

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

**Coding of moving pictures and audio**

**Convenorship: UNI (Italy)**

**ISO/IEC JTC 1/SC 29/WG 11 N19146**

**Document type: Approved WG 11 document**

**Title: Call for Incremental NNR Test Materials**

**Status: Approved**

**Date of document: 2020-01-31**

**Source: Video**

**Expected action:**

**No. of pages: 3**

**Email of convenor: leonardo@chiariglione.org**

**Committee URL: mpeg.chiariglione.org**

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

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

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11/N19146**

**January 2020, Brussels, Belgium**

|  |  |
| --- | --- |
| **Source** | **Video** |
| **Status** | **Approved** |
| **Title** | **Call for Incremental NNR Test Materials** |

# 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. This is addressed by the activity on Neural Network Compression for Multimedia Content Description and Analysis.

There are requirements to extend the technology to cover the compressed, interpretable and interoperable representation of updates of trained neural networks (see updated use cases and requirements document [1]). In this case, the compression method uses 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 rates.
* 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 known so far fall into the following categories:

|  |  |  |
| --- | --- | --- |
| **Use cases** | **Parameter updates** | **Structure changes** |
| Update of a network after refining/adding new training data, e.g. in federated learning | yes | no |
| Update of a network after transfer learning | yes | possible (e.g., different number of target classes) |
| Update of a network with higher precision/less compression | yes | yes, if sparsity/pruning methods were applied |

# Neural networks in update use cases

MPEG is calling for trained neural networks that can serve as test data in use cases related to updates/incremental representation such as those described above. Such data is expected to be a **set of pairs of base and updated neural networks**, where multiple (or all) such pairs could share the same base network, or use another updated network as their base network.

It is desired to provide for each model versions with **varying amounts of changes** (e.g., retrained base model after different number of epochs, different levels of quantisation applied).

For working with the established environments and test cases, and easier comparison with the compression performance on complete models, proposing test data based on variants of the models already included in the evaluation framework for neural network compression [2] is desirable. However, for any use cases that cannot be well covered with these models, test data based on any other model architecture may be proposed.

Here are some possibilities how such data sets could be generated.

* The case of refining after updating could be addressed by creating versions of the models trained on partial data or retrained with additional data.
* For transfer learning without structure changes, exchanging one or few classes could be considered.
* For transfer learning with structure changes, the following example could be considered.
  + Image classification with adding additional classes or training for different target classes. It is likely that variants of the networks already in use trained for other classes are publicly available.
  + Object detection networks based on an image classification network as backbone.

# Responding to this call

Proposals for test materials should be submitted as input documents to the 130th MPEG meeting, containing

* A description of the test materials, including a description how the updated models relate to the base models
* The base/updated model pairs in the format of a common NN framework or exchange format, or a link where to download them
* A link to the data set(s) used for (re)training

# Copyright notice

Content owners should provide a copyright notice along with the dataset to inform MPEG about copyright and usage restrictions.

# Contact:

* Werner Bailer, [werner.bailer@joanneum.at](mailto:werner.bailer@joanneum.at)
* Sungmoon Chun, [smchun@insignal.co.kr](mailto:smchun@insignal.co.kr)
* Wei Wang, [rickwang101@outlook.com](mailto:rickwang101@outlook.com)

# References

[1] N18770 Use cases and requirements for neural network compression for multimedia content description and analysis, Geneva, Oct. 2019.

[2] N18575 Evaluation Framework of Compression of Neural Networks for Multimedia Content Description and Analysis, Gothenborg, Jul. 2019.