

Crowd Counting 경량화를 위한 Knowledge Distillation 적용 연구

홍연주^{1,†}, 전해령^{2,†}, 김유연^{1,†}, 강현우^{1,†}, 박민균^{1,†}, 이경준³
^{1,†}동국대학교 컴퓨터공학전공 학부생, ^{2,†}연세대학교 응용통계학과 학부생,
³고려사이버대학교 정보관리보안학과 외래교수, [†]공동 1저자
 yeonjoo9928@dgu.ac.kr, jjhr0113@yonsei.ac.kr, 1rladbdus@dgu.ac.kr,
 tngkr7410@dgu.ac.kr, pkg0330@dgu.ac.kr, kjune83@cuk.edu

Research on apply to Knowledge Distillation for Crowd Counting Model Lightweight

Yeon-Joo Hong^{1,†}, Hye-Ryung Jeon^{2,†}, Yu-Yeon Kim^{1,†}, Hyun-Woo Kang^{1,†},
 Min-Gyun Park^{1,†}, Kyung-June Lee³

^{1,†}Dept. of Computer Science and Engineering, Dongguk University,

^{2,†}Dept. of Statistics and Data Science, Yonsei University,

³Dept. of Information Management and Security, the Cyber University of Korea

[†]These authors contributed equally to this work

요 약

딥러닝 기술이 발전함에 따라 모델의 복잡성 역시 증가하고 있다. 본 연구에서는 모델 경량화를 위해 Knowledge Distillation 기법을 Crowd Counting Model에 적용했다. M-SFANet을 Teacher 모델로, 파라미터수가 적은 MCNN 모델을 Student 모델로 채택해 Knowledge Distillation을 적용한 결과, 기존의 MCNN 모델보다 성능을 향상했다. 이는 정확도와 메모리 효율성 측면에서 많은 개선을 이루어 컴퓨팅 리소스가 부족한 기기에서도 본 모델을 실행할 수 있어 많은 활용이 가능할 것이다.

1. 서론

딥러닝 모델은 다양한 분야에서 성공적으로 활용되고 있고, 성능과 정확도 역시 개선됨에 따라 모델의 복잡도가 증가해 계산량 역시 증가하고 있다. 대규모의 딥러닝 모델은 수많은 파라미터와 연산으로 인해 모델 배포 및 실행에 부담이 된다. 특히 모바일, 임베디드, IoT 시스템의 경우 컴퓨팅 리소스가 부족해 모델 실행에 어려움이 있다. 예를 들면, 지하철 내의 혼잡도 판별을 위해 캠을 여러개 설치한 뒤 이미지를 서버로 전송해 혼잡도를 판별하는 것은 네트워크 용량 한계로 구현이 어렵다. 그러나 경량화된 모델로 임베디드 기기에서 혼잡도를 판별하고, 그 결과만 서버에 전송하여 구현이 가능하다.

본 논문에서는 이에 활용되는 모델 경량화를 위해 Knowledge Distillation[1] 기법을 Crowd Counting Model에 적용하고자 한다.

2. 실험 방법

2.1 Knowledge Distillation 및 적용 모델 선정

선행연구에 따르면, 경량화 기법 중 Knowledge Distillation 방법이 정확도 저하(Acc drop)와 연산량

감소(FLOPs drop) 에서 가장 우수한 성능을 보였다 [2]. 이를 근거로 Knowledge Distillation 경량화 방법을 채택해 실험을 진행하였다. 17가지 Crowd Counting 모델을 조사했으며 그 중에서 널리 쓰이는 ShanghaiTech A 데이터셋[3]에 대해 파라미터수가 적고 성능이 낮은 MCNN[3]을 Student 모델로 선정하였다. MCNN은 기하학적으로 적응형 커널을 사용하여 density map을 생성한다. Teacher 모델로는 출력하는 결과값 형태가 동일하면서 성능이 높은 M-SFANet[4]을 선정했다. 이는 ASPP와 CAN 모듈이 추가된 모델로, ASPP는 다중 스케일 특징을 추출하고, CAN은 컨텍스트의 스케일 변동을 적응적으로 처리한다.

