임베디드 NPU를 위한 Global Downsampling 구조 개선 기반 초경량 얼굴 인식 모델 최적화 연구

남현우*, 김동재

부경대학교 정보통신공학전공

nhwzzang@pukyong.ac.kr, kdj6306@pknu.ac.kr

Optimization of Ultra-Lightweight Face Recognition Model via Global Downsampling Redesign for Embedded NPUs

Nam Hyun Woo, Kim Dong Jae

Major of Information and Communication Eng., Pukyong National University

요 약

본 논문은 임베디드 NPU 환경에서 실시간 동작이 가능한 초경량 얼굴 인식 임베딩 모델의 최적화 방법론을 제안한다. 기존의 모바일 최적화 모델들이 소프트웨어적 연산량(MACC) 감소에 집중한 것과 달리, 본 연구에서는 실제 NPU 하드웨어의 연산자 지원 여부와 메모리 대역폭을 고려한 'NPU 친화적 네트워크 설계'를 제안한다. 주요 제안 사항으로, MobileFaceNet backbone의 핵심 병목 구간인 Global Downsampling 영역을 세 가지 방식(GDConv, 3×3 DWConv Stack, GAP)으로 재설계하여 비교 분석한다. 이 후 실험을 통해 Global Downsampling 구조의 단순화가 NPU Offload 비율을 극대화하고, 추론 지연시간과 가중치 및 메모리 사용량을 유의미하게 절감할 수 있음을 입증한다. 특히 최종 제안 모델(v3)은 기준 구조 대비 하드웨어 가속 효율을 최적화하여 100%의 NPU offload rate를 달성하고, 메모리 자원이 극히 제한된 임베디드 환경에서도 안정적인 실시간 추론이 가능함을 확인하였다.

Ⅰ. 서 론

얼굴 인식은 출입 인증, 단말 잠금 해제, 엣지 기반 사용자 식별 등 다양한 분야에서 활용된다. 최근에는 개인 정보 보호 및 네트워크 비용 문제로 인해 클라우드 대신 단말(임베디드)에서 직접 추론하는 온디바이스(on-device) 방식이 확대되고 있다. 그러나 임베디드 환경은 연산량(MACC), 메모리(RAM/Flash), 전력 제약이 크며, 특히 NPU가 탑재된 플랫폼에서는 모델이 가속 지원 연산자로 구성되는지에 따라 추론 속도 및 NPU offload 비율이 크게 달라지고, 이는 지연시간과 에너지 효율에 직접적인 영향을 미친다.

본 연구는 초경량 얼굴 임베딩 모델을 대상으로 임베디드 NPU에서 안정적으로 offload되는 연산 그래프를 구성하고, Global Downsampling 구간을 재설계하여 지연시간 및 메모리 사용량을 개선하며, PTQ 기반 INT8 양자화 및 임베디드 포팅 시 성능 변동 요인을 최소화하는 파이프라인을 제시하는 것을 목표로 한다.

Ⅱ. 본 론

2.1 데이터셋

본 연구는 AIHub 한국인 얼굴이 포함된 9개의 데이터셋을 전처리한 자체 구축 데이터셋을 사용하여 학습을 수행하였다. MS1M 데이터셋을 모델로 Insightface 얼굴 검출 모델을 사용하여 입력 이미지에서 양눈, 코끝, 입꼬리 랜드마크 5점을 검출하여 Arcface 템플릿에 정렬 후 112×112 크기로 crop하였다. 총 ID 18,249개, ID 당 평균 약 35장으로 이루어진 한국인 얼굴 데이터셋이 학습에 사용되었다.[1]

2.2 학습 설정

본 연구의 기반 모델(Backbone model)은 초경량 얼굴 인식에 특화된 MobileFaceNet을 사용하였다. 모델 학습은 ArcFace 손실 함수를 적용한 분류 헤드(head)를 통해 수행되었으며, 하이퍼파라미터인 Scale(s)과 Margin(m)은 각각 64.0과 0.50으로 설정하였다. 최종 임베딩 차원은 128로 고정하였다.

