DQ-ResUNet: 의료 영상 분할의 효율성 개선을 위한 동적 양자화 기반 최적화

홍인표[01] 조영완[2] 안성현[2] 김은지[2] 권세인[2] 박상현[†2]
[1]가천대학교 컴퓨터공학과
[2]연세대학교 컴퓨터과학과
hip9863@gachon.ac.kr, {jyy1551, skd, kejh66, seinkwon97, sanghyun}@yonsei.ac.kr

DQ-ResUNet: Optimization Based on Dynamic Quantization for Improving the Efficiency of Medical Image Segmentation

Inpyo Hong[01] Youngwan Jo[2] Sunghyun Ahn[2] Eunji Kim[2] Sein Kwon[2] Sanghyun Park[†2]
[1]Department of Computer Engineering, Gachon University
[2]Department of Computer Science, Yonsei University
hip9863@gachon.ac.kr, {jyy1551, skd, kejh66, seinkwon97, sanghyun}@yonsei.ac.kr

요 약

의료영상 분할을 위한 인공지능 연구는 활발히 연구되고 있으나, 의료분야의 인공지능 상용화 단계에서는 모델 추론속도 및 컴퓨팅 자원 사용량 관련 한계가 존재한다. 본 연구에서는 동적 양자화 기반 최적화 기법을 활용하여 분할 성능은 유지하면서, 모델 추론속도 및 컴퓨팅 자원 사용량 관련 한계를 개선하는 DQ-ResUNet을 제안한다. DQ-ResUNet은 합성곱 신경망의 32bit 실수형 가중치를 8bit 정수형 가중치로 양자화하며, 활성화 함수의 값 또한 동적으로 양자화하는 구조를 지닌다. Kvasir-SEG 벤치마크 데이터셋으로 실험 및 검증을 진행한 결과, DQ-ResUNet은 양자화를 진행하지 않은 ResUNet과 분할 성능은 동일하게 유지하면서 6.83 millisecond (13.37%)의 추론속도 절감 및 382.19 megabyte (208.55%)의 메모리 사용량 절감을 확인하였다. 본 연구를 기반으로 의료영상의 분할, 정합, 변환 등 다양한 의료영상 분석에서 인공지능 상용화 연구가 더욱 활발히 수행되기를 기대한다.

1. 서 론

의료분야의 영상 데이터는 환자의 개인정보 유출, 데이터 접근성 제한 등의 한계로 일반 영상 데이터보다 인공지능 학습에 추가적인 제약조건이 존재한다. 하지만, 이러한 한계에도 불구하고 의료영상 분할(medical image segmentation)의 성능 향상을 위해 U-Net[1]을 기반으로 ResUNet[2], TransU-Net[3] 등 다양한 구조의 모델들이 제안되며 활발한 연구가 수행 중이다. 특히, 최근 의료영상 분야는 데이터 증강(data augmentation), 전이 학습(transfer learning) 등을 바탕으로 의료 데이터의 한계를 해결할 수 있는 점진적인 연구가 수행되고 있으며, 고도화된 성능을 요구하는 의료분야에서도 적합한 인공지능 모델들이 제안되고 있다[4]. 하지만 이러한 발전에도 불구하고 인공지능 기술을 의료분야에 바로 상용화하기에는 어려움이 존재한다. 의료영상 분석의 고도화를 위한 알고리즘이 점차 복잡해짐에 따라 컴퓨터 자원 사용량 또한 증가하고 있기 때문이다. 의료분야의 IT기기는 컴퓨팅 성능이 매우 한정적이며, 환자의 데이터셋 수집으로 많은 비율의 자원을 사용하기 때문에 크기가 큰 인공지능 모델까지 활용하기 어려운 실정이다. 따라서, 의료영상을 다루는 인공지능 모델은 성능은 유지하면서 컴퓨팅 자원은 최소화하는 효율적인 컴퓨팅 기술이 요구된다.

