**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

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

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11/N17338**

**January 2018, Gwangju, Korea**

**Title: Use cases and requirements for coded representation of neural networks**

**Source Requirements Subgroup**

**Status: Approved**

**Editors: Werner Bailer**

# Introduction

NNR aims to define an efficiently coded, interpretable and interoperable representation for trained neural networks. The scope of existing exchange formats (NNEF, ONNX) is the interface between the framework used for training and the acceleration library/optimisation engine for a specific platform.

If we consider a use case of deploying a trained network to a large range of target devices (e.g., mobile phones, signal processors in a vehicle), the process is likely to have these steps (see Figure 7):

* Training of the network with deep learning framework *L*,resulting in trained neural network *T*
* Export to an exchange format (neural network *T’*)
* Optimisation for 1,..,*n* target platforms, using acceleration/optimisation libraries *A*1, …, *An*, resulting in optimised neural network *O*1, …, *On*
* Distribute the networks *Ok* to the terminal devices, where they are executed with specific (software or hardware) inference engines *Ik*



Figure 1: Deployment/inference process.

We can make the following observations about the two interfaces:

* Compression based on a standard format would target *T’*, which is not where the volume of data is distributed
* Compression would be most beneficial for *Ok*, where networks are distributed to a large number of devices running the inference engine.
* *Ok* may already have a pruned network structure and quantised weights, so that compression is less (in the worst case not) effective
* *Ok* may not be easy to represent in a standard representation, and the format coming out of a specific acceleration library *Ak* may not be documented at all

Some of the NNR use cases would benefit from compression on the first interface, while for other only a compressed representation for the second interface would be beneficial.

A processing chain for neural network compression is shown in Figure 2. Some of the steps change (simplify) the model structure, while others affect only the weights. Several of these operations are also performed by the acceleration/optimisation libraries for specific target platforms.

