Journal of the KNST

2026; Vol.9, No.1; pp.181-188

ISSN: 2635-4926 https://doi.org/10.31818/JKNST.2026.3.9.1.181

Check for updates

Received: 2025/12/10
Revised: 2025/12/23
Accepted: 2026/03/06
Published: 2026/03/31

*Corresponding Author:
Hyunbae Chang
5th floor, 231, Pangyoyeok-ro, Bundang-gu, Seongnamsi, Gyeonggi-do, Republic of Korea
Tel: +82-31-5178-4403
E-mail: hyunbae.chang@lignex1.com

Abstract

수중기뢰탐지는 해양 안전확보와 군사작전에서 핵심적인 임무로, 최근에는 수중 소나 영상 데이터를 딥러닝 기법으로 분석하여 기뢰 탐지를 수행하는 연구가 활발히 진행되고 있다. 특히 자율무인잠수정 및 무인수상정에 이러한 모델을 실시간으로 탑재-운용하기 위해서는 제한된 연산자원과 전력환경에 적합한 경량화 기법이 필수적이다. 본 논문은 먼저 수중기뢰 탐지 분야에서의 딥러닝 활용 동향을 개관하고, 대표적인 경량화 기법들을 소개한다. 또한 소나 영상 기반 탐지에 적용된 사례들을 정리하고, 성능, 효율 개선 효과를 분석한다. 마지막으로, 이러한 기법들의 실전 운용시 기대효과와 향후 연구방향을 제안한다.

Underwater mine detection is vital for maritime safety and naval operations. Deep learning has recently been applied to sonar imagery for this task. However, real-time deployment on Automated Underwater Vehicles (AUVs) and Unmanned Surface Vehicles (USVs) requires lightweight and efficient models. This paper reviews current deep learning-based sonar detection methods, outlines key model compression techniques, and summarizes representative applications. We also examine their impact on accuracy and efficiency and discuss implications for field deployment and future research.

Keywords

딥러닝(Deep Learning), 양자화(Quantization)
모델 경량화(Model Compression), 탐지(Detection)
수중기뢰탐지(Underwater Mine Detection)

Acknowledgement

이 논문은 2023년도 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임(No. KRIT-CT-23-035-03, AI기반 수중 기뢰 탐지 기술(기뢰탐지용 무인잠수정 군집 운용 기술)).

AI기반 수중기뢰 탐지를 위한 AI 모델 경량화 기법

Survey: Model Compression for AI-Based Underwater Mine Countermeasure

장현배[1*], 김가연[2], 김윤호[1], 박석준[3]

[1]LIG넥스원 해양연구소 선임연구원
[2]LIG넥스원 해양연구소 연구원
[3]LIG넥스원 해양연구소 수석연구원

Hyunbae Chang[1*], Kayeon Kim[2], Yoonho Kim[1], Seokjoon Park[3]

[1]Senior research engineer, Maritime R&D Center, LIG Nex1
[2]Research engineer, Maritime R&D Center, LIG Nex1
[3]Chief research engineer, Maritime R&D Center, LIG Nex1

1. 서론

수중기뢰탐지(Mine Countermeaure, MCM)는 해양 작전 및 항로 안전 확보에서 매우 중요한 역할을 한다. 기존에는 전통적인 신호처리 기반 탐지 알고리즘이나 전문가의 수동 판독에 크게 의존했으나, 최근에는 딥러닝의 발전과 함께 측면주사소나(Side Scan Sonar, SSS)와 합성개구소나(Synthetic Aperture Sonar, SAS) 영상으로부터 자동으로 기뢰를 탐지/분류하는 연구가 활발히 진행되고 있다. 그러나 AUV(Autonomous Underwater Vehicle)와 USV(Unmanned Surface Vehicle)에 딥러닝 모델을 실시간으로 탑재하기 위해서는 제한된 전력, 연산능력, 메모리 환경을 고려해야 한다. 단순히 정확도가 높은 모델을 만드는 것을 넘어, 경량화 기법을 통한 효율적인 모델 설계가 필수적인 연구과제로 대두되고 있다. 따라서 본 논문은 원활한 수중기뢰 탐지를 위하여, AUV에 AI모델이 탑재 및 실시간 운용 될 수 있도록 AI 모델 경량화 기법들을 소개하려 한다. 또한 기법들의 실제 적용 사례를 정리하여 개선효과를 분석하고, 향후 연구방향을 제안하려 한다.