효율적인 컴퓨팅을 위한 인공지능 모델 측면의 최적화(model compression) 방안으로는 가지치기(pruning)[5], 지식증류(knowledge distillation)[6], 양자화(quantization)[7] 등이 제안되고 있다. 가지치기의 경우 모델 학습 시 중요 파라미터는 유지하며, 중요하지 않은 파라미터의 경우 반영치를 줄이거나 제거하는 기법이다. 가지치기는 파라미터가 줄어들기 때문에 추론 속도가 향상된다는 장점이 존재하지만, 정보 손실이 발생할 수 있다는 한계 또한 존재한다. 지식증류 기법은 높은 성능을 가지는 대규모 모델(teacher model)의 지식(knowledge)을 실제 추론에 활용되는 경량 모델(student model)로 전이(transfer)하는 기법이다. 지식증류는 모델의 경량화 및 최적화에 매우 효과적인 방안이지만, 성능적으로 우수한 대규모 모델이 학습에 필요하다는 한계가 있다. 양자화 기법은 가지치기 기법과 지식 증류의 한계를 개선할 수 있는 기법으로, 모델 성능에는 부정적인 영향을 끼치지 않으면서 신경망(neural network)의 32bit 부동소수점 가중치 매개변수(floating point weight parameter)를 8bit 정수형(integer type)으로 변환하는 기법이다. 이를 통해 모델에 사용되는 매개변수를 작은 bit로 재표현(representation)함으로써 추론(inference) 속도와 메모리 사용량(memory usage)을 절감시킬 수 있다. 이러한 다양한

* 이 논문은 2024년도 정부(과기정통부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00229822).
† 교신저자: sanghyun@yonsei.ac.kr

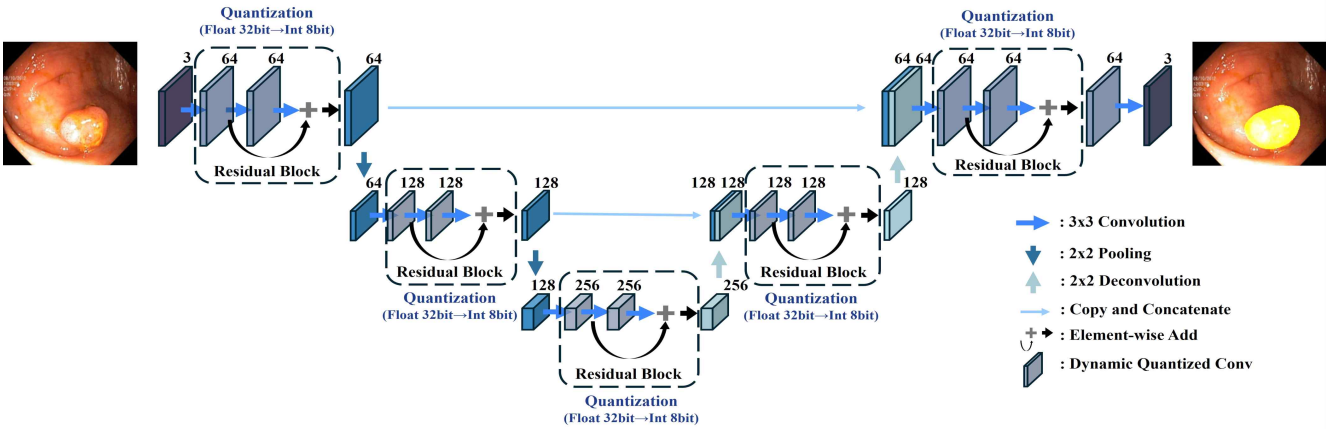

그림 1. DQ-ResUNet 모델구조

기법들을 인공지능 모델에 적용함으로써 최적화(optimization) 및 경량화(lightweight)에 기여할 수 있다. 하지만, 의료 인공지능 분야는 이처럼 다양한 최적화 기법을 적용한 연구가 상대적으로 미비하며, 모델의 성능과 컴퓨팅 자원 사용량 간 trade-off 관계 개선을 위한 연구가 필요한 실정이다.

이에 본 연구에서는 최적화된 의료영상 분할(medical image segmentation)을 위해 동적 양자화(dynamic quantization) 기반의 ResUNet 분할모델 (dynamic quantized ResUNet: DQ-ResUNet)을 제안한다. DQ-ResUNet은 양자화 전 기존의 분할 성능은 효과적으로 유지하면서, 데이터 추론 속도 및 메모리 사용량을 효과적으로 절감시킨다. Base-line 모델로 사용하는 ResUNet[2]과 본 연구에서 제안하는 DQ-ResUNet을 비교평가 한 결과, dice score는 동일하게 유지하면서 16개의 batch-size 당 13.71%의 추론 속도 향상을 보였으며, 208.55%의 메모리 사용량 절감 성능을 확인하였다.

2. 본 론

본 연구에서는 동적 양자화 기법을 ResUNet의 convolution layer에 적용하는 DQ-ResUNet을 제안하며, 제안하는 모델은 그림 1과 같다. DQ-ResUNet은 모델 최적화 성능을 극대화시키기 위해 convolution layer의 가중치(weight)에 대한 양자화를 진행한다.