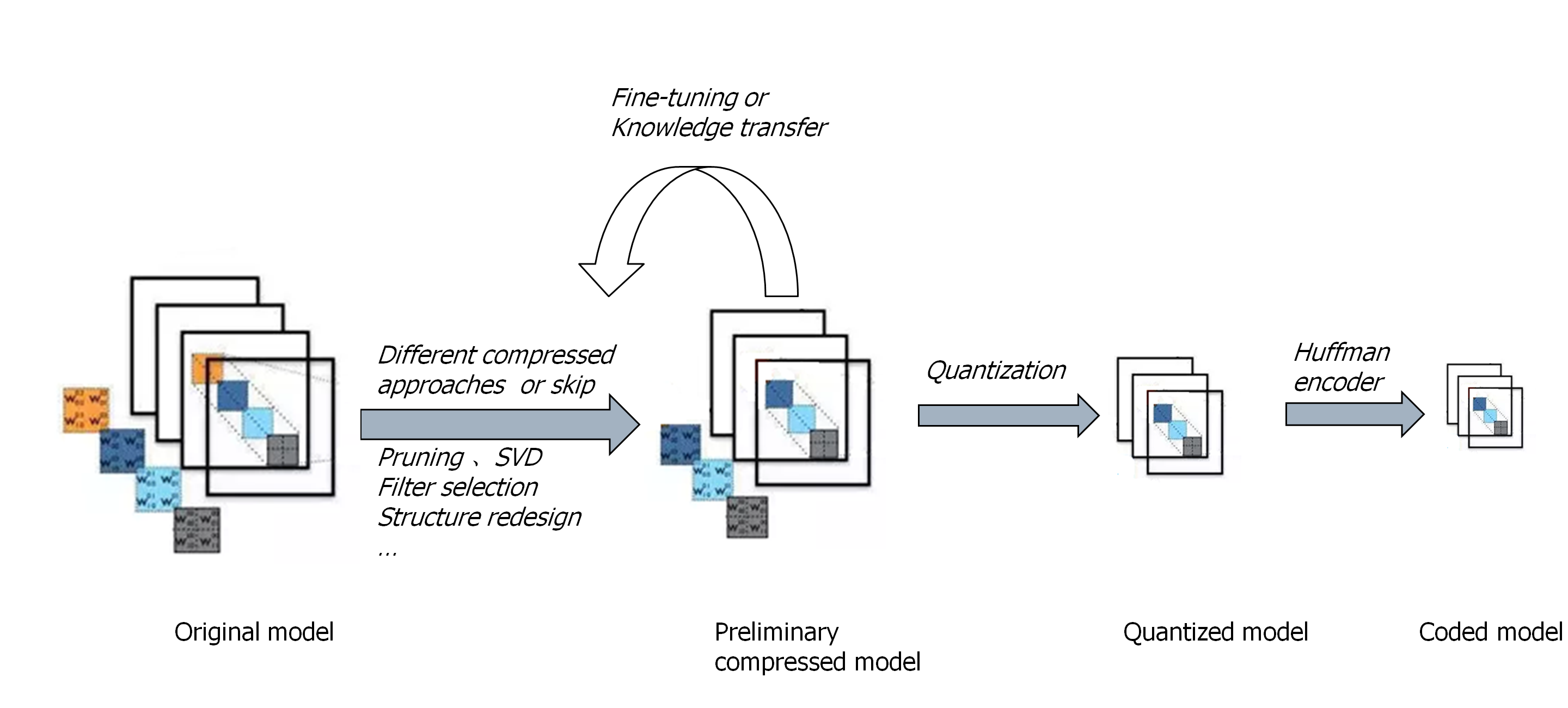


Figure 2: Framework for neural network compression.

The remainder of this document groups the proposed use cases and provides an overview of the requirements from these use cases.

# Use cases on NN distribution/deployment

|  |
| --- |
| UC1 Installing NN-based applications |
| In an extension of the current app model, users install Neural Network-based applications on their devices. The NN-based core of the application can be compressed |
| **Overlap with other use cases** |
| General description of use cases for distribution/deployment, other UCs in this category can be seen as specialisations |
| **Required  features** |
| 1. Lossy compression 2. Scalability (if end user stops downloading compressed NN, a level of performance is still guaranteed) |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| The device converts the compressed representation and obtains a NN capable of processing data.  We need to explore whether the compressed representation can process data at what cost |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)? What is the estimated size of these neural networks?** |
| At steady state this scenario would envisage tens of millions downloads every day globally, possibly growing to hundreds of millions.  The size of neural networks will be application dependent |

|  |
| --- |
| UC2 Camera app with object recognition |
| Recent smart phones include cameras, that adjust their automatic mode based on scene/object recognition results. For example, the Huawei Mate 10’s camera app classifies 13 scene/object types[[1]](#footnote-1).  In order to add/improve classification, the trained NN needs to be updated in the app. There are probably also cases where users are interested in adding types of objects/scenes, training some on their own and transferring models from one phone to another. |
| **Overlap with other use cases** |
| Specific example of UC1 and UC5, due to limitations of mobile device UC6 is relevant  Overlap with UC10 if retraining with local images is considered |
| **Required features** |
| * Significant size reduction of the trained network to speed up download of updated NNs. * If lossy compression is used, the performance of using a compressed representation of the network must be similar to that of using an uncompressed network. * (optional) The use of the network should be possible without inflating it to its original size in memory. |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| Currently, the app and the trained classes come from the same provider. Huawei has stated that there are plans to add more object classes and use user data to train classes. Third parties could provide trained NNs, and exchange of trained NNs would be necessary to avoid lock-in of users who invested in training their own models. |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| * Updates are expected to be infrequent (maybe a few times per year), but potentially to a large number of devices. * For ImageNet (1,000 classes), trained NN sizes are in the rage of 50-500 MB. * For mobile devices, the download bandwidth may vary considerably due to the network type. * Latency is an issue when a user downloads a new model to use it right away. |

|  |
| --- |
| UC3 Translation app |
| Some translation apps (such as e.g. Microsoft Translator[[2]](#footnote-2)) make use of NNs for language recognition and/or synthesis. Those apps can be extended with a set of languages, that can be downloaded within the app. The data for a language contains the trained networks for a language, but also other data (e.g. dictionaries). |
| **Overlap with other use cases** |
| Specific example of UC1, due to limitations of mobile device UC6 is relevant |
| **Required features** |
| * Significant size reduction of the trained network to speed up download of additional languages. * If lossy compression is used, the performance of using a compressed representation of the network must be similar to that of using an uncompressed network. * (optional) The use of the network should be possible without inflating it to its original size in memory. |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| Currently, the app and the trained language models come from the same provider, and are designed to run on a specific set of mobile platforms.  A standard representation would allow third parties to offer additional, or specifically trained models (dialects, terminology) for the translation app. |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| * Updates are expected to be infrequent (maybe a few times per year), but potentially to a large number of devices. * The size of the data per language (not only the NN) is currently around 200MB. * For mobile devices, the download bandwidth may vary considerably due to the network type. * Latency is an issue when a user downloads a language to use it right away. |

