Propainter Development

Yanglin ZHANG

2024-10-08

Task 1: Deploy gradio

Development Environment

Clone the repository by running the following command:

```
git clone git@github.com:lucky9-cyou/ProPainter.git
```

Download the propainter checkpoints and SAM checkpoints. For SAM, we use the `sam_vit_h_4b8939.pth` checkpoint.

Install the development environment by running the following commands:

```
# create new anaconda env
conda create -n propainter python=3.8 -y
conda activate propainter

# install pytorch
conda install pytorch torchvision torchaudio pytorch-cuda=11.8 -c pytorch -c
nvidia

# intall tensortrt for cuda 11.8
wget https://developer.nvidia.com/downloads/compute/machine-learning/tensorrt/
10.5.0/local_repo/nv-tensorrt-local-repo-ubuntu2204-10.5.0-cuda-11.8_1.0-1_amd
64.deb
dpkg -i nv-tensorrt-local-repo-ubuntu2204-10.5.0-cuda-11.8_1.0-1_amd64.deb
sudo cp /var/nv-tensorrt-local-repo-ubuntu2204-10.5.0-cuda-11.8/nv-tensorrt-
local-EE22FB8A-keyring.gpg /usr/share/keyrings/
sudo apt update
sudo apt install tensorrt
python3 -m pip install --upgrade tensorrt-cu11 --extra-index-url https://pypi.
nvidia.com

# install python dependencies
pip3 install -r requirements.txt

# install tensorrt model optimizer and some cuda dependencies
pip install cupy-cuda11x
pip install cuda-python
pip install "nvidia-modelopt[all]~=0.17.0" --extra-index-url https://pypi.
nvidia.com


# install web dependences
pip install -r web-demos/hugging_face/requirements.txt
```

Run the Gradio Application

Run the following command to start the Gradio application:

```
cd web-demos/hugging_face/
python3 app.py
```

The Gradio application will be available at 'http://127.0.0.1:6006/' by VSCode port forwarding or 'http://101.126.90.71:50183'.

Task 2: Invoke the Gradio Application

You can use `client.py` to invoke the Gradio application. The following is an example of how to use the client to invoke the Gradio application:

```
python client.py --video inputs/sample/sample.mp4 --pose weights/vitpose.pt
```

The inpainted video will be saved to `outputs/sample.mp4`. If you want to change the output path, you can use the `--output` option.

Task 3: Optimization inference speed

Time Analysis

Current command:

```
/usr/src/tensorrt/bin/trtexec --onnx=raft.onnx --saveEngine=raft-fp8.engine --
fp8 --verbose  --minShapes='gtlf_1:1x3x640x360','gtlf_2:1x3x640x360' --
optShapes='gtlf_1:12x3x640x360','gtlf_2:12x3x640x360' --
maxShapes='gtlf_1:12x3x640x360','gtlf_2:12x3x640x360'  --
dumpOptimizationProfile --builderOptimizationLevel=5 --useSpinWait --
sparsity=enable > raft-fp8.log
```

All the time is based on the `sample.mp4` video. The video resolution is 640x360 (360p), and the video length is 1032 frames.

	VOS tracking	Raft time	Complete flow time fp16	Image propagation fp16	Feature Propagation fp16
Time	24090.20447 ms	58275.726223 ms	6067.899583 ms	1963.095136 ms	86457.671271 ms

RAFT Optimization

The RAFT model is composed of three parts: `feature block`, `context block` and `update block`. The following is the optimization strategy for each block:

- Use TensorRT Model Optimizer to convert the PyTorch model to ONNX format.
- Using tensorrt `best` mode to optimization.

