자료조사(출처: 소이넷 홈페이지)

SoyNet

- 인공지능의 성능 최적화 및 고속 추론이 실현 되도록 하는 솔루션을 주력으로 하는 기업이다
- 대표적인 당사의 솔루션 SoyNet은 C/C++ 기반으로 개발된 고속 추론(실행) 엔진으로, AI 추론 속도를 높이고 AI 시장 출시 시간을 단축할 수 있도록 설계되었다.

  SoyNet이라는 솔루션으로 해결할 것을 기대할 수 있는 문제들을 다음과 같다.

  1. AI 경량화

  AI 모델 자체가 많은 용량을 차지한다. 특히 딥러닝 모델의 경우에 그것이 심하기 때문에 사물인터넷에 적용하기가 어렵다.

  2. 적용이 어렵다

  결국은 AI를 어플리케이션과 같은 부분에 적용을 시켜야 실용성을 가질 수 있는데 AI 모델을 다른 서비스에 적용을 시키는 통합과정이 매우 어렵다.

  3. AI 모델의 속도

  연구실 서버에서는 빠르다. 다만, 이것은 그 서버가 계산에 쓸 수 있는 자원이 많아서인데, 실제로는 계산할 것이 매우 많아 AI 모델 자체는 느리다고 봐야 한다. (예를 들어 자율주행차와 같은 경우에 적용할 때 실시간 계산이 그리 빠르지 않아 아직은 문제가 있다)

  4. AI 모델의 발전

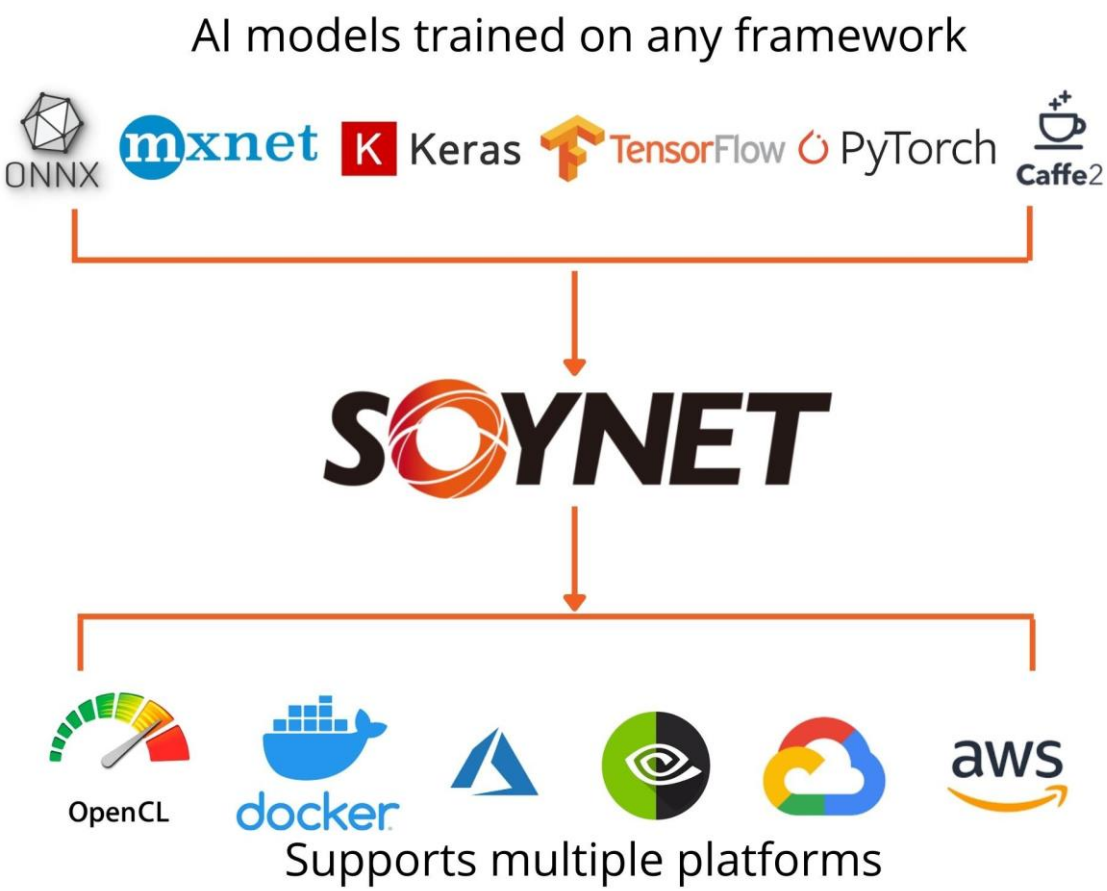
  AI 모델 자체가 굉장히 빠르게 변화하고 있고 이에 따라 새로운 모델들이 많이 나온다