|  |
| --- |
| UC4 Large-scale public surveillance |
| |  | | --- | | In recent years, surveillance cameras in public area are increasing and it is becoming a challenging problem to realize automatic surveillance systems for public areas such as stations, large-scale parks, exhibition halls, stadiums and shopping malls. Video analysis using deep neural network is becoming one of essential core functions for automatic surveillance systems.  Fig 1 shows one of potential automatic surveillance systems. Area 1 to area M have local automatic surveillance system composed of surveillance cameras which have a function to extract metadata like the output of object detection, object tracking, action recognition, etc. by deep neural network and a local server as “surveillance AI” to recognize the situation of the corresponding area by the metadata from the cameras, respectively. A cloud server periodically re-trains the neural network for the metadata extraction on the surveillance cameras and provide it to the cameras. Since the automatic surveillance system can be connected to a huge number of cameras[[3]](#footnote-3), it will be desirable to compress the neural network when it is transferred to the cameras. Note that it may be necessary to transmit different networks to different areas, e.g., depending on the relevant tasks for each area.  Moreover, since it is expected that it will be realized to collect/generate the training dataset automatically for re-training of the neural network in the future, the update cycle of the neural network will become shorter. In consideration of such situation, the compression of the neural network will be more important to reduce the traffic over the network.    Figure 3: An example of future public surveillance system. | |
| **Overlap with other use cases** |
| Specific case of UC1 |
| **Required features** |
| * Low computational complexity for decoding the compressed neural network model * Low memory consumption for decoding the compressed neural network model * “High efficient” (means small memory sized) representation for decoded neural network model * High error robustness for compressed neural network model * Ability to detect the manipulation of the compressed neural network model |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| * Interfaces benefitting from the standard: Surveillance cameras * Who: Providers of the automatic surveillance systems |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| * Need to update the neural network once a month or less in near future and once a week or more in further future. * Over hundreds million surveillance cameras will be connected to the surveillance systems. * Latency of download the neural network is less important compared with that of execution of actual process by the neural network |

|  |
| --- |
| UC5 Visual pattern recognition (VPR) |
| Visual pattern recognition (VPR) techniques are an important component of intelligent system and are used for many application domains. Visual pattern recognition is one of the most effective applications to which deep neural network can be applied, and excellent recognition results have been reported in many recent studies. In particular, visual analysis and pattern recognition applications in mobile and IoT devices are becoming more popular. In order to apply DNN in a mobile and IoT environment in which computing capability is limited, a system configuration cooperating with a PC or server is considered as a practical way. In other words, a networks is trained in a PC or a server, then pattern recognition is performed using the trained network in a mobile/IoT device.  Fig.1 shows an example such case of VPR systems using DNN in which object detection is performed. A set of neural networks with different algorithms are trained and provided by PC/Server. A user can select an appropriate network for the given application of pattern recognition, then the network model and the trained results of the selected network are delivered to a mobile/IoT device in which pattern recognition is performed.  In addition, a user can also optionally report the evaluation of the pattern recognition result, which is feedback to the server in which network will be retrained/updated.    Fig. 1 An example of pattern recognition using DNN in an mobile/IoT device |
| **Overlap with other use cases** |
| Specific case of UC1, similar to UC2, due to the target device UC6 applies |
| **Required features** |
| * High efficient representation of the neural network representation and trained weights * (optional) The use of the network should be possible without inflating it to its original size in memory * (optional) Scalability of representation in terms of leaving choice of some parameters open |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| * Interface between a mobile/IoT device and a server can be supported in an interoperable way to deliver trained network for pattern recognition * Implementation of pattern recognition in a mobile/IoT environment with limited computational power. * a DNN based pattern recognition device can be provided by different vendors * Who: service providers (DNN provider), vendors |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| * Updates are expected to be infrequent (incremental training, different pattern recognition methods), but potentially to a large number of devices. * The size of the training set and the method used to build it are often dependent on the selected pattern classification technique. * Latency is in most cases not critical, except for cases where a user downloads the mobile app to use it right away. |

