**== REPORT ==**

**< Lab 4 >**

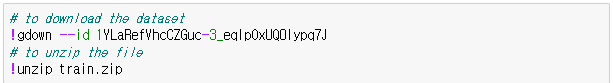
Deep Learning Model On-device Deployment

|  |  |
| --- | --- |
|  |  |
| 과 목 | 머신 러닝 |
| 담당 교수 | 정용진 교수님 |
| 제출 일자 | 2021-12-15 |
| 전공 | 전자통신공학과 |
| 학번 | 2019707044 |
| 이름 | 전영우 |

**- 목 차 -**

1. **Model Implementation**

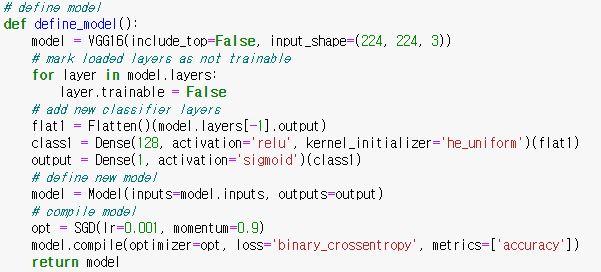
**2. On-device Deployment**

1. **Model Inference**
2. **Application Result**
3. **Conclusion**
4. **Model Implementation**
   1. **Download dataset**

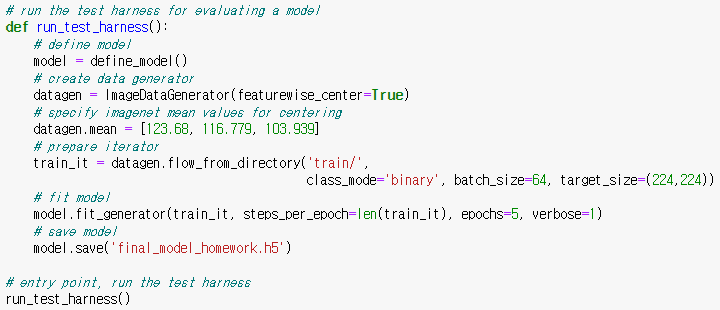
****

****

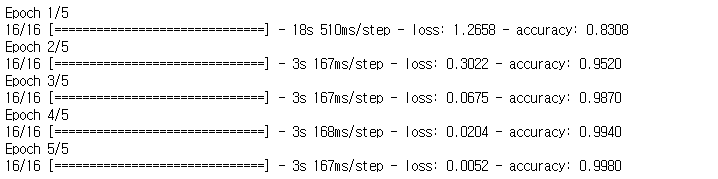
* 모델을 훈련시킬 데이터셋을 다운로드한다.
* 해당 데이터셋은 dandelion과 grass 클래스에 해당하는 이미지 총 999개로 이루어져 있다.
  1. **Define model**

****

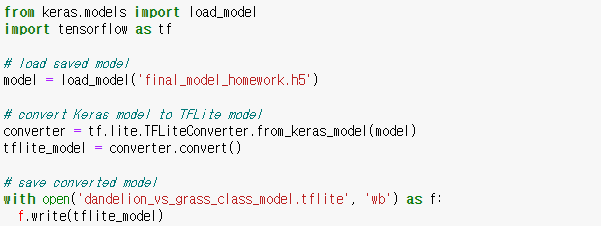
* 학습된 VGG16 모델을 분류기 부분을 제외하고 가져온다.
* Trainable 속성을 false로 설정하여 Feature Extraction 방법의 전이 학습을 사용한다.
* 새로운 분류기 층을 추가하고 컴파일 옵션을 준 뒤 생성된 모델을 반환한다.
  1. **Train model**

****

****

****

* 앞에서 다운로드한 데이터셋으로 모델을 훈련시킨다.
  1. **Save model and convert to TFLite**

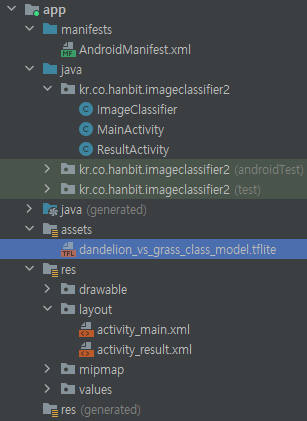
****

****

* 학습이 완료된 모델을 h5 파일 포맷으로 저장한다.
* 저장된 h5 파일을 불러와 TFLiteConverter로 tflite 포맷 파일로 변환한다.

