https://doi.org/10.7840/kics.2025.50.2.253

지능형 감시 시스템에서 최적 신뢰도 임계값을 활용한 단말-엣지 협력 추론 방식

최 현 호◆

Collaborative Inference Method Using Optimal Confidence Thresholds in Intelligent Surveillance Systems

Hyun-Ho Choi◆

요 약

본 연구에서는 지능형 감시 서비스를 위한 엣지 장치와 엣지 서버 간의 새로운 협력적 추론 방식을 제안한다. 제안 방식에서는 단말은 작은 신경망을 사용하는 대신 두 개의 신뢰도 임계값을 설정하여 모호한 입력 이미지를 필터링하고, 이를 엣지 서버로 전달하여 큰 신경망으로 재평가한다. 모의실험 결과, 최적의 신뢰도 임계값을 사용하는 제안된 협력 추론 방식이 요구 정확도를 만족시키면서 종단 간 지연 시간을 크게 줄일 수 있음을 확인하였다.

Key Words : Cooperative inference, mobile edge computing, confidence threshold, surveillance.

ABSTRACT

This study proposes a novel collaborative inference method between edge device and edge server for intelligent surveillance services. In the proposed method, the device uses a small neural network (NN) and filters ambiguous input images by setting two confidence thresholds. These filtered images are then forwarded to the edge server for re-evaluation with a larger NN. Simulation results show that the proposed collaborative inference method, using optimal confidence thresholds, significantly reduces latency while meeting the required accuracy.

Ⅰ. 서 론

인공지능(AI)의 발전으로 컴퓨터 비전(computer vision, CV)의 성능이 대폭 향상되면서, 엣지 카메라는 단순히 비디오를 전송하는 역할을 넘어 영상을 지능적으로 해석하는 방향으로 발전하고 있다[1]. 그러나 단말 장치에 AI 기반 CV 기능을 탑재하는 데에는 제한된 연산 능력과 메모리 용량의 한계가 존재한다. 이를 해결하기 위해 모바일 엣지 컴퓨팅(mobile edge computing, MEC)이 도입되었다 [2]. 이에 따라 단말은 계산 집약적인 작업을 고성능 서버가 탑재된 기지국으로 오프로딩하고, 엣지 서버가 deep neural network (DNN) 연산을 수행하여 높은 정확도와 빠른 처리 속도를 제공한다. 그러나 이 과정은 업링크 트래픽을 유발하여 무선 채널의 오류 및 간섭, 다수의 서비스 요청 등으로 인해 추가적인 지연을 초래할 수 있다.

이와 같은 단말 또는 서버의 단독 추론의 한계를 극복하기 위해 협력적 추론 방식이 연구되어 왔다[2-4]. 엣지 단말, 엣지 서버, 클라우드 센터로 구성된 MEC 네트워크에서 협력적 아키텍처와 이를 지원하는 기술들이 제시되었다[2]. 또한, 단말과 서버에서 크기가 다른 DNN을 사용하여 추론 결과를 결합함으로써 검출 정확도를 향상시키는 방식이 제안되었다[3]. 아울러, 엣지 장치는 캡처된 이미지를 초기 분석하고, 이후 클라우드 서버가 의심스러운 이미지 내에서 객체를 검출하는 AI 기반 위협 감시 시스템이 개발되었다[4].

기존에 다양한 협력 컴퓨팅 방법들이 제시되었으나, 분류 및 검출 문제의 정확도를 결정하는 중요한 파라미터인 신뢰도 임계값(confidence threshold)에 대해서는 고려하지 않았다. 일반적으로 신뢰도 임계값은 기본값으로 설정되거나 운영자에 의해 경험적으로 조정되어 왔다[2-4]. 또한, 대부분 신뢰도 임계값은 검출해야 하는 positive 이미지와 유사하여 혼동될 수 있는 negative 이미지를 같이 고려하지 않고 설정되었다. 이로 인한 오탐은 사용자에게 불편을 초래하고 시스템의 신뢰성을 저하시킬 수 있다.

