Neural Network Compression (NNC, ISO/IEC 15938-17)

Werner Bailer
AI and Multimedia Workshop, 2022-01-18
Outline

- Context & state of the art
- Standard: need and design considerations
- Coding tools
- Performance
- Ongoing work and conclusion
Context

- Artificial neural networks are widely used, e.g. in multimedia
  - visual and audio content recognition and classification
  - speech and natural language processing
- Deep learning makes use of very large networks
  - many layers with nodes and connections
  - parameters/weights attached to each of them (e.g. convolution operations)
**Context**

- **Training**
  - learn parameters from data
  - typically once, on powerful infrastructure
  - updates or adaptations in target environment may be necessary

- **Inference**
  - use the trained network for prediction
  - needs network with all its parameters, i.e., large amount of data to be transmitted and processed
    - *focus: small size to be transmitted*
  - often used on resource constrained devices (mobile phones, smart cameras, edge nodes, …)
    - *focus: low memory and computational complexity during inference*
SotA in NN Compression

- typically three steps
- reduction of parameters, e.g.
  - eliminating neurons (pruning)
  - reducing the entropy of a tensor
  - decomposing/transforming a tensor
- reducing the precision of parameters (i.e. quantization)
- performing entropy coding

Han et al., ICLR 2016
Relation to Network Architecture Search (NAS)

- finding an alternative architecture and train
- architecture search is computationally expensive
- training is then done using e.g. knowledge distillation, teacher-student learning
- Faster NAS methods have been proposed, e.g. Single-Path Mobile AutoML [Stamoulis et al., IEEE JSTSP 2020]
- needs also access to full training data, while fine-tuning could be done on partial data or application specific data
Relation to target hardware

- Optimising for target hardware
  - supported operations (e.g., sparse matrix multiplication, weight precisions)
  - relative costs of memory access and computing operations
- training a network, derive network for particular platform
  - first approaches [Cai, ICLR 2020]
  - no reliable prediction of inference costs on particular target architecture
  - in particular, prediction of speed and energy consumption
  - cf. autotune in inference of DL frameworks
**Need for a standardised interface**

- **Network training** in framework L
  - Trained network $T$, (or exported in exchange format $T'$)
  - Standardised format is crucial, compression useful in some cases

- **Accelerator Libraries**
  - Supporting multiple backends
  - OneAPI

- **Inference engines** $I_1$ and $I_k$
  - Optimised network $O_1$
  - Optimised network $O_k$
  - Compression is crucial
Standardization in MPEG (ISO/IEC JTC1 SC29)

- Develop interoperable compressed representation of neural networks
- Leverage the know-how in the MPEG on compression of various types of (multimedia) data
- Enable multimedia applications to benefit from the progress in machine learning using deep neural networks
- Cover a broad set of relevant use cases
  - selected image classification, visual content matching, content coding and audio classification as applications in which technology is validated
Design Considerations

- Interoperability with exchange formats (NNEF, ONNX) and formats of common DL frameworks
- Reuse existing approaches for representing topology
- Agnostic to inference platform and its specificities
- Different types of networks, applications, … may be best served by different compression tools
Evaluating compression technologies

- Compression ratio
- Reconstruction of original parameters (cf. PSNR for multimedia data)
  - not a useful metric
  - performance in target application (e.g., image classification) needs to be measured (cf. perceptual quality metrics for multimedia)
  - requires models and data sets for each target application
- Runtime/memory consumption
  - of the encoding/decoding process
  - inference using the resulting model
Evaluating compression technologies

- Compression ratio of stored/transmitted model (format must be considered)
- Compression ratio of model in memory (depends on platform)
- Performance for task (with/without finetuning)
Standard as a Toolbox

- Could be represented in any NN format
- Representation for storage/transmission
- Representation for inference
Parameter Reduction

- Sparsification
  - General sparsification
  - Micro-structured sparsification
- Pruning
  - estimate importance of weights to decide about pruning neurons
- Low-rank decomposition
  - approximate tensor as product of decomposition result (limiting number of parameters)
- Unification
  - generalisation of micro-structured sparsification (values other than 0)
- Batchnorm folding, local scaling
  - store batchnorm parameters, and apply to weights (better compressability of weights)
  - scaling factor per row (no additional parameters if used together with BN folding)
Quantisation