2. 수중기뢰탐지 분야 딥러닝 적용 현황

수중 환경은 저대비, 강한 잡음, 다양한 배경 패턴으로 인해 물체 탐지가 어렵다. 이러한 문제를 해결하기 위해 YOLO, Faster R-CNN,

© 2026 Korea Society for Naval Science & Technology

transformer 기반 탐지기 등 일반적인 객체 탐지 모델이 소나영상에 적용되어 왔다. 특히 YOLO 계열 모델은 속도와 정확도의 균형 덕분에 AUV/USV 실시간 탐지 연구에서 선호된다. 최근에는 소형 표적 탐지에 최적화된 변형 YOLO, CNN-transformer 하이브리드 네트워크, 합성 데이터셋을 활용한 학습 등 다양한 접근이 보고되고 있다.

3. 경량화 기법의 필요성 및 종류

딥러닝 모델은 높은 정확도를 얻기 위해 수백만 개 이상의 파라미터와 막대한 연산량을 필요로 한다. 그러나 AUV나 USV같은 해양 자율 플랫폼은 연산 자원이 제한적이며, 배터리 기반으로 동작하기에 전력효율이 매우 중요하다. 또한, 실제 작전 중에는 통신 지연이 크거나 네트워크 연결이 불가능한 상황이 많기 때문에 탑재형(on-board) 실시간 처리가 요구된다. 이러한 제약 속에서 복잡한 딥러닝 모델을 직접 운용하기는 어렵다. 따라서 모델 경량화(model compression) 기술은 수중기뢰탐지 시스템의 실시간성과 지속 운용성을 보장하기 위한 핵심 요소로 부각되고 있다.

경량화 기법은 크게 네 가지 범주로 나눌 수 있다. 양자화(quantization), 프루닝(pruning), 지식 증류(knowledge distillation), 저랭크 근사(low-rank approximation), 그리고 최근 각광받는 신경망 구조 탐색(Neural Architecture Search, NAS)이 그 대표적인 접근이다.

3.1 양자화(Quantization)

양자화는 모델의 가중치와 활성화 값의 표현 정밀도를 줄여 연산량과 메모리 사용량을 감소시키는 방법이다. 일반적으로 학습된 모델은 32비트 부동 소수점(FP32) 연산을 사용하지만, 이를 16비트(FP16), 8비트(INT8), 4비트(INT4) 정수 연산으로 변환함으로써 모델 크기를 줄일 수 있다.

양자화는 크게 사후 양자화(Post-Training Quantization, PTQ)와 양자화 인식 학습(Quantization-Aware Training, QAT)으로 나뉜다. PTQ는 이미 학습된 모델을 학습된 모델을 변환하는 방식으로 간단하고 빠르지만, 저비트로 변환 시 정확도 저하가 발생할 수 있다. 반면 QAT는 학습 단계에서부터 양자화 연산을 모사하여 정밀도 손실을 최소화할 수 있다. 구체적으로, 가중치 w를 양자화 하려면 다음과 같은 식을 이용할 수 있다.

$$w_q = Clip(\partial(\frac{w}{s}) + z) \quad (1)$$

위의 수식에서 w_q는 양자화된 가중치 w이며, s는 실제값을 원하는 범위 안으로 축소시키는 축척 인자(scale factor), z는 가중치들의 기준점이 될 영점(zero point)을 나타낸다. 이러한 방식으로 수십자리 단위의 데이터와 그 연산을 2자리 수준(INT4 기준)으로 변환하여 연산량과 모델의 크기를 줄이는 것이 핵심이다.

특히 수중기뢰 탐지의 경우, 소나영상의 회색조 대비가 약하고 배경잡음이 많은 특성이 있어, 단순 PTQ보다는 QAT가 선호되지만, 최근의 연구에서는 PTQ만으로도 INT8 모델이 FP32 모델 대비 거의 동일한 정확도를 유지가 가능해져 PTQ 방식이 더 선호되고 있다[1]. 다만 대표 보정 데이터(해역/거리/각도/기뢰 형상 등)를 확보하고 양자화 스케일과 제로 포인트 설계를 면밀히 해야 정밀도 손실을 최소화 할 수 있다.

3.2 프루닝(Pruning)

프루닝은 모델 내에서 중요도가 낮은 가중치나 채널을 제거함으로써 파라미터 수 및 연산량을 줄이는 기법이다. 크게 두 가지 방식이 존재하며, 하나는 비구조적 프루닝으로 개별 가중치를 0으로 만드는 방식이며, 다른 하나는 구조적 프루닝으로 가중치를 채널이나 필터 단위로 제거하는 방식이다.

