ONNX client for Acumos

ONNX community meeting

03/25/2021

Philippe Dooze / Orange

Bruno Lozach / Orange

ONNX client for Acumos

Agenda

- ONNX & Acumos.
- Main requirements to on-board ONNX in Acumos.
- Onnx4acumos client.
    - ✓ Dump mode
    - ✓ Test Onnx model
    - ✓ Push mode

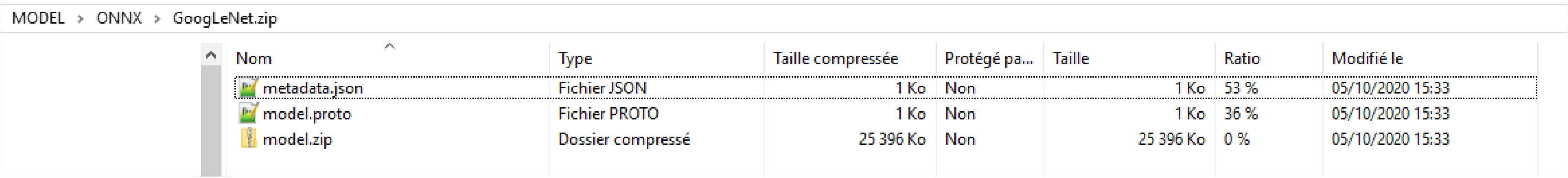
ONNX & ACUMOS

- ONNX is an open format built to represent machine learning models. ONNX defines a common set of operators - the building blocks of machine learning and deep learning models - and a common file format to enable AI developers to use models with a variety of frameworks, tools, runtimes, and compilers. https://onnx.ai/

- Acumos AI is a platform and an open source framework that makes it easy to build, share, and deploy AI apps. Acumos standardizes the infrastructure stack and components required to run an out-of-the-box general AI environment. This frees data scientists and model trainers to focus on their core competencies and accelerates innovation. https://www.acumos.org/

Main requirements to on-board ONNX in Acumos

- **Acumos needs some specific material (protobuf signature, meta data) in addition to the model itself to be able to manage it and to provide some specific features like Micro-service generation and design studio.**



MODEL › ONNX › GoogLeNet.zip

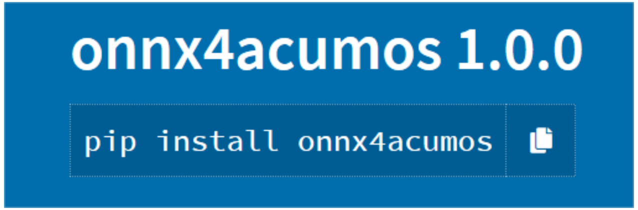
Nom	Type	Taille compressée	Protégé pa...	Taille	Ratio	Modifié le
metadata.json	Fichier JSON	1 Ko	Non	1 Ko	53 %	05/10/2020 15:33
model.proto	Fichier PROTO	1 Ko	Non	1 Ko	36 %	05/10/2020 15:33
model.zip	Dossier compressé	25 396 Ko	Non	25 396 Ko	0 %	05/10/2020 15:33

model.proto

metadata.json

Onnx4acumos client

- **Based on the existing Acumos python client, we developed the onnx4acumos client able to create a model bundle with all the required materials required by Acumos. (available on https://pypi.org/)**



