엣지 컴퓨팅을 위한 효율적인 SSM 기반 군중 집계

이현빈[1], 이장호[2]
[1]인천대학교 컴퓨터공학부 석사과정
[2]인천대학교 컴퓨터공학부 교수
{fedora, ubuntu}@inu.ac.kr

Efficient SSM-based Crowd Counting for Edge Computing

Hyeonbeen Lee[1], Jangho Lee[2]
[1,2]Dept. of Computer Science and Engineering, Incheon National University

요 약

실시간 군중 집계(Crowd Counting)는 군중 붕괴 등의 안전사고 예방을 위해 필수적인 기술이다. 지연 시간을 낮추고 현장에서 즉각적인 대응을 가능하게 하려면 종단 기기(Edge Device)에서 직접 추론을 하는 것이 효과적이다. 본 연구는 기존 합성곱 신경망(CNN, Convolutional Neural Network)이 주도하는 이 분야에 상태 공간 모델(SSM, State Space Model) 기반의 Vision Mamba(Vim) 아키텍처를 적용하여 새로운 가능성을 탐색한다. ShanghaiTech와 UCF-QNRF 데이터셋을 이용한 정확도 비교 실험에서, 제안 모델은 CNN 모델과 대등한 정확도를 보이면서도 추론 속도는 약 21% 단축했으며, 메모리 사용량은 27% 수준으로 대폭 감소시키는 등 뛰어난 효율성을 달성했다. 이는 SSM이 실시간 군중 집계 작업에 있어 강력한 잠재력과 실효성을 가졌음을 입증한다.

1. 서 론

군중 집계는 이미지나 영상 내 인원의 수를 추정하는 컴퓨터 비전 기술로, 군중 밀집으로 인한 붕괴 사고 예방, 도시 계획, 차량 흐름 제어 등 공공 안전 및 사회 기반 시설 관리의 핵심 요소로 자리 잡고 있다. 전통적으로 서버 환경에서 수행되던 군중 집계는 지능형 감시카메라(Intelligent CCTV)와 같은 종단 기기에 직접 탑재될 때 그 효용성이 극대화된다. 엣지 컴퓨팅(Edge Computing) 환경은 네트워크 의존성을 줄이고 데이터 전송 지연 없이 현장에서 즉각적인 분석과 대응을 가능하게 하여, 재난 상황에서 신속한 의사결정을 지원하는 데 효과적이다.

하지만 종단 기기는 서버보다 연산 능력, 메모리, 전력 등의 컴퓨팅 자원이 현저히 제한된다. 따라서 엣지 컴퓨팅에 적합한 군중 집계 모델은 높은 정확도를 유지하면서도 적은 메모리 사용량과 빠른 추론 속도를 가져야 하는 상충 관계(Trade-off)를 조율해야 한다. 이러한 문제를 해결하기 위해 기존에는 CNN 기반 모델의 경량화 연구가 활발히 진행되었다.

본 연구는 여기서 한 걸음 나아가, 최근 CNN과 Vision Transformer가 주도하는 파운데이션 모델의 대안으로 주목받고 있는 SSM을 군중 집계에 적용하는 새로운 접근법을 제안한다. 특히 긴 스퀀스 데이터를 선형 복잡도로 처리하며 뛰어난 효율성을 보이는 Mamba[1]와 이를 비전 작업에 맞게 발전시킨 Vision Mamba[2] 아키텍처에 주목한다. Vim은 메모리 사용률을 절감하고 높은 연산 효율을 확보하면서도 기존의 CNN, Transformer 기반의 비전 모델 중 SOTA(State-of-the Art) 모델들의 성능에도 근접한

