Deepstream 과 가중치 가지치기를 활용한 사람 자세추정 어플리케이션 가속화에 대한 연구

이혁상, 허수웅, 이성민, 조의현, 이상훈
연세대학교

{melungl, heartshape, lseong721, elf, slee}@yonsei.ac.kr

Accelerate Pose Estimation by using Deepstream and Weight pruning

Lee Hyucksang, Heo Suwoong, Lee Seongmin, Jo Uihyeon, Lee Sanghoon
Yonsei Univ

요 약

딥러닝을 활용한 사람 자세 추정 기술은 행동인지, 게임, 헬스케어 등 다양한 분야에서 활용되는 주목받고 있는 기술이다. 하지만 이러한 딥러닝 기반 사람 자세 추정과정에는 다양한 데이터 전처리 및 후처리 작업이 필요하고, 딥러닝 모델의 복잡도에 의해 전반적인 Application 의 실시간 성능이 떨어지게 된다. 본 논문에서는 딥러닝 기반 실시간 Application 개발 라이브러리인 Deepstream 을 통해 사람 자세 추정 Application 에서 생기는 전반적인 병목 현상을 완화하고, 객체 탐지 모델의 가중치 가지치기를 통해 자세추정 모델의 전처리모델을 경량화하여 기존 자세 추정 Application 보다 더 빠른 Application 을 제안한다.

Ⅰ. 서 론

최근 컴퓨터 비전 분야는 딥러닝의 발달로 많은 분야에서 이를 활용한 어플리케이션이 대세를 이루고 있다. 그 중에서 특히 사람의 자세 추정 분야는 현실의 사람의 모습을 가상세계로 옮길 수 있는 기반 기술로써 많은 관심을 받고 수많은 정확한 모델이 나오고 있다. 그리고 이는 실시간으로 가상세계에서 상호작용이 가능한 메타버스와 맞물려 포즈 추정 모델의 실시간 성능 또한 점차 중요성이 부각되고 있다.

딥러닝을 활용한 자세 추정 모델은 크게 하향식 방법과 상향식 방법이 존재한다. 이때 자세 추정의 정확도를 올리기 위해 사람이 위치한 영역을 먼저 찾고 그 안에서 자세를 추정하는 하향식 방법이 사용된다. 하지만 사람의 영역을 찾기 위해 사용되는 프로세스에 의해 그만큼의 수행시간이 늘어나 실시간 성능이 떨어지게 된다. 따라서 실시간 포즈 추정 어플리케이션의 실시간 성능을 올리기 위해서는 이러한 전처리 과정의 시간을 단축시킬 필요가 있다. 또한 위와 같이 추가적인 모델이 사용될 경우, 각 딥러닝 모듈 입력이 원하는 데이터의 형식을 맞춰주기 위한 전처리 및 후처리 작업이 늘어나게 되고, 이에 따라 CPU 와 GPU 메모리간 데이터 이동이 증가하면서 속도를 크게 떨어뜨리게 된다.

이에 따라 본 논문에서는 하향식 방법의 자세 추정 모델[1]의 전처리 과정인 객체 포착 딥러닝 모델[2]을 가중치 가지치기[3]를 통해, 딥러닝 모듈의 경량화를 진행하였다. 추가적으로 NVIDIA 의 실시간 딥러닝 Application 개발 라이브러리인 Deepstream[4]을 활용하여, Application 에서 진행되는 데이터의 흐름을 최적화시키고, 딥러닝 모델 또한 NVIDIA 그래픽 카드에 최적화시킨 TensorRT[5] 딥러닝 모델 형식으로 변환하여 보다 빠른 추론을 통해 전반적인 Application 의 실시간 성능을 올리고자 한다.

Ⅱ. 본론

본논문에서는 딥러닝 모델의 경량화를 진행하기 위해 가중치 가지치기 방법을 사용한다. 가중치 가지치기 방법은 크게 2 가지로 가지치기할 가중치를 0 으로 바꾸는 비구조적 가지치기와 가지치기할 가중치를 아예 지워 모델의 구조를 바꾸는 구조적 가지치기가 존재한다. 비구조적 가지치기의 경우 모델의 구조가 바뀌지 않기 때문에 기존 모델과 같은 메모리를 차지하여, 경량화에 따른 효과적인 성능향상을 보기 힘들다. 따라서 본 논문에서는 모델의 가중치를 아예 지워 그 모델의 구조를 바꾸는 방식인 구조적 가지치기를 진행하였다. 또한 모델 정확도에 보다 적은 영향을 주는 가중치를 가지치기하기 위해 배치 정규화 과정에서의 크기를 정하는 가중치가 작을수록, 다음 layer 에 영향을 적게 준다는 논문[3] 결과를 기반으로 가중치 가지치기를 진행하였다.

딥러닝의 발전으로 그에 따른 딥러닝 개발 프레임워크 또한 많이 발생하였다. 대표적으로 pytorch, tensorflow, caffe 등이 존재한다. 하지만 이러한 포맷들은 서로 호환되지 않고, 각 그래픽 카드에는 최적화되지 않았다는 단점이 존재한다. 따라서 본 논문에서는 GPU 카드 회사인 NVIDIA 에서 자체적으로 개발한 딥러닝 모델형식인 TensorRT[5]로 딥러닝 모델을 변환하여, 각 GPU 카드에 최적화된 연산을 지원받아 모델의 추론 속도를 향상시켰다.

