타일링을 활용한 CompressAI 기반 딥러닝 네트워크 피쳐맵 압축

이민석, 이성배, 김규헌*

경희대학교

qsibmini@khu.ac.kr, rhee@khu.ac.kr, *kyuheonkim@khu.ac.kr

Application of Tiling Method on CompressAI based Deep Learning Network Feature Map Compression

Lee Minseok, Rhee Seongbae, Kim Kyuheon*

Kyung Hee Univ.

요 약

본 논문은 파이토치(PyTorch) 기반의 객체 검출(object detection) 네트워크의 태스크 수행 과정에서 생성되는 피쳐맵을 타일링(tiling) 기법 및 CompressAI[1]를 활용하여 효율적으로 압축하는 방안을 제안한다. 이를 위해, CompressAI 플랫폼에서 지원하는 모델 중 bmshj2018-factorized의 압축 네트워크를 활용하였으며, 압축 대상은 Detectron2[2]에서 제공하는 Faster R-CNN X101-FPN의 P 레이어(P Layer)에서 추출된 피쳐맵이었다. 본 논문에서 제안하는 압축 방안은 MPEG(Moving Picture Experts Group) VCM(Video Coding for Machines) 표준의 기존 앵커(anchor) 기술과 비교해 더욱 높은 압축률을 선보인다.

Ⅰ. 서 론

최근 딥러닝 네트워크를 통한 인공지능의 태스크 수행 능력은 급속도로 발전하고 있으며, 이에 따라 데이터 크기 및 연산량 또한 함께 증가하는 추세이다 [3]. 이러한 고비용의 딥러닝 네트워크는 연산 속도 및 메모리가 부족한 저성능의 엣지 디바이스에서 활용하기 어렵다는 단점이 존재한다. 이에 따라 저성능 장치에서도 고비용의 딥러닝 네트워크를 활용할 수 있도록 하는 연구가 여럿 진행 중이며, 기계를 위한 영상 부호화가 그 중 하나다. 기계를 위한 영상 부호화는 현재 국제 표준화 기구인 MPEG(Moving Picture Experts Group)에서 VCM(Video Coding for Machines)이라는 명칭으로 표준화가 진행중이며 [4], 사람이 아닌 기계가 이해할 수 있는 태스크 수행용 영상을 압축하여 전송함으로써 엣지 디바이스의 연산 및 메모리 부담을 줄이고자 하는 목표를 가진다.

현재 MPEG VCM의 표준화 과정이 진행됨에 따라 VVC(Versatile Video Codec)[5]를 활용한 피쳐맵 압축 등 다양한 압축 기술이 개발되고 있다. 그러나 MPEG VCM의 두 번째 트랙인 이미지 압축 기술에 비해 압축률 대비 태스크 수행 능력이 부족하다. 이러한 한계를 극복하고자 본 논문에서는 CompressAI 플랫폼의 bmshj2018-factorized 모델 및 타일링 기법을 활용하는 압축 방안을 제안한다.

Ⅱ. 본론

그림1은 본 논문에서 제안하는 압축 방안의 구조도이며, 서버(server) 단에는 태스크 수행, 피쳐맵 추출, 타일링, 그리고 신경망 기반 부호화 모듈이 있으며, 클라이언트(client) 단은 신경망 기반 복호화 모듈, 타일 분해, 그리고 완전 연결 계층(fully connected layer) 기반의 태스크 수행 모듈로 구성된다.

