

1000 и 1 боль при конвертации модели

Андрей Шадриков

Октябрь 2020

Зачем конвертировать?

- Фреймворки в исследовании:



- Фреймворки в продакшене:



Откуда возникают проблемы?

- Не полная поддержка слоёв.
- Разный формат слоёв (NonMaximumSuppression, Upsample).
- Конвертаторы не поддерживают последние версии фреймворков.
- Не всегда можно оценить скорость работы модели в проде.
- Некорректная сериализация моделей.

Есть же tutorиалы!

- У каждого фреймворка есть свои способы конвертировать модели.
- Часто рассматриваются только базовые случаи.
- Могут требовать старых версий фреймворков для стабильной работы.

Open Neural Network Exchange



- Инициатива от Facebook и Microsoft для конвертации между PyTorch, Caffe2, CNTK.
- Позже поддержали другие команды.
- Статический граф через protobuf.
- Версионность используемых операторов.

Что хорошо?

- Большая поддержка конвертации между фреймворками.
- Большое коммьюнити, поддерживающее редкие конвертаторы.
- Активное развитие.
- Версии opset'ов для хранения старых моделей.

Когда не надо использовать

- Разработка и прод в одной экосистеме (TensorFlow, MxNet).
- Конвертация через ONNX требует больше усилий, чем конвертация напрямую.
- Модель оптимизировалась под конкретный фреймворк (и это не ONNX).

Что хотелось бы

- Отладка моделей (вывод типов, размерностей).
- Мета-поля для слоёв.
- Более удобные инструменты работы с сериализованными моделями.

XKCD #927

КАК МНОЖАТСЯ СТАНДАРТЫ:

(СМ.: ЗАРЯДНЫЕ УСТРОЙСТВА, КОДИРОВКИ, МГНОВЕННЫЕ СООБЩЕНИЯ И Т.Д.)



Альтернативы


TensorFlow

- Очень популярный, большая поддержка.
- Можно не выходить за рамки экосистемы.
- Мало конвертаций.

MMdnn

- Старее ONNX.
- Поддерживается очень малым числом участников.
- Собственный внутренний формат не стандартизован.

PyTorch в ONNX

 sfzhang15 / ATSS

Watch 22

Star 740

Fork 116

[Code](#) [Issues 2](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)


master ▾

Go to file

Add file ▾

Code ▾

About

 **sfzhang15** fix some bugs ... on Apr 22 🕒 15

atss_core	fix some bugs	5 months ago
configs	Initial commit	10 months ago

Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection, CVPR, Oral, 2020

Не всё так просто

```
/home/andrey/tmp/atss/ATSS_tmp/atss_core/modeling/rpn/atss/inference.py:63: UserWarning: This
overload of nonzero is deprecated:
  → nonzero()
Consider using one of the following signatures instead:
  → nonzero(*, bool as_tuple) (Triggered internally at
  /pytorch/torch/csrc/utils/python_arg_parser.cpp:766.)
  per_candidate_nonzeros = per_candidate_inds.nonzero()[top_k_indices, :]
/home/andrey/tmp/atss/ATSS_tmp/atss_core/modeling/backbone/fpn.py:62: TracerWarning: Converting a
tensor to a Python integer might cause the trace to be incorrect. We can't record the data flow of
Python values, so this value will be treated as a constant in the future. This means that the trace
might not generalize to other inputs!
  last_inner, size=(int(inner_lateral.shape[-2]), int(inner_lateral.shape[-1])),
/home/andrey/tmp/atss/ATSS_tmp/atss_core/structures/bounding_box.py:21: TracerWarning:
torch.as_tensor results are registered as constants in the trace. You can safely ignore this
warning if you use this function to create tensors out of constant variables that would be the same
every time you call this function. In any other case, this might cause the trace to be incorrect.
  bbox = torch.as_tensor(bbox, dtype=torch.float32, device=device)
/home/andrey/tmp/atss/ATSS_tmp/atss_core/structures/bounding_box.py:26: TracerWarning: Converting a
tensor to a Python boolean might cause the trace to be incorrect. We can't record the data flow of
Python values, so this value will be treated as a constant in the future. This means that the trace
might not generalize to other inputs!
  if bbox.size(-1) != 4:
/home/andrey/.local/lib/python3.8/site-packages/torch/tensor.py:452: RuntimeWarning: Iterating over
a tensor might cause the trace to be incorrect. Passing a tensor of different shape won't change
the number of iterations executed (and might lead to errors or silently give incorrect results).
  warnings.warn('Iterating over a tensor might cause the trace to be incorrect. ')
/home/andrey/tmp/atss/ATSS_tmp/atss_core/modeling/rpn/atss/inference.py:61: TracerWarning:
Converting a tensor to a Python index might cause the trace to be incorrect. We can't record the
data flow of Python values, so this value will be treated as a constant in the future. This means
that the trace might not generalize to other inputs!
  per_box_cls, top_k_indices = per_box_cls.topk(per_pre_nms_top_n, sorted=False)
/home/andrey/tmp/atss/ATSS_tmp/atss_core/modeling/rpn/atss/inference.py:107: TracerWarning:
Converting a tensor to a Python index might cause the trace to be incorrect. We can't record the
data flow of Python values, so this value will be treated as a constant in the future. This means
that the trace might not generalize to other inputs!
  number_of_detections = len(result)
/home/andrey/tmp/atss/ATSS_tmp/atss_core/modeling/rpn/atss/inference.py:116: TracerWarning:
```