|  |
| --- |
| UC6 NN representation for devices with limited memory and bandwidth |
| This use case covers a scenario where the device that will run a neural network has very limited memory, computational capability and bandwidth. An example of such an application is object recognition in memory, computation and bandwidth limited IoT devices which are capable of media processing.  It is challenging for such devices to load large neural networks and do inference. It is also not always possible to compress a neural network so that it fits into the memory of the IoT device. Moreover, such a demanding compression usually has an impact on the inference performance (i.e. significant loss of accuracy). Hence, it is required that a mechanism is defined to handle loading and inference of such neural networks that run on devices with very limited memory, computation capability and connectivity bandwidth. |
| **Overlap with other use cases** |
| Specific case of UC1 (restricting the type of target platform), related to UC2, UC3 and UC5 |
| **Required features** |
| * A mechanism for the communication of NN models to memory and/or bandwidth and/or computation limited devices. * A communication protocol to facilitate this kind of mechanism. * A mechanism for guiding the execution of NN models. |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| The use case is aimed for low-memory, low-bandwidth devices.  Required components are involved in mobile devices and memory and bandwidth limited IOT devices. |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| * The rate of neural networks transmission depends on the application use cases. * The amount of information to be communicated is determined by the bandwidth of the device that the neural network will be uploaded on. * Estimated size of the neural network can vary from hundreds of KBs to hundreds of MBs. * The expected bandwidth of the distribution channel from the minimal of 100Kbs. * The maximum latency is application dependent, in the range of mili-seconds to seconds. |

|  |
| --- |
| UC9 Efficient re-use of neural networks among different media applications |
| Many applications utilize neural networks derived from the same neural network. This is the case for example for mobile phone apps. The common procedure consists of taking a neural network pre-trained on a large dataset (such as ImageNet), which is able of extracting high-quality generic visual features, and deriving from it a new model for a specific down-stream task. The derivation may consist of one or more of the following options:  - fine-tuning all layers;  - freezing some layers/weights and fine-tuning only the remaining layers/weights;  - freezing or fine-tuning some layer, and adding new layers and branches which are trained.  This means that several apps would need to store on the device the same base neural network multiple times. This is a waste of both storage and inference speed. In fact, inference speed may become very low if multiple models need to be run on the same device at the same time, e.g., when taking a picture or video the camera may run a camera parameter tuning neural net, a person detection neural net, a style-transfer neural net, etc.  There is a need to efficiently share & reuse a network model among multiple applications and tasks.  Another need related to this use case is the update (versioning) of the neural network that an application use. The versioning may be needed to upgrade the network to a better performing one. |
| **Overlap with other use cases** |
| Extension of UC1 (partial updates of models), also applicable to other UCs in this group |
| **Required features** |
| * Signaling for negotiating the re-use of NN models among applications and between application providers and users. |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| The interfaces which will benefit will be the media consumption devices, which would require less bandwidth because less layers will be transmitted. Also, the devices will save storage as less neural networks will be stored on device. |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| Frequency  - Every time a new model for a certain application or a new version of that model needs to be transmitted. For mobile phones, the frequency may be from every day to every month for each user.  Size  - From few KB to several hundreds MB  Bandwidth  - The bandwidth may depend on the specific distribution channel. One example is 2G, 3G and 4G networks.  Maximum latency  - Can be from few milliseconds to few seconds. So, it is application-dependent. |

