djl

DJL - ONNX Runtime engine implementation

Overview

This module contains the Deep Java Library (DJL) EngineProvider for ONNX Runtime.

It is based off the ONNX Runtime Deep Learning Framework.

We don't recommend developers use classes within this module directly. Use of these classes will couple your code to the ONNX Runtime and make switching between engines difficult.

ONNX Runtime is a DL library with limited support for NDArray operations. Currently, it only covers the basic NDArray creation methods. To better support the necessary preprocessing and postprocessing, you can use one of the other Engine along with it to run in a hybrid mode. For more information, see Hybrid Engine.

Documentation

The latest javadocs can be found on the djl.ai website.

You can also build the latest javadocs locally using the following command:

```
# for Linux/macOS:
./gradlew javadoc

# for Windows:
..\..\gradlew javadoc
```

The javadocs output is generated in the `build/doc/javadoc` folder.

System Requirements

Read the System Requirements for the official ONNX Runtime project.

Installation

You can pull the ONNX Runtime engine from the central Maven repository by including the following dependency:

- ai.djl.onnxruntime:onnxruntime-engine:0.16.0

```
<dependency>
    <groupId>ai.djl.onnxruntime</groupId>
    <artifactId>onnxruntime-engine</artifactId>
    <version>0.16.0</version>
    <scope>runtime</scope>
</dependency>
```

This package by default depends on com.microsoft.onnxruntime:onnxruntime

Install GPU package

If you want to use GPU, you can manually exclude com.microsoft.onnxruntime:onnxruntime and add com.microsoft.onnxruntime:onnxruntime_gpu
to your project.

Maven:

```
<dependency>
    <groupId>ai.djl.onnxruntime</groupId>
    <artifactId>onnxruntime-engine</artifactId>
    <version>0.16.0</version>
    <scope>runtime</scope>
    <exclusions>
        <exclusion>
            <groupId>com.microsoft.onnxruntime</groupId>
            <artifactId>onnxruntime</artifactId>
        </exclusion>
    </exclusions>
</dependency>
<dependency>
    <groupId>com.microsoft.onnxruntime</groupId>
    <artifactId>onnxruntime_gpu</artifactId>
    <version>1.11.0</version>
    <scope>runtime</scope>
</dependency>
```

Gradle:

```
implementation("ai.djl.onnxruntime:onnxruntime-engine:0.16.0") {
    exclude group: "com.microsoft.onnxruntime", module: "onnxruntime"
}
implementation "com.microsoft.onnxruntime:onnxruntime_gpu:1.11.0"
```

Enable TensorRT execution

ONNXRuntime offers TensorRT execution as the backend. In DJL, user can specify the followings in the Criteria to enable:

```
optOption("ortDevice", "TensorRT")
```