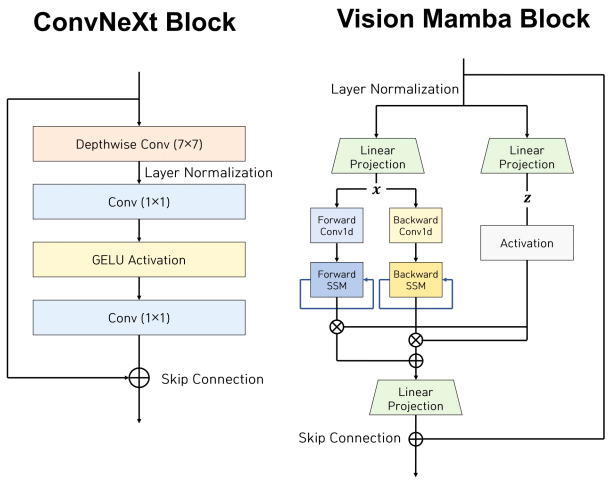

그림 1 ConvNeXt 블록과 Vision Mamba 블록. ConvNeXt 블록은 7×7 크기의 합성곱 커널을 사용한다. Vision Mamba 블록은 양방향 SSM을 통해 전역적 맥락을 파악하고 게이트로 핵심 정보를 추출한다

표 1 실험에 활용된 데이터셋의 통계

데이터셋	평균 해상도	이미지 수	최대 군중 수	최소 군중 수	평균 군중 수	총 머리 주석의 개수
SHHA	589 × 868	482	3139	33	501	241,677
SHHB	768 × 1024	716	578	9	124	88,488
UCF-QNRF	2013 × 2092	1535	12865	-	815	1,251,642

결과를 입증한 바가 있다.

본 논문에서는 Vision Mamba를 기반으로 한 군중 집계 모델과 ConvNeXt를 기반으로 한 모델과의 성능을 비교 분석한다. 이를 통해 SSM이 엣지 컴퓨팅 환경에서 요구되는 정확성과 효율성 간의 균형을 효과적으로 달성하여, 차세대 지능형 감시 시스템을 위한 강력한 대안이 될 수 있음을 입증하고자 한다.

2. 관련 연구

2.1 기계학습을 통한 군중 집계

군중 집계 작업은 감지 기반(Detection-based) 방식, 회귀 기반(Regression-based) 방식, 밀도 지도 추정 기반(Density Map Estimation-based) 방식으로 분류할 수 있다.

감지 기반 방식[3]은 보행자의 신체 전체 또는 머리 등 특정 부분을 객체 탐지 알고리즘으로 식별한 후 그 수를 세는 가장 직관적인 방법이다. 심층 학습 기반 탐지기의 발전으로 낮은 밀도의 군중에서는 높은 정확도를 보이지만, 군중 밀도가 높아져 폐색(occlusion)이 심해지면 성능이 급격히 저하되고 회귀 기반 방식에 비해 상대적으로 연산 비용이 높다는 한계가 있다.

회귀 기반 방식[4]은 이미지의 전경, 질감 등의 저수준 특징을 추출하여 프레임과 사람 수의 회귀적 관계를 학습하는 방식이다. 이 방식은 감지 기반 방식보다 고밀도 환경에서 상대적으로 강건하고 연산 비용이 적은 특징이 있다.

밀도 지도 추정 방식[5]은 회귀 기반 방식의 일종으로 이미지 내 각 개인의 머리 위치 좌표를 기반으로 가우시안 커널 등을 적용하여 연속적인 확률 분포 형태로 밀도 지도를 생성한다. 이후 CNN과 같은 심층 신경망을 입력 이미지에 해당하는 밀도 지도를 추정하도록 학습시키고, 지도의 전체 픽셀값을 적분하여 최종 인원수를 산출한다. 이 방식은 공간적 정보를 보전하면서 고밀도의 및 폐색 환경에 효과적으로 대응할 수 있어서 현재 군중 집계 연구에서 주류인 방식이다.

2.2 심층 학습에 적용된 상태 공간 모델

SSM은 전통적인 제어 이론에서 유래한 개념으로, Gu가 제안한 S4[6]가 긴 스퀀스 데이터 모델링에서 Transformer의 대안이 될 가능성을 제시하며 주목받기 시작했다. S4 등의 초기 연구들은 SSM이 시퀀스 길이에 대해 선형적인 연산 확장성을 가지면서도 장거리 의존성 포착에 효과적임을 보여주었다.

