붙임1

연구논문/작품 제안서

2025년도 제 2학기

논문/작품	○논문( ) ○작품(✓) ※ 해당란에 체크
제목	스마트폰을 위한 경량 ResNet 기반 사진 자동 태깅 시스템
GitHub URL	https://github.com/ohsj3781/FinalProject
팀원명단	오승재 (학번: 2020314916)

2025년 9월 18일

지도교수 : 신동군 서명

ㅇ 페이지 번호는 필수입니다.
ㅇ 그림은 절대로 남의 그림을 copy 하지 않습니다. 본인이 직접 그림을 그립니다. (단, 화면 capture는 그 수를 제한해서 넣습니다.)
ㅇ 그림에 fonts는 size와 fonts 이름이 전체 report를 통해 일치해야 합니다.
ㅇ 최대 그림 사이즈는 반 페이지로 제한합니다.
ㅇ 그림과 표에 caption은 필수입니다. 단, 표는 위에 그림은 아래에 caption을 답니다.
ㅇ 제안서 분량은 10페이지 내외로 작성합니다.

1. 과제의 필요성 (3페이지 내외)

Abstract

스마트폰과 SNS의 활성화로 이미지 데이터가 급격히 증가하면서 사진 관리를 위한 자동 태깅 기술의 필요성이 대두되고 있습니다. 기존 서버 기반의 이미지 분석 기술은 높은 정확도를 보이지만, 모바일 기기에서의 사용은 성능 및 배터리 소모의 한계가 있으며, 외부 API 사용 시 개인정보 유출 위험이 있습니다. 이를 극복하기 위해 스마트폰 내부에서만 동작하는 온 디바이스(On-Device) AI 기술이 강조되고 있으며, 본 연구는 낮은 비트 정밀도의 양자화 기술과 정수 연산 기반의 추론 방식을 결합하여 모바일 환경에서도 빠르고 안전하게 작동 가능한 ResNet 기반 사진 자동 태깅 시스템을 제안합니다. 이 시스템은 빠른 처리 속도와 보안성을 제공하여 사용자 편의성 향상과 함께 산업적 활용 가능성이 큽니다.

현대 사회는 스마트폰의 급속한 보급과 소셜 네트워크 서비스(SNS)의 활성화로 인해 이미지 데이터가 폭발적으로 증가하고 있으며, 이러한 데이터를 효과적으로 관리하고 활용하기 위한 기술적 필요성이 커지고 있다. 사람들은 일상생활 속에서 스마트폰 카메라로 다양한 순간을 손쉽게 촬영하고 공유하지만, 급증하는 이미지 데이터를 정리하거나 원하는 사진을 빠르게 찾기 위해서는 사진에 태그를 붙이는 과정이 필수적이다. 그러나 대다수의 사용자는 수많은 사진에 일일이 수동으로 태그를 붙이는 것이 매우 번거롭고 시간 소모적인 작업으로 여겨져, 사실상 제대로 된 사진 관리를 하지 못하고 있는 상황이다.
이러한 배경 속에서 이미지의 내용을 자동으로 분석하고 해당 내용을 기반으로 태그를 생성하여 효율적으로 사진을 분류하고 검색할 수 있도록 지원하는 자동 이미지 태깅 기술에 대한 수요가 증가하고 있다. 특히 스마트폰과 같은 모바일 환경에서는 제한된 하드웨어 자원과 배터리 소모를 고려해야 하기 때문에 고성능의