2.1. 동적양자화

동적 양자화는 모델 추론 시 input의 실수형 변수를 정수형 변수로 변환함으로써 동적으로 양자화를 진행하며, 이를 위한 수식은 수식 1과 같다.

$$f_q(x,s,z) = clip(round(\frac{x}{s}+z)) \qquad (1)$$

수식 1에서 $f_q(x,s,z)$는 quantized value를 의미하며, clip()은 정수형 범위로 values를 clip하는 것을 의미한다. float 32형태의 input값은 x에 해당하며, s는 scale값, z는 zero-point integer로 scaling 시의 기준값을 의미한다. 양자화 과정은 반올림 함수인 round()를 통해 값을 반올림한 후, scaling으로 실수형 범위 값을 정수형으로 변환하는 과정을 통해 수행된다. 동적 양자화는 수식 1을 기반으로 clipping range를 동적으로 결정함으로써 양자화를 수행하는 과정을 거친다.

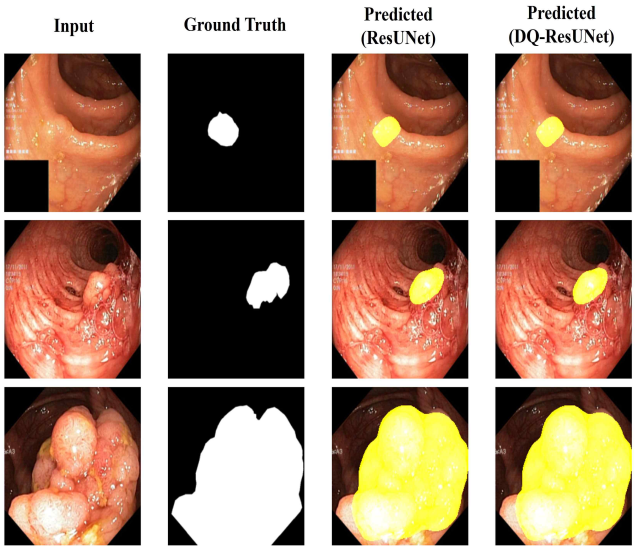

그림 2. input에 대한 모델 예측 결과

2.2. DQ-ResUNet

DQ-ResUNet은 convolution layer에 동적 양자화를 적용한 ResUNet[2] 기반 모델이다. ResUNet은 U-Net[1]과 residual learning[8]을 결합한 모델로 gradient vanishing 문제를 해결함과 동시에 효과적인 정보전달을 가능하게 한다. 본 연구에서 제안하는 DQ-ResUNet은 ResUNet의 residual block에서 사용되는 convolution layer들에 대해 가중치 동적 양자화를 진행하였다.

3. 실 험

3.1. 실험환경

DQ-ResUNet의 성능을 평가하기 위해 의료영상 분할 데이터셋인 Kvasir-SEG[9]를 활용하였다. Kvasir-SEG 데이터셋은 분할 마스크가 포함된 1,000개의 내시경 데이터셋으로 332x487에서 1920x1072 pixel까지 다양한 해상도로 구성되어 있다. 본 연구에서는 모든 input에 대해 224x224 해상도로 전처리하였으며, DQ-ResUNet의 비교평가를 위한 base-line 모

표 1. 동적 양자화 실험결과

Model / Metrics	ResUNet	DQ-ResUNet (Proposed)	Improvement Rate (%)
Dice Score	0.8162 (±0.0058)	0.8162 (±0.0052)	-
IoU (0.5)	0.8480 (±0.0044)	0.8479 (±0.0047)	-0.011
Latency (ms)	56.64 (±1.491)	**49.81** (±0.995)	**13.71**
Parameter	32.52×10^6	32.52×10^6	-
Memory Usage (MB)	565.45 (±1.099)	**183.26** (±0.405)	**208.55**

델로 ResUNet을 활용하였다. 모델의 학습 및 검증을 위해 8대2의 비율로 train, test 데이터셋을 분리하였으며, validation은 k-fold validation (k=5)을 활용하였다. 학습조건으로는 각 fold 당 epoch=100, adam optimizer, step lr(lr:1e-4), dice loss를 활용하였으며, 훈련데이터에 한해 50%의 확률로 random rotation 및 horizontal flip의 augmentation을 진행하였다.

실험에 사용된 평가지표는 모델의 분할 성능과 경량화 성능을 중점으로 선정하였다. 모델의 분할 성능을 측정하기 위해 dice score와 intersection over union(IoU)를 활용하였으며, IoU의 경우 0.5의 임계값을 기준으로 측정하였다. 모델의 경량화 성능 비교를 위해 latency와 매개변수(parameter), 메모리 사용량(memory usage)을 활용하였다. Latency의 경우 cpu기준으로 16개의 배치크기 당 millisecond(ms)속도를 측정하였으며, 메모리 사용량의 경우 megabyte(MB)를 기준으로 측정하였다.