# Use cases on NN (re)training

|  |
| --- |
| UC7 Deep NN Factory |
| The scientific and technical advancements witnessed in the last 5 years in the field of machine learning based on deep neural networks is making a whole new business area grow at an incredible pace. Main cloud providers are offering increasingly engineered services that can be used by their customers to train specialized networks on demand. If on the one hand training such complex networks in a reasonable amount of time requires the availability of a non-trivial amount of dedicated hardware, which can be affordable only by said providers, on the other hand the execution (inference) phase can be accommodated on much simpler infrastructure, in most cases affordable by the customers themselves. Thus, the opportunity to “download” pre-trained networks and use them on site starts to be a meaningful use case in this domain. The necessity to retrain/refine networks as long as the statistics of the data evolve or new classes are needed is an additional element making this scenario even more realistic. A “Deep NN Factory” is therefore a system implementing such training/refinement service and producing pre-trained (or refined) deep networks on demand of its customers. Delivered NN by the Factory can be compressed using a lossless or lossy technique, depending on the requested delivery latency. This use case is limited to those factories operating in the multimedia domain, and in particular specialized in multimedia classification tasks. |
| **Overlap with other use cases** |
| Relation to UC10 (specific case of distributed training) |
| **Required  features** |
| * The Deep NN Factory shall accept training data for each specific classification task in a standard format; * The Deep NN Factory shall provide a standard representation of the trained NN so that the customer can run the network on his preferred execution infrastructure; * The Deep NN Factory shall be able to compress the delivered NN using a standard lossless or lossy compression technique before delivery; * The decompression of compressed NN shall be fast enough; * The Deep NN Factory shall provide a unique identification of the trained NN and support ways to identify an instance of a trained NN based on its configuration (e.g. taking into account topology, weights); * The Deep NN Factory shall be able to refine or retrain an existing NN provided by a customer in a standard format (compressed or not, even if the original NN was not pre-trained by it), together with additional data for the training if needed. |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| I/O interfaces of the Factory are the main interfaces benefitting from a standard NN representation. The calling components may be either mobile devices or more complex systems. |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| The update rate is not expected to be high, since the lifetime of a pre-trained classification model is expected to be relatively long. The expected size of each trained network (uncompressed) is measurable in several Gbytes. The expected bandwidth of the uplink (when transferring classification data for the training or refining) and of the downlink (when getting the resulting NN) is that appreciable from current state-of-the-art cloud services (several hundred Mbps) although this may be capped by the capacity of the mobile network when these operations are made in a mobile environment. The maximum acceptable latency is highly depending on the target application. For mobile applications, that need to be highly responsive to the user, the latency should be within few seconds. Other cases (such as content management applications), acceptable latency can be considerably higher (until few minutes). |

|  |
| --- |
| UC8 Personalized machine reading comprehension (MRC) application |
| MRC is an advanced deep learning technique in natural language processing. For the user’s question, it finds the answer in the natural language text using the neural network (NN). In order to answer user’s questions based on personal e-mail text and SMS messages, the MRC NN model should be learned from the personal questions and the personal texts. However, because of the privacy protection, it is difficult to collect and learn the personal data to send to the central cloud server. This use-case describes the case where the user’s edge device (i.e. smart phone) learns personalized MRC NN model using the personal MRC usage record. The personalized MRC NN model learned at the user’s edge device is sent to the central cloud server and used to build an updated version of shared MRC NN model. Fig. 1 shows an overall architecture for personalized MRC application.    **Fig.1 An overall architecture for personalized machine reading comprehension application** |
| **Overlap with other use cases** |
| Relation to UC10 (specific case of distributed training) |
| **Required features** |
| * The NN model transmitted to the device should include algorithm and parameter information for NN learning. * The NN model transmitted to the cloud server should include training metadata information such as number of training instances. * Significant size reduction of the neural network representation and trained weights. * (optional) The use of the network should be possible without inflating it to its original size in memory. |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| The NN building and distribution can be implemented in an interoperable way by different vendors of a backend system (e.g., operated by a broadcaster, service operator) and by client software for different devices.  This includes a platform independent compressed representation of a neural network, that can be used regardless of the platform and software frameworks used to implement personalized MRC. |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| * Once the initial NN building is done, updates are expected to be infrequent, but potentially to a large number of devices. * The learning is performed when each individual device is in an unused and battery charging state. * Each device infrequently transmits an updated NN model to the cloud server. * The size of uncompressed NN model is about 500 and 1000 MB, depending on the model * For mobile devices, the download bandwidth may vary considerably due to the network type. * Latency is in most cases not critical, except for cases where a user downloads the mobile app to use it right away. |