비구조적 프루닝은 모델 압축률은 높지만 하드웨어(GPU나 임베디드 장치)에서의 실질적인 속도 향상은 제한적이다. 반면 구조적 프루닝은 제거단위가 명확하여 실제 연산량 감소에 직접적으로 기여하므로 임베디드 환경에 더 적합하다. 이와 관련하여, 저전력 딥러닝 및 컴퓨터 비전 응용을 대상으로 한 리뷰에서는 파라미터 양과 연산량을 크게 줄이면서도 정확도 저하를 억제하는 기법 군을 주요 범주로 삼고 있다[2].

최근 연구에서는 네트워크 내 가중치들의 중요도를 자동으로 평가하는 방법이 다양하게 제시되었다. 예를 들어 BN 스케일 기반 프루닝은 배치 정규화 계

수의 크기를 이용하여 중요 채널을 선택하여 모델 크기를 최대 20배, 연산량을 5배 줄이면서도 정확도를 유지한 바 있다[3]. 또한 의존성 그래프 기반 구조적 프루닝은 채널 간의 종속성을 고려해 효율적으로 구조를 줄일 수 있다[4].

3.3 지식증류(Knowledge Distillation)

지식증류는 대형 모델(teacher)의 학습된 정보를 소형 모델(student)에 전달하여, student 모델이 경량화된 상태에서도 높은 성능을 유지하도록 돕는 기법이다. 전통적인 지식증류 기법은 대형 모델이 출력하는 soft-label 확률 분포를 소형 모델이 모방 가능하도록 학습시키는 방식이며, 이에 필요한 지식증류 loss는 다음과 같은 수식으로 표현할 수 있다.

$$L_{KD} = aL_{CE}(y, y_x) + (1-a)L_{KL}(p_t, p_s) \quad (2)$$

L_{CE}는 정답과의 매칭을 위한 cross-entropy loss이고, L_{KL}은 대형 모델(teacher model)과 소형 모델(student model) 분포 간의 차이(divergence)를 측정할 수 있는 Kullback-Leibler divergence 이다. a는 이 두 항간의 1 이하의 가중치이며, y는 정답, y_s는 소형 모델의 출력 값이다. p_t의 경우 대형 모델의 확률 분포, p_s는 소형 모델의 확률 분포이다.

또한 최근에는 중간 특성맵(feature map)을 상호 모방하는 특성 증류(feature distillation)나 어텐션 맵(attention map)을 모방하는 어텐션 증류(attention distillation) 등이 제안되어 잡음이 많은 소나 영상 환경에서도 소형 모델이 더욱 안정적으로 작동할 수 있게 되었다.

지식 증류를 실제 연구에 적용한 사례로는 YOLOX-ViT 구조의 대형 모델로부터 소형 모델을 증류하여 거짓 양성(false positive) 오류를 20% 이상 감소시킨 바 있다[5].

3.4 저랭크 근사(Low-Rank Approximation)

딥러닝 모델의 가중치 행렬은 종종 중복된 정보나 상관관계가 높은 정보가 포함되어 있다. 저랭크 근사 기법은 이러한 행렬을 보다 작은 차원으로 분해하여 연산량을 감소시키는 방식이다.

최근 연구에 따르면, 저랭크 구조만으로도 모델 크기 및 연산량을 크게 줄일 수 있으며, Ou et al.은 저랭크 근사 + 희소화(sparsity) + 구조 탐색을 통합한 방법을 제시했다[6]. 또한 Sainath et al.은 TSVD, CUR 분해를 통해 다양한 분류모델에서 좋은 모델 경량화 성과를 보였다[7].

하지만 수 최적랭크 선정이 어렵고, 분해 후 재학습(fine-tuning) 필요성이 높다는 단점으로 인해 수중기뢰 탐지처럼 잡음이 많고, 저대비인 소나영상을 사용하는 환경에는 저랭크 근사의 적용이 제한된다. 따라서 저랭크 근사는 다른 경량화 기법(양자화, 프루닝, 증류)과 병합하여 전략적으로 사용하는 것이 바람직하다.

