<주요 Q&A>

Embedded Deep Learning
(딥러닝 경량화와 최적화 기술을 통한 Embedded AI 구현)

Q1. 딥러닝을 경량화하기 위해 SW 사이즈를 줄이면 어떤 문제점이 발생하게 되는지요? 이에 대한 중요 해결 방안은 어떻게 되는지요?

질문하신 SW 사이즈의 의미가 모델 사이즈라고 생각을 하고 답변 드리겠습니다. 일반적으로 모델 사이즈를 줄이면 정확도 성능이 떨어지게 됩니다. 그러나 최근에는 성능저하 없이 딥러닝 모델 사이즈를 줄이는 방법으로 다양한 경량 네트워크 설계 기술과 양자화/Pruning 기술 등이 공개되고 있습니다.

Q2. 센서에 프로세싱을 할 수 있는 프로세서를 장착하는 개념인가요?

이미지센서에는 ARM Cortex-M4 MCU 등이 내장되어있습니다.

Q3. 텐서플로우 라이트도 임베디드 AI 프레임워크이겠죠?

tensorflow-lite도임베디드 AI F/W입니다.

Q4. 이미지센서에는 arm이 장착되어 있지만 현재 보편화되어 있지는 않지 않나요? 그 보급이 과 표준이 문제일 것 같은데 이런 문제도 점점 해결되고 있는 것인가요?

현재 일반적으로 이미지센서에는 MCU가 내장되어있습니다. 그러나 MCU의 컴퓨팅 역량이 높지 않아서. AI 스마트센서를 만들기 위해서는 더 많은 컴퓨팅을 제공하는 프로세서가 추가될 필요는 있습니다.

Q5. 모바일향 생체인증 기술을 제품에 탑재하려는 경우 딥러닝 경량화와 최적화 기술을 최적으로 적용하고 활용하는 방법에 대해서 질문 드립니다.

실제로 모바일향 생체인증을 상용화하는 과정에서 개발한 경량화와 최적화 기술에 대해서 설명을 드렸습니다. 발표자료를 참고하세요.

Q6. 최근 이미지 크기가 점점 커지는데 그럼 그때마다 mcu의 성능을 업그레이드 해야 하는데 그렇지 않으면 효과성이 좀 떨어지지 않을까요?

하드웨어 측면에서는 MCU와 더불어 TinyNPU 등을 포함하는 방법도 있지만, 소프트웨어 측면에서도 경량화/고속화 연구를 더욱 발전시키는 게 필요합니다.

Q7. 값이 0에 가까워진다면 0으로 두고 연산은 skip한다는 뜻이었나요?

그럴 경우에 성능저하가 발생할 수 있기 때문에, 0이 아닌 값을 재 학습을 해서 정확도 복원이 가능합니다. 그리고 실제 Inference 에서는 0인 값인 Weight은 연산을 Skip합니다.

Q8. mac 연산과 bit wise 연산을 그 때 바꿔가면서 사용하는 면 좋을 것 같은데 이런 경우 변환을 해야 하므로 부하가 더 크고 속도저하가 더 크지 않을까요??

실제 레이어 별로 학습 과정에서 선택된 양자화 비트에 따라서...
Inference 처리를 레이어별 MAC연산과 Bit-wise연산 방식으로 구현하면 됩니다.

Q9. 다양한 임베디드 산업환경하에서 임베디드 AI가 제대로 자리잡으려면 제한된 하드웨어 사용을 고려한 경량화, 최적화된 산업분야별 맞춤형 응용엔진들이 필수적으로 요구 될 거라 생각하는데, 경량의 AI 응용 엔진들은 서로 기능이나 역량이 달라서 산업별 용도와 수요에 따라 인식, 추적, 예측에 필요한 적절한 엔진들을 선택하는데 어려움이 예상됩니다. 귀사에서는 이에 도움이 될만한 적절한 응용엔진들 또는 플랫폼 개발도 함께 진행되고 있는지 문의 드립니다.

동작 하드웨어 별(서버, 스마트폰, 센서 등) 요구되는 컴퓨팅과 메모리에 따라서 최적인 딥러닝 네트워크 설계하는 NAS 기술 등을 개발하고 있으며, 경량화 모델을 재 학습하고 임베디드 코드를 자동 생성하는 F/W를 개발하고 있습니다.

Q10. SOLVE의 기능들은 혹시 Open Source로 공개 할 계획은 없으신가요?

현재는 내년쯤에 사내 오픈 소스를 진행 예정입니다. 아직은 사외 오픈은 고려하고 있지는 않습니다.

Q11. 학습 최적화를 예약하거나 자동으로 바로 바로 실행하고 설정할 수 있고 그 결과를 바로 바로 피드백을 해줄 수 있나요?

기술원에서 개발중인 AI SW F/W은 학습된 Python 모델을 입력 받아서 모델 경량화와 실행코드 생성을 제공하는 솔루션입니다.

모델 재 학습 (양자화/Pruning)을 빠르게 처리하기 위하여, Early Stop 기능을 지원하는 자동화된 방식의 학습 매개변수 탐색 (Automatic Hyper-parameter Search) 방법을 사용하여 탐색 시간을 단축할 수 있습니다.
그리고 설정 범위 안에서 학습 Hyper-parameter를 변경하면서 최적 조합을 탐색하고 중간중간 결과를 보여줄 수 있습니다.

Q12. SOLVE 플랫폼을 이용하는 경우 아웃풋이 경량화된 모델인지? 아님 실제 타켓 기기에서 실행할 수 있는 엔진(코드)까지 포함되어 있는 것인지 궁금합니다.

SOLVE F/W의 입력은 사전 학습된 Pytorch 모델과 사용자 설정 세부 옵션을 주면, 경량화 모델을 만들고 대상 하드웨어에 적합한 최적 코드 (ARM NEON, AVX2, OpenCL, RISC-V)를 생성한 후에 성능 검증 결과 화면을 보여주고, 최종적으로 디바이스에 탑재 가능한 Library를 만들어줍니다.

Q13. Embedded 딥러닝 기술이 가까운 미래에 가장 효과적으로 적용되고 기여할 수 있는 산업분야에 대해서 문의 드립니다.

현재 모바일향 AI 기술 등이 점차 확대되고 있으며, 미래에는 자율주행이나 AIoT에서는 Edge-Device인 센서에서 동작하는 AI 기능이 필요해질 것을 생각합니다. 그래서 최근에는 TinyML이라는 기술 분야에 대한 관심이 많아지고 있습니다.(Tiny Machine Learning은 최신 임베디드 소프트웨어 기술로 엣지에서 컴퓨팅을 더 비용은 저렴하게 성능은 우수하게 인공지능으로 예측 가능하게 만드는 것입니다.)

Q14. Embedded 딥러닝 기술을 인공지능에 적용하고 활용하는 경우 현재 상황에서 극복해야 할 기술적인 한계나 제한사항에 대해서 질문 드립니다.

딥러닝기술은 일반적으로 많은 메모리와 컴퓨팅을 필요로 합니다.

영상 처리 분야 Image-to-image 응용 (디블러 처리, ISP 등)에서는 센서의 Pixel 사이즈 증가 (예, 2억화소) 등으로 더욱더 연산량이 증가하여 On-device에서 구현하기가 어렵습니다.

또한 경량화 (양자화/Pruning) 적용을 하면 화질성능 (PSNR)이 저하됩니다. 따라서 이러한 응용에 적합한 효율적인 경량화 기술개발이 필요합니다.