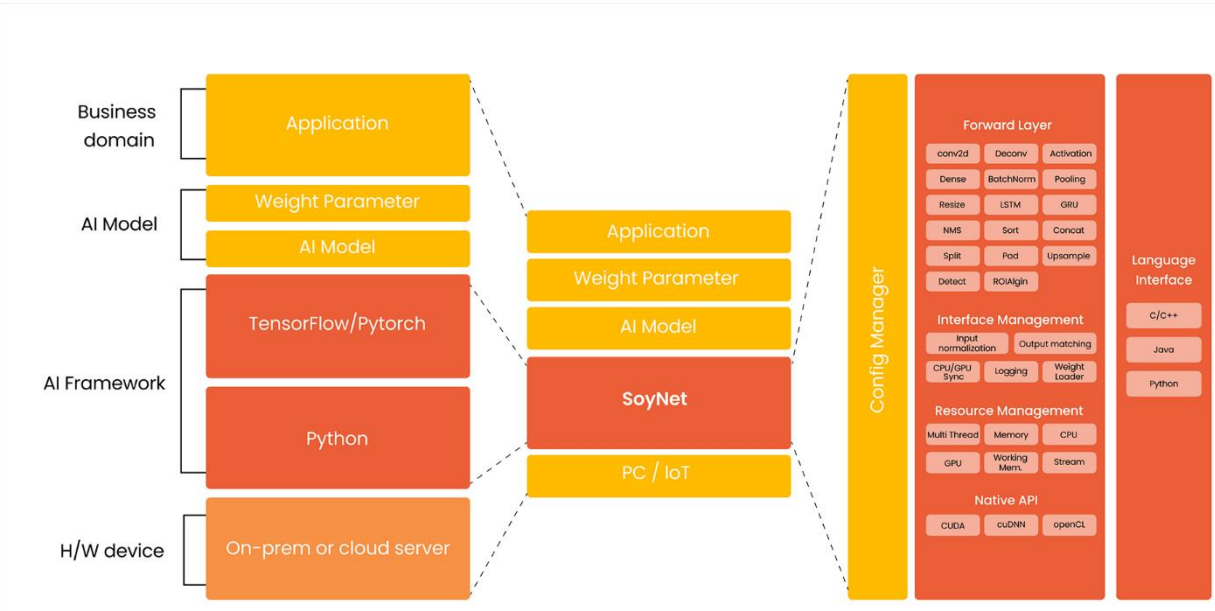
유니버설 액셀러레이터

SoyNet은 교육에 사용되는 TensorFlow / PyTorch / Caffe와 같은 공용 프레임워크에서 개발된 인공지능 모델을 지원한다.

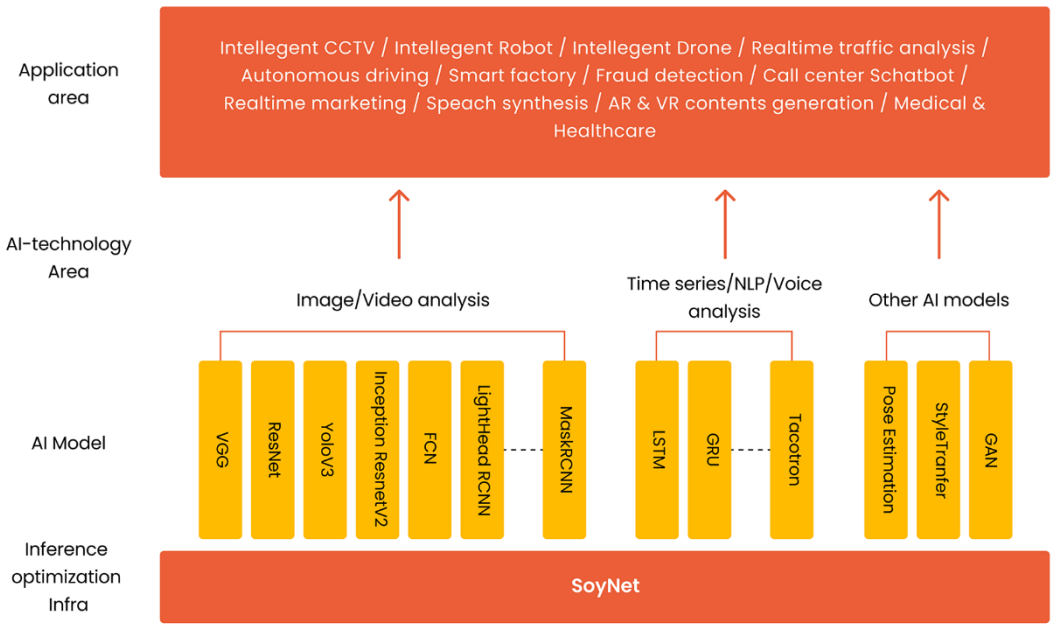
SoyNet은 또한 OpenCL 및 NVIDIA의 CUDA에서 모델 실행을 지원한다.

