Nama : Arfara Yema Samgusdian
NIM : 1103202004
Kelas : TK-44-04

Zero to Mastery Learn PyTorch for Deep Learning

09. PyTorch Model Deployment

Model deployment machine learning adalah proses membuat model machine learning Anda dapat diakses oleh orang atau sesuatu yang lain. Orang bisa menjadi seseorang yang dapat berinteraksi dengan model Anda, sedangkan sesuatu bisa berupa program, aplikasi, atau bahkan model machine learning lain yang berinteraksi dengan model machine learning Anda.

Model deployment sama pentingnya dengan melatihnya. Meskipun Anda dapat mendapatkan ide bagus tentang cara model Anda akan berfungsi dengan mengevaluasinya pada set uji yang baik atau memvisualisasikan hasilnya, Anda tidak pernah benar-benar tahu bagaimana performanya sampai Anda melepaskannya ke lingkungan yang sebenarnya.
Memungkinkan orang yang belum pernah menggunakan model Anda berinteraksi dengannya dan mengungkapkan kasus uji yang tidak terpikirkan selama pelatihan.
Jenis Penyelenggaraan Model Machine Learning:

On-device (edge/in the browser):

Keuntungan: Bisa sangat cepat (karena tidak ada data yang meninggalkan perangkat), menjaga privasi (tidak ada data yang harus meninggalkan perangkat), tidak memerlukan koneksi internet (kadang-kadang).
Kerugian: Keterbatasan daya komputasi (model yang lebih besar memerlukan waktu lebih lama untuk dijalankan), keterbatasan ruang penyimpanan (ukuran model yang lebih kecil diperlukan), memerlukan keterampilan khusus perangkat.
On cloud:

Keuntungan: Daya komputasi hampir tak terbatas (dapat diskalakan sesuai kebutuhan), dapat mendeploy satu model dan menggunakannya di mana saja (melalui API), terhubung dengan ekosistem cloud yang ada.
Kerugian: Biaya bisa meningkat (jika batasan penskalaan yang tepat tidak diberlakukan), prediksi bisa lebih lambat karena data harus meninggalkan perangkat dan hasil prediksi harus kembali (latensi jaringan), data harus meninggalkan perangkat (mungkin menimbulkan kekhawatiran privasi).

Cara Deploy Model Machine Learning

On-Device Deployment:
Google's ML Kit: Cocok untuk perangkat Android dan iOS.
Apple's Core ML dan coremltools Python package: Digunakan pada semua perangkat Apple.
Cloud Deployment:

Amazon Web Service's (AWS) Sagemaker: Layanan cloud.
Google Cloud's Vertex AI: Layanan cloud.
Microsoft's Azure Machine Learning: Layanan cloud.

Hugging Face Spaces: Layanan cloud.
API Deployment (Cloud/Self-hosted server):

API with FastAPI: Framework Python untuk membuat API cepat.
API with TorchServe: Server untuk menyajikan model PyTorch.
General Deployment:

ONNX (Open Neural Network Exchange): Format yang dapat digunakan untuk menyimpan dan menyelenggarakan model machine learning secara independen.

0. Persiapan:
Unduh kode yang telah ditulis untuk digunakan kembali.

1. Unduh Data:
Dapatkan dataset pizza_steak_sushi_20_percent.zip agar kita bisa melatih model terbaik sebelumnya pada dataset yang sama

2. Rencana Eksperimen Penyelenggaraan Model FoodVision Mini:
Meskipun di proyek ketiga ini, kita masih akan menjalankan beberapa eksperimen untuk melihat model mana (EffNetB2 atau ViT) yang mencapai kinerja terbaik sesuai dengan metrik tujuan.

3. Membuat Ekstraktor Fitur EffNetB2:
Replikasi model EfficientNetB2 sebagai kandidat untuk penyelenggaraan model.

4. Membuat Ekstraktor Fitur ViT:
Replikasi model Vision Transformer (ViT) sebagai kandidat untuk penyelenggaraan model, bersama dengan EffNetB2.

5. Melakukan Prediksi dengan Model Terlatih dan Mengukur Waktunya:
Melakukan prediksi dengan model EffNetB2 dan ViT yang telah dilatih, serta melacak hasilnya.

6. Membandingkan Hasil Model, Waktu Prediksi, dan Ukuran:
Membandingkan model untuk menentukan model mana yang mencapai kinerja terbaik sesuai dengan tujuan.

7. Membuat Demo FoodVision Mini dengan Gradio:
Mengubah salah satu model yang unggul menjadi aplikasi demo berfungsi.

8. Mengubah Demo Gradio FoodVision Mini menjadi Aplikasi yang Dapat Dideploy:
Menyiapkan aplikasi demo Gradio untuk penyelenggaraan.

9. Penyelenggaraan Demo Gradio ke HuggingFace Spaces:
Menyajikan FoodVision Mini secara daring agar dapat diakses oleh publik.