**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

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

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11/N18162**

**January 2019, Marrakech, MA**

**Title Updated Call for Proposals 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 and 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 (or 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 aims to define a compressed, interpretable and interoperable representation for trained neural networks. The representation 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.)
* allow inference without performing full reconstruction of the original network, in order to enable faster inference than with the original 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. The scope of this call is a self-contained representation of the compressed parameters/weights of a trained network, complementing the description of the network structure/architecture in existing (exchange) formats for neural networks.

MPEG has identified a set of relevant use cases and related requirements [1], including applications of neural networks in multimedia and beyond. A call for evidence has been issued [2], and the responses have shown that there is significant potential for using compressed networks in these use cases, which still perform comparably to the original uncompressed networks.

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

# Scope

The scope of this CfP is technology to reduce the size of trained neural networks, i.e., the representation of its weights/parameters. The technology shall provide a complete representation of the parameters/weights of the neural network that contains all information to correctly interpret the compressed parameters/weights. The description of the network structure/topology itself is not in the scope of the call, but such a description may be provided along the compressed information. The proposed representation shall enable integration into existing neural network exchange formats (e.g., NNEF, ONNX).

In this call we define the scope of compression as a method using a provided pretrained model as a starting point. The following constraints apply.

* Searching for an optimal model architecture for the particular problem and training from scratch is considered out of scope.
* Modification of the architecture is acceptable, if it follows an automated and deterministic procedure. The model architecture is defined in terms of a sequence of linear and non-linear operations, where the number of non-linear operations is expected not to increase during the compression operation. The kernel size and number of neurons (non-linearity operators) per layer and the number of connections between layers is expected not to increase during the compression operation. Introducing additional hyperparameters is considered in scope.
* Examples of operations considered in scope are
  + dropping connections
  + dropping layers
  + replacing convolutions with lower dimensional ones
  + changing stride in convolutions without increasing output size
* Retraining denotes an actual fine-tuning or retraining step, but does not refer to compression methods that perform data-dependent transformations (e.g., pruning).

The starting point is a trained neural network for one of the following applications:

* Visual object classification
  + UC2 Camera app with object recognition
  + UC4 Large-scale public surveillance
* UC11 Compact Descriptors for Video Analysis (CDVA)
* Image/Video Compression
  + UC12A Tool-by-tool use case
  + UC12B End-to-end use case
* UC16A Acoustic scene classification

Proponents can convert the provided models for each of the applications to a framework of their choice, and use this implementation as the reference uncompressed model. This includes models with the same architecture trained with other frameworks. In this case, all metrics shall be reported with respect to the uncompressed model in the format actually used in the experiments.

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 applications is limited to testing the compressed representation for inference in these applications (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 three of the applications, but preferably results should be provided for all use cases.

In the future, MPEG may issue additional calls in this area, for example related to incremental representation of neural networks.

# Timeline

|  |  |
| --- | --- |
| 2018/10/31 | Availability of initial set of neural networks, test data and detailed description for the respective applications |
| 2019/01/25 | Availability of all neural networks, test data and detailed description for the respective applications |
| 2019/02/15 | Registration deadline |
| 2019/03/04 | If evaluation process option B (see below) is used: Deadline for electronic submission of compressed (and reconstructed) neural network models |
| 2019/03/18 | Deadline for submission of descriptions (MPEG input contribution) of approaches and evaluation results (for both evaluation process options A and B) |
| 2019/03/23-28 | Evaluation of responses at the March MPEG meeting. |

# 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 during/after compression is permitted. In any case the results for the reconstructed (but not retrained) 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:

* **Application independent metrics**: Compression efficiency, runtime complexity and memory consumption of compression/decompression (measurement is independent of the application)
* **Application specific performance metrics**, comparing the performance of inference using the reconstructed network after compression[[1]](#footnote-1) with using the original network. For the application specific performance metrics, specific frameworks/tools may need to be used (in particular, for the CDVA and video compression applications), which require building and configuration in order to be used. We thus offer two options for evaluating the application 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 126th MPEG, and the compressed (and reconstructed) neural networks 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](mailto:smchun@insignal.co.kr)
    - Audio classification: [Sergiu.Gordea@ait.ac.at](mailto:Sergiu.Gordea@ait.ac.at), Alexander.Schindler@ait.ac.at

# Submission Requirements

The following steps are envisioned for the participation in the call for proposals:

* All proposals shall be prepared in accordance with the requirements provided in Annex A.
* All proposals will be evaluated according to the procedure described in Section 5.
* It is expected that proponents produce results by using tools and procedures described in the evaluation framework. Proposals bringing partial results, or results produced in manner that is different from the described evaluation procedure, may also be considered and evaluated as part of the core experiment process.
* In order to participate and get access to the evaluation framework and test material, proponents will be required to register their intent to participate.
* Proponents are required to subscribe to AHG reflector

<https://lists.aau.at/mailman/listinfo/mpeg-nnr>

The following material is to be submitted electronically. The material shall also be brought to the 126th MPEG meeting.