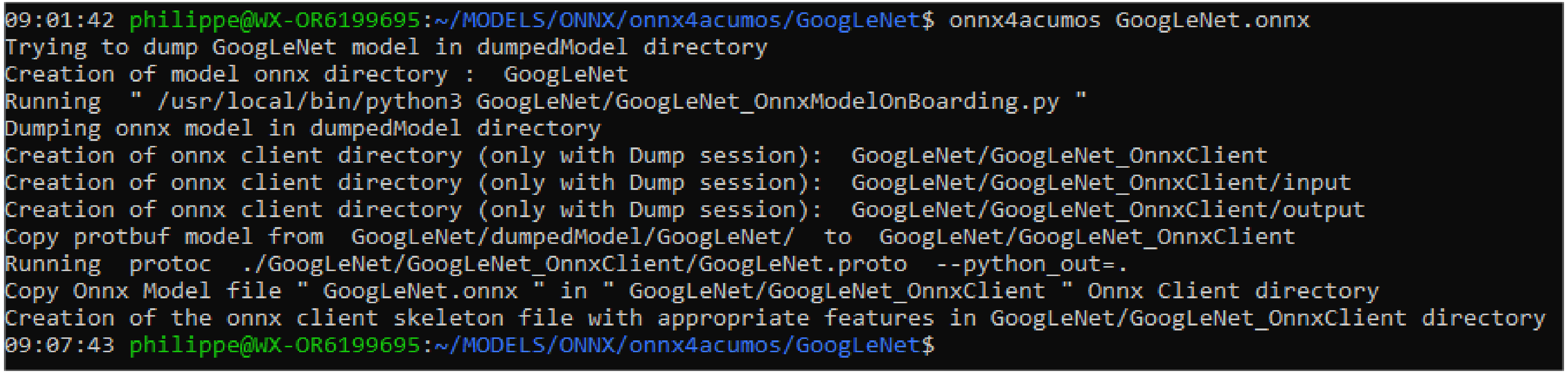
onnx4acumos 1.0.0

```
pip install onnx4acumos
```

- **The main python requirements are the following :**
  - **acumos (acumos python client)**
  - **onnx, onnxruntime, onnxruntime.backend**

Onnx4acumos client / Dump mode

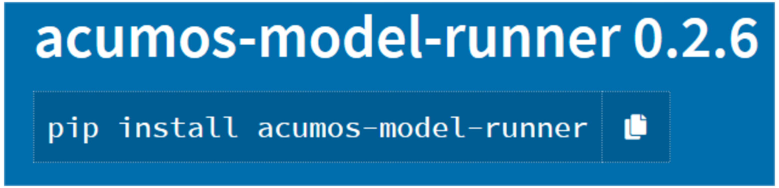
- **You can use onnx4acumos in a “Dump” mode for local test and later onboarding**
  - Dump mode : Create the model bundle and save it locally (for local test and later onboarding)



```
09:01:42 philippe@WX-OR6199695:~/MODELS/ONNX/onnx4acumos/GoogLeNet$ onnx4acumos GoogLeNet.onnx
Trying to dump GoogLeNet model in dumpedModel directory
Creation of model onnx directory :  GoogLeNet
Running  " /usr/local/bin/python3 GoogLeNet/GoogLeNet_OnnxModelOnBoarding.py "
Dumping onnx model in dumpedModel directory
Creation of onnx client directory (only with Dump session):  GoogLeNet/GoogLeNet_OnnxClient
Creation of onnx client directory (only with Dump session):  GoogLeNet/GoogLeNet_OnnxClient/input
Creation of onnx client directory (only with Dump session):  GoogLeNet/GoogLeNet_OnnxClient/output
Copy protbuf model from  GoogLeNet/dumpedModel/GoogLeNet/  to  GoogLeNet/GoogLeNet_OnnxClient
Running  protoc  ./GoogLeNet/GoogLeNet_OnnxClient/GoogLeNet.proto  --python_out=.
Copy Onnx Model file " GoogLeNet.onnx " in " GoogLeNet/GoogLeNet_OnnxClient " Onnx Client directory
Creation of the onnx client skeleton file with appropriate features in GoogLeNet/GoogLeNet_OnnxClient directory
09:07:43 philippe@WX-OR6199695:~/MODELS/ONNX/onnx4acumos/GoogLeNet$
```

Onnx4acumos client / Test ONNX model

- **If you have dumped your model locally, you can test it thanks to :**
  - The acumos_model_runner



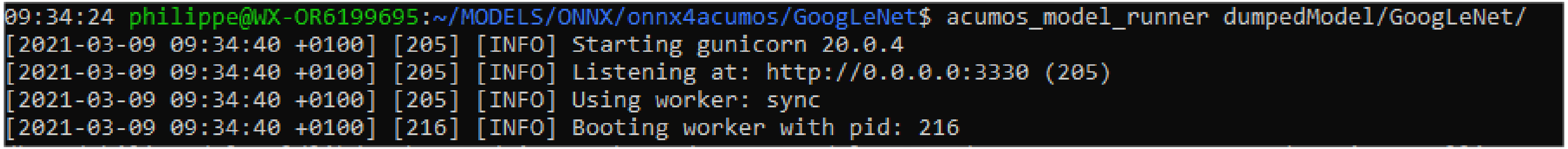
acumos-model-runner 0.2.6

```
pip install acumos-model-runner
```