즉, 범용성이 있다.

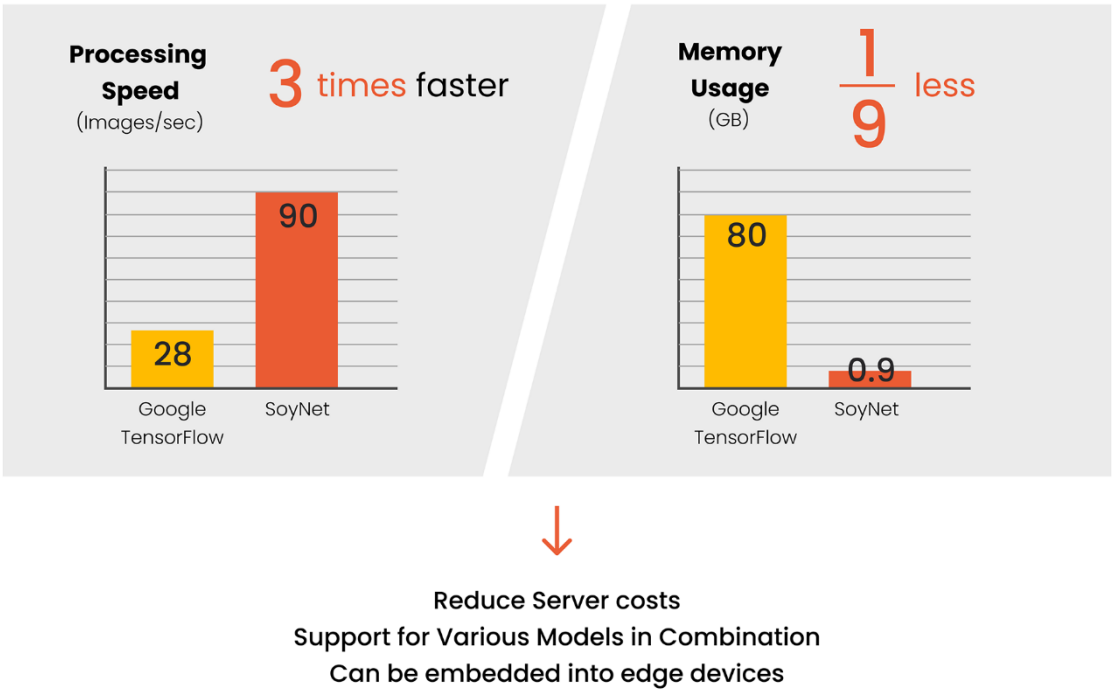




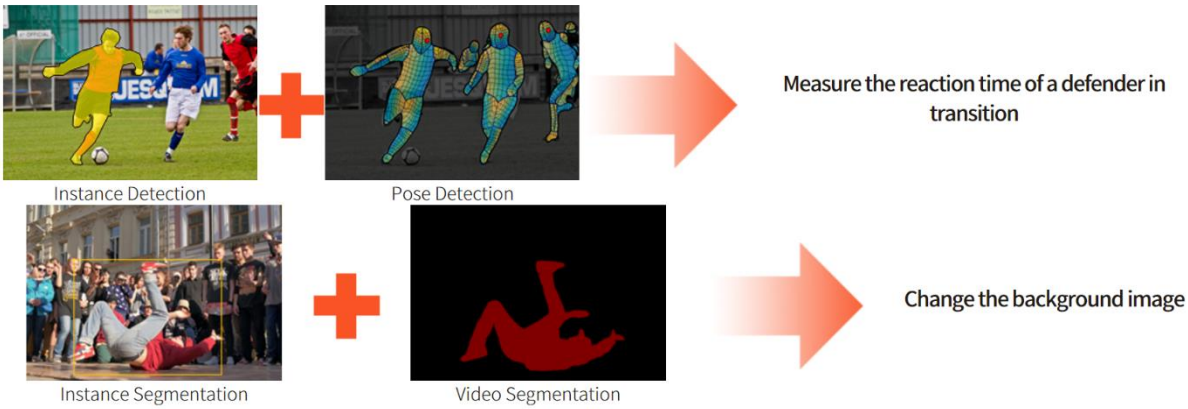
위 그림과 같이 SoyNet은 AI 모델이나 어플리케이션이 아니라 AI 프레임워크와 연관 있음을 보여준다. AI모델과는 별개로 성능향상을 지원한다.



