**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2018/N17757**

**July 2018, Ljubljana, SI**

**Title Call for Evidence on Neural Network Compression**

**Source MPEG Requirements**

**Status: Approved**

# Introduction

Artificial neural networks have been adopted for a broad range of tasks in multimedia analysis and processing, media coding, data analytics and many other fields. Their recent success is based on the feasibility of processing much larger and complex neural networks (deep neural networks, DNNs) than in the past, and the availability of large-scale training data sets. As a consequence, trained neural networks contain a large number of parameters (weights), resulting in a quite large size (e.g., several hundred MBs). Many applications require the deployment of a particular trained network instance, potentially to a larger number of devices, which may have limitations in terms of processing power and memory (e.g., mobile devices or smart cameras). Any use case, in which a trained neural network (and its updates) needs to deployed to a number of devices could thus benefit from a standard for the compressed representation of neural networks.

The MPEG activity on Compressed Representation of Neural Networks (NNR) aims to define a compressed, interpretable and interoperable representation for trained neural networks. NNR shall be able to

* represent different artificial neural network types (e.g., feedforward networks such as CNN and autoencoder, recurrent networks such as LSTM, etc.)
* enable efficient incremental updates of compressed representations of NNs
* enable scalability, i.e. NNs of different performance can be obtained from subsets of the complete compressed representation
* inference with compressed network
* enable use under resource limitations (computation, memory, power, bandwidth)

The scope of existing exchange formats (e.g., NNEF, ONNX) is the interface between the framework used for training and the acceleration library/optimisation engine for a specific platform. However, these exchange formats do not yet take features such as scalable and incremental updates and compression into account.

MPEG NNR has identified a set of relevant use cases and related requirements [1], including applications of neural networks in multimedia and beyond. A call for test data (including neural networks, evaluation frameworks and data) has been issued [2], and contributions covering some of the use cases have been received. The contributions include also instances of compressed neural networks, showing that there is significant potential for using compressed networks in these use cases, which still perform comparably to the original uncompressed networks.

MPEG NNR is thus calling for evidence on compression technology for neural networks, which is applicable to the networks used in the different use cases.

# Scope

The scope of this CfE is technology to reduce the size of trained neural networks, i.e. the representation of the network topology and all its weights/parameters. The starting point is a trained neural network for the use cases for which data is available.

These use cases are:

* Visual object classification
  + UC2 Camera app with object recognition
  + UC4 Large-scale public surveillance
* UC11 Compact Descriptors for Video Analysis (CDVA)
* UC12A Image/Video Compression – Tool-by-tool use case

The size reduction will impact the size of the serialized/stored network and/or the memory footprint of the reconstructed network used for inference. The complexity of compression and particularly of decompression needed for inference shall be taken into account, as well as the impact of the applied compression technology on the complexity of inference.

The performance assessment in the use cases is limited to testing the compressed representation for inference in these use cases (e.g., not taking incremental training into account).

Proponents are required to submit complete results for at least one network for each of at least two of the use cases, but preferably results should be provided for all use cases.

# Timeline

|  |  |
| --- | --- |
| 2018/07/27 | Availability of neural networks, test data and desription for the respective use cases |
| 2018/09/01 | Registration deadline |
| 2018/09/14 | If evaluation process option B (see below) is used: Deadline for electronic submission of compressed (and reconstructed) neural network models |
| 2018/10/01 | Deadline for submission of descriptions (MPEG input contribution) of approaches and evaluation results (for both evaluation process options A and B) |
| 2018/10/07-12 | Evaluation of responses (the October MPEG meeting) |

MPEG may or may not initiate a standardization activity after evaluating the responses.

# Test Conditions

Given the provided trained neural networks for the different use cases (see [3] for details about the data), proponents are asked to test one or more approaches for neural network compression on these trained networks.

Retraining the network after compression is permitted. In any case the results for the compressed network must be reported. Results for the compressed and additionally retrained network may be reported in addition.

# Evaluation Methods and Procedures

The evaluation procedure and metrics are described in [3]. The metrics consist of two parts:

* Compression efficiency, runtime complexity and memory consumption of compression/decompression (measurement is independent of the use case)
* Use case specific performance metrics, comparing the performance of inference using the reconstructed network after compression[[1]](#footnote-1) with using the original network. For the use case specific performance metrics, specific frameworks/tools may need to be used (in particular, for the CDVA and video compression use cases), which require building and configuration in order to be used. We thus offer two options for evaluating the use case specific metrics:
  + Option A – Proponents perform the entire evaluation themselves. They obtain the frameworks/tools as described in [3], build them themselves, and run them both with the original and the compressed (and reconstructed) neural network. The results must be reported in an input document to MPEG 124, and preferably the compressed (and reconstructed) neural network should be provided.
  + Option B – Proponents provide a compressed and reconstructed neural network, which is evaluated by the contributor of the respective use case / test data. The reconstructed NN model must be provided in the same format, in which the uncompressed NN model for this use case has been provided. The responsible contacts for evaluation in the use case frameworks are:
    - Visual object classification: haoji\_hu@zju.edu.cn
    - CDVA: wzziqian@pku.edu.cn
    - Video coding: [hcmoon@kau.kr](mailto:hcmoon@kau.kr), smchun@insignal.co.kr

# Submission Requirements

The following material is to be submitted electronically. The material shall also be brought to the 124th MPEG meeting.

The submission must contain:

* at least one of
  + compressed network
  + evaluation results with the specific metrics for the use cases (required for evaluation option A)
* compression efficiency, runtime complexity and memory consumption measurements
* preferably the reconstructed network used for inference (in the same model format as the uncompressed input network, ONNX or NNEF)
* preferably a description of the compression approach, including the parameterization used
* indication whether retraining has been performed, and a reference to the data set used for retraining

Proponents are required to submit complete results for at least one network for each of at least two of the use cases, but preferably results should be provided for all use cases.

# Participation fee

Participation in the call will not be associated with any fee.

# Logistics

Prospective contributors of responses to the Call for Evidence should contact the following people:

Jörn Ostermann (MPEG requirements chair)

Leibniz Universität Hannover.

Institut für Informationsverarbeitung

Tel. +49-5117625316, email ostermann@tnt.uni-hannover.de

Werner Bailer

JOANNEUM RESEARCH

Tel. +43 316 876 1218, email werner.bailer@joanneum.at

Expressions of interest to submit a response shall be made by contacting the people above on or before 2018/09/01. Interested parties are kindly invited to express their intent as early as possible.

Details on how to format and submit documents, bitstreams, and other required data will be communicated directly to those who express an interest of participation.

Details for access to the test data and tools for evaluation can be found in [3], for futher questions contact one of the above individuals.

# References

1. N17740, Use cases and requirements for compressed representation of neural networks, Ljubljana, SI, July 2018.
2. N17752, Call for Test Data for Compressed Representation of Neural Networks, Ljubljana, SI, July 2018.
3. N17750, Draft Evaluation Framework for Compressed Representation of Neural Networks, Ljubljana, SI, July 2018.

1. Depending on the compression methods applied, the compressed network may not be directly usable for inference, but decompression must be applied in order to obtain a reconstructed network, that is used for inference. [↑](#footnote-ref-1)