SAFEAI 2022

BEYOND TEST ACCURACY – THE EFFECTS OF MODEL COMPRESSION ON CNNS

ADRIAN SCHWAIGER, KRISTIAN SCHWIENBACHER, KARSTEN ROSCHER

MOTIVATION

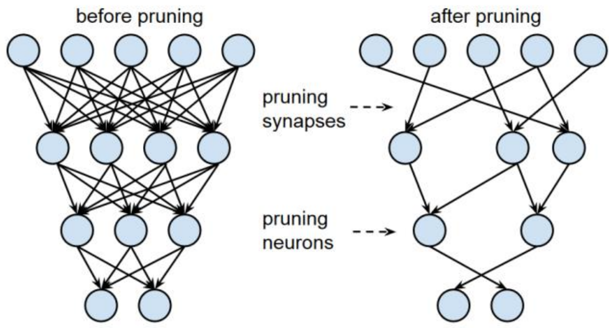
- Model compression
  - Enables deploying CNNs to low-power devices
  - Reduces model size and inference time significantly while maintaining accuracy

MOTIVATION

- Model compression
  - Enables deploying CNNs to low-power devices
  - Reduces model size and inference time significantly while maintaining accuracy
- We analyze **how model compression changes CNNs "under the hood"**
  - How is the predictive quality influenced on a class and sample level?
  - How is the attention of the models affected?

COMMON MODEL COMPRESSION TECHNIQUES

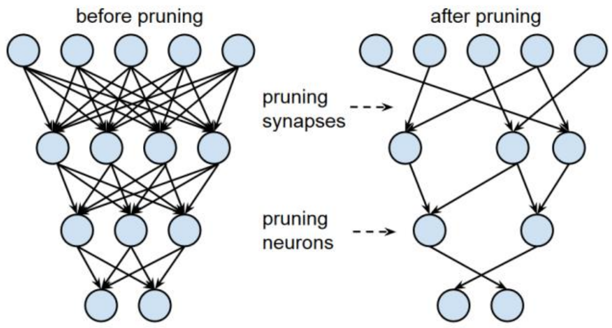
- Pruning
  - **Induce sparsity by removing neurons or connections**
  - Structured vs. unstructured pruning
  - Global vs. local pruning



Source: Han et. al: Learning both Weights and Connections for Efficient Neural Networks

COMMON MODEL COMPRESSION TECHNIQUES

- Pruning
  - **Induce sparsity by removing neurons or connections**
  - Structured vs. unstructured pruning
  - Global vs. local pruning



Source: Han et. al: Learning both Weights and Connections for Efficient Neural Networks

- Quantization
  - **Reduce number of bits required to represent model parameters**
  - Post training quantization vs. quantization aware training

EXPERIMENTAL SETUP

- **Networks**
  - **LeNet-5** (~62k parameters), **SqueezeNet** (~750k), **ResNet-18** (~11m)

EXPERIMENTAL SETUP

- **Networks**
  - **LeNet-5** (~62k parameters), **SqueezeNet** (~750k), **ResNet-18** (~11m)
- **Datasets**
  - **CIFAR-10** and German Traffic Sign Recognition Benchmark (**GTSRB**)

EXPERIMENTAL SETUP

- **Networks**
  - **LeNet-5** (~62k parameters), **SqueezeNet** (~750k), **ResNet-18** (~11m)
- **Datasets**
  - **CIFAR-10** and German Traffic Sign Recognition Benchmark (**GTSRB**)
- **Model compression**
  - **Global unstructured pruning** with L1 as scoring function
  - **Post-training quantization with 8-bit** for weight and activation precision
  - **Post-training quantization with 4-bit** for weight and 8-bit for activation precision
  - **Combination** of global unstructured pruning and 8-bit post-training quantization

GENERAL IMPACT OF MODEL COMPRESSION

- Decrease in accuracy after compression of < 1pp
  - Exception: 4-bit quantization half the time caused severe degradation, we ignored these instances