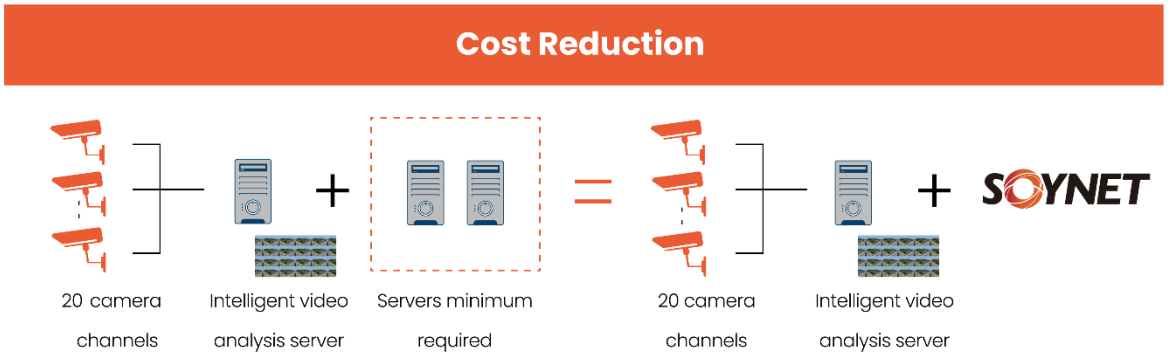
위 그림은 소이넷이 다양한 모델들에 전부 지원됨을 보여준다.



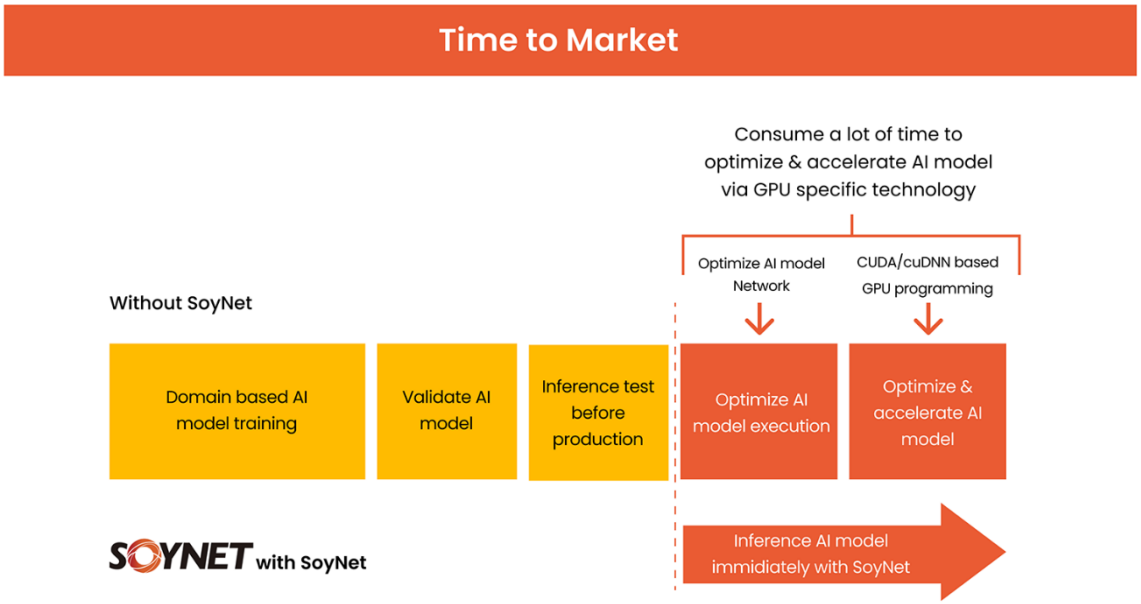
SoyNet에서 지원하는 성능향상의 정도이다. 메모리 사용과 시간에서 성능향상을 보였는데 이에 따라 비용절감도 할 수 있다.