Mamba[1]는 이러한 SSM 연구의 최신 성과로, 입력 데이터 따라 동적으로 파라미터를 조정하는 선택적 메커니즘(Selective Mechanism)과 하드웨어에 최적화된 병렬 스캔 알고리즘을 도입했다. 이를 통해 Mamba는 스퀀스 길이에 대해 선형 복잡도 $O(n)$을 달성하며, 기존 Transformer의 어텐션 메커니즘이 가지는 이차 복잡도 $O(n^2)$에 비해 복잡성을 대폭 완화할 수 있었다. 이를 통해 Mamba는 자연어 처리 분야에서 Transformer에 비견되는 성능과 월등한 효율성을 보이며 차세대 파운데이션 모델의 가능성을 열었다.

Vision Mamba[2]는 Mamba의 성공을 컴퓨터 비전 분야로 확장한 연구이다. 비전 데이터는 텍스트와 달리 2차원 공간 정보가 중요하므로, Vim은 이미지를 패치 스퀀스로 변환한다. 양방향 SSM(Bidirectional SSM)과 위치 임베딩(Positional Embedding)을 도입하여 공간적, 전역적 맥락을 효과적으로 학습하도록 설계되었다. Vim은 기존 Vision Transformer(DeiT) 대비 13.2% 수준의 GPU 메모리를 사용하면서도 2.8배 빠른 추론 속도와 더 높은 성능을 달성했다. 이러한 압도적인 효율성은 종단 기기와 같이 자원이 제한된 환경에서 군중 집계와 같은 컴퓨터 비전 작업을 수행하는 데 있어 SSM 기반 아키텍처가 매력적인 선택지가 될 수 있음을 시사한다.

3. Vision Mamba를 활용한 군중 집계

표 2 ViMCC와 CNXCC의 비교 결과

모델명	SHHA	SHHB	UCF-QNRF	평균 응답 속도 (ms)	GFLOPs	GPU 메모리 사용량 (MB)	Params
VimCC	**100.3**/173.8	24.0/42.4	163.9/**280.6**	**13.07**	**69.57**	**156.81**	32,642,417
CNXCC	101.3/**172.0**	**20.2**/**35.3**	**156.2**/284.8	16.50	358.07	576.14	38,123,489

본 연구에서는 Vision Mamba를 인코더로 활용하는 밀도 지도 추정 방식의 군중 집계 모델 ViMCC(Vision Mamba for Crowd Counting)를 설계했다. 이 모델은 SSM의 선형 복잡도 특성을 계승하여, 제한된 컴퓨팅 자원 내에서 정확성과 효율성 간의 균형을 달성하는 것을 목표한다. 모델의 전체 구조는 인코더-디코더 형태로 구성된다. 인코더 부분에는 Vim 아키텍처를 채택하여 입력 이미지로부터 깊이 있는 특징을 추출한다. 디코더는 UperNet 스타일의 업샘플러를 공통적으로 사용하여 구조를 통일했다. 밀도 지도 추정 성능에 인코더 아키텍처가 미치는 영향을 공정하게 비교하기 위함이다.

ViMCC의 성능을 CNN 기반 구조와 비교하기 위해서, 현대적인 CNN 구조인 ConvNeXt[7]를 인코더로 사용한 모델인 CNXCC(ConvNeXt for Crowd Counting)를 비교군으로 설정했다. 두 모델은 유사한 파라미터 규모를 갖도록 설계하였으며, 동일한 UperNet 디코더를 연결하여 아키텍처 자체의 성능을 순수하게 비교할 수 있도록 실험 환경을 구성했다.

4. 실험 결과

4.1 실험 환경 설정

데이터셋 본 연구는 널리 사용되는 세 가지 공개 군중 집계 벤치마크 데이터셋을 사용하여 모델의 성능을 평가했다. 고밀도 군중 환경을 포함하는 ShanghaiTech Part A(SHHA)와 상대적으로 저밀도 환경인 ShanghaiTech Part B(SHHB), 그리고 매우 넓은 화각과 다양한 밀도를 가진 도전적인 UCF-QNRF 데이터셋을 활용했다. 실험에 사용된 데이터셋에 대한 자세한 통계 지표는 표 1에 기술하였다.

