This repository Search Pull requests Issues Gist

songhan / **Deep-Compression-AlexNet**

Watch 24 Star 135 Fork 69

Code Pull requests 0 Projects 0 Wiki Pulse Graphs

Deep Compression on AlexNet

10 commits **2** branches **0** releases **1** contributor

Branch: master New pull request Create new file Upload files Find file Clone or download

songhan Update README.md 3 Latest commit `990c9c2` on 17 May 2016

AlexNet_compressed.net	init	a year ago
README.md	Update README.md	11 months ago
bvlc_alexnet_deploy.prototxt	init	a year ago
decode.py	init	a year ago

README.md

Deep Compression on AlexNet

This is a demo of Deep Compression compressing AlexNet from 233MB to 8.9MB without loss of accuracy. It only differs from the paper that Huffman coding is not applied. Deep Compression's video from ICLR'16 best paper award presentation is available.

Related Papers

Learning both Weights and Connections for Efficient Neural Network (NIPS'15)

Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding (ICLR'16, best paper award)

EIE: Efficient Inference Engine on Compressed Deep Neural Network (ISCA'16)

If you find Deep Compression useful in your research, please consider citing the paper:

```
@inproceedings{han2015learning,
  title={Learning both Weights and Connections for Efficient Neural Network},
  author={Han, Song and Pool, Jeff and Tran, John and Dally, William},
  booktitle={Advances in Neural Information Processing Systems (NIPS)},
  pages={1135--1143},
  year={2015}
}


@article{han2015deep_compression,
  title={Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman
  author={Han, Song and Mao, Huizi and Dally, William J},
  journal={International Conference on Learning Representations (ICLR)},
  year={2016}
}
```

A hardware accelerator working directly on the deep compressed model:

```
@article{han2016eie,
  title={EIE: Efficient Inference Engine on Compressed Deep Neural Network},
  author={Han, Song and Liu, Xingyu and Mao, Huizi and Pu, Jing and Pedram, Ardavan and Horowitz, Mark A ar
  journal={International Conference on Computer Architecture (ISCA)},
  year={2016}
}
```

Usage:

```
export CAFFE_ROOT=$your caffe root$

python decode.py bvlc_alexnet_deploy.prototxt AlexNet_compressed.net $CAFFE_ROOT/alexnet.caffemodel

cd $CAFFE_ROOT

./build/tools/caffe test --model=models/bvlc_alexnet/train_val.prototxt --weights=alexnet.caffemodel --iter
```

Test Result:

```
I1022 20:18:58.336736 13182 caffe.cpp:198] accuracy_top1 = 0.57074
I1022 20:18:58.336745 13182 caffe.cpp:198] accuracy_top5 = 0.80254
```