Squeezenet with Deep Compression

a 660KB model, AlexNet accuracy, fully fits in SRAM cache, embedded system friendly

View on GitHub Download .zip Download .tar.gz

Introduction

This is the 660KB compressed SqueezeNet, which is 363x smaller as AlexNet but has the same accuracy as AlexNet.

(There is an even smaller version which is only 470KB. It requires some effort to materialize since each weight is 6-bits.)

Usage

```
export CAFFE_ROOT=$your_caffe_root

python decode.py /ABSOLUTE_PATH_TO/SqueezeNet_deploy.prototxt /ABSOLUTE_PATH_TO/compressed_SqueezeNet.net /ABSOLUTE_PATH_TO/decompressed_SqueezeNet.caffemodel

note: decompressed_SqueezeNet.caffemodel is the output, can be any name.

$CAFFE_ROOT/build/tools/caffe test --model=SqueezeNet_trainval.prototxt --weights=decompressed_SqueezeNet.caffemodel --iterations=1000 --gpu 0
```

Related SqueezeNet repo

SqueezeNet

SqueezeNet-Deep-Compression

SqueezeNet-Generator

SqueezeNet-DSD-Training

SqueezeNet-Residual

Related Papers

SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5MB model size

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 SqueezeNet and Deep Compression useful in your research, please consider citing the paper:

```
@article{SqueezeNet,
  title={SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5MB model size},
  author={Iandola, Forrest N and Han, Song and Moskewicz, Matthew W and Ashraf, Khalid and Dally, William J and Keutzer, Kurt},
  journal={arXiv preprint arXiv:1602.07360},
  year={2016}
}

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

@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{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 and Dally, William J},
  journal={International Conference on Computer Architecture (ISCA)},
  year={2016}
}
```

Squeezenet with Deep Compression is maintained by songhan. This page was generated by GitHub Pages using the Cayman theme by Jason Long.