경량화된 이미지 처리 모델의 개발이 필수적이다.
기존 이미지 인식 및 분류 기술은 대부분 고성능 서버 환경에서 고정밀도의 부동 소수점 연산을 기반으로 한 대형 심층 신경망 모델을 활용하여 이루어졌으며, 모바일 기기에서는 계산량과 메모리 용량의 제한으로 인해 현실적으로 적용하기 어려웠다. 대표적으로 ResNet과 같은 고성능 심층 신경망은 ImageNet과 같은 대규모 데이터셋에서 매우 높은 정확도를 보이지만, 모바일 환경에서 이를 그대로 사용하는 것은 속도와 에너지 효율성 측면에서 큰 한계가 존재했다.
또한 이미지 데이터를 외부 API 연결을 통해 분석 및 태그를 붙일 경우, 사용자의 개인정보가 외부 서버로 전송될 수 있어 개인정보 유출에 대한 심각한 위험이 존재한다. 이에 따라 데이터 보안 및 프라이버시 보호를 위한 스마트폰 내부에서 모든 처리를 수행하는 온 디바이스(On-Device) AI 기술의 필요성이 강조되고 있다.
최근 인공지능 분야의 양자화 기술(Quantization)은 부동 소수점으로 표현된 신경망 파라미터를 낮은 비트의 정수로 표현하여 연산 속도 향상과 메모리 사용량 절감을 가능하게 하고 있다. 특히 "Learned Step Size Quantization(LSQ)"[1] 기술은 양자화 과정에서 각 신경망 계층의 양자화 간격(step size)을 학습 가능한 파라미터로 설정하여, 2~4비트의 낮은 정밀도에서도 원래의 높은 정확도를 유지하는 혁신적 방법을 제공한다. 이는 모바일 환경에서도 충분히 효율적인 성능과 정확도를 보장할 수 있는 가능성을 제시하고 있다.
아울러 Google의 "Integer-Arithmetic-Only Inference"[2] 기술은 TensorFlow Lite를 통해 신경망의 추론 과정을 완전히 정수 연산만으로 구현하여, 실제 모바일 프로세서 환경에서 효율적으로 작동할 수 있음을 보여주었다. ARM 기반의 프로세서(예: Qualcomm Snapdragon)를 탑재한 스마트폰 환경에서 부동 소수점 연산 대신 정수 연산을 사용하면 에너지 소비를 현저히 감소시키고, 계산 속도를 획기적으로 높일 수 있다는 점을 증명하였다.
최근 스마트폰은 NPU(Neural Processing Unit)와 같은 AI 전용 연산 유닛을 탑재하여 과거에 비해 더욱 성능이 뛰어난 온 디바이스 AI를 구현할 수 있게 되었다. 이에 따라 스마트폰 내부에서 작동하는 고성능의 자동 이미지 태깅 시스템의 효용성이 더욱 커지고 있다. 또한 개인당 사용하는 모바일 기기의 수가 지속적으로 증가하면서, 각각의 스마트폰에서 자체적으로 이미지 데이터를 처리하고 관리할 수 있는 자동 태깅 프로그램의 중요성은 더욱 부

각되고 있다.
이와 더불어, 애플(Apple)이 자사 제품군에 "애플 인텔리전스"라는 이름의 온 디바이스 AI를 도입하고, 삼성(Samsung) 역시 갤럭시 스마트폰에 온 디바이스 AI를 통한 이미지 수정 및 검색 기능을 탑재하는 등, 주요 글로벌 기업들이 경쟁적으로 온 디바이스 AI 기술을 상용화하고 있다는 점은 이러한 기술의 사회적 관심과 중요성이 점점 더 높아지고 있음을 방증한다
온 디바이스 AI는 처리 속도가 빠르고, 네트워크 연결이 불필요하며, 데이터가 외부로 전송되지 않아 개인정보 보호 측면에서 큰 장점을 가진다. 사용자는 외부 네트워크 상황에 구애받지 않고 안정적인 서비스 이용이 가능하며, 개인 이미지 데이터가 스마트폰 외부로 노출되지 않아 보안성을 높일 수 있다.
본 과제는 이와 같은 양자화 기술과 정수 기반 추론 기술을 결합하여 ResNet 모델의 성능을 모바일 환경에서도 거의 손실 없이 유지하면서, 스마트폰 내부에서 빠르게 작동하는 온 디바이스 자동 이미지 태깅 시스템을 구현하는 것을 목표로 한다. 이를 통해 사용자는 별도의 노력을 들이지 않고 사진 촬영과 동시에 즉각적인 태그 생성을 통해 사진 정리와 검색 효율을 극대화할 수 있게 된다.
이러한 자동 이미지 태깅 기술의 도입은 개인 사용자뿐만 아니라 다양한 산업 분야에서도 실질적인 이점을 제공한다. 예를 들어, 전자상거래 분야에서는 자동으로 상품 이미지를 분석하여 관련 태그를 생성함으로써 사용자가 원하는 제품을 쉽게 찾을 수 있게 하며, 미디어 콘텐츠 분야에서는 방대한 이미지 데이터베이스의 효율적인 관리 및 신속한 콘텐츠 추천을 가능하게 한다.
결론적으로, 모바일 환경에서 실시간으로 동작할 수 있는 양자화 및 정수 기반 추론 기술을 활용한 온 디바이스 AI 기반 ResNet 사진 자동 태깅 시스템의 개발은 사용자 편의성 향상뿐만 아니라 다양한 산업 분야의 데이터 관리 효율성 증대와 개인정보 보호를 위해 매우 필요한 과제로서 높은 연구 및 개발 가치를 지닌다.

2. 선행연구 및 기술현황 (2페이지 내외)

1. 과제 제안 목표 및 방향

본 과제의 주된 목표는 AI 기반의 사진 태그 프로그램을 개발하여 효율적인 이미지 분석과 정확한 태그 생성을 통해 사용자에게 유용한 기능을 제공하는 것이다. 이를 위해서는 고성능의 AI 모델이 필요하며, 동시에 모바일 및 임베디드 환경에서도 효율적으로 동작해야 한다는 조건을 충족해야 한다. 따라서 본 연구는 신경망 양자화(Neural Network Quantization)를 핵심 기술로 삼아, 정확성을 최대한 유지하면서 계산량과 메모리 사용량을 크게 줄이는 방향으로 진행된다.