2.2 성능 평가 지표

성능 평가지표는 MAE(Mean Absolute Error)와 RMSE(Root Mean Squared Error)를 사용한다.

$$MAE = \frac{1}{N} \sum |(pred_i - target_i)|$$

$$RMSE = \sqrt{\frac{1}{N} \sum (pred_i - target_i)^2}$$

3. 구현

3.1 Model

Model	Size (MB)	Loss	
		MAE	RMSE
M-SFANet (Teacher)	366.6	57.55	94.48
MCNN (Student)	0.532	110.2	173.2
MCNN with KD	0.543	90.86	172.99

<표 1> Knowledge Distillation에서 Teacher 모델과 Student 모델의 성능 비교

Teacher 모델의 경우 MAE가 57.55, RMSE가 94.48로 가장 높은 예측 성능을 보였으나 모델 크기가 366.6MB로 매우 크다. Student 모델의 경우 비교적 낮은 MAE(110.2)와 MSE(173.2) 값을 가지며, Teacher 모델 대비 예측 성능이 낮다. Student 모델에 Knowledge Distillation을 적용한 결과 MAE는 90.86, MSE는 172.99로, 일반 MCNN 대비 더 낮은 오차 값을 보였으며, 크기는 단지 0.011MB만 증가해 Teacher Model 대비 저용량 특성을 지속적으로 유지한다는 것을 확인할 수 있었다.

3.2 Training

학습 과정은 크게 두 단계로 구성하였다. 첫 번째 단계에서 Teacher Model인 M-SFANet을 미리 학습시켰다. 이후, 학습된 모델을 활용하여 Student 모델인 MCNN의 학습에 이용할 손실 함수를 설계했다. 손실 함수는 두 부분으로 이루어져 있는데, 하나는 Ground Truth 값과 Student 모델의 출력 간의 MSE(original_loss)이고 다른 하나는 Student 모델과 Teacher 모델의 출력 간의 MSE(distill_loss)이다. 이 두 손실 값을 결합할 때는 하이퍼 파라미터인 α 를 사용하여 두 손실의 비중을 조절한다. 즉, 최종 손실 함수 Loss는 $(1-\alpha) \times \text{original_loss} + \alpha \times \text{distill_loss}$ 로 정의된다.

기존 Knowledge Distillation 방식은 출력값에 T(temperature) 값을 나누어 소프트맥스 함수를 적용하지만, 본 연구의 모델은 소프트맥스를 사용하지 않아 T를 사용하지 않고 실험을 진행하였다. 또 모델 성능 향상을 위해 Cosine Annealing Warm Restarts[5] 스케줄링 기법을 사용하였다.

3.3 Experiments (MCNN with KD)

Loss	α (Alpha)			
	0.1	0.3	0.5	0.7
MAE	90.86	95.85	105.55	110.85
RMSE	172.99	177.71	217.68	202.47

<표 2> 알파(Alpha) 값 변화에 따른 성능 평가 지표(MAE와 RMSE)의 영향

Knowledge Distillation를 적용한 실험 결과, α 를 0.1로 설정했을 때 가장 좋은 성능을 보였다. 이 실험 결과, 제안된 모델은 기존 모델에 비해 99.85%의 모델 크기 감소를 이루면서도 정확도는 17.55%의 성능 향상을 보였다.

4. 결론

본 연구에서는 임베디드 실시간 환경에서 Crowd Counting을 처리하기 위해 경량화된 모델을 성공적으로 제안하였다. 이 모델은 M-SFANet 대비 저용량 특성을 유지하면서도 기존 MCNN 모델보다 성능이 향상되었기 때문에, 컴퓨팅 리소스가 제한된 임베디드 컴퓨팅 환경 등에서 사용에 적합하다. 추후 연구에서는 보다 다양한 환경과 조건에서 경량화된 Crowd Counting 모델의 최적화와 그 효용성을 더욱 검증할 필요가 있다.

※ 본 프로젝트는 과학기술정보통신부 정보통신 창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." NIPS 2014 Deep Learning Workshop. 2015.
 [2] 서기성, 딥러닝 경량화를 위한 구조, 가지치기, 지식중류 비교. 전기학회논문지, 2021, 70(12), pp. 1934-1939
 [3] Y. Zhang, D et al, "Single-Image crowd counting via multi-column convolutional neural network", Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 589 - 597.
 [4] P. Thanasutives, K. Fukui, M. Numao and B. Kijisirikul, "Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting," in 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 2382-2389.
 [5] I. Loshchilov, F. Hutter, "SGDR: Stochastic gradient descent with warm restarts". ICLR 2017 conference paper. 2016.