

도입



Bentoml: Unified Model Serving Framework

- BentoML은 다양한 환경에 쉽게 공유하고 배포할 수 있는 형식으로 기계 학습 모델을 패키징하기 위한 오픈 소스 프레임워크





BentoML vs Flask

- BentoML은 기계 학습 모델 서빙을 위한 프레임워크이며, Flask는 웹 애플리케이션 개발을 위한 마이크로 웹 프레임워크이다.

- BentoML은 모델 패키징, 서빙, 모니터링을 통합하여 제공합니다. 반면, Flask는 다양한 웹 애플리케이션을 개발하기 위한 유연한 도구입니다.
- BentoML은 다양한 모델 타입(PyTorch, TensorFlow, XGBoost 등)을 지원한다.

Installation

 BentoML is distributed as a Python Package available on PyPI. Install BentoML alongside with whichever deep learning library you are working with, and you are ready to go! BentoML supports Linux/...

 <https://docs.bentoml.org/en/latest/installation.html>

특징

특징 및 장점

- 다양한 머신러닝 프레임워크를 지원합니다.
- 대규모 모델 배포를 위한 클러스터링을 지원합니다
 - Online API Serving(실시간 처리)
 - Offline Batch Serving(Batch 처리)
- 모델 패키징, 서빙 및 모니터링을 통합하여 제공합니다.
- 다양한 모델 타입 (PyTorch, TensorFlow, XGBoost 등)을 지원합니다.
- 클라우드나 온프레미스 환경에서 유연하게 사용 가능합니다.
 - Containerization 가능
- BentoML API를 통해 쉽게 모델 서빙할 수 있습니다.

BentoML 을 활용하여 딥러닝 모델 API 서빙하기

'투자증 뉴스 카테고리 분류 딥러닝 모델'을 BentoML 로 패키징하여 서빙한 경험에 대해 공유합니다.

[Z https://zuminternet.github.io/BentoML/](https://zuminternet.github.io/BentoML/)



Installation

```
pip install git+https://github.com/bentoml/bentoml # Install from Source
# pip install "bentoml[all]"
```

Migration

1.0 Migration Guide

BentoML version 1.0.0 APIs are backward incompatible with version 0.13.1. However, most of the common functionality can be achieved with the new version. We will guide and demonstrate the migration...

<https://docs.bentoml.org/en/latest/guides/migration.html>

분류 모델 서비스 정의

`@artifacts()` 는 예측 서비스에 필요한 모델을 정의한다.

`@env()` 는 예측 서비스에 필요한 종속성 및 환경 설정을 지정한다.

`@api` 는 예측 서비스에 액세스하기 위한 endpoint이다. Inference API를 정의한다.

Tutorial

Tutorial: Intro to BentoML

time expected: 10 minutes In this tutorial, we will focus on online model serving with BentoML, using a classification model trained with scikit-learn and the Iris dataset. By the end of this tutor...

 <https://docs.bentoml.org/en/latest/tutorial.html>

```
docker run -it --rm -p 8888:8888 -p 3000:3000 -p 3001:3001 bentoml/quickstart:latest
```

Triton Inference Server

time expected: 10 minutes NVIDIA Triton Inference Server is a high performance, open-source inference server for serving deep learning models. It is optimized to deploy models from multiple deep le...

 <https://docs.bentoml.org/en/latest/integrations/triton.html>