GENERAL IMPACT OF MODEL COMPRESSION

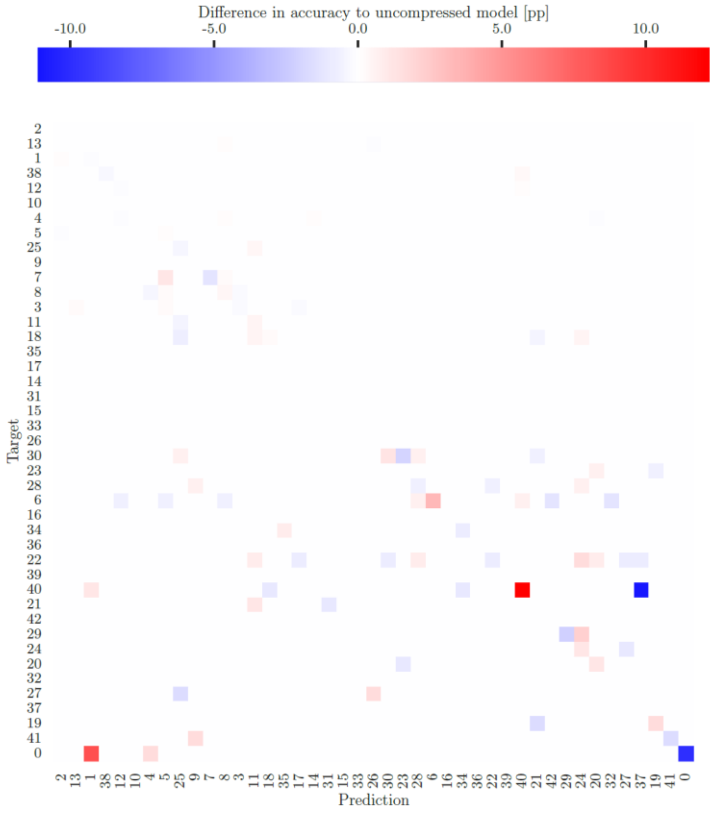
- Decrease in accuracy after compression of < 1pp
  - Exception: 4-bit quantization half the time caused severe degradation, we ignored these instances
- No difference between the two datasets
- ResNet affected the least, SqueezeNet the most
- Amount pruned up to 72.5%

GENERAL IMPACT OF MODEL COMPRESSION

- Decrease in accuracy after compression of < 1pp
  - Exception: 4-bit quantization half the time caused severe degradation, we ignored these instances
- No difference between the two datasets
- ResNet affected the least, SqueezeNet the most
- Amount pruned up to 72.5%
- **Difference in classifications after pruning of up to 7.5%**

CHANGES REGARDING CLASS CONFUSION AND ACCURACY

- Pruning and quantization can introduce **significant changes at the class level**
- **No pattern in the classes affected** by any of the applied compression methods
- Combination of pruning and quantization didn't show any peculiarities



CHANGES REGARDING MODEL CONFIDENCE

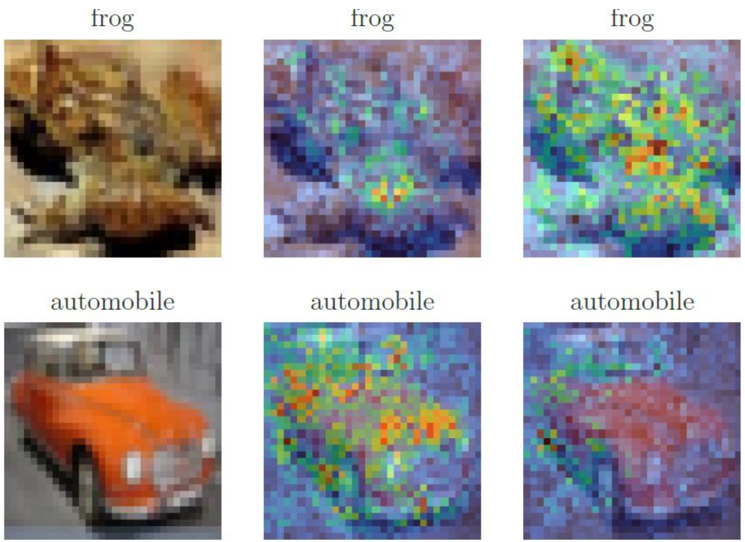
- Compression methods can **significantly change the prediction confidence**, especially for differently classified samples

