
ONNX Simplifier

license Apache License v2.0

license Apache License v2.0

license Apache License v2.0

PRs welcome

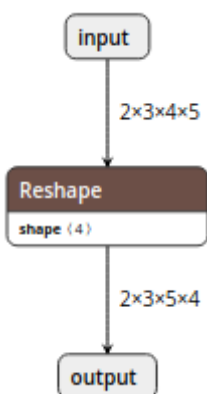
ONNX is great, but sometimes too complicated.

Background

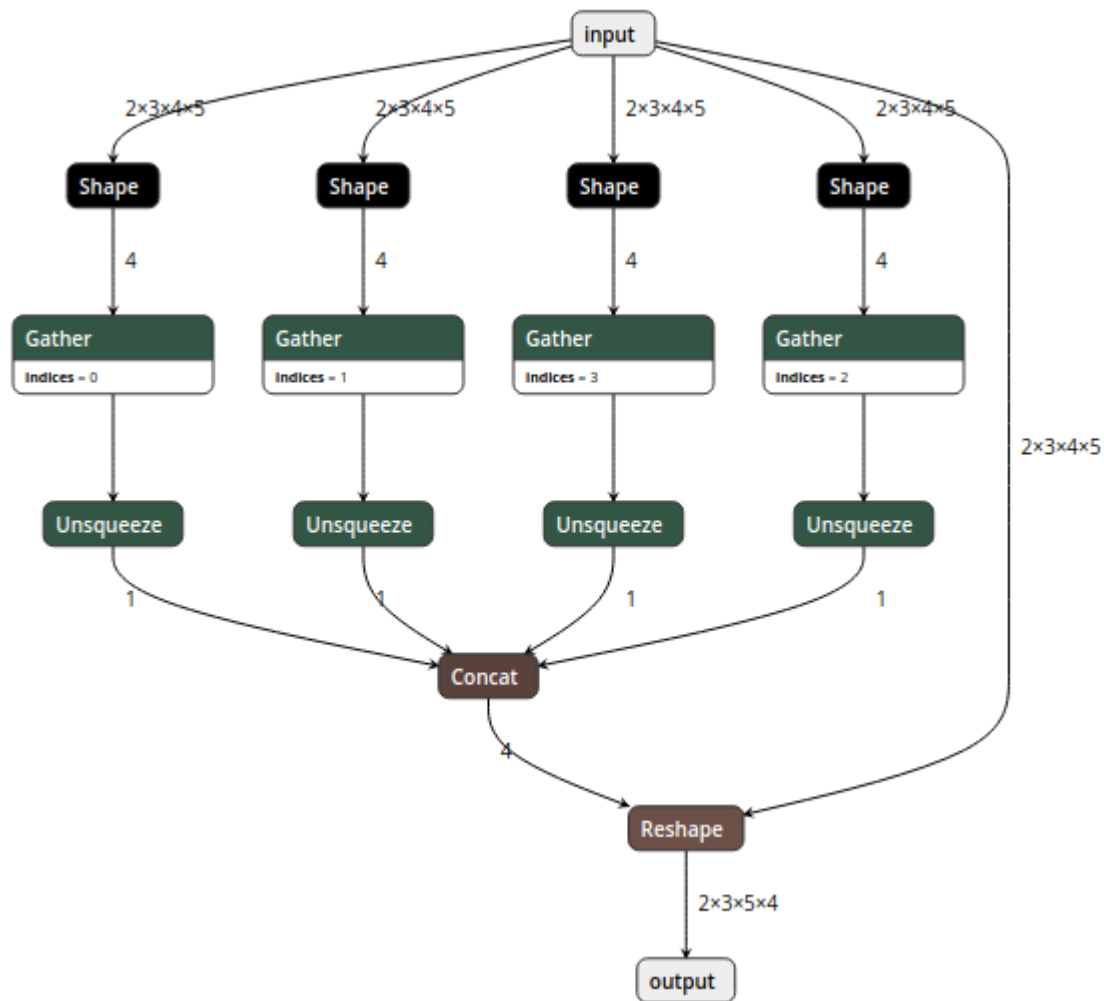
One day I wanted to export the following simple reshape operation to ONNX:

```
1 import torch
2
3
4 class JustReshape(torch.nn.Module):
5     def __init__(self):
6         super(JustReshape, self).__init__()
7
8     def forward(self, x):
9         return x.view((x.shape[0], x.shape[1], x.shape[3], x.shape[2]))
10
11
12 net = JustReshape()
13 model_name = 'just_reshape.onnx'
14 dummy_input = torch.randn(2, 3, 4, 5)
15 torch.onnx.export(net, dummy_input, model_name, input_names=['input'],
16                   output_names=['output'])
```

The input shape in this model is static, so what I expected is



However, I got the following complicated model instead:



Our solution

ONNX Simplifier is presented to simplify the ONNX model. It infers the whole computation graph and then replaces the redundant operators with their constant outputs (a.k.a. constant folding).

Web version

We have published ONNX Simplifier on convertmodel.com. It works out of the box and **doesn't need any installation**. Note that it runs in the browser locally and your model is completely safe.

Python version

```
1 pip3 install -U pip && pip3 install onnxsim
```

Then

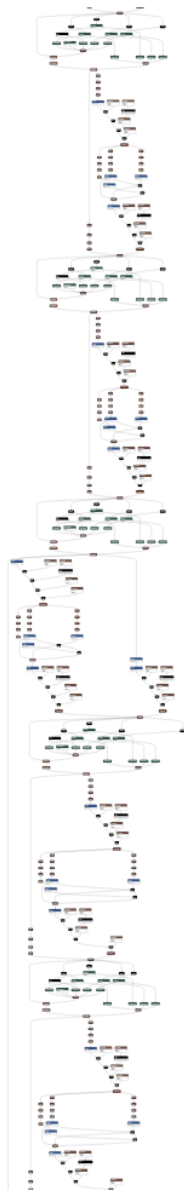
```
1 onnxsim input_onnx_model output_onnx_model
```

For more advanced features, try the following command for help message

```
1 onnxsim -h
```

Demonstration

An overall comparison between a complicated model and its simplified version:



Old model
(3.4M)



New model
(1.9M)

In-script workflow

If you would like to embed ONNX simplifier python package in another script, it is just that simple.

```
1 import onnx
```

```
2 from onnxsim import simplify
3
4 # load your predefined ONNX model
5 model = onnx.load(filename)
6
7 # convert model
8 model_simp, check = simplify(model)
9
10 assert check, "Simplified ONNX model could not be validated"
11
12 # use model_simp as a standard ONNX model object
```

You can see more details of the API in `onnxsim/onnx_simplifier.py`

Projects Using ONNX Simplifier

- MXNet
- MMDetection
- YOLOv5
- ncnn
- ...

Chat

We created a Chinese QQ group for ONNX!

ONNX QQ Group (Chinese): 1021964010, verification code: nndab. Welcome to join!

For English users, I'm active on the ONNX Slack. You can find and chat with me (daquexian) there.