- Uniform Nearest Neighbor Quantization
- Codebook Quantization
- Dependent Scalar Quantization
  - Trellis-coded quantisation
  - two scalar quantisers, and procedure for switching between them (state-machine with 8 states)
Entropy Coding

- DeepCABAC
  - Adaptation of Context-adaptive Binary Arithmetic Coding (CABAC)
- Binarization
- Context-modelling
  - separate models for each of the flags
  - select from a fixed set of models
- Arithmetic coding to regular and bypass bins
Decoding

- Output of encoded tensor
  - Integer or floating point
  - Block: set of combined tensors (e.g., components of decomposed tensor)
- Parallel decoding of parts of a tensor is supported
  - Option to specify entry points for the parts during encoding
Interoperability with exchange formats

- Include network topology in encoded bitstream
  - ONNX, NNEF, Tensorflow, PyTorch
  - Supports encoding just some of the tensors
  - Compatibility with quantisation formats supported in those formats

- Include encoded tensors in exchange format
  - Recommended approach for NNEF and ONNX
## Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>$c_r$ in %</th>
<th>top-1 / top-5 acc. reconstr.</th>
<th>top-1 / top-5 acc. original</th>
<th>Orig. size (bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG16</td>
<td>2.98</td>
<td>70.51 / 89.54</td>
<td>70.93 / 89.85</td>
<td>553.43 M</td>
</tr>
<tr>
<td>ResNet50</td>
<td>6.54</td>
<td>74.42 / 91.80</td>
<td>74.98 / 92.15</td>
<td>102.55 M</td>
</tr>
<tr>
<td>MobileNetV2</td>
<td>12.18</td>
<td>71.13 / 90.06</td>
<td>71.47 / 90.27</td>
<td>14.16 M</td>
</tr>
<tr>
<td>DCase</td>
<td>4.12</td>
<td>58.15 / 92.35</td>
<td>58.27 / 91.85</td>
<td>467.26 k</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>$c_r$ in %</th>
<th>PSNR / SSIM reconstructed</th>
<th>PSNR / SSIM original</th>
<th>Orig. size (bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC12B</td>
<td>17.34</td>
<td>29.98 / 0.954</td>
<td>30.13 / 0.956</td>
<td>304.72 k</td>
</tr>
</tbody>
</table>
Performance

- VGG16 Pretrained: 3.38% (2.96% LS / 5.5% BL)
- VGG16 Low Rank: 1.82% (LS / BL)
- MobileNetV2 Pretrained: 9.34% (8.44% LS / 9.68% BL / 14.67% BL)
- MobileNetV2 Low Rank: 9.53% (8.38% LS / 9.92% BL / 15.94% BL)

Colours:
- Blue: Dependent Quantization + DeepCABAC + Local Scaling Adaptation (LS) + Batch-norm Folding + QP-Optimization
- Orange: Dependent Quantization + DeepCABAC + Local Scaling Adaptation (LS) + Batch-norm Folding
- Green: Dependent Quantization + DeepCABAC + Local Scaling Adaptation
- Red: BL: Dependent Quantization + DeepCABAC
- Purple: Low Rank Decomposition
Performance

Top-1 Accuracy in %

Pretrained

PSNR in dB

image classification
audio classification
image encoding
Ongoing work: incremental compression

- Use cases that need to send updated models
  - e.g. deploy to mobile devices, federated learning
- Encode model w.r.t. base model
  - support tensor updates and structural changes, e.g. transfer learning with different number of output classes
- Initial results
  - updates in distributed training can be represented at <1% of the base model size
Conclusion

- standard for compressing NN parameters
- compresses to less than 10% without performance loss
- interoperability with exchange formats

Status
- compression standard (ISO/IEC 15938-17) going to FDIS ballot
- reference software (ISO/IEC 15938-18) under CD ballot
- work on incremental compression ongoing (to become 2nd ed. of pt. 17)
This work has received funding from the European Union's Horizon 2020 research and innovation programme, under grant agreement no. 951911 AI4Media (https://ai4media.eu).

Werner Bailer
Co-chair of MPEG AhG on Neural Network Compression
werner.bailer@joanneum.at