|  |
| --- |
| UC10 Distributed training and evaluation of neural networks for media content analysis |
| This use-case concerns the distributed training and evaluation of neural networks for applications such as object detection and data generation performed on mobile devices and other devices. The open issues related to centralized solutions are: 1) power in-efficiency for centralized training; 2) training data communication; and 3) the deployment of the learned models.  In distributed (or federated) training, there is a central server which orchestrates the distributed training on several end devices (training devices) for the same (or similar) task.  The training devices are assumed to have media capture capabilities, or other means for utilizing content for training. Each training device uses captured or other media to perform a local model update on the local version of the neural network.  In such a framework, it is necessary to have good neural network representation and evaluation capabilities, where the NN representation can be generated and shared among central server and the training devices.  The estimated model update is not communicated to the central entity and/or to the other devices until it is validated. In fact, for each training-device performing a partial training, it will be extremely important to evaluate the quality and integrity of the partially-trained model, in order to decide whether the model-update estimated by a certain training-device should be taken into account (i.e., communicated to a central entity and/or to all other training devices) or be temporarily ignored. If it is successfully validated, the model update is then taken into account. This may mean that the model update is readily shared with all other training devices or it will be further processed and then shared.  Examples of aspects to be evaluated will be robustness to adversarial examples, performance based on the specific task (e.g., object detection, data-generation), etc.  The goal is to standardize the signaling content and format needed to evaluate and exchange neural networks. Signaling content may include data for evaluating the model updates and data representing the actual weight updates. |
| **Overlap with other use cases** |
| Generalization of UC7&8, which describe specific cases of distributed training |
| **Required features** |
| * Validation of model updates in distributed NN from individual training devices * Communication of the whole (or partial) NN models between entities * The NN representation, for this use case, should support different types of neural networks, such as discriminative (for which input data is usually mapped to a classification or regression output), and generative (for which input data is transformed to data in the same domain as the input or to a different domain). Also, generative models may be conditional (on a specific input data) and unconditional (where the input is a sample from a latent space). * An efficient way to represent model updates. * Efficient compression schemes of NN models to be communicated * Robustness to adversarial attacks * Detection of malicious model updates * Interpretability of trained models to explain inference resultsTime-stamping of the model updates. |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| The interface which benefits from a standard neural network representation is the central entity which is interested in obtaining a high quality trained model.  In addition, in the case of federated training, another interface is the training devices themselves, as their partially-trained model will improve based on the weight updates sent by the central entity. |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| Frequency  - The neural network model updates are expected to be communicated in an adjustable manner (e.g., based on end device settings), and only when a device’s model update has been validated.  - Also, at the server side, the model updates from multiple devices can be shared with the devices only if some criteria are met  Size  - From few KB to several hundreds of MB  Bandwidth  - It depends on the distribution channels, such as 2G, 3G, 4G, 5G mobile networks etc.  Maximum latency  - When the model update made by a certain device has been validated, the update is sent to the central entity and the maximum acceptable latency is application & implementation dependent which could be from ms to seconds or even longer.  - When the central entity sends a weight update to all devices, the max acceptable latency is also application & implementation dependent. |

# Use cases on image/video processing and coding

|  |
| --- |
| UC11 Compact Descriptors for Video Analysis (CDVA) |
| MPEG CDVA specifies a compact descriptor for video segments, aiming to support search for specific objects across video collections. The standard describes the bitstream representation of such a descriptor, as well as those parts of the extraction process that are required to ensure interoperability of the resulting descriptors. Features extracted from deep convolutional neural networks (often named deep features) have been shown to be applicable for a wide range of computer vision tasks. CDVA makes use of features extracted from CNNs, which have been shown to provide good performance and are complementary to traditional features such as those included in CDVS.  There are two levels of conformance: With strict conformance, the neural network is exactly defined, and thus the issue of transmitting the network is only one of the initial deployment of the application. With loose conformance the network may be exchanged, but the same network representation needs to be shared between all parties to establish interoperability. In this case, an update of the neural network may need to be updated more frequently.  Intended use cases of CDVA include descriptor extraction on devices such as smart phones or set top boxes. |
| **Overlap with other use cases** |
| Related to UC13 |
| **Required features** |
| * Significant size reduction of the neural network representation and trained weights. * If lossy compression is used, the performance of using a compressed representation of the network must be similar to that of using an uncompressed network. * (optional) The use of the network should be possible without inflating it to its original size in memory. * (optional) Matching descriptors extracted using a compressed network with descriptors extracted using an uncompressed network results in only small performance loss compared to matching two descriptors extracted using uncompressed networks. |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| The aim of CDVA is to decouple descriptor extraction and matching/retrieval, so that these two steps can be implemented in an interoperable way by different vendors of a backend system (e.g., operated by a broadcaster, service operator) and by client software for different devices.  This includes a platform independent compressed representation of a neural network, that can be used regardless of the platform and software frameworks used to implement CDVA. |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| * Updates are expected to be infrequent (maybe a few times per year), but potentially to a large number of devices. * The size of the uncompressed network is about 500-600 MB. * For mobile devices, the download bandwidth may vary considerably due to the network type. For set top boxes this would have to be done with as part of a regular software update, thus the size is an issue. * Latency is in most cases not critical, except for cases where a user downloads the mobile app to use it right away. |

