2 개로, 활성화 함수를 소프트맥스로, 그리고 손실 함수를 categorical cross-entropy 로 구성하였다.

기존 딥러닝 방식과의 비교를 위하여, 일반적인 영상 분류를 위한 딥러닝 모델과 본문에서 제안하는 모델의 구조와 훈련 방식을 차이점 위주로 서술하겠다. 본문에서는 데이터의 입력 크기는 128 x 128 로 설정하였다. mVGG16 은 VGGNet16 에서 모든 컨볼루션층과 최대 풀링층 사이 배치 정규화층을 추가하였다. 객관적인 비교를 위해 mVGG16 과 ResNet 모델 모두 출력층의 노드 수와 활성화 함수, 그리고 손실 함수를 통일하여 훈련을 진행하였다. 아틀라스 영상의 경우 그렇지 않은 영상과는 다른 특성이 있을 것으로 추측하고 기존의 선 학습된 가중치(pre-trained weights)를 사용하지 않았음을 알린다.

표 1. mVGG16 과 ResNet 모델 비교

	정확도	모델 크기	학습시간 (1 epoch)
mVGG16	0.9552	65,080,642	64 s
ResNet	0.9516	23,591,810	43

이렇게 구성된 모델들의 학습을 진행한 결과, 표 1 과 같은 각 모델의 성능과 모델 크기, 학습시간에 대한 결과를 얻을 수 있었다.

3. 결론

본문에서는 딥러닝 모델을 사용하여 실감미디어 영상의 특수함을 분류하는 문제를 해결하고자 수정된 VGG16 모델을 제안하였다. 모델을 구동하기 여유로운 자원이 있다면 더 높은 정확도를 위해 mVGG16 모델을 사용하는 것을 추천한다.하지만 훈련 시간과 파라미터 수를 감안했을 때 한정적인 자원에서 모델을 구동해야 하는 경우에는 mVGG16 보다는 ResNet 을 사용하는 것이 조금 더 효과적일 수 있다. 반면, 본 데이터셋의 특성상 이진분류문제의 경우 상위의 전결합층 크기가 대폭 줄어듦에 따라, 일반적으로 알려진 ResNet 의 모델 크기에 대한 VGG16 대비 효용성이 본 기법에서는 상대적으로 줄어들었음을 알 수 있다.

4. 참고 문헌

[1] ISO/IEC JTC 1/SC 29/WG 11, “Committee Draft of MPEG Immersive Video,” Document ISO/IEC JTC 1/SC 29/WG 11/N19482, Jul. 3, 2020.

[2] High Efficiency Video Coding (HEVC), *Recommendation ITU-T H.265 and ISO/IEC 23008-2*, ITU-T and ISO/IEC JTC 1, Jan. 2013 (and subsequent editions)

[3] Versatile Video Coding (VVC), *Recommendation ITU-T H.266 and International Standard ISO/IEC 23090-3:2020*, ITU-T and ISO/IEC JTC 1, Aug. 2020.

[4] J. Jeong, S. Lee, D. Jang and E. Ryu, “Towards 3DoF+ 360 Video Streaming System for Immersive Media,” *IEEE Access*, vol. 7, pp. 136399-136408, 2019.

[5] H. Park, S. Park, and J.-W. Kang, “6 자유도 전방위 몰입형 비디오의 압축 코덱 개발 및 성능 분석,” 방송공학회논문지, vol. 24, no. 6, pp. 1035‑1052, Nov. 2019.

[6] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *Proc. ICLR*, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual

Learning for Image Recogntion," *Proc. CVPR*, 2016.
[8] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *Proc. ICML*, 2015.