3.5 신경망 구조 탐색(Neural Architecture Search, NAS)

NAS는 사람이 직접 설계하지 않고, 자동화된 탐색 알고리즘이 모델의 구조(예: 블록 수, 채널 수, 필터 크기, 스킵 연결 등)를 최적화 하는 방법이다. 탐색 대상에는 모델성능(accuracy)뿐 아니라 파라미터 수, 연산량(FLOPs, TOPs), 추론 지연(latency) 등이 포함될 수 있어, 경량화 및 실시간성에 매우 적합하다. 최근에는 경량화를 위한 백본(backbone) 네트워크로서 GhostNet, ShuffleNet, MobileNetV3 등이 활발히 사용되고 있다.

특히 MobileNetV3와 같은 경우는 하드웨어 인지 NAS에 수기 설계를 통합하여 개발한 기법으로, 기존 대비 분류 지연시간을 15% 감소, 탐지 지연시간을 25% 감소하였다[8]. 이는 모바일/임베디드 디바이스에서의 NAS의 실측 지연 개선을 보여주는 대표적인 사례다.

3.6 종합 비교 및 시사점

이처럼 각 기법별로 장점과 단점이 혼재되어있으며, 단일 기법만을 적용하는 것보다는 프루닝 + 양자화 + 지식증류의 복합 적용이 수중기뢰 탐지 시스템의 실시간 탐지환경에서 가장 실효적인 전략이라 할 수 있다.

Table 1. Model compressions comparison

No.	Technique	Advantage	Fault
1	Quantization	Speed, Memory Efficiency	Accuracy Loss
2	Pruning	FLOPS Parameters	Accuracy Loss
3	K.D.	Accuracy Robustness	Teacher Dependency
4	Low-rank Approx.	HD Image Processing	Fine-tuning Cost
5	NAS	Real-time	Time/Computation Cost

4. 심화 연구 사례

최근 몇 년간, 수중기뢰탐지 또는 유사 해양 객체 탐지 응용분야에서는 경량화 기법들이 단순한 이론적 제안에 그치지 않고 실제 탐지 네트워크 및 임베디드 플랫폼에 적용된 사례가 늘고 있다.

이 절에서는 대표적인 다섯 가지 경량화 적용 연구를 사례별로 살펴보고, 각 기법의 핵심 전략과 성능지표, 그리고 수중기뢰 탐지 분야에 대한 적용 가능성을 분석한다.

4.1 프루닝과 증류를 복합 적용하여 탐지 모델 압축

Chen et al.에서는 YOLOv7 모델을 대상으로 파라미터 그룹간 종속성을 고려한 자동의존성 그래프 기반 채널/필터 프루닝을 수행하고, 이후 출력/특징맵 기반 지식증류를 결합하는 방식으로 모델을 압축하였다[9].

먼저 네트워크 내 각 채널의 상호 의존도를 분석하여 중복 또는 기여도가 낮은 채널을 자동으로 제거하였다. 이후 남은 파라미터를 INT8 형식으로 양자화하여 추론속도를 향상시켰고, 추가적으로 지식 증류를 결합하여 정확도 손실을 보정하였다.

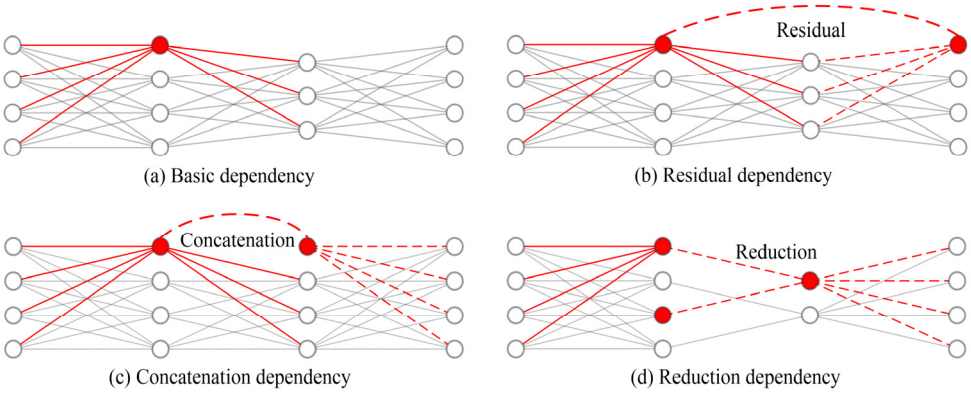

Fig. 1. Pruning criteria[9]