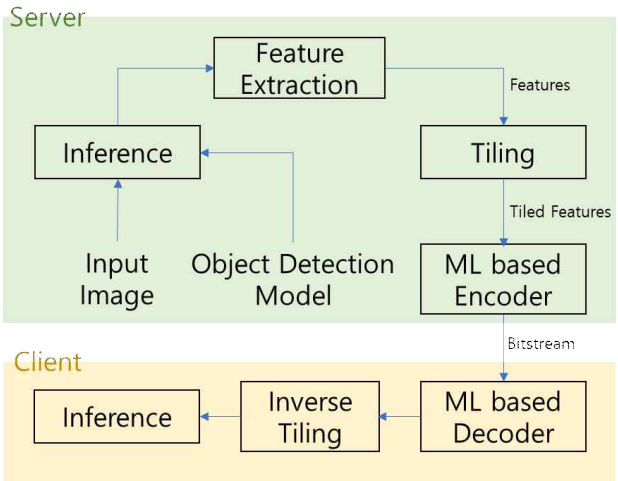


그림 1 제안 방안 구조도

A. 피쳐맵 추출

딥러닝 네트워크가 태스크를 수행할 시, 입력 영상에 대해 사전 학습된 컨볼루션 커널(convolutional kernel)을 기반으로 여러 피쳐맵이 생성된다. 각 컨볼루션 계층(convolutional layer) 별로 여러 채널의 피쳐맵이 생성되며, 최종적으로 이러한 피쳐들의 정보를 완전 연결 계층에서 활용하여 객체 검출 등의 태스크를 수행하게 된다. Detectron2에서 제공하는 객체 검출 모델인 Faster R-CNN X101-FPN의 경우, 완전 연결 계층 직전의 컨볼루션 계층을 P 레이어(P layer)라 부르며, 그림2와 같이 총 5가지의 계층크기를 가지는 각 256 채널의 패쳐맵이 생성된다. 이러한 피쳐맵을 추출하여 클라이언트에게 전송할 시, 클라이언트는 마지막 단계인 완전 연결 계층만을 활용하여 태스크를 수행할 수 있다. 이로써 컴퓨팅 파워가 부족한 클라이언트의 연산량에 대한 부담을 줄일 수 있다.

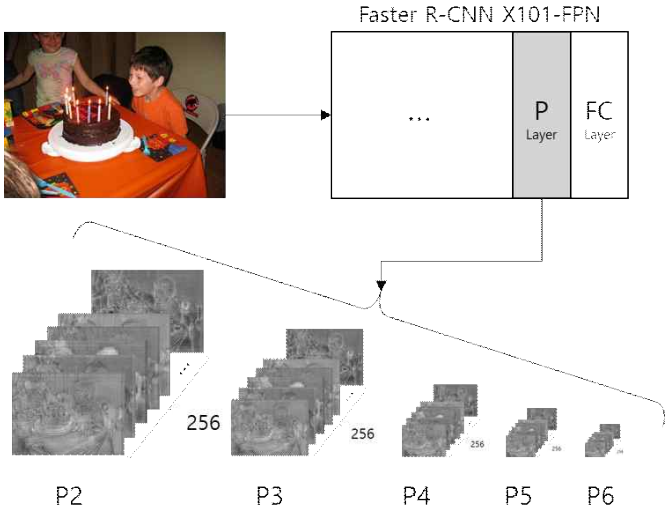


그림 2 피처맵 추출 예시

B. 타일링

타일링(tiling)은 그림3과 같이 여러 채널로 구성된 피처맵을 하나의 2D 프레임으로 만드는 작업으로써, MPEG VCM의 피처맵 압축 기술 앵커(anchor) 방법은 더 나은 압축 성능을 위해 타일링 기법을 쓰고 있다. 본 논문에서 제안하는 피처맵 압축 방법 또한 타일링 기법을 활용하며, 이는 딥러닝 기반 부호화기의 경우 차원 축소를 활용하기에 크기가 작은 입력 영상에 대해서는 복원 영상의 왜곡이 크기 때문이다.

P 레이어의 모든 피처맵을 하나의 영상으로 타일링하는 MPEG VCM의 방법과 달리 본 논문에서는 각 256채널을 따로 타일링하여 총 4장의 2D 프레임을 생성하며, 가장 큰 프레임은 절반으로 다운스케일링(down-scaling), 가장 작은 프레임은 업스케일링(upscaling)한다. 이때 프레임의 스케일링은 기본적으로 이중 선형보간법(bi-linear interpolation)을 활용하나, 가장 높은 정확도를 달성한 지점에서는 슈퍼레졸루션(super resolution) 기술인 EDSR(Enhanced Deep Super-Resolution Network)[6]을 활용하였다.