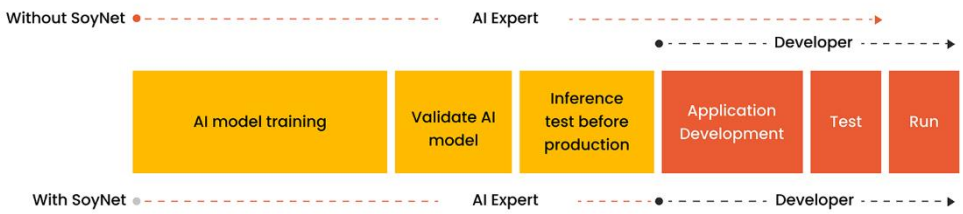
AI 모델도 사용에 따라 여러 모델이 나오는데 용도가 다른 두 모델을 결합해서 사용할 수 있게 해준다. 즉, 두모델을 같이 사용해서 내야하는 결과에 대해 이 두 모델을 따로 한번 씩 사용하는 것이 아닌 같이 사용하여 그보다 매우 효율적인 결과를 내주는 듯 하다.(두 모델의 결합에 대해서도 최적화를 지원한다)



소이넷의 서비스를 적용하면 최적화로 인해 자원을 아낄 수 있고 이를 CCTV에 적용하여 인공 감시망을 구성하면 대략 3배 정도로 많은 CCTV를 사용할 수 있게 된다.



실제 개발 과정에서 최적화와 가속을 실행할 때 많은 시간이 걸리는 것을 SoyNet을 이용하면 단순히 그것을 적용하는 것으로 쉽고 빠르게 할 수 있다.



소이넷 모델 적용은 쉽기 때문에 AI 전문가의 일을 줄여줄 수 있음을 강조할 수 있다

다음은 소이넷 홈페이지에 기록된 장점들이다.

SoyNet's Benefits

Reduced Costs

SoyNet supports larger processing tasks using the same equipment, lowering high-end server costs for service execution.

Support for High-Speed Processing

In various environments requiring high-speeds with limited resources, such as autonomous driving, SoyNet offers handy solutions to resolve execution speed issues.

Workable with No Internet Connection

Unlike existing Cloud-based inference processing, SoyNet can be embedded into edge devices, offering inference processing at the edge without Internet connection, thereby supporting businesses to expand their service scope.

Greater Value from Existing Developer Resources

SoyNet APIs enable businesses to develop AI services with their own existing application developer resources instead of highly-trained AI specialists, helping them harness their existing developer pool more efficiently.

우리 팀이 해야할 과제 개요는 다음과 같다:
1. 과제 개요

딥러닝 모델을 SoyNet에 포팅, 성능 최적화하는 업무 수행

-필요 기술에 대한 교육

Convolution, FC, BathNorm, Activation 등 연산자에 대한 알고리즘 이해

실무에 사용되는 딥러닝 모델 구조 이해, C/C++ 교육

병렬 처리를 위한 CUDA/OpenCLS/SIMD 교육

-딥러닝 모델 포팅

Inference 관점의 딥러닝 모델 분석

TensorFlow/PyTorch 구현을 SoyNet의 Config 방식으로 전환 포팅

Python으로 구현된 로직을 C/C++로 변환하며 일부는 CUDA/OpenCL을 사용하여 변환

-포팅된 모델에 대한 속도, 메모리 사용량 분석 및 최적화

1. 어떤 모델을 분석하는가?

2. 우리 팀에 원하는 구체적인 작업의 개요

3. 구체적인 학습에 대해서(ML, 멀티쓰레드, ...)

4. C/C++ 에 대해 우리가 해야할 것은? (실무적인 기능을 교육하는 것이 있다고 들었는데 어떻게 진행되는지)

5. SoyNet의 config방식에 대한 설명요구

6. 대략적인 일정은? 예를 들어 딥러닝 모델 포팅은 언제까지 하는가와 같은 것.

7. 우리 결과물의 기업에서 기대하는 효과는?

8. 결과물에 대한 활용

9. 주제를 선정하면서 어떤 것을 기대하였는가. 어떤 방향성을 가지고 활동하는 것이 바람직할까.