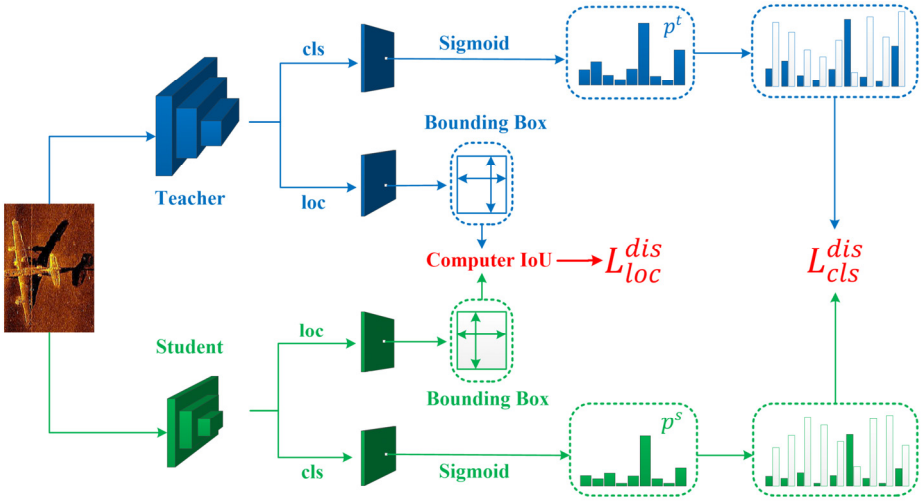

Fig. 2. Knowledge distillation example[9]

실험 결과, 모델 파라미터 수는 약 75% 감소, 연산량(FLOPs)은 약 70% 감소하였으며, 모델 크기는 기존 약 72 MB에서 15 MB 수준으로 기존의 1/4 수준으로 압축 되었다. 탐지 정확도는 mAP@50 기준으로 1% 이내의 손실만 발생하였으며, NVIDIA Xavier NX 보드에서 초당 12 fps에서 23 fps로 약 1.9배 향상된 추론 속도를 보였다.

이러한 결과는 프루닝과 양자화를 결합하면 정확도 저하를 최소화하며 리소스 효율을 극대화할 수 있음을 보여준다. 또한, 수중기뢰 탐지에서도 이와 같은 조합은 탐지 모델 탑재 AUV의 전력소모를 감소시키고, 탐색 범위를 확대하는 데 유효할 것으로 예상된다.

4.2 지식증류를 이용한 소형 모델의 성능 유지

Aubard et al.에서는 SSS 데이터를 대상으로 YOLOX-Large-ViT의 CNN-트랜스포머 결합 모델을 Teacher 모델로, YOLOX-Nano-Vit와 같은 소형 모델을 student 모델로 사용하여 지식증류를 수행하였다[5]. 이는 지식증류가 복잡한 수중영상 환경에서도 효과적임을 검증한 최초의 연구 중 하나이다.

Teacher 모델과 student 모델은 유사한 구조의 백본을 가지지만, teacher 모델은 640×640의 이미지를 입력으로 받으며, student 모델은 416×416의 이미지를 입력으로 받고 백본의 크기 또한 teacher 모델에 비해 더 적은 layer를 사용한다. 이러한 두 모델 간의 지식증류를 이용해 연산량 및 크기를 감소시키면서도 탐지성능은 최대한 확보할 수 있도록 하였다. 두 모델 사이의 출력 확률 분포 및 피처맵을 동시에 전달하여 학습시키는 하이브리드 증류 전략을 통해 단순 단일 데이터셋에서의 탐지 성능뿐이 아닌 네트워크 자체의 성능을 증류할 수 있도록 하여 두 모델간의 간극을 최소화하였다.

이 방법을 통해 SSS 탐지에서 거짓 양성(false positive)은 약 20% 감소, precision 및 recall 유지율은 teacher 모델 대비 1% 이내의 저하를 보이며 연산량과 모델 크기는 줄이면서도 성능은 유지 혹은 더 개선시키는 성과를 획득하였다.

4.3 NAS/지식증류 결합 경량화 기법 실시간 임베디드 적용

Fan et al.에서는 신경망 구조 탐색(NAS)과 지식증류를 결합하여[10], YOLOv8 기반 경량 탐지망을 자동으로 설계하였다. 해당 연구는 임베디드 환경에서의 탐지 성능 최적화를 목표로 하며, teacher 모델로 YOLOv8-L, student 모델로는 NAS를 사용하여 선정된 별개의 경량 구조 AI모델을 사용하였다.