최적화 알고리즘으로는 SGD(momentum=0.9)를 사용하였고, 배치 크기는 128, 총 에폭(Epoch)은 50으로 설정하였다. 초기 학습률은 0.1에서 시작하여 Cosine decay 스케줄을 적용해 점진적으로 감소시켰으며, 학습의 안정성을 위해 초기 일정 에폭 동안 Warm-up을 적용하였다. [2]

2.3 평가 방식 설정

학습된 backbone은 추론 최적화를 위해 임베딩 추출 전용 모델(ONNX/TFLite)로 변환되었으며, PTQ(Post-Training Quantization)를 통해 INT8로 양자화되었다.

임베디드 성능 평가는 STM32N6 하드웨어와 Neural-ART 런타임을 타깃으로 수행하였으며, ST AI Developer Cloud를 활용하여 모델 컴파일 및 프로파일링을 진행하였다. 하드웨어 효율성 검증을 위해 추론 지연시간(Latency), NPU offload 비율, 모델 크기(Weights), 그리고 런타임 메모리(Flash/RAM) 점유량을 측정하였다.

얼굴 인식 성능은 전체적인 분류 성능을 나타내는 ROC_AUC와 보안 환경에서의 성능 지표인 TPR@FPR(10^{-3}, 10^{-4}), 그리고 일반화 성능 검증을 위한 LFW(Labeled Faces in the Wild) 벤치마크 정확도를 통해 다각도로 평가하였다.

2.4 Backbone 구조(공통)

입력 이미지 $112 \times 112 \times 3$에 대하여, 초기 특징 추출을 위한 Stem Block(3×3 Conv stride 2, DWConv, PWConv)과 MobileNetV2 스타일의 Inverted Residual Block을 기본 Backbone으로 사용한다. 전체 네트워크는 특징 맵의 해상도 변화에 따라 총 5개의 스테이지로 구성되며, 상세 채널 및 해상도 전이 과정은 다음과 같다: 56^2(64ch)→28^2(128ch)→14^2(128ch)→14^2(128ch)→7^2(128ch), 최종 스테이지 이후 1×1 Conv를 통해 채널을 512로 확장하여 $7 \times 7 \times 512$ 크기의 특징 맵을 생성하며, 이는 Global Downsampling 구간의 입력으로 사용된다.[3]

2.5 v1/v2/v3 구조 차이(Global Downsampling)

본 연구의 핵심 차별점은 $7 \times 7 \times 512$ 특징 맵을 $1 \times 1 \times 128$ 임베딩 벡터로 압축하는 Global Downsampling 구간의 설계 방식에 있다. NPU의 가속 효율과 메모리 사용량을 최적화하기 위해 다음과 같이 세 가지 버전을 설계하였다.

v1 (Global Depthwise Conv 기반): MobileFaceNet의 표준 구조로, 7×7 DWConv(valid padding)를 사용하여 해상도를 1×1로 즉시 축소한다. 이후 두 차례의 1×1 Conv를 통해 128차원의 최종 임베딩을 생성한다. 대형 커널(7×7) 사용으로 인해 일부 NPU 런타임에서 하드웨어 가속이 제한될 수 있다.

v2 (Stacked 3×3 DWConv 기반): v1의 대형 커널 문제를 해결하기 위해 NPU 친화적인 3×3 DWConv(stride 2)를 3회 중첩하여 $7 \rightarrow 4 \rightarrow 2 \rightarrow 1$로 단계적으로 해상도를 축소한다. 하드웨어 최적화 연산자를 활용하여 NPU Offload 비율을 극대화하도록 설계하였다.

v3 (Global Average Pooling 기반): 연산 효율을 극대화하기 위해 해상도 축소를 Global Average Pooling(GAP)으로 대체하고, 중간 512×512 투사층을 제거하였다. GAP 이후 단일 1×1 Conv ($512 \rightarrow 128$)만을 배치하여, 인식 성능을 유지하면서도 파라미터 수와 메모리 사용량(Activation footprint)을 최소화하였다.

Ⅲ. 실험 및 결과

3.1 실험 환경 및 설정

제안 모델의 성능 평가를 위해 112×112 해상도의 이미지를 입력으로 사용하였다. 임베디드 벤치마크는 STM32N6 하드웨어 환경에서 Neural-ART 런타임을 통해 측정되었으며, 모든 모델은 INT8 PTQ(Post-Training Quantization)를 적용하였다. 알고리즘 정확도는 자체 구축 한국인 얼굴 데이터셋과 LFW(Labeled Faces in the Wild) 공개 데이터셋을 사용하여 검증하였다.