Не всё так просто


```
Traceback (most recent call last):
  File "convert.py", line 87, in <module>
    ...convert()
  File "convert.py", line 67, in convert
    ...sess = rt.InferenceSession(output_onnx, sess_options)
  File "/home/andrey/.local/lib/python3.8/site-packages/onnxruntime/capi/session.py", line 158, in
    __init__
    ...self._load_model(providers or [])
  File "/home/andrey/.local/lib/python3.8/site-packages/onnxruntime/capi/session.py", line 177, in
    _load_model
    ...self._sess.load_model(providers)
onnxruntime.capi.onnxruntime_pybind11_state.InvalidGraph: [ONNXRuntimeError] : 10 : INVALID_GRAPH :
This is an invalid model. Type Error: Type 'tensor(int64)' of input parameter (3925) of operator
(Clip) in node (Clip_2738) is invalid.
```

Особенности конвертации из PyTorch

PyTorch при конвертации использует собственное внутреннее представление через `trace`:

- Операции должны выполняться над объектами `torch.Tensor`.
- In-place операции скорее всего создадут неверный граф (о чём предупреждает конвертатор).
- Control flow посчитается один раз.
- Проверять корректность модели лучше на других данных, что использовались при конвертации.

MxNet в TensorFlow

 **zhreshold / mxnet-ssd**

Watch ▾ 50

Star 764

Fork 351

<> Code

Issues 94

Pull requests


Actions

Projects

Wiki


Security

Insights


 master ▾



Go to file

Add file ▾

 Code ▾

About

 **zhreshold** Fix docker hub name ... on Jun 25, 2018 ⌚ 155

 config	Fix more Python 3 compatibilities	3 years ago
 data/demo	finalize structure	4 years ago

MXNet port of SSD: Single Shot MultiBox Object Detector. Reimplementation of <https://github.com/weiliu89/caffe/tree/ssd>

Конвертатор из MxNet поддерживает мало

Официальная документация

Prerequisites

To run the tutorial you will need to have installed the following python modules: - **MXNet** **>= 1.3.0** - **onnx** v1.2.1 (follow the install guide)

Note: MXNet-ONNX importer and exporter follows version 7 of ONNX operator set which comes with ONNX v1.2.1.

В последней стабильной версии ONNX v1.7.0 версия операторов 12-я, и не все имеют forward-совместимость.

Можно обойтись тем, что есть

- Нам нужна на самом деле 10-я версия операторов.
- Почти все операторы имеют forward-совместимость.
- Слой для якорей и декодинга нужно переписать.
- Остаётся дело за малым — нужно выкрутиться с NonMaximumSuppression.

Можно обойтись тем, что есть

- Нам нужна на самом деле 10-я версия операторов.
- Почти все операторы имеют forward-совместимость.
- Слой для якорей и декодинга нужно переписать.
- Остаётся дело за малым — нужно выкрутиться с NonMaximumSuppression.
- Сконвертируем модель без него, а затем добавим в граф TensorFlow.

Что хотелось показать?

- Задача конвертации возникает достаточно часто.
- Неожиданные моменты могут потребовать много работы.
- Обсуждайте итоговые особенности продакшена заранее.
- ONNX хорошо использовать для сериализации моделей.
- Иногда проще менять граф в нужном фреймворке.

Что за рамками

- Редкие специфичные модели (Grid Sample).
- Рекуррентные модели.
- Не нейронки.
- Отладка и проверка корректности моделей.
- Мобильные и другие платформо-специфичные фреймворки.

Спасибо за внимание!



@vuvko



@vuvko



vuvko



vuvko@fmap.me



vuvko@fmap.me