결과적으로 모델 파라미터는 44M에서 9M으로 약 80% 감소하였으며, 연산량은 81% 감소, 탐지성능을 나타내는 지표인 mAP@50은 97.8%에서 96.5%로 감소하여 약 1.3% 정도의 성능저하만을 유지하였다. 또한 NVIDIA의 Jetson Xavier XT와 같은 상용 AI SOM 보드에 탑재 및 운용하여 실제 엣지 디바이스에서도 사용이 가능함을 확인하였다.

이 연구는 수중 객체 탐지 분야에서 경량화된 탐지

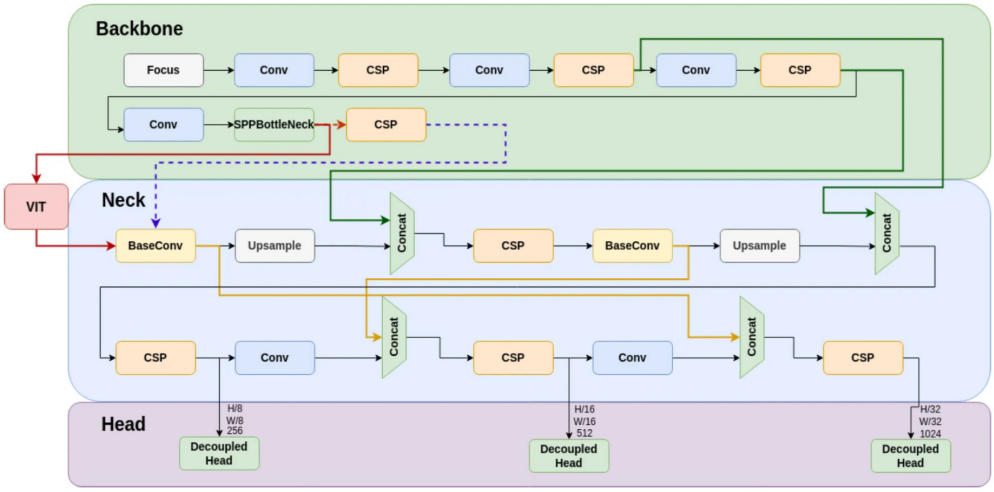

Fig. 3. YOLOX-ViT architecture[5]

모델의 설계가 가능하다는 사실에 의의가 있다. 특히, NAS와 지식증류를 복합 사용할 시, 탐지성능을 유지하면서도 모델 크기와 연산량을 대폭 감소시키고, 해당 모델을 임베디드 디바이스에 탑재 및 운용함으로써 향후 수중객체탐지 분야에서 이와 같은 전략이 매우 유망하다는 사실을 보여준 데에 의의가 있다.

4.4 PTQ 단독 적용을 통한 INT8 양자화

Zhou et al.은 라이다 기반 3D 객체 탐지 모델에 PTQ만을 적용하여, 추가 재학습(QAT) 없이 모델 크기 및 연산량은 줄이면서도 정확도 유지를 달성하였다[1]. 해당 연구는 희소성 기반 보정(sparsity-based calibration) 및 태스크 지향 글로벌 양자화 손실(task-guided global quantization loss)을 도입하여 정확도 손실을 최소화하였다.

그 결과, PTQ가 적용된 INT8 모델은 기존 FP32 모델 기준 mAP 손실은 0.3% 이내, 추론속도는 약 3배 향상, 모델크기는 75% 감소하여 PTQ 단독 적용으로도 탐지 성능을 유지하면서 경량화를 달성할 수 있음을 보여주었다. 이는 수중객체탐지 모델에서도 시간과 자원을 소모하는 QAT 적용 없이 단순 PTQ만으로도 실시간 탐지할 수 있다는 가능성을 시사한다.

4.5 객체탐지 모델 PTQ 세부 최적화

Niu et al.은 기존 PTQ 기법의 한계를 보완하기 위해 task loss-guided metric을 도입한 탐지 모델 양자화 기법(DetPTQ)을 제안하였다[11]. 기존 PTQ는 단순히 분포 기반으로 양자화에 사용할 스케일을 결정하여 정확도 손실이 유의미하게 발생했으나, 이 연구에서는 정확도 저하를 감소하기 위해 탐지 손실(object detection output loss)을 직접 활용하여 층별로 적응형 스케일링을 적용하였다.