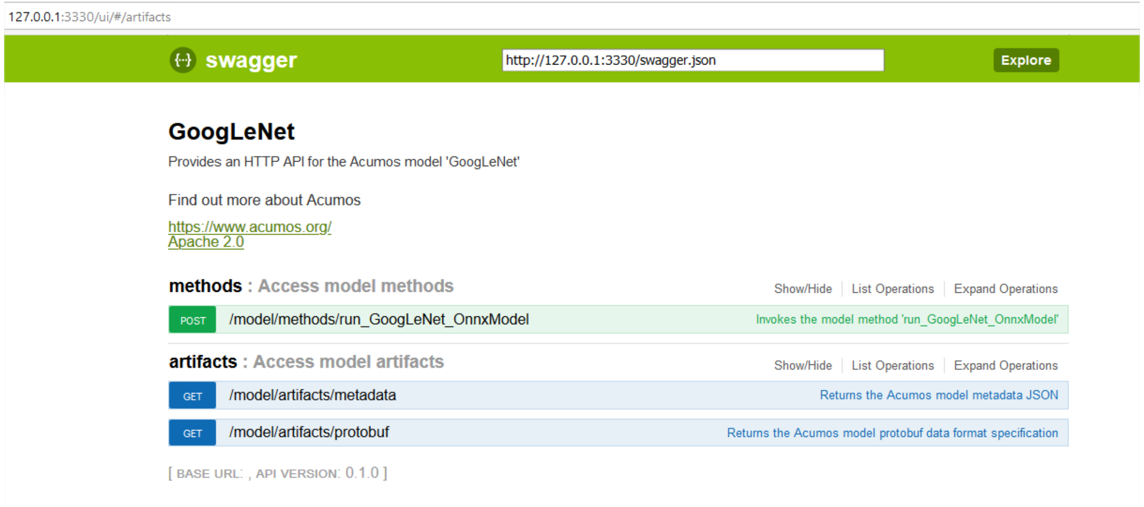
  - A skeleton python script that must be filled with Pre and Post processing data methods. This skeleton python script is provided by onnx4acumos in the following folder : *“ModelName”/”ModelName”_OnnxClient*

Onnx4acumos client / Test ONNX model

- **Acumos_model_runner**



```
09:34:24 philippe@WX-OR6199695:~/MODELS/ONNX/onnx4acumos/GoogLeNet$ acumos_model_runner dumpedModel/GoogLeNet/
[2021-03-09 09:34:40 +0100] [205] [INFO] Starting gunicorn 20.0.4
[2021-03-09 09:34:40 +0100] [205] [INFO] Listening at: http://0.0.0.0:3330 (205)
[2021-03-09 09:34:40 +0100] [205] [INFO] Using worker: sync
[2021-03-09 09:34:40 +0100] [216] [INFO] Booting worker with pid: 216
```


127.0.0.1:3330/ui/#/artifacts
swagger
http://127.0.0.1:3330/swagger.json
Explore

GoogLeNet

Provides an HTTP API for the Acumos model 'GoogLeNet'

Find out more about Acumos

https://www.acumos.org/
Apache 2.0

methods : Access model methods

Show/Hide | List Operations | Expand Operations

POST /model/methods/run_GoogLeNet_OnnxModel — Invokes the model method 'run_GoogLeNet_OnnxModel'