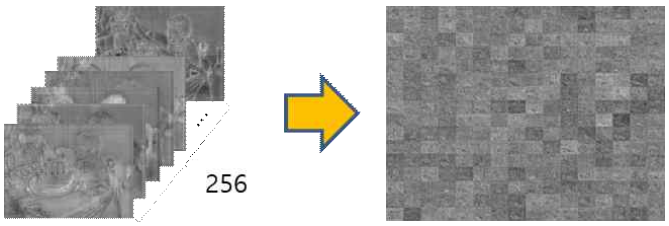
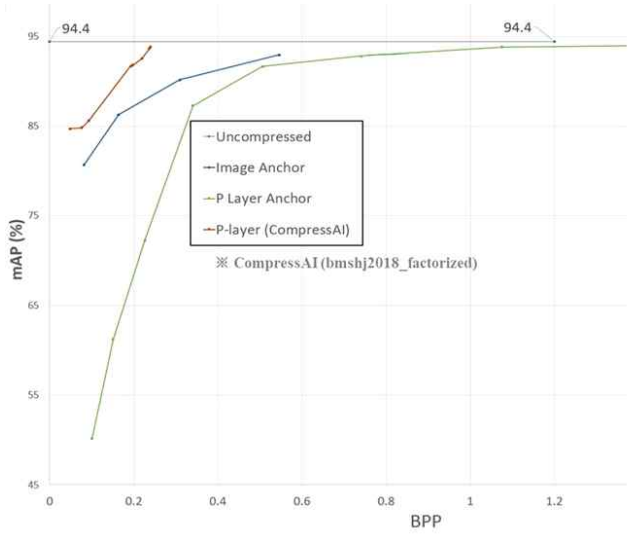


그림 3 피처맵 타일링 예시

Ⅲ. 실험 결과

실험은 그림1의 구조도를 따랐으며, 결과는 그림4와 같이 압축된 피처맵 비트스트림의 크기 대비 mAP(mean Average Precision)를 그래프로써 표현하였다. 여기서 mAP는 클라이언트 단의 객체 검출 태스크 수행의 정확도를 수치적으로 표현한 값이다. Image Anchor는 일반적인 정지 영상을 VTM-12.0[5]을 통해 압축하고 딥러닝 네트워크에 입력한 경우를 나타내며, P Layer Anchor는 모든 피처맵을 하나의 정지 영상에 타일링하여 VTM-12.0을 통해 압축한 경우를 나타낸다. 객체 검출 모델로 Detectron2에서 제공하는 Faster R-CNN X101-FPN을 사용하였으며, 신경망 기반 압축 모델로는 CompressAI 플랫폼의 사전 학습된 bmshj2018-factorized 모델을 사용하였다. 실험 데이터는 OpenImageV6 validation set[6] 중 임의로 선별한 100개의 이미지였다.



Ⅳ. 결론

본 논문에서는 타일링 및 CompressAI 플랫폼의 bmshj2018-factorized를 활용하여 피처맵의 압축을 구현하였다. 제안한 기술을 기반으로 피처맵을 압축할 시 기존 피처맵 압축 기법으로 달성하기 어려웠던 이미지 압축 기술의 Image Anchor보다 성능이 높아짐을 확인할 수 있다. 그러나 이는 일반 이미지 기반 사전 학습된 bmshj2018-factorized 모델을 활용한 결과이며, 타일링된 피처맵 데이터셋을 통해 학습시킨 압축 모델을 활용한다면 더욱 높은 압축률을 달성할 수 있을 것으로 사료된다.

ACKNOWLEDGMENT

This research was supported by the This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-0-02046) supervised by the IITP(Institute for Information & communications Technology Promotion).

참 고 문 헌

[1] CompressAI, https://github.com/InterDigitalInc/CompressAI

[2] Detectron2, https://github.com/facebookresearch/detectron2

[3] M. Lee, S. Rhee and K. Kim, "Variance based Averaging and Standard Normal Distribution for Improvement of SBT Quantization in Neural Network Coding," 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS), Ise, Japan, 2022, pp. 1-6, doi: 10.1109/SCISISIS55246.2022.10001990.

[4] S. Rhee, M. Lee, and K. Kim, "Analysis of Feature Map Compression Efficiency and Machine Task Performance According to Feature Frame Configuration Method," Journal of Broadcast Engineering, vol. 27, no. 3, pp. 318 - 331, May 2022.

[5] VVC, https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM

[6] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," 2nd NTIRE: New Trends in Image Restoration and Enhancement workshop and challenge on image super-resolution in conjunction with CVPR 2017

[7] https://storage.googleapis.com/openimages/web/download_v6.html