RetinaNet-ResNet18 모델을 INT4 형식으로 양자화한 결과, 객체 탐지에 널리 사용되는 데이터셋인 COCO 데이터셋 기준 0.6%의 정확도 손실만을 기록하였으며, 이는 기존의 INT4 PTQ 기법들이 평균적으로 3% 이상의 정확도 손실을 내는 것을 고려하면 INT4 PTQ도 INT 8 수준의 정확도 보전이 가능하다

Algorithm 1 LiDAR-PTQ quantization

Input: Pretrained FP model with N layers; Calibration dataset D^c, iteration T.
Output: quantization parameters of both activation and weight in network, i.e., weight scale s_w, weight zero-point z_w, activation scale s_a, activation zero-point z_a and adaptive rounding value for weight θ.

Optimize only weight quantization parameters s_w and z_w to minimize Eq 16 in every layer using the grid search algorithm;
input D^c to FP network to get the FP final output O_{fp}
for $L_n = \{L_i | i = 1, 2, ...N\}$ **do**
  Optimize only activation quantization parameters s_a and z_a to minimize Eq 16 in layer L_i using the grid search algorithm;
  Collect input data I_i to the FP layer L_i;
  Input I_i to quantized layer L_i^q and FP layer L_i to get quantized output $\hat{A}_i$ and FP output A_i;
  Input $\hat{A}_i$ to the following FP network to get output $\hat{O_{par}}$ of partial-quantized network;
  for all $j = 1, 2, \ldots, T$-iteration **do**
    Check quantized output $\hat{A}_i$ and FP output and calculate L_{local} using Eq 11;
    Check partial-quantized network output $\hat{O_{par}}$ and FP final output O_{fp} to calculate L_{TGPL} using Eq 9;
    Optimize quantization parameters s_w, z_w, s_a, and z_a, θ of layer L_i to minimize L_{total} using Eq 12;
  end for
  Quantize layer L_i with the learnable quantization parameters s_w, z_w, s_a, and z_a, θ;
end for

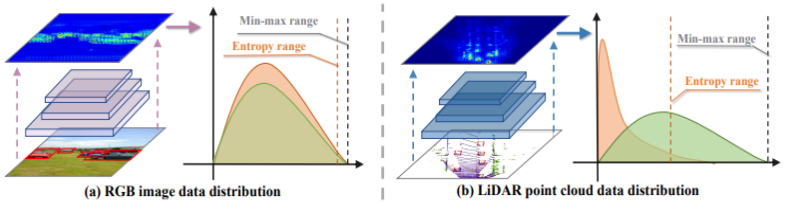

Fig. 4. LiDAR-PTQ algorithm and data distribution[1]

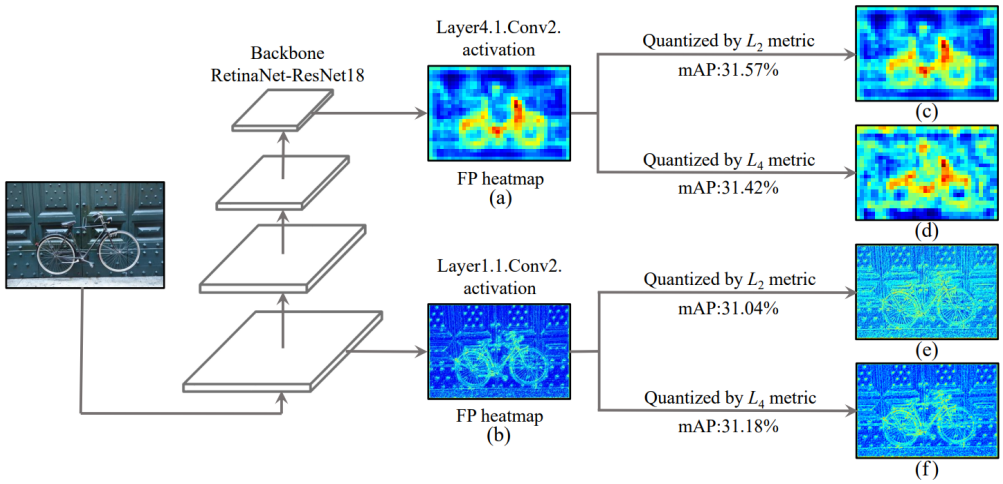

Fig. 5. 4-bit quantization applied RetinaNet-ResNet18 architecture[11]

는 것을 증명한다.