A submission must contain:

* evaluation results for the specific metrics for the applications (measured according to the evaluation framework, reported in the attached template)
* evaluation results for the generic metrics: runtime (and optionally energy consumption) of compression, decompression and inference, memory consumption of compression and decompression (measured according to the evaluation framework, reported in the attached template)
* the compressed/reconstructed network(s) used for inference (in the same model format as the uncompressed input network, ONNX or NNEF)
* the compressed bitstream of the neural network(s)
* the binaries used to decode the submitted compressed bitstream(s)
* a description of the compression approach, including the parameterization used, as well as hyperparameters, seeds etc. used for training in order to enable reproducability of the training process
* a description whether external data is needed for the compression (or included retraining), and whether (part of) the original training data or other data is used for this purpose

Proponents are required to submit complete results for at least one network for each of at least three applications, but preferably results should be provided for all applications.

Proponents are encouraged (but not required) to allow other committee participants to have access, on a temporary or permanent basis, to their source code.

Proponents are encouraged to submit a statement about the programming language in which the software is written, e.g. C/C++, the frameworks used (e.g., Tensorflow, PyTorch) and the platform(s) on which the binaries were compiled.

Proponents are advised that, upon acceptance for further evaluation, it will be required that certain parts of any proposed technology be made available in source code format to participants in the core experiments process and for potential inclusion in the prospective standard as reference software. When a particular technology is a candidate for further evaluation, commitment to provide such software is a condition of participation. The software shall produce identical results to those submitted to the test. Additionally, submission of improvements (bug fixes, etc.) is strongly encouraged.

# Participation fee

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

# IPR

Proponents are advised that this call is being made subject to the patent policy of ISO/IEC (see [ISO/IEC Directives Part 1](http://isotc.iso.org/livelink/livelink?func=ll&objId=4230455&objAction=browse&sort=subtype), Appendix I) and other established policies of the standardization organization.

# Logistics

Prospective contributors of responses to the Call for Proposals 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 2019/02/15. Interested parties are kindly invited to express their intent as early as possible.

Further 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. N17924, Use cases and requirements for compressed representation of neural networks, Macau SAR, CN, October 2018.
2. N17757, Call for Evidence on Compressed Representation of Neural Networks, Ljubljana, SI, July 2018.
3. N18129, Evaluation Framework for Compressed Representation of Neural Networks, Marrakech, MA, January 2019.

# Annex A. Information Form

**Information form**

1. Title of the proposal
2. Organization name
3. Applications addressed by proposal
4. Indication whether retraining has been performed during/after compression, and a reference to the data set used for retraining
5. Availability of software modules needed for evaluation of the proposal
6. Information on additional functionality supported by the proposal
7. Information on parts of the proposal that must be defined as normative to ensure interoperability

**Requirements**

| **Requirement** | **Description** | **Fulfilment information** |
| --- | --- | --- |
| Efficient representation of the network | The size needed to represent the compressed network should be lower than 30% of the size of the original network. | Measured as described in the evaluation framework. |
| Support representation of different types of artificial neural networks | The compression method shall be applicable to any type/architecture of neural network, and not specific to particular types (e.g., CNNs). | Supported types or any limitations to be described. |
| Self-contained representation of parameters and weights | The representation of the compressed neural network shall contain all required information for decoding the parameters and weights (i.e., not require external information for their interpretation). | Y/N |
| Performance of reconstructed network comparable to original network | The use of the reconstructed network after decoding shall result in a performance comparable to the original uncompressed network in the specified applications. | Measured as described in the evaluation framework.  The best performance achievable with a particular method should be reported. If the proposed method supports lossy compression, additional working points trading performance against compression efficiency should be reported. |
| Inference with compressed network | It is desirable that the compressed network can be directly used for inference without complete reconstruction of the original network. Some methods (e.g., pruning, quantisation) can result in a reduced network that can still be directly used for inference, while others may require decoding/reconstruction of the network in order to perform inference. | Y/N, any required decoding steps (if applicable) |
| Use of additional training data | It must be described whether the method uses (optional) retraining or fine-tuning, and which type and amount of data is required. | steps for which training data is used (if applicable), indication whether training is required or optional |
| Low computational complexity and memory consumption of decoding | The computational complexity and memory consumption of the decoding process needs to be suitable to support use on devices with limited capabilities (e.g., mobile phones, smart cameras). | Measured as described in the evaluation framework. It shall be reported, whether layers of the network can be decoded independently. |

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)