artifacts : Access model artifacts

Show/Hide | List Operations | Expand Operations

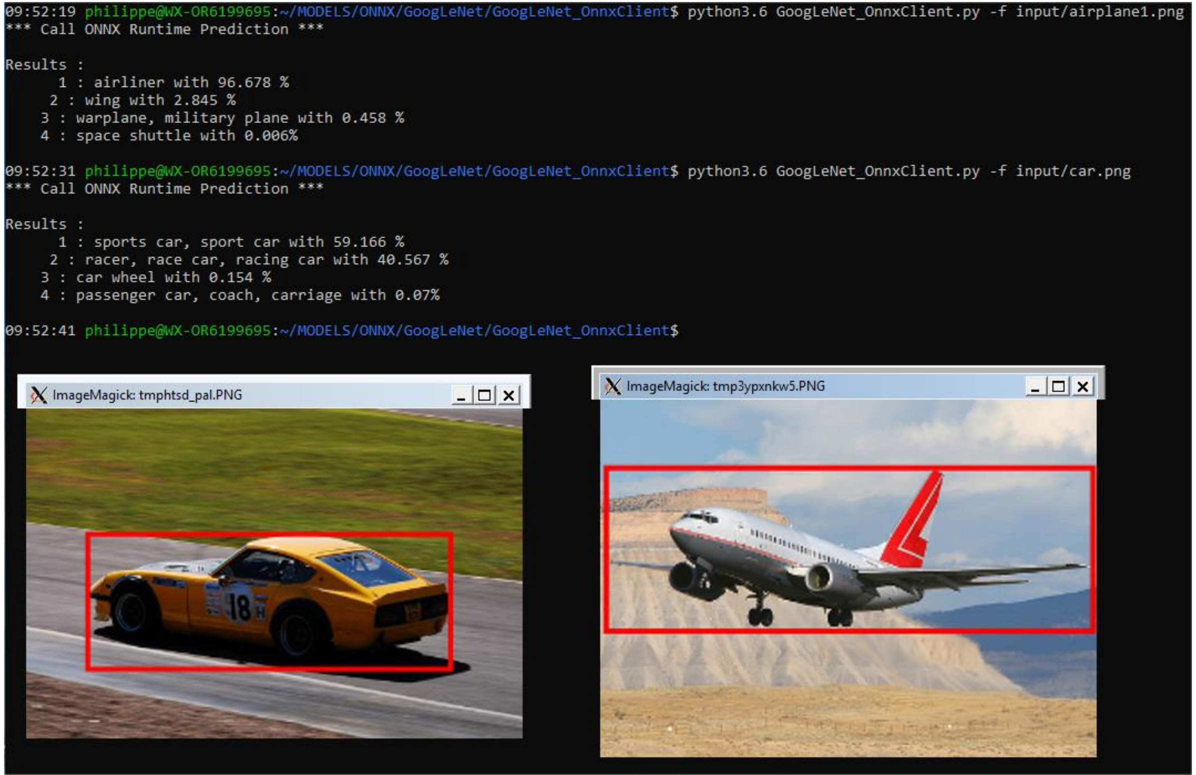
GET /model/artifacts/metadata — Returns the Acumos model metadata JSON

GET /model/artifacts/protobuf — Returns the Acumos model protobuf data format specification

[BASE URL: , API VERSION: 0.1.0]

Onnx4acumos client / Test ONNX model

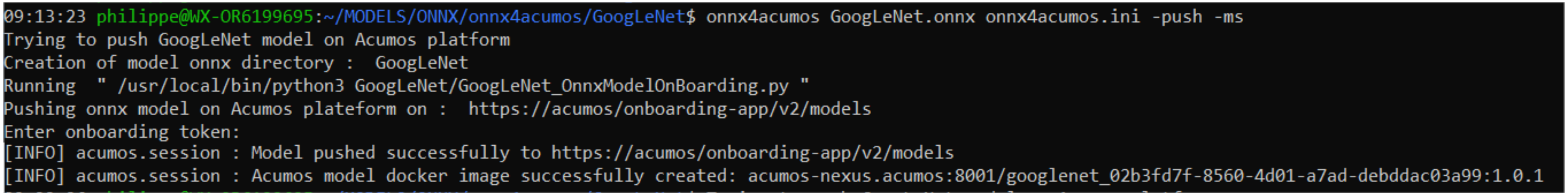
- **Use of Python script**



```
09:52:19 philippe@WX-OR6199695:~/MODELS/ONNX/GoogLeNet/GoogLeNet_OnnxClient$ python3.6 GoogLeNet_OnnxClient.py -f input/airplane1.png
*** Call ONNX Runtime Prediction ***

Results :
      1 : airliner with 96.678 %
     2 : wing with 2.845 %
    3 : warplane, military plane with 0.458 %
    4 : space shuttle with 0.006%

09:52:31 philippe@WX-OR6199695:~/MODELS/ONNX/GoogLeNet/GoogLeNet_OnnxClient$ python3.6 GoogLeNet_OnnxClient.py -f input/car.png
*** Call ONNX Runtime Prediction ***

Results :
      1 : sports car, sport car with 59.166 %
     2 : racer, race car, racing car with 40.567 %
    3 : car wheel with 0.154 %
    4 : passenger car, coach, carriage with 0.07%

09:52:41 philippe@WX-OR6199695:~/MODELS/ONNX/GoogLeNet/GoogLeNet_OnnxClient$
```