3.2 임베디드 성능 분석

표 1은 각 버전별 하드웨어 자원 점유량 및 추론 지연시간을 비교한 결과이다.

표 1. 버전별 임베디드 성능 비교(INT8 기준)

Model	MACC (M)	Latency (ms)	Weights (KB)	Act (KB)	Flash (KB)	RAM (KB)	NPU rate (%)
v1	109	6.266	799.09	820.75	983.51	822.42	77.8
v2	109	5.093	871.94	820.75	1030	820.76	100
v3	109	3.609	498.83	624.75	593.32	624.76	100

실험 결과, 7×7 GDConv를 사용한 v1은 NPU Offload 비율이 77.8%에 그쳐 CPU 폴백에 의한 지연시간이 발생하였다. 이를 3×3 DWConv 스택으로 재구조화한 v2는 NPU rate 100%를 달성하며 지연시간을 약 18.7% 단축하였다. 최종 제안 모델인 v3는 GAP를 적용하고 중간 프로젝션 레이어를 제거함으로써 V1 대비 지연시간을 약 42.4% 개선(3.609ms)하였으며, 특히 가중치(Weights)와 Flash 메모리 사용량을 획기적으로 낮추어 임베디드 환경에서의 효율성을 극대화하였다.

3.3 얼굴 인식 성능 평가

모델 구조 변경 및 최적화에 따른 인식 정확도 변화를 확인하기 위해 ROC_AUC 및 TPR@FPR 지표를 분석하였다(표 2 참조).

표 2. 모델 버전별 인식 성능 비교

version	ROC_AUC	TPR@FPR=1e-3	TPR@FPR=1e-4	LFW
v1	0.9996	0.9808	0.9453	0.9674
v2	0.9996	0.9791	0.9247	0.9660
v3	0.9996	0.9784	0.9332	0.9656

3.4 결과

세 모델 모두 ROC_AUC 0.9996으로 매우 높은 판별 성능을 유지하였다. 하드웨어 최적화가 가장 많이 적용된 v3의 경우에도 LFW 정확도 하락이 v1 대비 0.18%p에 불과하여, 모델 경량화 및 NPU 최적화 과정에서도 얼굴 인식의 핵심적인 특징 추출 성능이 견고하게 유지됨을 확인하였다. 특히 v3는 v2 대비 TPR@FPR=$1e^{-4}$ 지표에서 오히려 소폭 우세한 성능을 보여, Global Average Pooling이 초경량 모델의 일반화 성능 유지에 효과적임을 입증하였다.

Ⅳ. 결 론

본 논문은 임베디드 NPU의 하드웨어 특성을 반영하여 초경량 얼굴 인식 모델의 추론 효율을 극대화하는 구조 개선 방법을 제안하였다. 특히 Global Downsampling 구간을 NPU 친화적 연산자로 재설계함으로써 연산 병목을 해소하였다.

연구 결과, 최종 제안 모델(v3)은 기존 모델 대비 정확도 하락을 0.2%p 미만으로 억제하면서도 100%의 NPU Offload와 약 42.4%의 지연시간 단축(3.609 ms)을 달성하였다. 또한 모델 가중치와 메모리 점유량을 대폭 낮추어 자원 제약이 큰 MCU급 시스템에서의 실용성을 확보하였다. 향후 양자화 인식 학습 및 지식 증류 기법을 적용하여 모델의 최적화와 인식 정확도를 동시에 개선하는 연구를 지속할 계획이다.

참 고 문 헌

[1] Yandong Guo. “MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition”, Jul 2016, (https://doi.org/10.48550/arXiv.1607.08221)

[2] H. Nam, "GhostFaceNets: Lightweight Face Recognition Model," Jan. 2024, (https://doi.org/10.48550/arXiv.2401.03462)

[3] Chen, S., Liu, Y., Gao, X., and Han, Z. “MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices,” arXiv preprint, arXiv:1804.07573v4, 2018.