2. 선행연구 분석 및 기술 현황

2.1 Learned Step Size Quantization (LSQ)[1]

『Learned Step Size Quantization』 논문은 신경망의 가중치 및 활성화를 저정밀도(2-bit, 3-bit, 4-bit)로 표현하면서도 높은 정확도를 유지할 수 있도록 스텝 사이즈(step size)를 학습하는 기법을 제안하였다. LSQ 방법은 기존의 고정된 양자화 방식과 달리, 네트워크 학습 과정에서 task loss를 최소화할 수 있는 방향으로 스텝 사이즈를 학습하여 최적화한다. 이 방법의 주요한 장점은:

- 간단한 구현 방식과 기존 학습 코드를 최소한으로 수정하여 적용 가능
- ImageNet 데이터셋에서 기존의 저정밀도 네트워크 방법 대비 현저한 정확도 향상
- 3-bit의 정밀도만으로도 완전 정밀도(full precision) 모델의 성능과 동등한 정확도를 달성할 수 있음을 최초로 입증 이러한 LSQ 접근 방식은 과제에서 목표로 하는 모바일 환경에서의 효율적인 추론 및 높은 정확도 유지를 위한 핵심 기술로 활용될 수 있을 것이다.

2.2 Quantization and Training for EfficientInteger-Arithmetic-Only Inference [2]

『Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference』 논문에서는 모바일 기기 및 임베디드 시스템에서의 효율적인 추론을 위해, 신경망의 가중치와 활성화를 8-bit 정수 연산만으로 수행하는 양자화 방식을 제안하였다. 이 방식의 핵심 특징은:

- 부동소수점 연산 대신 정수 연산만을 활용하여 모바일 및 임

베디드 하드웨어에서 높은 효율성 보장
- 정확성 손실을 최소화하기 위한 양자화 파라미터의 정교한 설계 및 학습 방법 제시
- MobileNet과 같은 효율적인 구조에서도 의미 있는 정확성 및 처리 속도의 개선
- 실제 ARM NEON 및 Qualcomm Hexagon과 같은 모바일 CPU 환경에서의 구체적인 성능 향상을 입증

이 기술은 과제의 실현 가능성을 높이고, 특히 모바일이나 임베디드 기기에서 효율적으로 사진 태그 프로그램을 구동하는 데 큰 기여를 할 것으로 예상된다

2.3. 결론 및 향후 연구 방향

본 과제는 LSQ의 학습 가능한 양자화 기법과 Integer-Arithmetic-Only Inference의 정수 연산 기반 효율성 향상 기법을 결합하여, 모바일 및 임베디드 환경에서도 고성능 사진 태그 AI 프로그램을 구축할 수 있는 현실적이고도 혁신적인 방향을 제안한다. 향후 연구는 이 두 기술의 결합을 통해 실제 모바일 하드웨어에서 최적화된 신경망 모델을 구현하고, 실환경 데이터셋을 이용한 검증과 추가적인 성능 개선 방법을 탐색하는 방향으로 이루어질 것이다. 이를 통해 효율성과 정확성을 모두 확보하는 실질적인 AI 응용 프로그램의 실현을 목표로 한다.

3. 작품/논문 전체 진행계획 및 구성 (2페이지 내외)

3.1 데이터 수집 및 전처리

모바일 환경에서 효율적인 이미지 분석을 위해 공개된 이미지 데이터셋인 COCO dataset을 활용하고, 추가적으로 스마트폰으로 촬영된 실제 환경의 사진 데이터를 수집하여 데이터셋을 구성한다. 이후 데이터의 크기 조정, 정규화 및 증강(Augmentation)을 포함한 전처리 과정을 수행하여 모델 학습에 최적화된 데이터셋을 구축한다.

3.2 기반 모델 선정 및 구현

ResNet 기반의 심층 신경망을 모바일 환경에 적합하도록 경량화하여 초기 기반 모델을 설계하고 구현한다. 모바일 환경에서의 연산 효율성을 극대화하기 위해 ResNet-18 또는 그 이하의 간소화된 구조를 활용하여 성능과 속도의 균형을 맞추는 작업을 수행한

다.