3.2. 실험결과

표 1 및 그림 3에 따르면, 제안모델인 DQ-ResUNet은 0.8162의 dice score와 0.8479의 IoU를 보여 비교모델인 ResUNet에 비해 분할 성능이 감소 되지 않았음을 확인하였다. 반면, latency와 메모리 사용량에서는 ResUNet에 비해 각각 13.71%의 추론 속도 절감 및 208.55%의 메모리사용량 절감의 성능을 보여 제안모델의 정량적 우수성을 입증하였다. 매개변수는 비교모델과 제안모델이 32.52 million(M)의 동일한 크기를 지니는데, 이는 양자화 방식이 모델의 매개변수를 직접적으로 줄이는 것이 아닌 매개변수의 type 변환작업을 통한 최적화 방식이기 때문이다. 따라서, 양자화가 모델에 적절히 수행되었음을 분석하였다. 또한, 그림 2의 용종(polyp) 예측에서 비교모델인 ResUNet과 제안모델인 DQ-ResUNet의 결과가 거의 동일하게 나타난 것을 정성적으로 확인할 수 있다. 이와 같은 평가를 통해 동적 양자화가 적용된 DQ-ResUNet이 분할 성능의 하락 없이 적절하게 최적화되었음을 분석하였다.

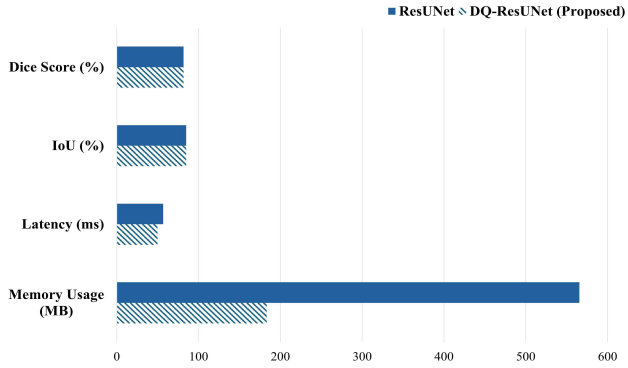

그림 3. 분할 성능 및 경량화 성능을 위한 성능지표 시각화

4. 결 론

본 연구에서는 의료분야의 인공지능 상용화를 위해 동적 양자화 기반의 최적화된 의료영상 분할모델을 제안한다. 제안모델인 DQ-ResUNet을 통해 내시경 분할 작업에서 우수한 최적화 성능을 입증하였으며, 기존 분할모델의 한계였던 모델 성능과 컴퓨팅 자원 사용량 사이의 균형(trade-off) 문제를 개선하였다. 해당 분석결과를 기반으로 추후연구에서는 동적 양자화가 아닌 quantization-aware training(QAT) 연구를 수행하고자 한다. 의료분야를 위한 인공지능 경량화 연구는 의료영상뿐만 아니라 신약개발연구와 같이 다양한 의료분야에 적용될 수 있다. 신약개발의 경우 경량화 모델의 적용을 통해 약물-표적 상호작용 예측 및 약물 설계 최적화 단계에서 신약개발의 전반적인 효율성이 증가할 수 있으며, 개발주기의 단축이 가능하다. 이와 같은 연구를 기반으로 의료분야의 인공지능 최적화 연구가 활발히 수행되기를 기대한다.

참 고 문 헌

[1] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical image computing and computer-assisted intervention-MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015.

[2] Diakogiannis, Foivos I., et al. "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data." ISPRS Journal of Photogrammetry and Remote Sensing 162 (2020): 94-114.

[3] Chen, Jieneng, et al. "Transunet: Transformers make strong encoders for medical image segmentation." arXiv preprint arXiv:2102.04306 (2021).

[4] Antonelli, Michela, et al. "The medical segmentation decathlon." Nature communications 13.1 (2022): 4128.

[5] Liu, Zhuang, et al. "Rethinking the value of network pruning." arXiv preprint arXiv:1810.05270 (2018).

[6] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

[7] Wu, Hao, et al. "Integer quantization for deep learning inference: Principles and empirical evaluation." arXiv preprint arXiv:2004.09602 (2020).

[8] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[9] Jha, Debesh, et al. "Kvasir-seg: A segmented polyp dataset." MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II 26. Springer International Publishing, 2020.