Some commands:

```
/usr/src/tensorrt/bin/trtexec --onnx=raft_fnet_quan.onnx --
saveEngine=raft_fnet_quan_best.engine --best --verbose  --
```

```
minShapes='x:2x3x640x360' --optShapes='x:24x3x640x360' --
maxShapes='x:24x3x640x360'  --dumpOptimizationProfile --
builderOptimizationLevel=4 --useSpinWait --sparsity=enable >
raft_fnet_quan_best.log

/usr/src/tensorrt/bin/trtexec --onnx=raft_cnet_quan.onnx --
saveEngine=raft_cnet_quan_best.engine --best --verbose  --
minShapes='x:1x3x640x360' --optShapes='x:12x3x640x360' --
maxShapes='x:12x3x640x360'  --dumpOptimizationProfile --
builderOptimizationLevel=4 --useSpinWait --sparsity=enable >
raft_cnet_quan_best.log

/usr/src/tensorrt/bin/trtexec --onnx=raft_update_block_quan.onnx --
saveEngine=raft_update_block_quan_best.engine --best --verbose  --
minShapes='net_in:1x128x80x45','inp:1x128x80x45','corr:1x324x80x45','flow:1x2x80x45'
--
optShapes='net_in:12x128x80x45','inp:12x128x80x45','corr:12x324x80x45','flow:12x2x80x45'
--
maxShapes='net_in:12x128x80x45','inp:12x128x80x45','corr:12x324x80x45','flow:12x2x80x45'
--dumpOptimizationProfile --builderOptimizationLevel=4 --useSpinWait --
sparsity=enable > raft_update_block_quan_best.log
```

Some commands for multi resolution:

```
/usr/src/tensorrt/bin/trtexec --onnx=raft_fnet_quan_res.onnx --
saveEngine=raft_fnet_quan_res_best.engine --best --verbose  --
minShapes='x:2x3x180x180' --optShapes='x:24x3x640x360' --
maxShapes='x:24x3x1280x1280'  --dumpOptimizationProfile --
builderOptimizationLevel=4 --useSpinWait --sparsity=enable >
raft_fnet_quan_res_best.log

/usr/src/tensorrt/bin/trtexec --onnx=raft_cnet_quan_res.onnx --
saveEngine=raft_cnet_quan_res_best.engine --best --verbose  --
minShapes='x:1x3x180x180' --optShapes='x:12x3x640x360' --
maxShapes='x:12x3x1280x1280'  --dumpOptimizationProfile --
builderOptimizationLevel=4 --useSpinWait --sparsity=enable >
raft_cnet_quan_res_best.log

/usr/src/tensorrt/bin/trtexec --onnx=raft_update_block_quan_res.onnx --
saveEngine=raft_update_block_quan_res_best.engine --best --verbose  --
minShapes='net_in:1x128x22x22','inp:1x128x22x22','corr:1x324x22x22','flow:1x2x22x22'
--
optShapes='net_in:12x128x80x45','inp:12x128x80x45','corr:12x324x80x45','flow:12x2x80x45'
--
maxShapes='net_in:12x128x160x160','inp:12x128x160x160','corr:12x324x160x160','flow:12x2x160
--dumpOptimizationProfile --builderOptimizationLevel=4 --useSpinWait --
sparsity=enable > raft_update_block_quan_res_best.log
```

Optimization results:

	Torch fp32	TensorRT best	Speedup
Time	58275.726223 ms	25342.446789 ms	2.2

Feature Propagation and Transformer Optimization

Model Inference Optimization

The feature propagation and transformer are the most time-consuming parts of the model. It is composed of `encoder`, `decoder`, `softsplit`, `softcomp`, `feat_prop` and `transformer`. The following is the optimization strategy for each part:

- Use TensorRT Model Optimizer to convert the PyTorch model to ONNX format.
- Using tensorrt `best` mode to optimization.
- Not consider `transformer` optimization.