평가 지표 모델의 정확도는 평균 절대 오차(MAE, Mean Absolute Error)와 평균 제곱 오차(MSE, Mean Squared Error)를 통해 측정했다. 모델의 효율성은 초당 기가 부동소수점 연산(GFLOPs), 추론에 소요되는 평균 시간, 그리고 GPU 메모리 사용량을 통해 다각적으로 평가했다.

학습 설정 두 모델은 동일한 하이퍼파라미터로 학습되었다. Optimizer는 AdamW, Learning rate는 1×10^{-4}, Batch size는 6으로 설정했다. 총 500 Epoch 동안 학습을 진행했으며, 300 Epoch부터 25 Epoch마다 0.99의 비율로 Learning rate를 감소시키는 Decay 스케줄을 적용했다. 추론 시간 측정은 RTX 4090 GPU 환경에서 1,000회 반복 실행한 후 평균값을 기록했다.

4.2 성능 비교 분석

정확도 비교 결과 실험 결과, ViMCC는 모든 데이터셋에서 CNXCC와 비슷한 성능을 보였다. 이 결과는 SSM 구조가 CNN 구조만큼 군중 집계에 적합하다는 것을 입증한다. 자세한 결과는 표 2에서 확인할 수 있다.

연산 효율성 비교 결과 ViMCC가 두각을 나타낸 부분은 효율성 측면이다. CNXCC와 비교하였을 때 GFLOPS는 19.4% 수준에 불과했으며, 평균 추론 시간은 약 20.8% 더 빨랐다. GPU 메모리 사용량은 CNXCC의 27.2% 수준으로 엣지 컴퓨팅에서의 주요 지표에서 ViMCC는 CNXCC에 비해 유의미한 격차를 보였다.

5. 결론 및 향후 연구

본 연구에서는 종단 기기에 최적화된 실시간 군중 집계 모델의 새로운 대안으로, 기존 CNN과 차별화되는 SSM 기반의 Vision Mamba 아키텍처를 활용하는 모델을 제시하였고 그 잠재력을 검증했다. 실험 결과, Vision Mamba 기반 모델은 ConvNeXt 기반 모델과 대등하거나 더 나은 정확도를 보이면서도, 추론 시간은 약 21% 단축했으며 메모리 사용량은 27% 수준으로 대폭 감소시켜 제한된 환경에서의 운용 가능성을 명확히 입증했다.

이러한 결과는 SSM 기반 구조가 정확도와 효율성 간의 상충 관계를 효과적으로 해결하며, 실시간 군중 집계 작업에서 CNN의 강력한 대안이 될 수 있음을 의미한다. 향후 SSM 기반 구조를 군중 집계에 더욱

적합하도록 조정하여, 최고 수준의 성능을 가지면서도 효율적인 모델을 제안하는 것을 목표로 한다.

사사문구

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학·석사연계ICT핵심인재양성 지원을 받아 수행된 연구임(IITP-2025-RS-2024-00437024).

참고문헌

[1] GU, Albert; DAO, Tri., Mamba: Linear-Time Sequence Modeling with Selective State Spaces, First Conference on Language Modeling, -, 2023, -.

[2] ZHU, Lianghui, et al., Vision mamba: efficient visual representation learning with bidirectional state space model, Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 2024, pp. 62429-62442.

[3] VIOLA, Paul; JONES, Michael; SNOW, Daniel, Detecting pedestrians using patterns of motion and appearance, Proceedings ninth IEEE international conference on computer vision, Nice, France, 2003, pp. 734-741.

[4] IDREES, Haroon, et al., Multi-source multi-scale counting in extremely dense crowd images, Proceedings of the IEEE conference on computer vision and pattern recognition, Portland, USA, 2013, pp. 2547-2554.

[5] LEMPITSKY, Victor; ZISSERMAN, Andrew, Learning to count objects in images, Advances in neural information processing systems, 23, -, pp. 1361-1369, 2010.

[6] GU, Albert; GOEL, Karan; RE, Christopher, Efficiently Modeling Long Sequences with Structured State Spaces, International Conference on Learning Representations, Virtual Event, 2022, -.

[7] LIU, Zhuang, et al., A convnet for the 2020s, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, USA, 2022, pp. 11976-11986.