 **ISO/IEC JTC 1/SC 29/ WG 11 N 19515**

**ISO/IEC JTC 1/SC 29/WG 11**

**Coding of moving pictures and audio**

**Convenorship: Japan (JISC)**

|  |  |
| --- | --- |
| **Document type:** | Approved WG 11 document |
| **Title:** | Draft Call for Proposals on Incremental Compression of Neural Networks for multimedia content description and analysis |
| **Status:** | Approved |
| **Date of document:** | 2020-07-17 |
| **Source:** | Convenor, ISO/IEC JTC 1/SC 29/WG 11 |
| **No. of Pages:** | 9 |
| **Email of acting convenor** | ostermann@tnt.uni-hannover.de |
| **Committee URL:** | <http://isotc.iso.org/livelink/livelink/open/jtc1sc29> |
|  |  |

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

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

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11/N19515**

**July 2020, Online**

**Title: Draft Call for Proposals on Incremental Compression of Neural Networks for multimedia content description and analysis**

**Source Requirements Subgroup**

**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. In addition, these trained network are often updated, or networks for different applications are derived from the same base network. This requires also considering the efficient representation of incremental updates.

Building on the earlier CfP for Compression of Neural Networks [3], this call aims at technology for the compressed, interpretable and interoperable representation of updates of trained neural networks. The representation shall be able to

* represent incremental updates different artificial neural network types (e.g., feedforward networks such as CNN and autoencoder, recurrent networks such as LSTM, etc.)
* represent incremental updates that do not require changes to the network structure, as well as those that do
* enable comparable performance to the original network in a set of use cases
* 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.

MPEG is thus calling for proposals on incremental 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 the representation of any required structure updates must be included. The proposed representation shall enable integration into existing neural network exchange formats (e.g., NNEF, ONNX).

In this call we define the scope of a compression as a method using two trained models as a starting point: a base network (i.e., an instance of a trained neural network for the particular use case) and an updated network, which represents an incremental update wrt. the base network. The updated model is typically the result of one of the following operations (this list is considered non-exhaustive):

* Retrain the base network with other data or parameters.
* The base network and the updated model are compressed versions of the same network with different compression ratio.
* The updated network is the result of applying transfer learning, starting from the base network.
* The updated network uses (part of) the base network in its structure, possibly with retraining (parts of) the base network.

The use cases considered so far fall into the following categories:

|  |  |  |  |
| --- | --- | --- | --- |
| **Solution categories** | **Parameter updates** | **Structure changes** | **Data** |
| Update of a network after refining/adding more training data, e.g. in federated learning | yes | no | same as base model |
| Update of a network after transfer learning/adapting to specific data | yes | no | different data set needed |
| Update of a network after transfer learning | yes | yes (e.g., different number of target classes) | different data set needed |
| Update of a network with higher precision/less compression | yes | yes, if sparsity/pruning methods were applied | same as base model |

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 pair of base and updated neural networks for one of the following applications:

* UC10 Federated training and evaluation of neural networks for media content analysis
* UC14A Federated Learning for Medical Applications

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.

# Timeline

|  |  |
| --- | --- |
| 2020/08/31 | Availability of initial set of neural networks, test data and detailed description for the respective applications |
| 2020/10/31 | Availability of all neural networks, test data and detailed description for the respective applications |
| 2021/01/10 | Registration deadline |
|  |  |
| 2020/04/18 | Deadline for submission of descriptions (MPEG input contribution) of approaches and evaluation |
| 2020/04/23-29 | Evaluation of responses at the 134th MPEG meeting. |

# Test Conditions

Given the provided trained neural networks for the different use cases (see [2] 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 [2]. 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. Proponents perform the entire evaluation themselves. They obtain the frameworks/tools as described in [2], 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 134th MPEG, and the compressed (and reconstructed) neural networks should be provided.

# 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 134th 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)
* 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 2021/01/10. 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 [2], for futher questions contact one of the above individuals.

# References

1. N18770, Use cases and requirements for neural network compression for multimedia content description and analysis, Oct. 2019, Geneva, CH.
2. N19516, Evaluation Framework for Compressed Representation of Neural Networks, July 2020, Online.
3. N18162, Updated Call for Proposals on Neural Network Compression, Jan. 2019, Marrakech, MA.

# 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 (partial) network update | The size needed to represent the compressed network should be lower than 10% of the size of the original (partial) network. | Measured as described in the evaluation framework. |
| Efficient incremental representation of neural networks. | The compressed representation shall allow updates of networks with the following characteristics:   * Updates of some or all parameters of a network without structure changes (e.g. after (partial) retraining or transfer learning) * Updates of networks adding weights for new neurons/layers. * Combinations of these two cases. | Tested as described in the evaluation framework. |
| Support updates for scalable compression. | The compressed representation shall allow updates of networks that provide higher precision versions of some or all parameters. |  |
| 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 update, together with the representation of the network to be updated, 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 updated network comparable to original network | The use of the reconstructed network after decoding and applying the update to the network to be updated 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. |
| 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)