이를 해결하기 위해 본 연구에서는 지능형 감시 서비

※ 이 성과는 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1A2C1011901).
◆ First Author : (ORCID:0000-0002-6785-2596) Hankyong National University, School of ICT, Robotics & Mechanical Engineering, hhchoi@hknu.ac.kr, 종신회원
논문번호 : 202411-266-A-LU, Received November 4, 2024; Revised November 29, 2024; Accepted November 29, 2024

스를 위한 엣지 장치와 엣지 서버 간의 새로운 협력적 추론 방식을 제안한다. 제안 방식에서 단말은 작은 신경망을 사용하는 대신 두 개의 신뢰도 임계값을 설정하여 모호한 입력 이미지를 필터링하고, 이를 엣지 서버로 전달하여 큰 신경망으로 재평가한다. 이때, positive 및 negative 이미지에서 도출된 신뢰도 점수(즉, 추론 결과에 대해 맞을 확률값)와 이에 대한 확률 분포를 고려하여 요구 정확도를 충족하면서 종단 간 지연 시간을 최소화할 수 있는 단말과 엣지 서버의 최적 신뢰도 임계값을 도출한다.

Ⅱ. 제안하는 협력 추론 방식

그림 1은 제안하는 협력 추론 방식의 동작 흐름을 신뢰도 점수(confidence score)의 확률 밀도 함수(PDF)와 함께 보여준다. 카메라 단말은 BS에 무선으로 연결되며, 엣지 서버는 BS 바로 옆에 배치된다. 클라이언트는 유선으로 BS에 연결되어 이벤트 발생 시 알림을 전달받는다. 단말은 프로세서 및 메모리 제한으로 인해 작은 신경망을 사용하는 반면, 서버는 이러한 제약 없이 큰 신경망을 사용한다.

단말은 두 개의 신뢰도 임계값 θ_{d1}과 θ_{d2}를 사용하여 추론 결과 신뢰도 점수 X_d가 θ_{d2}보다 크면, 서버의 추가 추론 없이 해당 알림 메시지를 클라이언트로 직접 전송한다(Case I). 반면, X_d 값이 θ_{d1}와 θ_{d2} 사이에 존재할 경우(즉, $\theta_{d1} < X_d \leq \theta_{d2}$), 단말은 신뢰도 점수가 애매하다고 판단하여 결정을 미루고 해당 이미지를 서버로 전송한다(Case II). 서버는 수신된 이미지를 큰 신경망을 사용해 다시 추론한다. 서버의 추론 결과, 신뢰도 점수 X_s가 θ_s보다 크면, 서버는 클라이언트에 알림 메시지를 전송한다(Case IV). 이 외에 단말에서 $X_d \leq \theta_{d1}$인 경우(Case III)와 서버에서 $X_s \leq \theta_s$인 경우(Case V)에는 해당 이미지에서는 이벤트가 발생하지 않았다고 판단하고 이를 폐기한다.

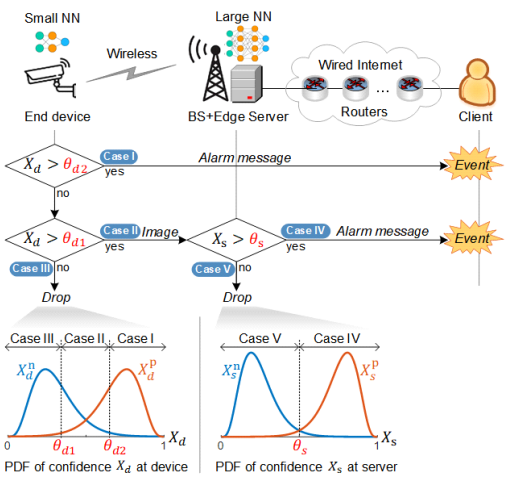


그림 1. 제안하는 협력 추론 방식의 동작 흐름
Fig. 1. Operation flow of proposed cooperative inference method.

제안 방식에서 사용하는 신뢰도 임계값들은 greedy search를 통해 최적화할 수 있다. 먼저 단말에서 정확도를 최대로 하는 θ_{d1}과 θ_{d2}를 찾고, 이후 두 임계값을 고정하고 서버에서 정확도를 최대로 하는 θ_s를 찾는다. 다시 θ_s를 고정하고 θ_{d1}과 θ_{d2}의 최적값을 찾는 이와 같은 과정을 임계치들이 변하지 않을 때까지 반복함으로써 최적의 임계값을 결정할 수 있다.