1. **On-device Deployment**
   1. **Create android project**

* Empty activity로 새로운 안드로이드 프로젝트를 생성한다.
* Lab guide에 주어진 대로 클래스 및 레이아웃 파일을 생성하고 코드를 입력한다.
  1. **Make assets folder**
* Tflite 모델 파일을 안드로이드 프로젝트에 포함시키기 위해 app 폴더 아래에 assets 폴더를 생성한다.
  1. **Include tflite file in assets folder**

****

* 앞에서 생성한 tflite 파일을 assets 폴더에 포함시킨다.

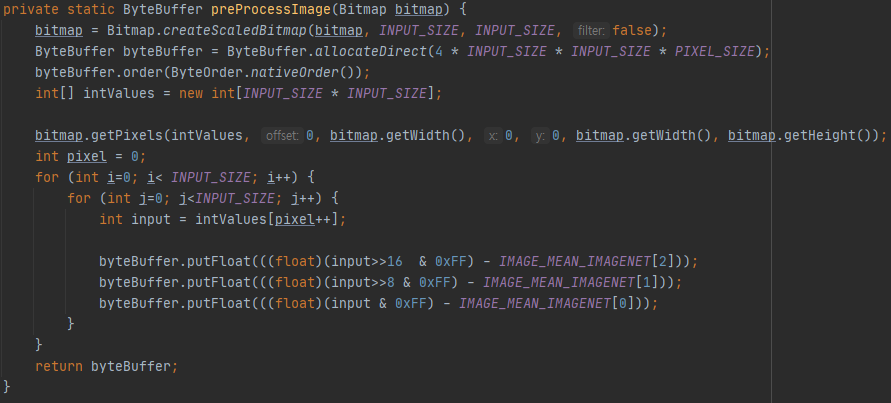
1. **Model Inference**
   1. **Initialize model**

****

* 모델을 초기화한다.
  1. **Model prediction**

****

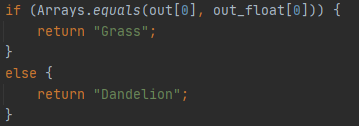
* 모델의 predict 메서드를 호출한다.
  1. **Preprocess input image**

****

* Predict 메서드가 호출되면 입력 이미지 전처리를 먼저 수행한다.
* 모델의 입력 요구 사이즈에 맞게 리사이징하고 이미지를 보간 한다.
  1. **Model inference**

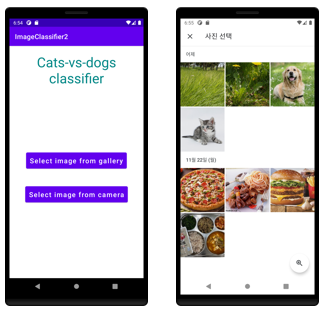
****

* 사용할 tflite 모델 파일의 Interpreter 객체의 run 메서드를 호출하여 모델 추론을 수행한다.
  1. **Return result**

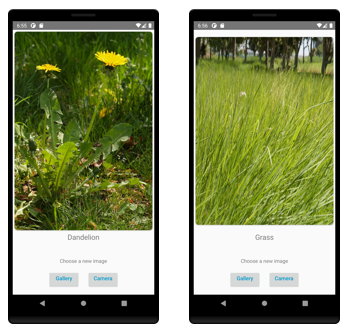
****

* 모델 추론 결과에 따라 결과 값을 반환한다.

1. **Application Result**
   1. **Image selection**

****

* 1. **Inference result**

****

1. **Conclusion**

* 지난 번 Lab2: Server-based model deployment에 이어서 이번 Lab4 에서는 On-device model deployment를 해보는 실습을 진행하였다. 서버-클라이언트 간 통신을 통해 모델을 훈련시키고 추론을 수행하는 서버 기반 모델 배포에 비해, 온 디바이스 모델 배포는 통신 과정이 필요 없어 더욱 간편하고 쉽게 모델을 사용할 수 있다.
* 또한 온 디바이스 모델 배포는 서버 비용이 불필요하고, 더욱 빠른 모델 추론이 가능하다.
* 하지만 서버 기반 모델이 입력 값으로 retrain을 수행할 수 있는 것에 비해, 온 디바이스 모델의 경우 사용자 기기 내에서 retrain은 아직까지 수행할 수 없는 것으로 알고 있다. TFLite 개발 팀 측에서 앞으로 사용자 기기 내에서 모델의 재 학습이 가능하도록 발전시킨다면, 더욱 강력한 수단이 될 것이라는 생각을 하였다.