Some commands:

```
/usr/src/tensorrt/bin/trtexec --onnx=inpainter_encoder.onnx --
saveEngine=inpainter_encoder_best.engine --best --verbose  --
minShapes='input:9x5x640x360' --optShapes='input:18x5x640x360' --
maxShapes='input:18x5x640x360'  --dumpOptimizationProfile --
builderOptimizationLevel=4 --useSpinWait --sparsity=enable >
inpainter_encoder.log

/usr/src/tensorrt/bin/trtexec --onnx=inpainter_decoder.onnx --
saveEngine=inpainter_decoder_best.engine --best --verbose  --
minShapes='input:6x128x160x90' --optShapes='input:11x128x160x90' --
maxShapes='input:11x128x160x90'  --dumpOptimizationProfile --
builderOptimizationLevel=4 --useSpinWait --sparsity=enable >
inpainter_decoder.log

/usr/src/tensorrt/bin/trtexec --onnx=inpainter_feat_back_deform_align.onnx --
saveEngine=inpainter_feat_back_deform_align_best.engine --best --verbose  --
dumpOptimizationProfile --builderOptimizationLevel=4 --useSpinWait --
sparsity=enable > inpainter_feat_back_deform_align.log

/usr/src/tensorrt/bin/trtexec --onnx=inpainter_feat_forw_deform_align.onnx --
saveEngine=inpainter_feat_forw_deform_align_best.engine --best --verbose  --
dumpOptimizationProfile --builderOptimizationLevel=4 --useSpinWait --
sparsity=enable > inpainter_feat_forw_deform_align.log

/usr/src/tensorrt/bin/trtexec --onnx=inpainter_feat_back_backbone.onnx --
saveEngine=inpainter_feat_back_backbone_best.engine --best --verbose  --
dumpOptimizationProfile --builderOptimizationLevel=4 --useSpinWait --
sparsity=enable > inpainter_feat_back_backbone.log

/usr/src/tensorrt/bin/trtexec --onnx=inpainter_feat_forw_backbone.onnx --
saveEngine=inpainter_feat_forw_backbone_best.engine --best --verbose  --
dumpOptimizationProfile --builderOptimizationLevel=4 --useSpinWait --
sparsity=enable > inpainter_feat_forw_backbone.log
```

```
/usr/src/tensorrt/bin/trtexec --onnx=inpainter_feat_fuse.onnx --
saveEngine=inpainter_feat_fuse_best.engine --best --verbose  --
minShapes='feat:6x258x160x90' --optShapes='feat:11x258x160x90' --
maxShapes='feat:11x258x160x90' --dumpOptimizationProfile --
builderOptimizationLevel=4 --useSpinWait --sparsity=enable >
inpainter_feat_fuse.log
```

Optimization results:

	Torch fp32 + fp16	TensorRT Encoder best	TensorRT Feature best	Speedup
Time	86457.671271 ms	79078.691251 ms	78972.896806	1.09

NOTE: TensorRT Model Optimizer will cause loss of accuracy for encoder and decoder. Most computation is in the `transformer` part, but the `transformer` part very complex and hard to optimize. It need more time to optimize.

Multi-thread Optimization

We can use multi-thread to optimize the inference feat propagation and transformer.

Optimization results:

	Torch fp32 + fp16	Multi Thread	Speedup
Time	86457.671271 ms	68060.301863 ms	1.27

Inpainting Optimization

Optimization results:

	Torch fp32	Torch fp32 + fp16	Final	Speedup
Time	227701.289064 ms	185057.978153 ms	122179.337429 ms	1.86

How to running

Normal branch is `main` branch, you can run the following command to start the Gradio application:

```
conda activate propainter
cd /root/ProPainter/web-demos/hugging_face/
python3 app.py
```

Optimization branch is `feat/tensorrt-model-opt` branch, you can run the following command to start the Gradio application:

```
conda activate propainter
cd /root/ProPainter/web-demos/hugging_face/
python3 app.py
```

Run the following command to invoke the Gradio application:

```
conda activate propainter
cd /root/ProPainter/
python client.py --video inputs/sample/sample.mp4 --pose weights/vitpose.pt
```