Ⅲ. 모의실험 결과

모의실험을 위해 표 1의 파라미터를 사용하였다. 객체 감지를 위해 YOLOv8을 고려하였으며, YOLOv8이 제공하는 다섯 가지 신경망 구조 중 단말에는 가장 작은 nano 모델을, 서버에는 가장 성능이 좋고 큰 xlarge 모델을 사용하였다[5]. 감시 이벤트로는 화재 검출을 설정하였으며, 이를 위해 YOLOv8의 nano 및 xlarge 모델을 화재 데이터셋으로 학습시켰다. 테스트에는 학습에 사용되지 않은 각각 500장의 화재와 연기 이미지로 구성된 positive 이미지와, 화재 또는 연기로 오인될 수 있는 태양, 차량 조명, 구름, 스모그 이미지 각각 250장으로 구성된 negative 이미지를 사용하였다. 무선 채널을 고려하여 전송 속도를 1~100 Mbps로 변화시키며,

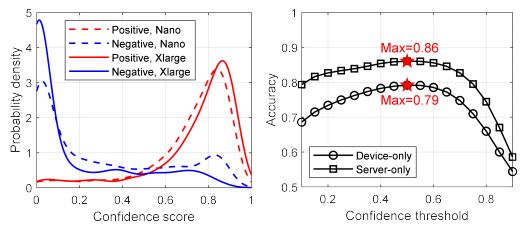
표 1. 파라미터 설정
Table 1. Parameter Setup

Description	Value
NN model in device	YOLOv8 nano
NN model in server	YOLOv8 xlarge
Number of positive images for test	1000 (500 fires and smokes each)
Number of negative images for test	1000 (250 suns, car lights, clouds, and smogs each)
Frame per second	10 fps
Size of alarm message	64 Kbps
Size of image data	256 Kilobytes
Tx. rate in wireless	R=1~100 Mbps (default=10)
Inference times	100 ms at device, 50 ms at server
Tx. time from server to client	20 ms

기본값은 10 Mbps로 설정하였다. 전력 소모, 컴퓨팅 능력, 비용 등을 고려하여 단말에는 Intel Core i9-10900X CPU를, 서버에는 NVIDIA RTX 4090 GPU를 사용하며, 각 장치에서 측정된 평균 처리 및 추론 시간을 적용하였다.

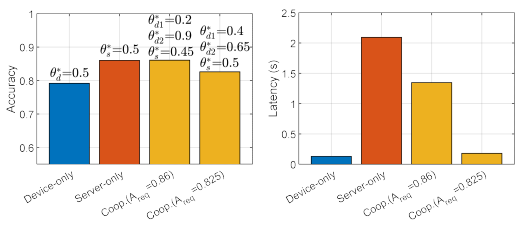
그림 2는 nano 및 xlarge 모델에서 positive 및 negative 이미지의 신뢰도 점수에 대한 PDF와 단말 단독(device-only) 방식과 서버 단독(server-only) 방식에서 신뢰도 임계값에 따른 정확도를 보여준다. Nano 모델에 비해 xlarge 모델의 분포는 양극단으로 더 넓게 퍼져 있음을 알 수 있다. 이는 xlarge 모델을 사용할 때 positive 이미지의 신뢰도 점수는 높아지고, negative 이미지의 신뢰도 점수는 낮아짐을 보여준다. 이러한 이유로 그림 2(b)에서 보듯이 server-only 방식은 device-only 보다 더 높은 정확도를 갖는다. 또한, 각 방식은 정확도를 최대화하는 최적의 신뢰도 임계값이 존재함을 확인할 수 있다. 이는 positive 및 negative 이미지의 신뢰도 점수를 기반으로 적절한 신뢰도 임계값을 설정하는 것이 정확도를 높이는데 매우 중요함을 의미한다.

그림 3은 고려된 세 가지 추론 방법의 정확도와 지연을 보여주며 이때 사용된 최적의 신뢰도 임계값(θ^*)을 보여준다. 여기에서 제안 협력 추론의 경우 요구 정확도(A_{req})로 server-only 방식과 같은 0.86과 두 단독 방식



(a) (b)

그림 2. (a) 신뢰도 점수의 PDF 및 (b) device-only 와 server-only 방식에서 신뢰도 임계값에 따른 정확도
Fig. 2. (a) PDFs of confidence scores and (b) Accuracy vs. confidence threshold in device-only and server-only.

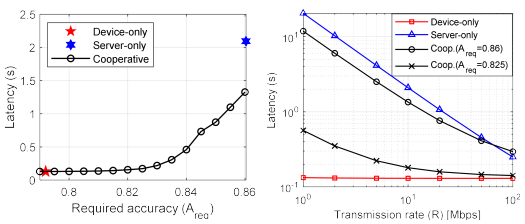


(a) (b)

그림 3. 추론 방식의 비교: (a) 정확도 및 (b) 지연
Fig. 3. Comparison of inference methods: (a) Accuracy and (b) Latency.