5. 기대효과 및 적용 전망

본 논문에서는 수중기뢰탐지 분야에서 딥러닝 모델을 AUV/USV 탑재 형태로 실시간 운용하기 위해 반드시 고려해야 할 경량화 기법들에 대하여 살펴보았다. 양자화, 프루닝, 지식증류, 저랭크 근사, 신경망 구조 탐색 등의 기법을 소개하고, 최근 탐지, 임베디드 응용 사례를 통해 그 실효성을 논하였다. 또한 이러한 기법들이 적용됨으로써 얻을 수 있는 다양한 기대효과를 제시하였다. 향후 연구방향으로는 다음이 예상된다.

- **합성데이터 기반 학습/증강 전략:** 수중기뢰 탐지는 데이터 수집이 어려우므로 합성/시뮬레이터 기반 학습이 중요하며, 이 과정에서도 경량화 모델이 유리하다.
- **다중센서 융합형 경량 모델:** 광학/소나/자기탐지 등 다중센서 데이터를 경량 딥러닝 모델로 융합하는 연구가 필요하다.
- **실적용 실험 및 벤치마크 마련 방안:** AUV/USV 실제 임무 환경에서의 경량모델 벤치마크 데이터셋 및 평가체계를 마련해야 한다.
- **온디바이스 학습 및 적응형 모델 압축:** 해역이나 수심 변화 등 운용환경이 바뀔 때 자동으로 적응하는 경량화 모델(온라인 프루닝, 동적 양자화 등) 연구가 요구된다.

결론적으로, 수중기뢰 탐지 분야는 제한된 자원을 가진 플랫폼상에서 정확도, 실시간성, 경량성의 세 가지 특성을 동시에 만족시켜야 하는 특수한 환경이다. 따라서 앞으로의 연구에서는 단순한 모델 경량화가 아닌 임무조건, 하드웨어 제약, 데이터 특성을 함께 고려한 통합적 설계가 필요하리라 여겨진다.

참고문헌

[1] Sifan Zhou, Liang Li, Xinyu Zhang, Bo Zhang, Shipeng Bai, Miao Sun, Ziyu Zhao, Xiaobo Lu, & Xiangxiang Chu, 'LiDAR-PTQ: Post-Training Quantization for Point Cloud 3D Object Detection,' arXiv preprint arXiv:2401.15865, 2024.
[2] Abhinav Goel, Caleb Tung, Yung-Hsiang Lu, & George K. Thiruvathukal, 'A Survey of Methods for Low-Power Deep Learning and Computer Vision,' in proceedings of 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), New Orleans, LA, 2020.
[3] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, Changshui Zhang, 'Learning Efficient Convolutional Networks Through Network Slimming,' Proceedings of the IEEE International Conference On Computer Vision (pp. 2736-2744), 2017.
[4] Minjie Y. Wang, 'Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs.' ICLR Workshop on Representation Learning on Graphs and Manifolds, 2019.
[5] Martin Aubard, László Antal, Ana Madureira, & Erika Ábrahám, 'Knowledge Distillation in YOLOX-Vit for Side-Scan Sonar Object Detection,' arxiv preprint arxiv:2403.09313, 2024.
[6] Xinwei Ou, Zhangxin Chen, Ce Zhu, & Yipeng Liu, 'Low Rank Optimization for Efficient Deep Learning: Making a Balance Between Compact Architecture and Fast Training,'

Journal of Systems Engineering and Electronics, VOL. 35, NO. 3, 2023, pp. 509-531.
[7] Tara N. Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, & Bhuvana Ramabhadran, 'Low-Rank Matrix Factorization for Deep Neural Network Training with High-Dimensional Output Targets,' in proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, 2013.
[8] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, & Vijay Vasudevan, 'Searching for MobileNetV3,' Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1314-1324), 2019.
[9] Chensheng Cheng, Xujia Hou, Can Wang, Xin Wen, Weidong Liu, & Feihu Zhang, 'A Pruning and Distillation Based Compression Method for Sonar Image Detection Models,' Journal of Marine Science and Engineering, VOL. 12, NO. 6, 2024, 1033.
[10] Yibing Fan, Lanyong Zhang, & Peng Li, 'A Lightweight Model of Underwater Object Detection Based on YOLOv8n for an Edge Computing Platform,' Journal of Marine Science and Engineering, VOL. 12, NO. 5, 2024, 697.
[11] Lin Niu, Jiawei Liu, Zhihang Yuan, Dawei Yang, Xinggang Wang, & Wenyu Liu, 'Improving Post-Training Quantization on Object Detection with Task Loss-Guided Lp Metric,' arxiv preprint arxiv:2304.09785, 2023.