3.3 양자화 기술 적용 및 최적화
모델 경량화를 위해 핵심 기술인 LSQ(Learned Step Size Quantization)의 아이디어를 차용하여 모델의 가중치와 활성화를 8비트 정수로 양자화한다. LSQ 기법의 스텝 사이즈(step size)를 학습 가능한 파라미터로 설정하여 8비트 정밀도 내에서 네트워크의 정확도를 최대한 유지하면서 메모리 사용량과 연산 속도를 최적화하도록 한다

.
3.4 정수 연산 기반 추론 구현
모바일 환경에서 연산 효율성을 높이기 위해 Google의 Integer-Arithmetic-Only Inference 기술을 활용하여 신경망 추론 과정을 정수 연산만으로 구현한다. 이를 위해 TensorFlow Lite와 같은 프레임워크를 사용하여 모바일 프로세서(ARM NEON, Qualcomm Hexagon 등)에서의 최적화를 수행하고 성능 향상을 검증한다.

3.5 모바일 디바이스 최적화 및 테스트
구현된 양자화 및 정수 연산 기반 모델을 실제 모바일 기기 환경에 탑재하여 성능 테스트를 진행한다. 갤럭시 S24+에 내장 되어 있는 Exynos 2400 프로세서의 CPU와s NPU(Neural Processing Unit)를 적극 활용하여 실제 환경에서 모델의 속도, 정확도 및 전력 효율성을 평가하고 최적화 과정을 반복적으로 수행한다.

3.6 성능 평가 및 개선
구축된 시스템의 최종 성능을 검증하기 위해 정확도, 처리 속도, 에너지 소비량 등을 포함한 다양한 지표를 통해 평가를 진행한다. 사용자 피드백과 평가 결과를 기반으로 지속적인 개선 방안을 도출하고 추가적인 최적화 작업을 수행하여 시스템의 성능과 효용성을 높인다

4. 기대효과 및 개선방향 (1페이지 내외)

본 과제를 통해 구현되는 ResNet 기반의 온 디바이스 사진 자동 태깅 시스템은 다음과 같은 기대효과를 지닌다.

1. 모바일 환경에서의 실시간 이미지 인식 성능 향상

Learned Step Size Quantization(LSQ)을 통한 양자화 및 정수 연산 기반 추론 기술을 결합함으로써, 기존 서버 기반 방식에 의존하지 않고도 모바일 기기 내에서 빠르고 정확한 이미지 분류와 태깅이 가능해진다. 이는 처리 속도와 에너지 효율성을 동시에 개선하며, 실시간 반응성과 지속 가능한 사용자 경험을 제공할 수 있다.

2. 개인정보 보호 및 보안성 향상
   본 시스템은 클라우드 서버에 데이터를 전송하지 않고 스마트폰 내부에서 모든 처리를 수행하기 때문에, 민감한 이미지 데이터의 외부 유출 가능성을 원천 차단할 수 있다. 이는 사용자 프라이버시에 대한 사회적 요구가 증가하는 현재의 기술 환경에서 매우 중요한 장점이다.

3. 사용자 편의성 증대
   사진 촬영과 동시에 자동으로 적절한 태그가 부여되므로, 사용자는 일일이 수작업으로 사진을 정리할 필요가 없으며, 원하는 이미지를 빠르게 검색할 수 있다. 이는 일반 사용자뿐만 아니라 콘텐츠 크리에이터, 디지털 마케터 등 태그 기반 검색이 중요한 직군에서도 활용도가 높다.

4. 산업적 확장성 및 응용 가능성
   본 기술은 전자상거래, 미디어 콘텐츠 추천, 디지털 아카이빙, 스마트 갤러리 등의 다양한 산업 분야로 확장 가능하다. 특히 상품 이미지 자동 태깅, 유사 이미지 클러스터링 등으로 이어지는 응용이 가능하여 높은 산업적 가치를 지닌다.

5. 하드웨어 효율 기반 개선 여지
   향후 최신 모바일 SoC의 NPU 최적화를 통해 추론 속도 및 에너지 효율을 더욱 높일 수 있으며, 추후 4비트 및 2비트 양자화 등으로의 확장을 통해 메모리 및 모델 크기 측면에서 추가적인 개선 가능성도 존재한다.

따라서 본 과제는 AI 기술의 경량화, 모바일 실시간 추론, 프라이버시 중심 설계라는 시대적 요구에 부응하는 방향성을 갖추고 있으며, 향후 지속적인 개선과 확장을 통해 다양한 실생활 문제를 해결하는 데 기여할 수 있다

5. 기타 (1페이지 내외)

5.1 팀원간의 역할분담

오승재 작품 개발 및 그 외

월별 / 내용	3 (9)	4 (10)	5 (11)	6 (12)	7 (1)	8 (2)	9 (3)	10 (4)	11 (5)
연구노트 작성 및 프로젝트 진행	←	—	—	—	—	—	—	—	→
논문/작품 중간보고서							↔ 월 초		
논문/작품 최종보고서, 학회발표자료									↔ 월 초
논문/작품 발표회									↔ 월 말

,

6. 참고문헌 (1페이지 내외)

참고문헌 인용은 다음과 같이 합니다.

[1] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020.

[2] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," *arXiv preprint arXiv:1712.05877*, 2017.