의 평균 정확도에 해당하는 0.825를 고려하였다. 각 방식이 해당 정확도를 달성할 때, 지연은 device-only 방식이 가장 짧고 server-only 방식이 가장 길다. 반면, 제안 협력 추론의 지연은 두 단독 방식의 지연 시간 사이에 위치한다. 특히 A_{req}가 0.86에서 0.825로 낮아질 때 지연이 크게 줄어드는데, 이는 A_{req}가 감소함에 따라 $\theta^*_{d2}-\theta^*_{d1}$ 값이 감소하여 서버로의 오프로딩 양이 줄어들기 때문이다. 따라서 제안 방식은 A_{req}에 따라 device-only와 server-only 사이에서 균형 잡힌 성능을 보이며, device-only보다 높은 정확도와 server-only보다 낮은 지연을 달성한다.

그림 4(a)는 요구 정확도(A_{req})에 따른 지연 성능을 보여준다. A_{req}가 증가함에 따라 초기에는 device-only와 유사한 낮은 수준의 지연을 유지하지만, A_{req}가 특정 값을 초과하면 지연이 증가한다. 이는 높은 A_{req}를 충족하기 위해서 서버로의 오프로딩 양이 크게 증가하기 때문이다. 그럼에도 불구하고 제안 방식은 server-only와 같은 정확도를 가지면서도 지연을 60% 수준으로 감소시킨다. 한편, 그림 4(b)는 전송 속도(R)에 따른 지연 성능을 보여준다. Device-only는 전송하는 알림 메시지의 크기가 상대적으로 작기 때문에 가장 낮고 거의 일정한 지연 시간을 보인다. 그러나 다른 방식들은 R이 증가할수록 지연 시간이 감소하는 경향을 보여준다. A_{req}=0.825일 때 제안 방식은 정확도를 약간 희생하는 대신 매우 낮은 지연을 가지며, A_{req}=0.86일 경우 R=70 Mbps에서 server-only의 지연 성능과 교차한다. 이는 채널이 매우 좋아져 R이 크게 증가하면 더 이상 단말에서는 추론하지 않고 서버에서만 하는 것이 더 유리함을 의미한다.



(a) (b)

그림 4. (a) 지연 대 요구 정확도(A_{req}) 및 (b) 지연 대 전송 속도(R)
Fig. 4. (a) Latency vs. A_{req} and (b) Latency vs. R.

Ⅳ. 결 론

본 연구에서는 단말과 엣지 서버 간 협력 추론 방식을 제안하고, 추론 정확도를 보장하면서 종단 간 지연을 최소화하기 위한 최적의 신뢰도 임계값을 도출하였다. 실험 결과, 제안 방식은 정확도 요구사항에 따라 device-only와 server-only 방식 간의 균형 잡힌 성능을 보여주며, device-only 방식보다 높은 정확도와 server-only 방식보다 낮은 지연 시간을 달성하였다. 또한, 무선 채널 상태에 따라 적절한 추론 방식을 선택할 필요가 있음을 밝혔다. 제안된 협력 추론 방식이 엣지 컴퓨팅을 활용한 지능형 감시 시스템에 효과적으로 적용될 것으로 기대된다.

References

[1] D.-H. Lim and D.-W. Park, "Artificial intelligence CCTV operation model applying road markers," *J. KICS*, vol. 48, no. 1, pp. 74-80, Jan. 2023. (https://doi.org/10.7840/kics.2023.48.1.74)

[2] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738-1762, Aug. 2019. (https://doi.org/10.1109/JPROC.2019.2918951)

[3] Y. Zhao, et al., "Deep learning based couple-like cooperative computing method for IoT-based intelligent surveillance systems," in *Proc. IEEE PIMRC*, pp. 1-4, Istanbul, Sep. 2019. (https://doi.org/10.1109/PIMRC.2019.8904229)

[4] A. A. Ahmed and M. Echi, "Hawk-eye: An AI-powered threat detector for intelligent surveillance cameras," *IEEE Access*, vol. 9, pp. 63283-63293, Apr. 2021. (https://doi.org/10.1109/ACCESS.2021.3074319)

[5] A. Chaurasia and G. Jocher, "Ultralytics yolov8 documentation," 2023. (https://docs.ultralytics.com)