|  |
| --- |
| UC12 Image/Video Compression |
| Recently, studies on still image and video compression based on deep neural network (DNN) have been actively conducted, and it has been reported that its performance can be comparable to the conventional image compression standard[[4]](#footnote-4). Image/video compression using DNN is widely thought to be a new compression method. RNN (Recurrent Neural Network) is mainly applied for image/video compression. In particular, for video compression, different approaches have been attempted to apply deep learning techniques on a tool-by-tool basis or on the entire codec.  DNN based image/video compression requires coded representation of neural network in the following two aspects. Image/video encoders and decoders are used at different locations in general applications codec. In other words, a feature vector, which is the output of the encoder to the input image/video, is stored or transmitted, and the original input image/video is reconstructed at the decoder. Therefore, the coded representation of the trained neural network of a codec is needed to be delivered to decoders. In addition, re-trained network is desirable to be updated periodically.  In addition, NNR may be needed in a specific application environment of the image/video codec. For example, there may be available different NNs customized to categories of image/video to be encoded. In this case, the coded representation of NN associated with a given input image/video may be required to support such application. |
| **Overlap with other use cases** |
| Related to UC11 and UC13 (using specific types of feature extraction/processing) |
| **Required features** |
| * High efficient representation of neural network including network model and trained weights * Low complex decoding (computation/memory) of the compressed representation of neural network * High error robustness for the coded representation of neural network |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| * DNN based image/video codec can be used in an interoperable way in diverse media services (broadcasting, mobile streaming, visual communications, etc.). * Implementation of a DNN based decoder can be supported in an interoperable way by different vendors * Who: service providers (broadcaster, IPTV, mobile streaming, etc.), network operators (mobile), vendors |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| * Once a DNN based decoder is initialized, updates of NN are expected to be infrequent (periodic update may be feasible depending on applications, for example, once a month), but potentially to a large number of devices. * The estimated size/distribution bandwidth of the presentation of DNN encoder/decoder may not be critical to be initialized infrequently. * For the representation of output of an encoder, the size/bandwidth should be smaller than that of the conventional codec. * After initialization of a decoder, latency should be comparable to the current latency in the applications of broadcasting, communication, and streaming. |

|  |
| --- |
| UC13 Distribution of neural networks for content processing |
| Recently, analysis and processing of media such as images and videos require the application of neural networks. One example of content processing is applying neural network based super resolution (NSR) on a video at the client side, so that the video can be transmitted to that client at reduced resolution and thus save bandwidth.  The neural network performing NSR may be trained to increase the resolution by a specific factor (e.g., 2), thus the sender needs to down-sample the content by the same factor and signal to the client the neural network to be used. This may consist of either signaling the neural network type (e.g., the super resolution factor) if the client already has that neural network or the actual neural network weights and topology.  This can be extended to other content processing techniques apart from super resolution, such as other content enhancing neural networks.  In this use case proposal, we use neural network and model interchangeably. |
| **Overlap with other use cases** |
| Relation to UC11 and UC12 |
| **Required features** |
| * Signaling is required between server and client in order for client and server to negotiate the relevant models or model update (topologies/architectures, and associated weights) to be used with the content. * When the model is not already present at client side, efficient representation and transmission of the model from server to client which performs the content processing. |
| **Which is/are the interface(s) that benefit from a standard neural network representation? Who is in control of the components (e.g., mobile devices, embedded processors) involved?** |
| The content retrieval and negotiation interfaces will benefit from a standardized signaling of neural network representations, from inter-operability perspective.  Moreover, the media consuming devices will benefit from a standardized neural network representation which is suitable to their AI capabilities. |
| **How often are neural networks expected to be transmitted/updated in this use case (per unit, in total)?**  **What is the estimated size of these neural networks?**  **What is the expected bandwidth of the distribution channel?**  **What is the maximum latency that is acceptable?** |
| Frequency  - The frequency may depend on how often the model is updated, and it may vary from once every few milliseconds to once every few days.  Size  - Sizes can range from few KB to several hundreds of MB.  Bandwidth  - The bandwidth may depend on the specific distribution channel. One example is 2G, 3G and 4G networks.  Acceptable latency  - In the order of few milliseconds to few seconds. |