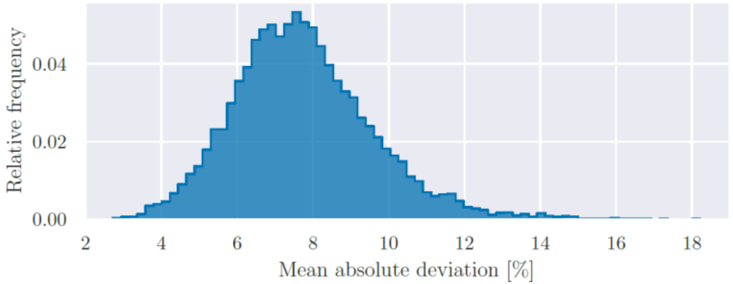
CHANGES REGARDING MODEL CONFIDENCE

- Compression methods can **significantly change the prediction confidence**, especially for differently classified samples
- *However, no explicit uncertainty quantification technique was employed*

CHANGES IN SALIENCY MAPS

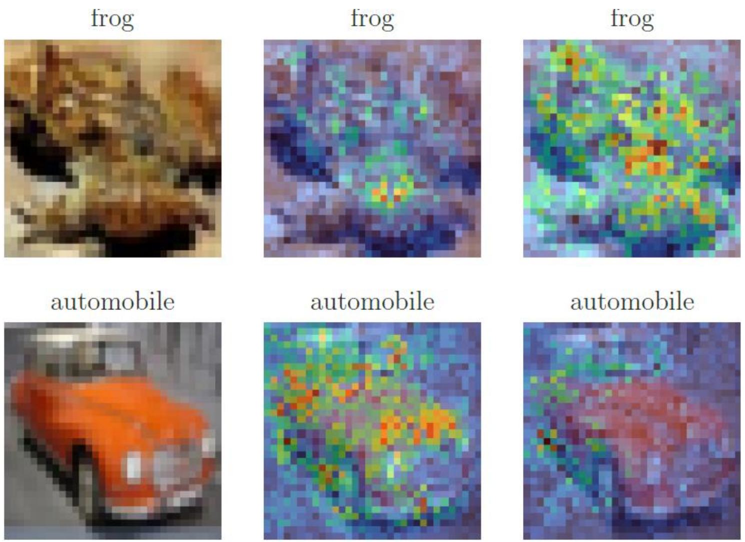
- Compression methods **don't systematically change the attention** of the model
- **Saliency maps change significantly after compression** with an average mean absolute deviation of ~7.5%
- Even for ResNet-18 on CIFAR-10 we saw a significant difference in the saliency maps

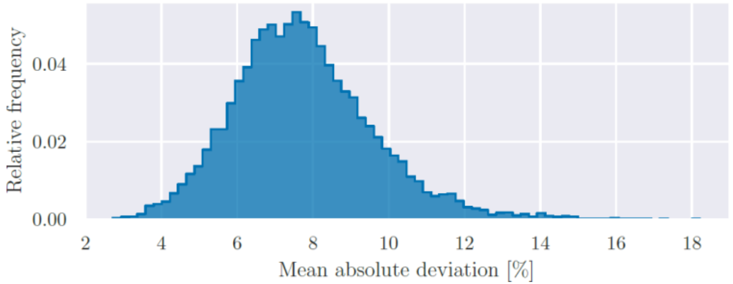




CHANGES IN SALIENCY MAPS

- Compression methods **don't systematically change the attention** of the model
- **Saliency maps change significantly after compression** with an average mean absolute deviation of ~7.5%
- Even for ResNet-18 on CIFAR-10 we saw a significant difference in the saliency maps
- *But also raises questions regarding the expressiveness of saliency maps generated with this method*





CONCLUSIONS

- **Model compression introduces significant changes that are not uncovered by superficial metrics**
  - Changes in predicted class of up to 7.5%
  - Changes in accuracy on the class level of up to 15pp
  - Drastic changes in the confidence scores in some cases
  - Significant differences in the observed input salience

CONCLUSIONS

- **Model compression introduces significant changes that are not uncovered by superficial metrics**
  - Changes in predicted class of up to 7.5%
  - Changes in accuracy on the class level of up to 15pp
  - Drastic changes in the confidence scores in some cases
  - Significant differences in the observed input salience

- **Future directions**
  - Investigation of further model compression techniques
  - Development of further methods to systematically analyze ML systems beyond current metrics
  - Research regarding continuous safety assurance to consider safety as integral part of ML development