Onnx4acumos client / Push mode

- **You can use onnx4acumos in a “Push” mode for instantaneous Acumos On-boarding**
  - Push mode : Create the model bundle and push it in Acumos



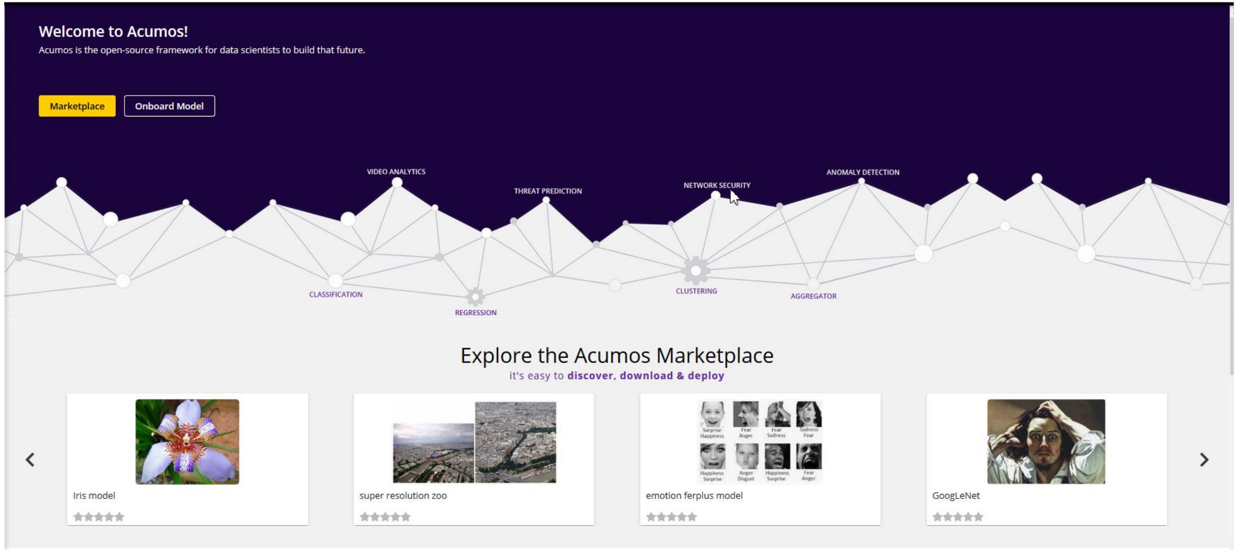
```
09:13:23 philippe@WX-OR6199695:~/MODELS/ONNX/onnx4acumos/GoogLeNet$ onnx4acumos GoogLeNet.onnx onnx4acumos.ini -push -ms
Trying to push GoogLeNet model on Acumos platform
Creation of model onnx directory :  GoogLeNet
Running  " /usr/local/bin/python3 GoogLeNet/GoogLeNet_OnnxModelOnBoarding.py "
Pushing onnx model on Acumos plateform on :  https://acumos/onboarding-app/v2/models
Enter onboarding token:
[INFO] acumos.session : Model pushed successfully to https://acumos/onboarding-app/v2/models
[INFO] acumos.session : Acumos model docker image successfully created: acumos-nexus.acumos:8001/googlenet_02b3fd7f-8560-4d01-a7ad-debddac03a99:1.0.1
```

  - You need to add an .ini file in the command line, that contains the Acumos push url, proxy settings if needed and certificates (optional)
  - Without “–ms” parameter, the model is Onboarded, but the serving model (microservice) is not created.

Onnx4acumos client

Once the model is on-boarded successfully in Acumos you can take benefits of all the Acumos features.

- Market place
- Versioning
- Sharing model
- Serving model
- Design studio
- Licensing



Welcome to Acumos!

Acumos is the open-source framework for data scientists to build that future.

Explore the Acumos Marketplace

it's easy to **discover, download & deploy**

Thank you