# Summary of Requirements

This table summarises the different requirements described, and the use cases that make reference to them.

|  | **distribution** | | | | | | | **(re)training** | | | **processing** | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** | **9** | **7** | **8** | **10** | **11** | **12** | **13** |
| **Exchange representation** | X | X | X | X | X | X | X | X | X | X | X | X | X |
| **Efficient representation of the network** | X | X | X | X | X | X | X | X | X | X | X | X | X |
| **Support lossless compression** |  | X | X |  |  |  |  | X |  |  |  |  |  |
| **Support lossy compression** | X | X | X |  |  |  |  | X |  |  | X |  |  |
| **Comparable performance of compressed network than original network** | X | X | X | X | X | X | X | X | X | X | X | X | X |
| **Scalable compression** | X |  |  |  |  |  |  |  |  |  |  |  |  |
| **Inference with compressed network** |  | X | X |  | X |  |  |  | X |  | X |  |  |
| **Low computational complexity decoding** |  |  |  | X |  | X |  |  |  |  |  | X |  |
| **Low memory consumption** |  |  |  | X |  | X |  |  |  |  |  | X |  |
| **Ability to detect manipulation of compressed network representation** |  |  |  | X |  |  |  |  |  | X |  |  |  |
| **Robustness to bit error of received compressed network** |  |  |  | X |  |  |  |  |  |  |  | X |  |
| **Mechanism to guide execution of NN models** |  |  |  |  |  | X |  |  |  |  |  |  |  |
| **Interface for training data for classification** |  |  |  |  |  |  |  | X | X |  |  |  |  |
| **Fast NN decompression** |  |  |  |  |  |  |  | X |  |  |  |  |  |
| **Unique identification and time-stamping of instance of trained NN** |  |  |  |  |  |  |  | X |  | X |  |  |  |
| **Represent training metadata** |  |  |  |  |  |  |  |  | X |  |  |  |  |
| **Signaling for negotiating the re-use of NN models** |  |  |  |  |  |  | X |  |  |  |  |  |  |
| **Represent partial NN models** |  |  |  |  |  |  |  |  |  | X |  |  |  |
| **Efficiently represent model updates** |  |  |  |  |  |  |  |  |  | X |  |  |  |
| **Support validation of model updates** |  |  |  |  |  |  |  |  |  | X |  |  |  |
| **Support discriminative, generative and unconditional network types** |  |  |  |  |  |  |  |  |  | X |  |  |  |
| **Robustness to adversarial attacks** |  |  |  |  |  |  |  |  |  | X |  |  |  |
| **Interpretability of trained models to explain inference results** |  |  |  |  |  |  |  |  |  | X |  |  |  |
| **Interoperability of components using compressed and uncompressed versions of the network** |  |  |  |  |  |  |  |  |  |  | X |  |  |
| **Signaling to negotiate relevant model updates** |  |  |  |  |  |  |  |  |  |  |  |  | X |

1. http://www.zdnet.com/article/huawei-mate-10-pro-camera-enhancing-auto-mode-through-artificial-intelligence/ [↑](#footnote-ref-1)
2. https://play.google.com/store/apps/details?id=com.microsoft.translator&hl=en [↑](#footnote-ref-2)
3. |  |
   | --- |
   | For example, over 400 million surveillance cameras will be installed in China by 2020,  http://www.bbc.com/news/av/world-asia-china-42248056/in-your-face-china-s-all-seeing-state |

   [↑](#footnote-ref-3)
4. G. Todericiet et al., “Full Resolution Image Compression with Recurrent Neural Networks,” In Proc. CVPR 2017, July 2017.  
   C. Kin, B. Coker, “Video Compression Using Recurrent Convolutional Neural Networks,” [↑](#footnote-ref-4)