일반적인 딥러닝 기반 Application 의 경우 딥러닝에 알맞은 입력을 넣어주기 위한 전처리 과정과 나온 딥러닝 결과를 시각적으로 보여주기 위한 후처리 과정이 필요하다. 일반적으로는 이러한 전처리와 후처리 작업은 CPU 에서 진행하게 되는데, 이럴 경우 전처리 작업과 후처리 작업에 관련한 데이터는 컴퓨터의 RAM 상에 딥러닝에 관련한 데이터는 GPU 메모리에 올리게 되면서 메모리간의 데이터 이동으로 인한 병목현상이 일어난다. 이러한 문제를 해결하기 위해 본 논문에서는 대부분의 과정을 GPU 메모리에서 해결할 수 있는 Deepstream[4] 라이브러리를 활용해 Application 의 전반적인 흐름을 최적화시켰다.

Ⅲ. 분석

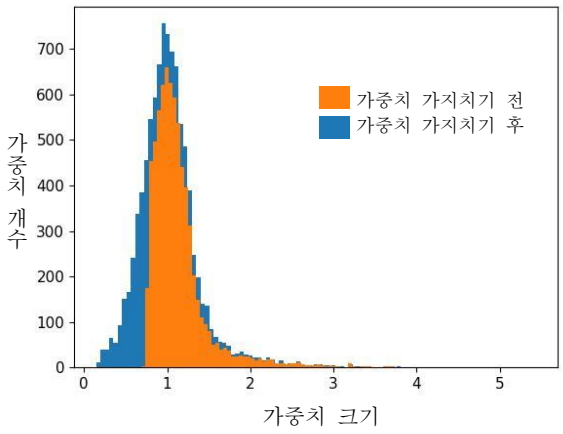

그림 1. Yolov5s 의 배치 정규화 가중치의 하위 20%를 가지치기한 히스토그램 결과

Yolov5s	기존	가지치기 후
가중치 수	7,235,389	4,456,115
파일 크기	14,462 KB	8,941 KB
실행 시간	22.3 ms	22.5ms

표 1. Yolov5s 의 가중치 가지치기를 진행 후 결과

실험을 위해 사용한 자세 추정 모델의 경우 하향식 방법의 회귀방법인 [1]를 사용하였고, 이에 대한 전처리과정인 객체 검출 모델은 Yolov5s[6]를 사용했다. 먼저 Yolov5s 의 하위 20%의 배치 정규화 가중치를 가지치기를 진행하여 그림 1, 표 1 과 같이 모델 경량화를 진행하였다. 이후 pytorch 기반 Yolov5s 모델과 가지치기한 Yolov5s 모델의 실시간 실시간 성능을 비교해보았다. 기존 모델의 경우 1 프레임당 처리시간이 22.3 밀리초이고, 가지치기한 모델의 경우 1 프레임당 처리시간이 22.5 밀리초로 속도 향상을 보지 못했다. 이는 가중치의 채널을 기반으로 가지치기를 진행하여, 병렬 연산량이 줄었지만, pytorch 라이브러리가 줄어든 딥러닝 연산을 효율적으로 적용하지 못한 것으로 추측된다.

표 2. 전처리 모델의 가중치 가지치기와 Deepstream 을 적용한 자세 추정 Application 성능 결과

Application method	Yolov5s (Pytorch)	Pruned Yolov5s (Pytorch)	Yolov5s (Deepstream)	Pruned Yolov5s (Deepstream)
전처리시간	22.3	22.5	4.12	3.56
자세 추정 시간	31.27	31.47	10.37	9.81

반면 Pytorch 기반 Yolov5s 를 TensorRT 로 변환하고 자세추정 Application 에 Deepstream 을 적용한 결과, 표 2 와 같이 전처리 과정인 yolov5s 의 실행 시간은 4.12 밀리초로 기존 pytorch 기반 yolov5s 보다 5.4 배의 속도 향상을 보여주었다. 또한 가중치 가지치기를 적용한 Yolov5s 를 TensorRT 로 변환하여 Deepstream 에 적용했을 경우에도 TensorRT 는 가지치기 한 가중치에 대해 딥러닝 연산을 효과적으로 최적화하여 속도 향상이 없었던 pytorch 기반에 비해 15%정도의 속도 향상이 있는 것을 확인하였다. 최종적인 실시간 자세 추정 Application 의 경우, 기존 31.27 밀리초에서 9.81 밀리초로 3.1 배의 속도 향상을 보여주었다.

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 딥러닝 자세추정 Application[1]의 속도 향상을 위해 전처리 과정인 객체 검출 모델[6]을 가중치 가지치기[3]로 경량화시키고, Deepstream[4]과 TensorRT[5]를 활용하여 전반적인 Application 의 흐름 최적화와 딥러닝 모델의 연산 최적화를 진행해 전처리 속도는 5 배 이상 전반적인 Application 속도는 3 배 이상 향상시켰다. 이후 실시간 행동인지 Application 과 같이 자세추정 모델 기반의 다양한 응용시스템에도 적용하여 활용될 수 있을 거라고 보인다.

ACKNOWLEDGMENT

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00352, 설명 가능한 감성 경험 예측 모델 기반 콘텐츠 평가 기술 개발 및 상용화)

참 고 문 헌

[1] Li, Jiefeng, et al. "Human pose regression with residual log-likelihood estimation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[2] REDMON, Joseph, et al. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 779-788

[3] Liu, Zhuang, et al. "Learning efficient convolutional networks through network slimming." Proceedings of the IEEE international conference on computer vision. 2017

[4] NVIDIA Corporation, DeepStream SDK [Online]. Available: https://developer.nvidia.com/deepstream-sdk

[5] NVIDIA Corporation, TensorRT SDK [Online]. Available: https://developer.nvidia.com/tensorrt

[6] Jocher, G. (2020). YOLOv5 by Ultralytics (Version 7.0) [Computer software]. https://doi.org/10.5281/zenodo.3908559